Laying Lexical Foundations for NLP: the Case of Basque at the *Ixa* Research Group

Xabier Artola-Zubillaga

Ixa NLP Research Group
Fac. of Computer Science – Univ. of The Basque Country
649 p.k., 20080 Donostia
jiparzux@si.ehu.es

Abstract

The purpose of this paper is to present the strategy and methodology followed at the *Ixa* NLP Group of the University of The Basque Country in laying the lexical foundations for language processing. Monolingual and bilingual dictionaries, text corpora, and linguists' knowledge have been the main information sources from which lexical knowledge currently present in our NLP system has been acquired. The main lexical resource we use in research and applications is a lexical database, EDBL, that currently contains more than 80,000 entries richly coded with the lexical information needed in language processing tasks. A Basque wordnet has also been built (it has currently more than 50,000 word senses), although it is not yet fully integrated into the processing chain as EDBL is. Monolingual dictionaries have been exploited in order to obtain knowledge that is currently being integrated into a lexical knowledge base (EEBL). This knowledge base is being connected to the lexical database and to the wordnet. Feedback obtained from users of the first language technology practical application produced by the research group, i.e. a spelling checker, has also been an important source of lexical knowledge that has permitted to improve, correct and update the lexical database. In the paper, doctorate research work on the lexicon finished or in progress at the group is outlined as well, as long as a brief description of the end-user applications produced so far.

1 Introduction

Basque is a language spoken on both sides of the west-end border between France and Spain by approx. 700,000 people (25% of the population). It is co-official in some regions of the country, and moderately used in administration instances. Its use in education, from the mother school up to the university, is growing since the early eighties. There is one TV channel and a newspaper is published daily in Basque. The standardization of written language is in progress since 1968. More information on the web can be obtained at http://www.euskadi.net/euskara.

The purpose of this paper is to present the strategy and methodology followed at the *Ixa* NLP Group of the University of The Basque Country in laying the lexical foundations for natural language processing tasks.

In section 2, the *Ixa* NLP Research Group is introduced. Section 3 is devoted to describe EDBL, the main lexical database. Next two sections illustrate the construction of *Euskal Wordnet*, a wordnet of Basque, and EEBL, a lexical knowledge base which links the database, the wordnet, and knowledge derived from a monolingual dictionary. Next, in section 6, research work carried out at the group on the field of the lexicon is outlined. Finally, and before the conclusions, some end-user products and applications are briefly presented in section 7.

2 The *Ixa* NLP Research Group at the University of The Basque Country

The *Ixa* Research Group on NLP (http://ixa.si.ehu.es) belongs to the University of The Basque Country and its research has been conducted from the beginning (1986/87) on the fields of computational linguistics and language engineering.

The main application language has been and currently is Basque, but research and applications involving English, Spanish or French have been carried out as well. The strategy of language technology development at the group has been from the beginning a bottom-up strategy (Agirre *et al.*, 2001a, Díaz de Ilarraza *et al.*, 2003), that is, our goal has been first to lay the processing infrastructure –basic resources and tools–, in order to then be ready to produce end-user applications. Even our first product, a spelling checker, was conceived upon a general-purpose morphological analyzer (Alegria *et al.*, 1996).

The group is interdisciplinary (computer scientists and linguists) and it is formed nowadays by around 40 people, between lecturers, senior researchers and grant-aided students.

Ixa maintains scientific relationships with universities in different countries, and funding comes mainly from the university, local and Basque Governments, Spanish Government and European institutions.

3 EDBL: building the main lexical database from scratch

EDBL (Aldezabal *et al.*, 2001) is the name of our main lexical database, which is used as a lexical support for the automatic treatment of the language.

EDBL is a large store of lexical information that currently contains more than 80,000 entries. It has been conceived as a multi-purpose lexical basis, i.e. a goal-independent resource for the processing of the language.

The need of such a lexical database arose when the design and implementation of the morphological analyzer was faced. It was evident that a store for words and their attributes was necessary, and so, we took a dictionary and picked up all the entries with their part-of-speech (POS) information: this was the seed of EDBL. In these years, the design of the database has been significantly modified and updated in two occasions, to arrive to the current conceptual schema described in section 3.3.

3.1 The lexical database within the stream of language processing tools

EDBL is fully integrated in the chain of language processing resources and tools (see fig. 1), and the information contained in it is exported when required to be used as input by the language analysis tools.

A customizable exportation procedure allows us to select and to extract the information required by the different lexicons or tools in the desired format (XML, plain text, etc.). The lexicons obtained in this way are subsequently used in tools such as a morphological analyzer, a spelling checker (Aduriz *et al.*, 1997), a tagger/lemmatizer (Aduriz *et al.*, 1996a), and so on.

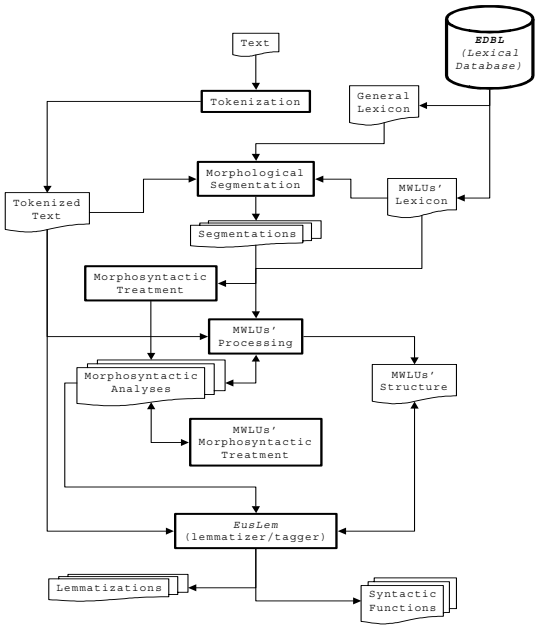

Figure 1: EDBL within the stream of language processing tools

3.2 Sources of knowledge used to populate EDBL

Different sources are used to populate the database: linguists and lexicographers' knowledge, monolingual and bilingual dictionaries, standard word lists regularly published by the Basque Language Academy (*Hiztegi Batua*: Euskaltzaindia, 2000), and the feedback given by the spelling checker (application and users) and other NLP tools such as the morphological analyzer or the lemmatizer.

When gaps in the database are detected, the lexicographer in charge of EDBL decides whether the entries are to be added or not, and fills the values for the required attributes. An especially conceived importation application facilitates this task to the lexicographer, allowing him or her to specify the input format, and making some deductions based on the POS of the entry, for example.

Apart from *Hiztegi Batua*, other dictionaries that have been used for this purpose are a small monolingual dictionary (Elhuyar, 1998), a Basque-Spanish/Spanish-Basque dictionary (Morris, 1998), a synonym dictionary (UZEI, 1999), and *Euskal Hiztegia* (Sarasola, 1996), a bigger monolingual explanatory dictionary.

3.3 Conceptual schema of the database

In this section, the Extended Entity-Relationship (EER) data model is used to describe the conceptual schema of the database (see fig. 2).

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. Every lexical unit in EDBL belongs to this data class. The units in it can be viewed from three different standpoints, giving us three total specializations (all units in EDBL belong to the three specializations). 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.

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

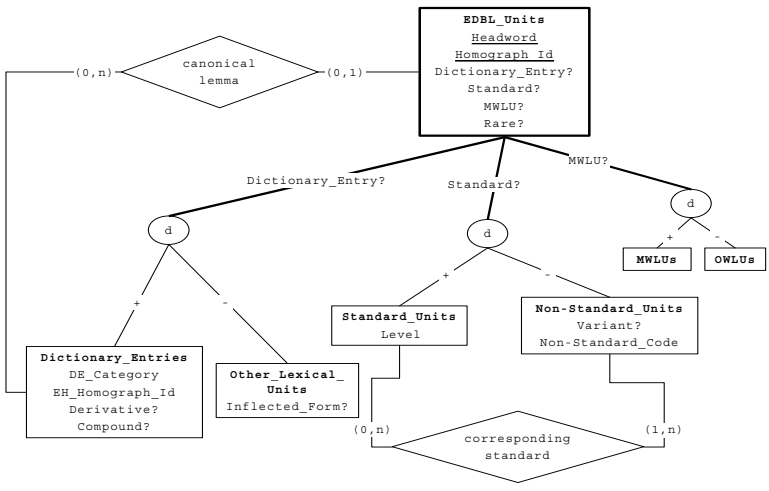

Figure 2: `EDBL_Units` and the three main specializations

3.3.1 Standard and non-standard lexical units

Basque is a language still in course of standardization; so, 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, because, in fact, a relatively large number of non-standard forms may still be found in written language.

This specialization divides all the lexical units in EDBL into standard and non-standard. The entries belonging to the `Non-Standard_Units` class can be either variant (mainly dialectal) forms, or simply non-accepted entries.

The relationship between standard and non-standard units allows us to relate the correct forms to the ones considered incorrect. Each non-standard unit must be related at least to one standard unit.

3.3.2 Dictionary entries and other lexical units

Another main specialization in EDBL is the one that separates `Dictionary_Entries` from `Other_Lexical_Units`.

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 POS. 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. Inflected forms are split up into verbal forms (auxiliary and synthetic verbs) and others (mostly irregularly inflected forms). Non-independent morphemes are affixes in general, which require to be attached to a lemma for their use inside a word form, and they are subdivided into different categories (graduators, declension morphemes, etc.).

Each class is characterized by different attributes. Nowadays these attributes are mainly of a morphosyntactic nature, although semantic features are already included in some cases.

3.3.3 One-word and multiword lexical units

The third total specialization of the main class classifies all the units in EDBL into One-Word Lexical Units (OWLUs) and Multiword Lexical Units (MWLUs). We consider an entry as OWLU if it has not any blanks in its spelling (hyphened forms and affixes included). Otherwise, it is taken as MWLU.

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). This information is encoded in the database following the Koskenniemi's (1983) two-level formalism. So, our lexical system consists currently of 80,625 OWLUs, grouped into 201 two-level sublexicons and 159 continuation classes, and a set of 24 morpho-phonological rules that describe the changes occurring between the lexical and the surface level.

On the other hand, the description of a MWLU within the lexical database includes two aspects: (1) its composition, i.e. which its 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.

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: a MWLU's component may appear in different positions in the text; and, finally, the components may either be inflected (accepting any of the allowed inflection morphemes or in a restricted way) or occur always in an invariable form. Moreover, some MWLUs are "sure" and some are ambiguous, since it cannot 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.

3.4 Linguistic contents

We will give now some figures on the linguistic contents actually stored in EDBL.

According to the classification into the three main specializations, EDBL contains: 60,940 dictionary entries and 20,939 other lexical units (20,591 inflected forms and 348 non-independent morphemes); 78,417 standard forms and 3,462 non-standard; 80,625 OWLUs and 1,254 MWLUs. Among dictionary entries there are 40,087 nouns, 9,720 adjectives, 6,533 verbs, and 3,448 adverbs, among others; non-independent morphemes include 192 declension morphemes 45 subordinating morphemes, and 37 lexical suffixes, among others.

3.5 Current status and future improvements

At the *Ixa* group, we have designed and implemented a plan to integrate the exploitation of the language processing chain, in such a way that a common data exchange XML encoding is used as an input and delivery format between the different tools. According to this format, the information in the database is exported and delivered from it as a collection of feature structures.

So, the conceptual schema of the relational database has been mapped into a hierarchy of typed feature structures (FS). The leaves of this hierarchy are 22 disjoint classes, and each one of them defines a different FS type. 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. Information exported from EDBL is currently used in every task requiring morphological and/or syntactic processing.

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 are not mainly computer scientists but linguists, stresses the reasons why we need a user-friendly, readily accessible and flexible interface. For that purpose, we designed and developed a graphic interface that gives help to the user based on context and that is accessible from the web (http://ixa2.si.ehu.es/edbl/). This GUI provides two levels of access to the database: one that lets common users only consult the data, and the second one that offers full reading and writing access, especially to the linguists in charge of the database.

4 *Euskal WordNet*: using bilingual and native dictionaries to construct a Basque wordnet

In EDBL, although homograph entries are separated, no semantic distinction between senses is made. As the group grew and the processing needs increased, semantics infrastructure became a must. As a point of departure, we decided to build *Euskal WordNet* (Agirre *et al.*, 2002), a Basque wordnet based on the English Wordnet (Fellbaum, 1998). Considering Wordnet as a *de facto* standard for the lexical-semantic representation for English, new wordnets in some other languages have been built, especially in the framework of the EuroWordNet project (EuroWN, http://www.illc.uva.nl/EuroWordNet/).

Euskal WordNet follows the EuroWN framework and, basically, it has been produced using a semi-automatic method that links Basque words to the English Wordnet (hereafter Wordnet). This section describes the current state of the Basque wordnet and the methodology we have adopted to ensure its quality in terms of coverage, correctness, completeness, and adequacy.

In order to ensure proper linguistic quality and avoid excessive English bias, a double manual pass on the automatically produced Basque synsets is desirable: a first concept-to-concept pass to ensure correctness of the words linked to the synsets, and then a word-to-word pass to ensure the completeness of the word senses linked to the words. By this method, we expected to combine quick progress (as allowed by a development based on Wordnet) with quality (as provided by a development based on a native dictionary). We have completed the concept-to-concept review of the automatically produced links for the nominal concepts, and are currently performing the word-to-word review.

4.1 Automatic generation and concept-to-concept review

In order to help the linguists in their task, we automatically generated noun concepts from machine-readable versions of Basque-English bilingual dictionaries (Morris, 1998; Aulestia & White, 1990). All English/Basque entry pairs in the dictionaries were extracted, and then were combined with Wordnet synsets; the resulting combinations were then analyzed following the class methods (Atserias *et al.* 1997). The algorithm produces triples like word - synset - confidence ratio. The confidence ratio is assigned depending on the results of the hand evaluation. The pairs produced by class methods with a confidence rate lower than 62% were discarded.

All the results of the previous process were validated by hand. The linguists reviewed the synsets that had a Basque equivalent one by one, checking whether the words were correctly assigned and adding new words to the synonym set if needed. This process led to the preliminary *Euskal WordNet 0.1* release.

4.2 Quantitative and qualitative analysis of *Euskal WordNet 0.1*

Table 1 reviews the amount of synsets, entries, etc. of the Basque wordnet compared to Wordnet 1.5 and the EuroWN final release (Vossen *et al.*, 2001). The first two rows show the number of Base Concepts, which were manually set. For nouns in the *Euskal WordNet 0.1*, the *Nouns (auto)* row shows the figures as produced by the raw automatic algorithm, and the *Nouns (man)* row the figures after the manual concept-to-concept review. The number of entries was manually reduced down to 50%, and the number of senses down to 15%. This high number of spurious entries and senses was caused primarily by a high number of orthographic and dialectal variants that were introduced by the older bilingual dictionary, which does not follow the standard current rules.

		Synsets	*No. of senses*	*Senses/ synset*	*Entries*	*Senses/ entry*
Euskal WordNet	**Nominal BC**	228	-	-	-	-
	Verbal BC	792	-	-	-	-
Euskal WordNet 0.1	**Nouns (auto)**	27641	291011	10.52	46164	6.3
	Nouns (man)	23486	41107	1.75	22166	1.8
	Verbs (man)	3240	9294	2.86	3155	2.95
Wordnet 1.5	**Nouns**	60557	107484	1.77	87642	1.23
	Verbs	11363	25768	2.27	14727	1.75
Dutch WordNet	**Nouns**	34455	54428	1.58	45972	1.18
	Verbs	9040	14151	1.57	8826	1.60
Spanish WordNet	**Nouns**	18577	41292	2.22	23216	1.78
	Verbs	2602	6795	2.61	2278	2.98
Italian WordNet	**Nouns**	30169	34552	1.15	24903	1.39
	Verbs	8796	12473	1.42	6607	1.89

Table 1: Figures for *Euskal WordNet 0.1* compared to Wordnet 1.5 and the final EuroWN release

The senses per entry figures are higher than those from Wordnet 1.5 and most of the wordnets, but similar to the Spanish WordNet. The fact that the nouns and verbs included are in general more polysemous can explain this fact. We also performed an analysis of the distribution for the variants in each synset and the number of word senses per entry.

All in all, the amount of synsets and entries for the *Euskal WordNet 0.1* is comparable to those for the wordnets produced in EuroWN, but lower than the Wordnet 1.5 release. The coverage of nominal concepts is 38% of those in Wordnet 1.5.

Somehow, we were not satisfied by the quantitative analysis and the results of the concept-to-concept review. On the one hand, the quantitative analysis only shows the state of the coverage of concepts and entries, as long as they are compared to reference figures from Wordnet (concepts) and Basque reference dictionaries (entries). It is rather difficult to assess the coverage of the number of word senses and synonyms, as these can only be compared to Wordnet, but there are no reference figures for the Basque wordnet itself. We think that the coverage of word senses and synonyms can be more reliably estimated measuring by hand the completeness of the word senses of a sample of words and the variants for a sample of concepts.

On the other hand, the concept-to-concept review only enforces the correctness and completeness of the variants in the synset. As the focus of the first stage was on quickly producing a first version, correctness was more important than completeness, and we were not completely satisfied with the completeness of the variants.

These are the correctness, completeness and adequacy requirements that were not covered by the quantitative analysis:

a) Correctness and completeness of the word senses of a word.
b) Correctness and completeness of the variants of a concept.
c) Adequacy of the specificity level for variants in a concept, i.e. all variants of a concept are of the same specificity level.
d) Adequacy of the specificity level for word senses, i.e. granularity of word senses.

In order to assess points a and d, we performed a manual comparison and mapping of the word senses given by *Euskal WordNet 0.1* with those of a monolingual dictionary and a bilingual dictionary. This assessment is presented in the next subsection.

We have also manually checked the correctness and completeness of the variants for a concept (b), using a synonym dictionary for this purpose. The results were highly satisfactory, but we decided to explicitly include the use of the synonym dictionary in all subsequent reviews and updates of the wordnet.

4.2.1 Manual mapping of word senses from the Basque wordnet and native dictionaries

The sense partition of any dictionary reflects a suitable native sense partition, and needs not to be of the same granularity as of Wordnet. In principle, both sense partitions could even be incompatible, in the sense that they could involve many-to-many mappings.

We chose to use the *Euskal Hiztegia* (EH) dictionary (Sarasola, 1996), a general purpose monolingual dictionary that covers standard Basque and that contains about 33,000 entries. One drawback of this dictionary is that it mainly focuses on literature tradition, and it lacks many entries and word senses which are more recent. For this reason, we decided to include also a bilingual Basque-English dictionary (Morris, 1998). Moreover, if the linguist thought that some other word sense was missing he/she was allowed to include it.

All in all, both bilingual and monolingual dictionaries contribute equally to the new senses. An average of 1.9 new senses are added for each word, which makes an average of 0.24 new senses for each existing sense. This makes an idea of the completeness of the word senses for words. All word senses were found to be correct. These figures can be interpolated to estimate that the coverage of word senses for the entries currently in *Euskal WordNet* is around 80%.

Regarding the mapping between the word senses of *Euskal WordNet* and the monolingual dictionary, most of the times it was one-to-one or many-to-one. The granularity of the word senses in *Euskal WordNet* is much finer. We have not found many-to-many mappings.

4.2.2 Adequacy of the specificity level of variants in synsets

As already mentioned in the quantitative analysis, we found out that some words had an unusually high number of senses. Quick hand inspection showed that for some concepts the variants were of heterogeneous specificity, and we suspected that some words were placed in too many concepts. In fact, a program that searches for words that have two word senses, one hypernym of the other, found out that there are 4,500 such pairs out of 41,107 word senses. This is a very high figure compared to Wordnet, and indicates that we need to check those word senses.

4.3 Conclusions of the quantitative and qualitative analysis and current status

We have presented here a methodology that tries to integrate the best of development methods based on the translation of Wordnet and development methods based on native dictionaries. We first have developed a quick core wordnet comparable to the final EuroWN release using semi-automatic methods that includes a concept-to-concept manual review, and later performed an additional word-to-word review based on native lexical resources that guarantees the quality of the wordnet.

As a summary of the quality assessment for the nominal part of the Basque wordnet, we can say that it contains 38% of the concepts in Wordnet 1.5, 25% of the entries (although it accounts for all the noun entries in EH), and 80% of the senses for the entries already in *Euskal WordNet*.

4.4 Word-to-word review and future work

Most of the shortcomings detected in the previous section can be overcome following an additional review

of the current *Euskal WordNet 0.1*. In this review we want to ensure that the coverage of word senses is more complete, trying to include the estimated 20% of word senses that are missing. In this case, the review is to be done studying each word in turn and taking attention to the following issues:

- Coverage of senses: add main word senses of basic words.
- Correctness of word senses of a word: delete inadequate word senses when necessary.
- Completeness of word senses of a word: add main word senses.
- Adequacy of the specificity level of word senses of a word: check that sense granularity is balanced.

The need to build a core wordnet led us to define a subset of the nominal entries to be covered: on the one hand, the top 400 words from a frequency analysis; on the other hand, the entries in a basic bilingual Basque-Spanish dictionary (Elhuyar, 1998) which defines a core vocabulary of Basque (13,000 nouns). The word senses are provided by the monolingual (EH) and the bilingual dictionaries. The bilingual dictionary includes modern words and word senses which are not in EH.

We are currently extending the coverage of the noun entries and word senses to those in a basic vocabulary of Basque. In the future we plan to apply the methodology to verbs and adjectives, and to extend the coverage to a more comprehensive set of nouns.

The current version of *Euskal WordNet* has about 25,400 entries and 52,500 senses that have been manually revised.

5 From the lexical database to a general-purpose lexical knowledge base: EEBL

A way to furnish EDBL with semantic content is to link it to other lexical resources such as machine-readable monolingual dictionaries (so providing it with definitions and related words), multilingual dictionaries (equivalents in other languages), etc.

This section describes a lexical-semantic resource under construction: EEBL, the Basque Lexical Knowledge Base, that constitutes the core of a research work currently in progress (Agirre *et al.*, 2003).

EEBL is a large store of lexical-semantic information that has been conceived as a multi-purpose and goal-independent resource for language processing tasks. It will be composed of three interlinked databases: EDBL, *Euskal WordNet*, and a dictionary knowledge base extracted from EH (see 5.1).

So, our aim here is to configure a general lexical-semantic framework for Basque language processing, linking EDBL entries with senses (definitions and examples) and related entries in the monolingual dictionary (derivatives, antonyms, hypernyms, hyponims, meronyms, etc.), synsets in *Euskal WordNet*, etc. On the other hand, this gives us the possibility to enrich the information contained both in the wordnet and in the dictionary knowledge base with the information contained in EDBL.

To start with, we decided to connect EDBL with the dictionary knowledge base and the last one with the wordnet. EDBL and *Euskal WordNet* have been already presented in this paper. In order to build a dictionary knowledge base from EH, word definitions in the dictionary have been semi-automatically analyzed to find and extract lexical-semantic relations among senses (see the next subsection). The results of such an analysis have been stored in the Concept Classification component of the EH Dictionary Knowledge Base (see fig. 3). 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.

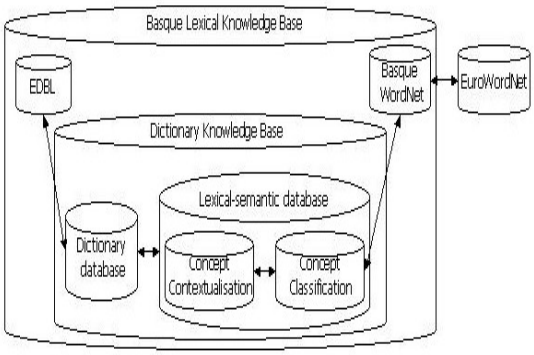

Figure 3: General architecture of the Basque Lexical Knowledge Base (EEBL)

The interrelation between EDBL and the EH Dictionary Knowledge Base allows us to manage lexical information of both grammatical and semantic nature, given that EDBL stores mainly grammatical information about words.

5.1 Exploiting a monolingual dictionary to build the EH Dictionary Knowledge Base.

The EH Dictionary Knowledge Base groups two different views of the dictionary data (see fig. 3). The Dictionary Database stores the dictionary itself in a conventional way whereas the Lexical-Semantic Database represents the lexical-semantic relations extracted from it in a semantic network-like fashion.

In a work currently in progress by Lersundi (Agirre *et al.*, 2000; Agirre & Lersundi, 2001), whose final goal is to enrich the lexical database with semantic information, EH (the monolingual dictionary) has been exploited to extract from it such kind of information. The work focuses on the extraction of the semantic relations that best characterize the headword, that is, those of synonymy, antonymy, hypernymy, and other relations marked by specific relators[1] and derivation.

All nominal, verbal and adjectival entries in EH have been parsed. Basque uses morphological inflection to mark case, and therefore semantic relations have to be inferred from suffixes rather than from prepositions. Our approach combines a morphological analyzer and

[1] We take as specific relators typical expression patterns used by the lexicographers when writing dictionary definitions. By means of these relators, some words in the definition text are linked to the headword in a special way, often determining the semantic relation that holds between them.

surface syntax parsing based on Constraint Grammar (Karlsson *et al.*, 1995), and has proven very successful for highly inflected languages such as Basque. Both the effort to write the rules and the actual processing time of the dictionary have been very low. At present we have extracted more than 40,000 relations, leaving only 9% of the definitions (mostly adjectives) without any extracted relation. The error rate is extremely low, as only 2.2% of the extracted relations are wrong.

The EH Dictionary Knowledge Base has been already supplied with the information extracted from EH. Namely, 33,102 dictionary units, 3,160 sub-entries (mainly multiword lexical units), and 45,873 senses with their corresponding relations are stored in the knowledge base.

In the future we plan to cover the semantic relations in the rest of the definition, that is, those relations involved in the part of the definition that is not the main defining pattern. For this we will use more powerful partial parsers (Aldezabal *et al.*, 1999). Besides, the coverage of derivational phenomena is also being extended, focusing specially in adjectival suffixes, in order to reduce the number of adjectives without any relation.

In order to include the extracted relations in EDBL (the lexical database), it is necessary to perform two disambiguation processes. On the one hand, there are some cases in which the surface relation extracted is ambiguous, that is, it could convey more than one deep semantic relation. On the other hand, the word senses of the words in the semantic relation have to be also determined. Anyway, some work aiming at the enrichment of EDBL based on the information extracted from EH has already been done. In particular, a method for semi-automatically assigning the animate feature to common nouns has been developed based mainly on the hyponym/hypernym relationships discovered in the dictionary (Díaz de Ilarraza *et al.*, 2002). The method obtains an accuracy of over 99% and a scope of 68,2% with regard to all the common nouns contained in a real corpus of over 1 million words, after the manual labelling of only 100 nouns. The results of this process have not yet been incorporated into EDBL.

5.2 Current status and future work

The level of integration between the lexical database and the EH Dictionary Knowledge Base can be summarized as follows: 80% of the total of entries in EH and 33% of the sub-entries have been satisfactorily linked to EDBL's entries. These links have been established automatically. In the case of derived entries, lemmatization has been used to establish links between roots, whenever it was not possible to link whole forms.

With respect to the lexical-semantic part of the knowledge base, the acquisition of relations from the dictionary is still in progress. Table 2 shows the number of relations that have been extracted from the dictionary and stored in the knowledge base so far. About 40,000 relations have been already stored. The difference between the number of extracted and stored relations is due, mainly, to the fact that some words occurring in definitions do not appear as entries. The other important reason is that some relations are duplicated because the morphological analyzer 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 analyses.

	Extracted relations	**Stored relations**	
Synonyms	19,809	16,949	85.6%
Hypernyms	20,658	18,331	88.7%
Spec. relators	5,386	4,169	77.4%
Overall	45,853	39,449	86%

Table 2: State of the DKB

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 relationships.
- Repeat the same process with a bigger monolingual dictionary (Elhuyar, 2000).
- Include relations extracted from other sources, such as corpora, as it is aimed in the MEANING project at which the group is participating (Atserias *et al.*, 2004).
- Incorporate information on named entities and classify them.

6 Research completed and in progress

In this section, we would like to outline research work on the field of the lexicon, and particularly, to present doctorate research work carried out at the group on this field. Some of these works have been already completed while others are nearly finished or just in progress. In many of them, enrichment and improvement of the knowledge contained in the lexical database, especially in that what concerns its semantic component, is one of the main goals pursued. Different approaches and methodologies have been used for that.

Research work already finished includes:

- Artola's work (Artola, 1993; Agirre *et al.*, 1997) on a small French dictionary, where sense definitions were analyzed to semi-automatically extract lexical-semantic relationships. He proposes a general framework for the representation of dictionary knowledge, which is then used in a prototype of the so-called Intelligent Dictionary Help System (*Hiztsua*), a dictionary system aimed at human users.
- Following this work, Arregi (Arregi, 1995; Agirre *et al.*, 2001b) exports the idea to a multilingual system called *Anhitz*. In this system the representation model is extended and enriched to cope with a multilingual dictionary architecture. Arregi carried out as well a comprehensive and in-depth research on the use of dictionaries in translation tasks, so expanding the functionality of the system to a great detail.
- Agirre (Agirre & Rigau, 1996; Agirre, 1998) tackles the problem of word-sense disambiguation (WSD), and proposes a method for the resolution of lexical ambiguity that relies on the use of the Wordnet taxonomy and the notion of conceptual

distance among concepts, captured by a Conceptual Density formula developed for this purpose. This fully automatic method requires no hand coding of lexical entries, hand tagging of text nor any kind of training process.

- Arriola's work (Arriola *et al.*, 1999; Arriola, 2000) is motivated by two considerations: (1) the use of existing lexical resources to contribute to the design of more complete lexical entries, and (2) the acquisition of basic subcategorization information of verbs to support NLP tasks. The examples in verbal entries of the EH monolingual dictionary are analyzed in his work using for that a Constraint Grammar parser (Karlsson *et al.*, 1995), and basic subcategorization patterns are obtained.
- Aldezabal (Aldezabal *et al.*, 2002; Aldezabal, 2004) follows the previous work in the sense that she also looks for verb subcategorization information, which is an urgent need in our lexical system if we want to be able of performing deep syntactic parsing of free texts. In her thesis, Aldezabal makes an in-depth analysis of Levin's work (1993), and tries to adapt it to the case of Basque. As a result of this work, the occurrences of 100 verbs in a corpus have been thoroughly examined, and the different syntactic/semantic patterns applicable to each of them have been encoded in a database.

Research work currently in progress includes:

- Urizar's work (Aduriz *et al.*, 1996b), which is focused on the representation and processing of Multiword Lexical Units and multiword expressions in general. He proposes a representation schema for MWLUs that, due to its expressive power, can deal not only with fixed expressions but also with morphosyntactically flexible constructions. It allows to lemmatize word combinations as a unit and yet to parse the components individually if necessary. This work must be placed in a general framework of written Basque processing tools, which currently ranges from the tokenization and segmentation of single words up to the syntactic processing of general texts, and is closely related to the work by Ezeiza (Ezeiza, 2002), who developed a parser of multiword expressions.
- Martinez (Martínez *et al.*, 2002) explores the contribution of a broad set of syntactically motivated features to WSD. This set ranges from the presence of complements and adjuncts, and the detection of subcategorization frames, up to grammatical relations instantiated with specific words. The performance of the syntactic features is measured in isolation and in combination with a basic set of local and topical features, and using two different algorithms. Additionally, the role of syntactic features in a high-precision WSD system based on the precision-coverage trade-off is also investigated in his work.
- Atutxa's thesis (Aldezabal *et al.*, 2003) deals with lexical knowledge acquisition from raw corpora. The main goal is to automatically obtain verbal subcategorization information, using for that a shallow parser and statistical filters. The arguments are classified into 48 different kinds of case markers, which makes the system fine grained if compared to equivalent systems that have been developed for other languages. This work addresses the problem of distinguishing arguments from adjuncts, being this one of the most significant sources of noise in subcategorization frame acquisition.
- Finally, an architecture for a federation of highly heterogeneous lexical information sources is proposed in a PhD work nearly finished by Soroa (Artola & Soroa, 2001). The problem of querying very different sources of lexical information lexical and dictionary databases, heterogeneously structured electronic dictionaries, or even language processing programs such as lemmatizers or POS taggers, using for that a unique and common query language, is addressed in this work from the point of view of the information integration research field. The so-called *local-as-view* paradigm is used for describing each lexical source as a view over a general conceptual model. A general conceptual model for describing lexical knowledge has been designed, as well as the way to describe each source in terms of the classes and relationships of this general model. Both the conceptual model and the sources are described and implemented using a description logic language.

7 Products and applications

A first by-product of the research work accomplished on the field is *Xuxen*, a morphological analysis based general-purpose spelling checker/corrector (Aduriz *et al.*, 1997) widely used nowadays.

Moreover, two dictionaries have been also integrated as plugins into *Microsoft Word*: a Basque-Spanish bilingual dictionary (Elhuyar, 1998) and a synonym dictionary (UZEI, 1999); in both cases on-the-fly lemmatization is performed when consulting them, allowing users a very handy lookup.

The forthcoming publication of a quite sophisticated electronic version of *Euskal Hiztegia* (Arregi *et al.*, 2003), a monolingual dictionary already mentioned several times in this paper, which has been parsed from its original RTF format and encoded into XML following the TEI guidelines, completes the panorama of end-user applications co-published by the group. This electronic version of the dictionary allows the user to search into the definitions and examples as in a fully lemmatized corpus, by posing complex queries based on lemmas and/or inflected forms, and using logical operators to construct the queries.

8 Conclusions

A language that seeks to survive in the modern information society requires language technology products. "Minority" languages have to make a great effort to face this challenge. Lesser-used language communities need, in our opinion, a long-term and well-thought strategy if they want to be able to produce language technology applications; good foundations in

terms of resources and basic tools are a must to get this goal.

At the *Ixa* NLP Group, the development of language technology has been faced from the very beginning in a bottom-up fashion, that is, laying first the infrastructure (resources and tools) in order to later be able to produce end-user applications. If anything, it is principally this conception of the strategy we have designed and developed that we could "export" to other "minority" languages as ours.

Based on our 15-year experience in NLP research, we can conclude that the combination of (semi-)automatic procedures and manual work warrants a moderately fast but reliable setting when building the lexical foundations needed in NLP. Common dictionaries constitute an obvious resource for NLP: lists of words, homographs and senses, basic grammatical information (POS, subcategorization, etc.), and, if further worked out, lots of implicit knowledge about words and their interrelationships may be extracted from them.

We have shown that work done for "bigger" languages has been sometimes very useful for our research: the use of the English Wordnet along with bilingual dictionaries has facilitated our work when building the Basque wordnet. However, if NLP research is conducted only on the "main" languages, there will be nothing we can do about the survival of our "minor" languages. Investigation on the language itself and on the application of general techniques to the processing of the language are needed as well, and, moreover, they contribute to general research in the sense that they provide a different and enriching point of view of the problems undertaken.

In the paper just our work on laying the lexical infrastructure for NLP has been presented. We are currently working as well on other areas of NLP, ranging from morphology to semantics, and tackling problems related to machine translation, computer-aided language learning, information retrieval and extraction, etc.

Basque is a minority language but we think that a substantial amount of work has already been done in the field, and that sound foundations have been established. As it has been said above, we firmly believe that the establishment of such an infrastructure is fundamental for language technologies to be developed, and our group is entirely devoted to this research since its inception in the late eighties. To finish, just to say that, apart from us, several other groups are also working on Basque automatic processing; we think that the cooperation between the different groups and sharing of the results should undoubtedly improve the development of the whole field in our country.

9 References

Publications by the Ixa Group

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

Aduriz I., Aldezabal I., Artola X., Ezeiza N., Urizar R. (1996b). Multiword Lexical Units in EUSLEM, a lemmatiser/tagger for Basque. Proc. of the 4th Conf. on Computational Lexicography and Text Research, COMPLEX'96, Linguistics Institute, Hungarian Academy of Science. Budapest (Hungary).

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 12/1, pp. 31-36. ALLC, Oxford (England).

Agirre, E., Rigau, G. (1996). Word Sense Disambiguation using Conceptual Density, Proceedings of COLING'96, pp. 16-22. Copenhagen (Denmark).

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Soroa A. (1997). Constructing an intelligent dictionary help system. Natural Language Engineering 2(3): 229-252. Cambridge Univ. Press. Cambridge (England).

Agirre, E. (1998). *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala. / Formalization Of Concept-Relatedness Using Ontologies: Conceptual Density*. PhD thesis. Dept. of Computer Languages and Systems, University of The Basque Country.

Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M., Martínez D., Sarasola K., Urizar R. (2000). Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. Proceedings of Euralex. Stuttgart (Germany).

Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sarasola K., Soroa A. (2001a). Developing Language Technology for a Minority Language: Progress and Strategy. Elsnews 10.1, pp. 4-5. ELSNET, Utrecht (The Netherlands).

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Soroa A. (2001b). MLDS: A Translator-Oriented Multilingual Dictionary System. Natural Language Engineering, 5 (4), pp. 325-353. Cambridge Univ. Press. Cambridge (England).

Agirre, E., Lersundi, M. (2001). Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In Proceedings of SEPLN 2001. Jaén (Spain).

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

Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M. (2003). A Conceptual Schema for a Basque Lexical-Semantic Framework. Proceedings of COMPLEX'03. Budapest (Hungary).

Aldezabal, I., Gojenola, K., Oronoz, M. (1999). Combining Chart-Parsing and Finite State Parsing, Proceedings of the European Summer School in Logic, Language and Information (ESSLLI) Student Session. Utrecht (The Netherlands).

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi, M. (2001). EDBL: a General Lexical Basis for The Automatic Processing of Basque. IRCS Workshop on Linguistic Databases. Philadelphia (USA).

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K. (2002). Learning Argument/Adjunct

Distinction for Basque. ACL'2002 SigLex Workshop on Unsupervised Lexical Acquisition. Philadelphia (USA).

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K. (2003). A unification-based parser for Basque and its application to the automatic analysis of verbs. *In* B. Oyharçabal ed., *Inquiries into the lexicon-syntax relations in Basque*. Supplements of ASJU no. XLVI.

Aldezabal I. (2004). *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz*. PhD thesis. Dept. of Basque Filology, University of The Basque Country.

Alegria I., Artola X., Sarasola K., Urkia M. (1996). Automatic Morphological Analysis of Basque. Literary and Linguistic Computing 11/4, pp. 193-203. ALLC, Oxford (England).

Arregi, X. (1995). *ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza*. PhD thesis. Dept. of Computer Languages and Systems, University of The Basque Country.

Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., García E., Lascurain V., Sarasola K., Soroa A., Uria L. (2003). Semiautomatic construction of the electronic *Euskal Hiztegia* Basque Dictionary. *In* Traitement automatique des langues. Les dictionnaires électroniques (ed. Michael Zock and John Carroll), 44-2, pp. 107-124. ATALA. Paris (France).

Arriola J.M., Artola X., Maritxalar A., Soroa A. (1999). A Methodology for the Analysis of Verb Usage Examples in a Context of Lexical Knowledge Acquisition from Dictionary Entries. Proc. of EACL'99. Bergen (Norway).

Arriola J.M. (2000). *Euskal Hiztegiaren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen analisia Murriztapen Gramatika baliatuz, azpikategorizazioaren bidean*. PhD thesis. Dept. of Basque Filology, University of The Basque Country.

Artola, X. (1993). *HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza. Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena. / Conception et construction d'un système intelligent d'aide dictionnariale (SIAD). Acquisition et représentation des connaissances dictionnariales, établissement de mécanismes de déduction et spécification des fonctionnalités de base*. PhD thesis. Dept. of Computer Languages and Systems, University of The Basque Country.

Artola X., Soroa A. (2001). An Architecture for a Federation of Highly Heterogeneous Lexical Information Sources. IRCS Workshop on Linguistic Databases. Philadelphia (USA).

Díaz de Ilarraza A., Mayor A., Sarasola K. (2002). Semiautomatic labelling of semantic features. Proc. of COLING'2002. Taipei (Taiwan).

Díaz de Ilarraza A., Sarasola K., Gurrutxaga A., Hernaez I., Lopez de Gereñu N. (2003) HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities. Workshop on NLP of Minority Languages and Small Languages (TALN). Nantes (France).

Ezeiza, N. (2002) *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua*. PhD thesis. Dept. of Computer Languages and Systems, University of The Basque Country.

Martínez D., Agirre E., Márquez L. (2002). Syntactic features for high precision Word Sense Disambiguation. Proc. of the 19th International Conf. on Computational Linguistics (COLING 2002). Taipei (Taiwan).

Other references

Atserias J., Climent S., Farreras J., Rigau G., Rodríguez, H. (1997). Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In Proceedings of Conference on Recent Advances on NLP. (RANLP'97). Tzigov Chark (Bulgaria).

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., Vossen P. (2004). The MEANING Multilingual Central Repository. Proc. of the 2nd Global WordNet Conference. Brno (Czech Republic).

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

Karlsson F., Voutilainen A., Heikkila J., Anttila A. eds. (1995). *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Berlin and New York: Mouton de Gruyter.

Koskenniemi K. (1983). *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki (Finland).

Levin B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation.* Chicago and London. The University of Chicago Press.

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

Dictionaries

Euskaltzaindia (2000). *Hiztegi Batua*. Euskera 5, Bilbo. http://www.euskaltzaindia.net/hiztegibatua/

Elhuyar (1998). *Elhuyar Hiztegi Txikia.* Elhuyar Kultur Elkartea, Usurbil.

Morris, M. (1998). *Morris Hiztegia*. Klaudio Harluxet Fundazioa, Donostia. http://www1.euskadi.net/morris/indice_e.htm

UZEI (1999). *Sinonimoen Hiztegia.* UZEI, Donostia. http://www.uzei.org/nagusia.cfm?hizkuntza=0&orria=online&atala=sinonimoak

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

Aulestia, G., White, L. (1990). *English-Basque Dictionary*. University of Nevada Press, Reno (USA).

Elhuyar (2000). *Hiztegi Modernoa.* Elhuyar Kultur Elkartea, Usurbil.