Lexical Information Extraction for Basque

Résumé

L'objectif de ce projet est l'utilisation de ressources lexicales existantes afin de faire plus complètes et détaillées les entrées lexicales de la Base de Données Lexicales du Basque (EDBL). Dans cet article on décrit la première étape dans le processus d'extraction de l'information à partir du dictionnaire Hauta-Lanerako Euskal Hiztegia (HLEH). La deuxième section est consacrée à étudier l'applicabilité de l'information du HLEH afin de l'utiliser pour enrichir EDBL. Dans la troisième on présente les caractéristiques principales et la structure générale du dictionnaire. Dans la section 4 on décrit la conversion du HLEH (version MRD) en une structure étiquetée: la "tokenization" du texte an moyeu d'un "tagger", la grammaire des entrées et la conversion de la structure étiquetée qui est encodée selon les recommandations du TEI. Finalement, on présente les résultats obtenus et les problèmes trouvés dans le processus d'analyse.

Abstract

The aim of this project is to use existing lexical resources in order to contribute to the design of more complete and detailed lexical entries for the Lexical Database for Basque (EDBL). This paper presents the first step in this process of information extraction from the dictionary Hauta-Lanerako Euskal Hiztegia (HLEH). The second section is devoted to examining the applicability of the information contained in HLEH in order to enrich EDBL. The third section presents the main characteristics and the general structure of the dictionary. Section 4 deals with the conversion of HLEH (MRD version) into a labelled structure: the tokenization of the text by means of a tagger, the grammar of entries and the conversion of the labelled structure which is encoded following the TEI guidelines. Finally, we present the results and the problems encountered in the analysis.

Lexical Information Extraction for Basque

* Arriola, J.M., Soroa, A.
Department of Computer Languages and Systems
(University of the Basque Country)
Computer Science Faculty, Apdo. 649 20080 Donostia
E-mail: jibaregj@si.ehu.es

1. Introduction

It is widely recognised that machine readable dictionaries are important sources of lexical data which, suitably analysed, extracted and formalised, could be used effectively in the construction of lexical components for natural language processing (NLP) applications (Castellón 93). 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 presently, and in the near future are as follows: morphological analysis, spell checking and correction, (semi-)automatic lemmatisation and tagging, syntactic analysis and analysis of textual corpora. Taking into account the large amount of information needed into EDBL, we take the Machine Readable Dictionary (MRD) of HLEH (Sarasola 84/95), to exploit the information contained in it and finally to integrate it into EDBL, which is the general basis for the multiple lexical needs that current and further work on the automatic treatment of Basque will have (Agirre et al., 95).

2. The applicability of the information contained in the HLEH in order to enrich EDBL

We consider the HLEH dictionary as an adequate source because it is a general purpose dictionary, monolingual and covers standard Basque. It also includes non-standard forms (marked by the symbol +, indicating with the *Ik.* code shows its corresponding standard form) and variants of the standard form entries. Variants and non-standard forms are important because we are involved in the automatic processing of written Basque which is yet to be normalised. These forms are stored in EDBL in a section devoted to what we have called typical errors and are already used in our recently commercialised spelling checker/corrector Xuxen (Aduriz et al., 93). Two other sources of information are the parts of speech and especially important the subcategorization tags that appear in the verbal entries. These subcategorization tags indicate the type of auxiliary demanded by the verb. All these types of information are explicit and can be acquired directly. But there is also implicit information, e.g. in definitions and examples, which is very interesting for EDBL. From the analysis of the sentences of the definitions in this dictionary, certain lexical-semantic information will be integrated into EDBL: taxonomic relationships, synonymy, meronymic

---

* Was supported by a grant from the Basque Government.

relations, and so on. The examples, that have been extracted from a very large corpus, can be considered as a specialised corpus and will be a good source to extract basic syntactic information. The integration of all this information will become a good testing bench for the design of the database itself.

There is also information that can be useful for lexical semantic disambiguation:

•In the relation field (synonym or antonym), a sense number is attached to the related entry, indicating the sense that is related with.

•In the definition texts, sometimes the genus is followed by a sense number which gives information about the sense of the genus (explicit lexical disambiguation).

There are also other types of information which could be of use to the EDBL, but this information is difficult to extract due to the way it has been specified. For instance, information about verb arguments is specified in an implicit way (i.e. the verb *adiskidetu* (to be reconciled)) indicates that this argument must be a person, but there is not explicit code that indicates it.

3. The structure of the information in HLEH

In general, the content of the HLEH dictionary is: headword; date; variants; part of speech; abbreviations (register labels, field labels, etc.); definition; relations; scientific names; examples; subentries and grammatical information.

Lexicographers who have worked on the HLEH explain the main codes used in the dictionary as follows:

•Senses: the criterion to ordenate the different senses is chronological. The codes are:

- •**A**, **B**, ... for grouping the senses which are analogous.
- •**1**, **2**, ... sense numbers.
- •a), b), ... or #1,#2, ... for subsenses.

•Relations:

- •*Ik.* for synonym relations.
- •*Ant.* for antonym relations.

•Some typographic conventions:

- •Bold is used to indicate: headword, homograph, sense number, subsenses, relations.
- •Italics are used to indicate: part of speech, different types of abbreviations, examples, participles of nominal verbs.

All these aspects were taken into account during the design of the dictionary grammar.

4. Dictionary grammar and the generation of a labelled structure

The MRD version is intended for human rather than machine interpretation. The lexicographer has used a text processor (Word Perfect, Word) to configure the structure of the entries, hence we have to deal with a text file in which the only available codes are of typographic and lexicographic nature. In order to generate a labelled structure the following two main tasks have been developed:

4.1. Tokenization

The objective of this process is the segmentation of the text in order to identify the different entries and determine their parts. Tokenization involves the identification of entry-internal field delimiters, their interpretation in context and mark-up of individual components by "begin-end" brackets. Entry components are bracketed by [Btag] and [Etag] to mark their beginning and end respectively.

A. The tags corresponding to lexicographic codes:

-BD/ED: beginning and ending of the date.
-BS/ES: beginning and ending of the sense number.
-BSS/ESS: beginning and ending of the subsense code (#1, #2, etc.).
-BSG/ESG: beginning and ending of the general sense group code.
-BSY/ESY: beginning and ending of the synonym relation code.
-BA/EA: beginning and ending of the antonym relation code.
-BH/EH: beginning and ending of the homograph identifier.

B. The tags corresponding to typographic codes.

The only codes available correspond to two style types: bold and italics.

-BB/EB: beginning and ending of bold.

Bold is used to indicate entries, the homograph identifier, and in lexicographic codes (sense number, subsenses, relations, etc.). No difficulties exist in distinguishing the different elements that appear in bold, because each lexicographic code belongs to a fixed and closed group.

-BI/EI: beginning and ending of italics.

Italics are used to indicate five different elements: part of speech, use abbreviations, thematic abbreviations, examples and participles of nominal verbs which have their own entry. The distinction between parts of speech, abbreviations and the rest is easy, because the parts of speech and abbreviations belong to a closed group. The distinction between examples and participles can be made taking into account the position in which the participles appear. The participles are always followed by a nominal verb.

Other tags:

-GUR: identifier of a non-standard entry.

-BC/EC: beginning and ending of capital letters correspond to subentries. Most of them are compounds or fixed combinations (Noun + Verb, Noun + adjective, etc.) and present the structure of an entry.

-BE/EE: beginning and ending of each entry using the paragraph delimiters.

The rest of the text will be tagged by default with the tag TX, which correspond to normal text. The definitions are always in normal style.

There are some common problems in establishing a workable field structure and in determining the boundary markers for those fields:

•The appearance of codes that had not been foreseen.

•The use of codes with several functions. For instance, the subsense codes (#1,#2, etc.) are used for distinct functions: they indicate the part of speech that changes when it is due to a sense change, or they can appear preceding the sense number when the new sense is due to a thematic change.

Another lexicographic code with several functions is *Ik.*, that is used to cross-relate different entries by means of their semantic relations or to indicate the corresponding standard form of a non-standard one.

In these cases we prefer not to increase the number of tags, and solve these problems in the grammar.

Let us see an example of the tokenization process:

- MRD entry:

lotsatu, lotsa, lotsatzen.[1] da-du *ad.* (1545). **1.** *Ipar.* Beldurtu. *Ene hezurrak eta ene zainak lotsatu dira, ikaratu dira ilunbean. Hotzak, beroak, deusek ez du lotsatuko.* **2.** (*1745, 1761). Lotsa sentitu edo sentierazi. Ik. **ahalketu.** *Bekatu hura aitortzeaz lotsatzen bada. Beren hizkuntzaz maiz lotsatzen direlako. Gauza apaletara makurtzeak lotsatzen ez gintuelako. Biluzik zeudelako lotsatu gabe. Ez zekiten zer erantzun eta lotsaturik gelditu ziren. Lotsaturik nago zu nere etxean ikusita. Lotsatuta utzi zuen. Lotsatzeko gauza.*

- Tagged entry:

[BE][BB]lotsatu, lotsa, lotsatzen. [EB][TX]da-du [BI]ad. [EI][BD](1545). [ED][BB]1. [EB][BI]Ipar. [EI] [TX]Beldurtu. [BI]Ene hezurrak eta ene zainak lotsatu dira, ikaratu dira ilunbean. Hotzak, beroak, deusek ez du lotsatuko. [EI][BB] 2. [EB][BD](*1745, 1761).[ED] [TX]Lotsa sentitu edo sentierazi. [BSY]Ik. ahalketu. [ESY][BI]Bekatu hura aitortzeaz lotsatzen bada. Beren hizkuntzaz maiz lotsatzen direlako. Gauza apaletara makurtzeak lotsatzen ez gintuelako. Biluzik zeudelako lotsatu gabe. Ez zekiten zer erantzun eta lotsaturik gelditu ziren. Lotsaturik nago zu nere etxean ikusita. Lotsatuta utzi zuen. Lotsatzeko gauza. [EI][EE]

4.2. The grammar of entries

The grammar of entries has been written as a Definite Clause Grammar (DCG) in Prolog. In order to illustrate the general structure of entries we will use an auxiliary metalanguage:

- <A>/<B> A or B , alternative nodes.
- <A> <B> A and B, A followed by B.
- [<A>] optional node.
- <*NULL> empty node.
- <*A> terminal node.

The syntax of the rules is:

<ELEMENT>=<EL1> <EL2>...<ELN>.

Here we have a little grammar to describe the general structure of the entries:

<ENTRY> =<HEADWORD> [<RELATIONS>] <CATEGORY>
    [<*DATE>] [<DEFINITION_EXAMPLES>] .
<HEADWORD> = [<HOMOGRAPH>][<NON-STANDARD_HEADWORD> /
    <STANDARD_HEADWORD>].
<HOMOGRAPH> = <*BH> <*NUMBER> <*EH>.
<NON-STANDARD_HEADWORD> = <*GUR> <*BB> <*HEADWORD> <*EB>.
<STANDARD_HEADWORD>= <*BB> <*HEADWORD> <*EB>.
<CATEGORY> = [<*SUBCATEGORY>] <CATEGORY>.
<CATEGORY> = <*BI> <*CAT> <*EI>.
<DEFINITION_EXAMPLES> = <DEFINITION> [<EXAMPLES>]
    <DEFINITION_EXAMPLES> / <*NULL>.

---

[1] to be ashamed

```
<DEFINITION>=[<SENSE_NUMBER>][<SENSE_GROUP>]<*DEFINITION>
                    [RELATIONS] .
<SENSE_NUMBER> = <*BS> <*NUMBER> <*ES>.
<SENSE_GROUP> = <*BG> <*SENSE_GROUP> <*EG>.
<RELATIONS> = [<SYNONYM_RELATION> / <ANTONYM_RELATION>]
                    <RELATIONS> [EXAMPLES]./<*NULL>.
<SYNONYM_RELATION> = <*BSY> <*SYNONYMS> <*ESY>.
<ANTONYM_RELATION> = <*BA> <*ANTONYMS> <*EA>.
<EXAMPLES> = <*BI> <*EXAMPLES> <*EI>.
```

5. The analysis

As a result of the analysis we get a labelled structure that represents the information of an entry from the MRD in a Lisp Structure. For example:

```
(entry
(entry(headword(lotsatu, lotsa, lotsatzen. ))
        (parts of speech(subcategory(da-du ))(category(ad.  )))
        (dates(date((1545). )))
        (definition_examples(sense_number(1. ))
              (def_adib1(abbreviations(Ipar. ))
               (definition( Beldurtu. ))
               (examples(Ene hezurrak eta ene zainak lotsatu dira, ikaratu dira ilunbean.
               ))
              (def_adib2(examples(Hotzak, beroak, deusek ez du lotsatuko. ))))
              (definition_examples(sense_number( 2. ))
              (dates(date((*1745, 1761).)))
              (def_adib1(definition( Lotsa sentitu edo sentierazi. ))
                (def_adib2(relations(synonym_relation(Ik. ahalketu. )))
                (def_adib3(examples(Bekatu hura aitortzeaz lotsatzen bada. Beren
                hizkuntzaz maiz lotsatzen direlako. Gauza apaletara makurtzeak lotsatzen
                ez gintuelako. Biluzik zeudelako lotsatu gabe. Ez zekiten zer erantzun eta
                lotsaturik gelditu ziren. Lotsaturik nago zu nere etxean ikusita. Lotsatuta
                utzi zuen. Lotsatzeko gauza. ))))))))))
```

6. HLEH encoded following TEI guidelines

We have adopted the Text Encoding Initiative (TEI) (Sperberg-McQueen et al., 1990) guidelines for encoding the dictionary taking as starting point the labelled structure shown in previous section. These guidelines comprise the only comprehensive attempt to provide a standard means to encode machine readable texts. This labelled structure is encoded automatically following the TEI guidelines.

For example:

```
<entry>
  <form><orth>ahalketu, ahalke, ahalketzen.</orth></form>
  <GramGrp><subc>da-du </subc><pos>ad. </pos></GramGrp>
  <notes>(1571)</notes>
  <usg type=geo>. Ipar. edo jas. </usg>
  <sense n = 'A.'>
    <GramGrp><subc> da </subc><pos>ad.</pos></GramGrp>
    <def>Lotsatu. </def>
     <eg>Ahalke zaitez, zerbitzari laxoa. Erori orduko ahalketu ziren, eta estali zituzten
     beren gorputzak piko-hostoz. </eg>
    <sense n = '#1'>
     <eg> Zerbaitez edo norbaitez ahalketu. Ahalketzen baita erokeria haiek ikusteaz eta
     entzuteaz.</eg>
    </sense>
  </sense>
  <sense n = 'B.'>
    <GramGrp><subc> du </subc><pos>ad.</pos></GramGrp>
    <xr type = syn>Ik. ahalkerazi.</xr>
    <sense n = '1. '>
     <notes>(1571).</notes>
     <def> Lotsatu. </def>
     <eg>Neure bihotzeko hasperenek ahalketzen ninduten.</eg>
    </sense>
    <sense n = '2. '>
     <notes>(XIX b. ).</notes>
     <def> Beldur erazi. </def>
     <eg>Mamu horrek txoriak ahalketzen ditu.</eg>
    </sense>
  </sense>
</entry>
```

7. Results of the analysis

We have analysed the whole dictionary (23.157 entries) recognising 85% of the entries. Out of the 85% that have been recognised, 90% have been analysed fully and correctly. It is important to point out that we have examined the reasons of this error level and 95% of them are caused by typographic errors or inconsistencies in the source dictionary. There are some entries that have not been correctly analysed due to their special nature. These entries contain very detailed grammatical information and constitute a particular type of grammar concerned with its use, syntax function, etc. of the entry being defined. An added difficulty is that the order of the elements is quite free, moreover if fine-grained analysis of the dictionary is made the analysis process becomes much more complex. From a practical point of view, the next task will be to build an environment so as to analyse semi-automatically the rest of the entries.

8. Conclusion

In this paper the first step to acquire lexical information from an MRD has been described. The source dictionary is oriented towards a human reader and much of the information is not well-structured. This is the reason why the full-automatic lexical acquisition is not possible, as we have already seen in the results obtained, and therefore it must be combined with manual or semi-automatic lexicographic labour. Nevertheless, the translation of the dictionary from MRD format to something more usable is an essential task in order for future research to be carried out. For example, analysis of definitions, examples, etc.

In addition, the ambiguities and inconsistencies that have been found in the dictionary reveal the need of a more systematic and consistent criteria for constructing dictionaries. From this perspective, there are some very helpful tools that aid the lexicographers in the systematisation and automatization of their work. Lexical data-bases as basis of dictionary construction are one of the ways to improve dictionary construction.

References

Agirre E., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Insausti J.M., Sarasola K. (1995). "Different issues in the design of a general-purpose Lexical Database for Basque" In Proc. of the First Workshop on Applications of Natural Language to Databases.(NLDB´95). Versailles, pp. 299-313.

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Da Costa A., Diaz de Illarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M. (1993). "A Morphological Analysis Based Method for Spelling Correction" In Proc. of the 6th Conference of the EACL, pp.463.

Castellón, I. (1993). Lexicografía Computacional: Adquisición automática de conocimiento léxico. Ph.D. Thesis.Univ. of Barcelona.

Sarasola, I. (1984/1995). Hauta-Lanerako Euskal Hiztegia. G.K. Donostia.

Sperberg-McQueen, C. M. and Burnard, L. (eds) (1990). "Guidelines for the Encoding and Interchange of Machine-Readable Texts". Document No. TEI P1. Draft: Version 1.1. Text Encoding Initiative, Chicago and Oxford.