A Methodology for the Analysis of Verb Usage Examples in a Context of Lexical Knowledge Acquisition from Dictionary Entries

Jose Mari Arriola, Xabier Artola, Aitor Maritxalar, Aitor Soroa

Department of Computer Languages and Systems

University of the Basque Country

Computer Science Faculty, P.O. Box. 649, E-20080 Donostia

jibaregj@si.ehu.es

Abstract

This paper presents the development of a methodology the last goal of which would be to help the linguist or lexicographer when defining the subcategorisation pattern of a verb. As a starting point we have applied this methodology over the verb entries examples of an ordinary dictionary in machine readable version. The methodology consists of the following steps: morphological analysis, resolution of morphological ambiguities, syntactic function assignment, detection of verb chains and phrasal units, and the design and implementation of an SGML data structure that represents the verb entries examples, and a query system to exploit the information contained in this data structure.

1 Motivation: Why Analyse Verb Examples?

The investigation reported in this article was motivated by two considerations: (1) the use of existing lexical resources in order to contribute to the design of more complete lexical entries for the Lexical Database for Basque (Agirre et al. 1995); and (2) the acquisition of a basic subcategorisation information of verbs to support our parsing tools.

The practical goal of our work is to enrich the information in verb entries information with their corresponding basic subcategorisation patterns. In that sense we think that our effort could be useful to increase the lexicographer's productivity and to help solving the problem of identifying predicate-argument structures of verbs.

It is widely recognised that verb subcategorisation represents one of the most important elements of grammatical/lexical knowledge for efficient and reliable parsing. Researchers in NLP have increasingly felt the need to construct computational lexicons dynamically from text corpora, rather than relying on existing 'static' lexical databases (Pustejovsky and Boguraev, 1994). Because of the lack of accurate verb subcategorisation information causing half of the parse failures (Briscoe and Carroll, 1993), attempts have been made to construct, from empirical data, lexicons that encode information about predicate subcategorisation that capture the valences of the verb and its structural collocations (cf. Brent, 1991; Manning, 1993; Briscoe and Carroll, 1997).

In our project we extract information from a machine readable dictionary (MRD) as a starting point to guide the lexical acquisition from corpora. We think that dictionaries and corpora should be combined in the acquisition of this kind of information.

2 Previous Work: from the MRD to a LDB

We considered the Euskal Hiztegia (EH) dictionary (Sarasola, 1996) an adequate source because it is a general purpose monolingual dictionary, and it covers standard Basque. The content of the EH dictionary is: headword; date; variants; part of speech; abbreviations (style and usage labels, field labels, etc.); definition; relations; scientific names; examples; subentries and grammatical information. It contains 33,111 entries and 41, 699 senses.

The previous work dealt with the conversion of EH (MRD version) into a labelled structure (for more details, see Arriola & Soroa, 1996). The MRD version was intended for human rather than machine interpretation. The lexicographer used a text-processor (Word Perfect, Word) to configure the structure of the entries, so we had to face a text file in which the only available codes were of typographic and lexicographic nature. In order to generate a structured representation of the information contained in the MRD the following three main tasks were developed: the tokenization of the text by means of a tagger, the grammar of entries that covered the general structure of the dictionary (it was written as a Definite Clause Grammar (DCG) in Prolog) and the conversion of the labelled structure which was encoded automatically following the Text Encoding Initiative (TEI) guidelines (Sperberg-McQueen et al. 1994).

As a result of this conversion process we recognised 98,49% of the entries with all the information contained in them, being the error rate of 3%. There were some errors referred to the date or some grammatical codes, but the part of speech, definition, examples and so on were correctly recognised.

3 Corpus of Verb Usage Examples

The corpus of verb examples that we have been able to analyse in the previous work is composed of 13,089 examples. These examples were extracted by the lexicographer when writing the dictionary from a very large corpus in order to show the usage of the verbs. So we can consider it a specialised corpus.

The average of words per example is 6,44. This implies that sentences are not too complex and we expect to get good results in the subcategorisation extraction process. However, sometimes the examples are just fragments of a sentence, they are not complete sentences, and in these cases we can not get a significative subcategorisation pattern.

4 A Methodology for the Analysis of Verb Usage Examples

In this section we describe the steps followed for the analysis of verb usage examples. The main basis in the analysis of the examples are the morphological analyser and the disambiguation grammar.

4.1 Morphological Analysis of Example Sentences

The two-level morphological analyser (Alegria et al. 1996) attaches to each input word-form all possible interpretations and its associated information. The result is the set of possible analyses of a word, where each morpheme is associated with its corresponding features in the lexicon: category, subcategory, declension case, number and definiteness, as well as the lexical level syntactic functions and some semantic features. The full output of the morphological analysis constitutes the input for the processes of context-based morphological disambiguation and syntactic function assignment.

4.2 Morphological Disambiguation and Assignment of Syntactic Functions

We chose the Constraint Grammar (CG) formalism (Karlsson et al. 1995) to disambiguate and analyse the examples syntactically. The Basque Constraint Grammar that currently contains 1,100 rules works on a text where all the possible interpretations have been assigned to each word-form by the morphological analyser. The rules are applied by means of the CG-2 rule compiler developed and licensed by Pasi Tapanainen (1996). On the basis of eliminative linguistic rules or constraints, contextually illegitimate alternative analyses are discarded. As a result we get almost fully disambiguated sentences, with one interpretation per word-form and one syntactic label. But there are word-forms that are still ambiguous morphologically and syntactically. At this point we are aware that there can also be analysis errors and, consequently, due to the remaining ambiguity and the errors, the results of the extraction process must be manually checked

```
/<lemma      ausiki, ausikitzen>/
/<Category    verb. >/
/<Type _of _Auxiliary  DU>/
/<Example>/
"<$.>"
   PUNT-PUNT
"<Basurdeek>"
   "basurde" NOUN COMMON ERG PL DEFINITE @SUBJ
"<ausikiko>"
   "ausiki" V SIMPLE PART PERFECTIVE DU @-FMAINVERB
   "ausiki" V SIMPLE PART S DEFINITE GEL ABS UNDEFINITE DU @<NCOMP @NCOMP> @ADVERBIAL @OBJ @SUBJ
@PRED
   "ausiki" V SIMPLE PART DEFINITE GEL S DEFINITE DU @<NCOMP @NCOMP> @ADVERBIAL
   "ausiki" NOUN COMMON S DEFINITE GEL ABS UNDEFINITE IWLP @<NCOMP @NCOMP>
   "ausiki" NOUN COMMON S DEFINITE GEL IWLP @<NCOMP @NCOMP>
"<gaituzte>"
   "*edun" AUXV PRESENT_OF_INDICATIVE  TRANSITIVE  1stPER_PL  3rdPER_PL  @+FAUXVERB
   "*edun" SYNTHETICV PRESENT_OF_INDICATIVE  TRANSITIVE  1stPER_PL  3rdPER_PL  @+FMAINVERB
"<gutxien>"
   "gutxi" ADJ GEN PL DEFINITE ABS UNDEFINITE @<NCOMP @NCOMP> @OBJ @SUBJ @PRED
   "gutxi" ADJ GEN PL DEFINITE GEN DEFINITE @<NCOMP @NCOMP>
   "gutxi" ADJ SUPERLATIVE ABS UNDEFINITE @OBJ @SUBJ @PRED
   "gutxi" ADJ SUPERLATIVE
```

"gutxi" DET ABS UNDEFINITE @OBJ @SUBJ @PRED
"gutxi" DET INDEFINITE
"<ustean>"
"uste" NOUN COMMON S DEFINITE INESIVE @ADVERBIAL
"<$.>"

Example 1. Example before the analysis process. "Basurdeek ausikiko gaituzte gutxien ustean" (English gloss: The wild boars will bite us when we least expect it.)

In the disambiguation process performed by the grammar, apart from the information of the output of the morphological analyser we use the information contained in the dictionary itself to help in the disambiguation process. We add in the morphological reading of the verb entries the tag corresponding to the type of auxiliary [1] that appears in the dictionary. This tag is useful to discard some interpretations that do not have concordances with the type of auxiliary.

Apart from that, a new tag is added for us as a result of the assumption that those readings of the verb under study which do not have the verb category in their interpretation have less probabilities to occur in an example: the tag IWLP (interpretation with less probabilities). This tag is only used for the disambiguation grammar in the case we have not enough linguistic information to discard this interpretation. In the example 1 we can see a verb entry example in which we have added the above mentioned tags to the verb entry interpretation before the analysis process.

4.3 Analysis of Verb Chains and Phrasal Units

At this stage we have the corpus syntactically analysed following the CG syntax which stamps each word in the input sentence with a surface syntactic tag. In this syntactic representation there are not phrase units. But on the basis of this representation, the identification of various kinds of phrase units such as verb chains and noun phrases is reasonably straightforward. For that purpose we base on the syntactic function tags designed for Basque (Aduriz et al. 1997). We can divide these tags into three types: main function syntactic tags, modifier function syntactic tags and verb function tags. The last ones are used to detect verb chains. This distinction of the syntactic functions is essential for the subgrammars which has been developed apart from the general grammar. These subgrammars are CG-style grammars which contain mapping rules.

[1] The verb in Basque is split up into two components: the main verb and the auxiliary. The lexical meaning and aspectual information is encoded in the main verb, while tense and mood are encoded in the auxiliary. Moreover, the auxiliary can exhibit up to three agreement morphemes corresponding to the absolutive, dative and ergative cases.

4.3.1 Subgrammar for Verb Chains

We use the verb function tags like as for example: @+FAUXVERB, @-FAUXVERB, @-FMAINVERB, @+FMAINVERB , etc.; and some particles: the negation particle and the modal particles, in order to detect verb chains. Based on these elements we are able to make explicit the continuous verb chains as well as those which are not continuous. The tags attached to mark-up the continuous verb chains are the following:

- %VCH: this tag is attached to a verb chain composed only by one element.
- %VCHI: this tag is attached to words with verb syntactic function tags that are linked to other words with verb syntactic function tags and constitute the initial element of a complex verb chain.
- %VCHE: this tag is attached to words with verb syntactic function tags that are linked to other words with verb syntactic function tags and constitute the final element of a complex verb chain.

The tags used in detecting the non continuous verb chains are:

- %NOVCHI: this tag is attached to the initial element of a non continuous verb chain.
- %NOVCHC: this tag is attached to the second element of a non continuous verb chain.
- %NOVCHE: this tag is attached to the final element of a non continuous verb chain.

As we can see in the next example (Example 2.) the maximum length of a non continuous verb chain is of three elements.

4.3.2 Subgrammar for Noun Phrases and Prepositional Phrases

Our assumption is that any word having a modifier function tag is linked to some word with a main syntactic function tag. And a word with a main syntactic function tag can by itself constitute a phrase unit. Taking into account this assumption we establish three tags to detect this kind of phrase units:

- %PHR: noun phrases or prepositional phrases; this tag is attached to words with main syntactic function tags that constitute a phrase unit by themselves.
- %PHRI: this tag is attached to words with main syntactic function tags that are linked

to other words with modifier syntactic function tags and constitute the initial element of a phrase unit.

- %PHRE: this tag is attached to words with main syntactic function tags that are linked to other words with modifier syntactic function tags and constitute the end of a phrase unit.

The aim of this subgrammar is to attach to each word-form one of those three tags in order to delimit the noun phrases and prepositional phrases. They make explicit the linking relations expressed by the syntactic functions and facilitate the phrase units recognition. Here are some examples of the analyses get after applying the above mentioned subgrammars:

```
"<$.>"
  PUNT-PUNT
"<Euriak>"
   "euri" NOUN COMMON ERG S DEFINITE @SUBJ  %PHR
"<ez>"
   "ez" PARTICLE  CERTAINTY  @PRT  %NOVCHI
"<du>"
   "*edun" AUXV PRESENT_OF_INDICATIVE  TRANSITIVE  3rdPER_S  3rdPER_S @+FAUXVERB %NOVCHC
"<ia>"
   "ia" ADVERB COMMON @ADVERBIAL  %PHR
"<kalea>"
   "kale" NOUN COMMON ABS S DEFINITE @OBJ  %PHR
"<busti>"
   "busti" V SIMPLE PART PERFECTIVE DU @-FMAINVERB  %NOVCHE
"<$.>"
```

Example 2. A non continuous verb chain and the correspondig syntagmatic units detected."Euriak ez du ia kalea busti" (English gloss: The rain has scarcely wetted the street.)

```
"<$.>"
   PUNT-PUNT
"<Harria>"
   "harria" NOUN COMMON ABS S DEFINITE @OBJ  %PHR
"<zortzi>"
   "zortzi" DET PL ABS @NOUN_DETERMINER>  %PHRI
"<aldiz>"
   "aldiz" NOUN COMMON INS UNDEFINITE @ADVERBIAL %PHRE
   "aldiz" LOT LOK @LOK
"<jaso>"
   "jaso" V SIMPLE PART PERFECTIVE DU @-FMAINVERB  %VCHI
"<du>"
   "*edun" AUXV PRESENT_OF_INDICATIVE  TRANSITIVE  3rdPER_S  3rdPER_S  @+FAUXVERB  %VCHE
"<minutu>"
   "minutu" NOUN COMMON @CASE_MARKER_MODIFIER>  %PHRI
"<batean>"
   "bat" DET INE S DEFINITE @ADVERBIAL  %PHRE
"<$.>"
```

Example 3. A continuous verb chain and the corresponding syntagmatic units detected. "Harria zortzi aldiz jaso du minutu batean" (English gloss: He picked the stone up eight times within a minute.)

4.4 Evaluation of the Analysis: Results

The results of the analysis are referred to the above mentioned subgrammars applied to the output of the disambiguation grammar. These figures are taken from the sentences in the corpus of verb examples (13,089) sentences. A sample of 300 sentences (extracted at random) has been manually checked and the proportion of sentences that have the phrase units and verb chains tags correctly assigned is 75%. We consider this result satisfactory as a first approach, even more if we take into account that the disambiguation work is still in progress and also that, as we have pointed out, the examples in some cases are not complete sentences.

Additionally, we have studied some samples to determine the different possible sources of error. As a result of this study we know that most errors are due to the disambiguation process.

4.5 An SGML Data Structure for the Exploitation of the Results

As a result of the steps described in the previous points, the corpus of verb examples contains very rich information. In order to exploit this information we design an SGML data structure in which we recover the verb usage examples classified by sense code and the type of auxiliary tag that appears in the MRD. We organise verb examples taking into account the sense code and the tag corresponding to the auxiliary type since we think it is interesting to study the impact of these factors in the argument structure. Figure 1 shows how the examples are organised.

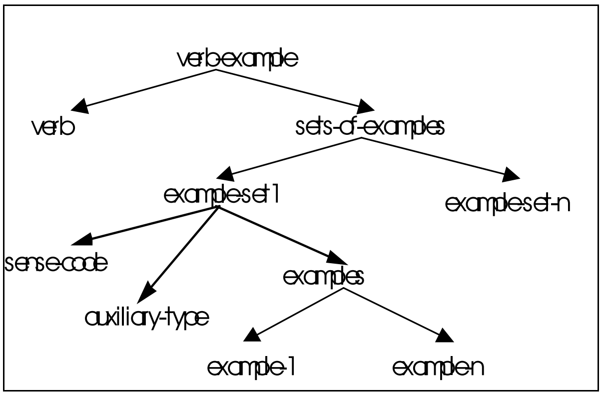

Figure1- Outline of the organisation of examples.

We adopt the SGML mark-up language format for all the corpus of verb examples. From this corpus we extract some pieces of information that we consider more important for verb argument extraction. We choose the verb entry that is object of study with the following information:

- the sense code and the type of auxiliary tag that appear on the MRD.
- the set of examples and the different phrase units that have been detected by means of the above described subgrammars.
- for the verb chains that have been detected, we distinguish between the verb chains that correspond to the verb entry and the other verb chains that can be associated or not with this verb entry. Anyway, for both kinds of verb chains the following information is offered: verb chain, type of auxiliary, syntactic function, person, aspect, modality, mood and time, and the subordinate relation.
- for phrase units we get these kind of information: the phrase unit chain, syntactic function, case, number, definiteness, and subcategorization in the case of nouns. This information is extracted from the last element of the phrase unit.

Apart from these features for each chain or phrase unit of the example, we know its position in the sentence. This is an important factor in order to study the relationship between the verb entry under study and the position in which the different phrase units appear. Those phrase units that are not close to the studied verb entry have fewer possibilities to be considered as arguments. Below we can see the third example (see Example 3.) represented in this way:

```
<Verb-Chain-Example>
   <Verb> jaso, jasotzen. </Verb>
   <Set-of-Examples>
      <Example-Set>
        <Sense-Code>A1.</Sense-Code>
        <Type-of-Auxiliary>DU</Type-of-Auxiliary>
        <Examples>
         <Example>
           <Example-Sentence>Harria zortzi aldiz jaso du minutu batean.</Example-
        Sentence>
           <Verb-Entry-Chain>
                <Chain>jaso du</Chain>
                <Position>3</Position>
                <Auxiliary-Verb>
                    <Base>*edun</Base>
                    <Syntactic-Function>@+FAUXVERB</Syntactic-Function>
                    <Chain>nuke</Chain>
                </Auxiliary-Verb>
                <Person>
                    <Nor>3rdPER_S</Nor>
                    <Nork>3rdPER_S</Nork>
                </Person>
                <Mood-Time>Present_of_Indicative</Mood-Time>
                <Main-Verb>
                    <Chain>jaso</Chain>
                    <Syntactic-Function>@-FINITEVERB</Syntactic-Function>
                </Main-Verb>
           </Verb-Entry-Chain>
           <Phrases>
                <Phrase>
                      <Chain>Harria</Chain>
                      <Position>1</Position>
                      <Part-Of-Speech>NOUN</Part-Of-Speech>
                      <Syntactic-Function>@OBJ</Syntactic-Function>
                      <Case>ABS</Case>
                      <Number>S</Number>
                      <Definiteness>DEFINITE</Definiteness>
                </Phrase>
                <Phrase>
                      <Chain>zortzi aldiz</Chain>
                      <Position>2</Position>
                      <Part-Of-Speech>NOUN</Part-Of-Speech>
                      <Syntactic Function>@ADVERBIAL</Syntactic-Function>
                      <Case>INS</Case>
                      <Definiteness>UNDEFINITE</Definiteness>
                </Phrase>
                <Phrase>
                      <Chain>minutu batean</Chain>
                      <Position>4</Position>
                      <Part-Of-Speech>DET</Part-Of-Speech>
```

```
                    <Syntactic Function>@ADVERBIAL</Syntactic-Function>
                    <Case>INE</Case>
                    <Number>S</Number>
                    <Definiteness>DEFINITE</Definiteness>
                </Phrase>
            </Phrases>
          </Example>
        </Examples>
      </Example-Set>
   </Set-Of-Examples>
</Verb-Chain-Example>
```

5 Querying the Data Structure towards Verb Argument Extraction

We have defined some queries that we think are useful for the detection of arguments. The main goal of these queries is to extract the basic arguments of the verb examples. For that purpose as a first approach we have designed 20 queries that can be divided into three main types:

1. One group of queries is used to ask for the number of arguments or possible arguments of the verb entry. In these queries we also inquire about the type of these arguments, distinguishing between phrase arguments and clause arguments.

2. Another type of queries is aimed for studying the relation between the syntactic function of a possible argument and its case.

3. In this group we include those queries that are related to the position of the elements recovered as candidates for the argument structure of a certain verb. In this case we want to discard those elements that appear in a position that is greater than the position of a verb chain that does not correspond with the verb entry.

For the moment we have grouped all the verb examples into five samples according to their type of auxiliary tag. We are testing these initial questions on these samples and, as a result of the use of the query mechanism, we are formulating new queries in order to reduce the amount of data which has to be browsed manually.

6 Future Work

One of the main aims of our future work is to get a classification of the verbs and its corresponding argument frames according to the criteria expressed by the queries and the integration of this information into the Lexical Data-base for Basque (EDBL). We have made a few experiments with some samples, but before applying the system to all the examples we want to improve the methodology presented in this paper, in order to deal with complex phenomena such as embedded clauses, coordination and the errors due to the part of speech tagger.

7 Conclusion

In this paper we have presented a methodology for helping linguists in the hard task of determining the basic argument structures of verbs. This approach performs the query-system as a tool to manipulate the full range of information contained in the examples, in order to derive the most reasonable argument structure. The basic type of information that we can extract from the examples is of syntactic nature, in particular phrase structure information as well as some pieces of information that we have chosen from the analysis.

The information extracted from these examples is not meant to be exhaustive. It will be updated when analysing other corpora and, most importantly, it can be helpful when searching in a corpus for relevant surface argument structure patterns.

8 References

Aduriz, I., Alegria, I., Arriola, J.M., Artola, X., Díaz de Illarraza, A., Gojenola, K., Maritxalar, M. Morphosyntactic Disambiguation for Basque Based on the Constraint Grammar Formalism, RANLP´97, Tzigov Charch, Bulgaria, 1997.

Agirre, E., Arregi, X., Arriola, J.M., Artola, X., Díaz de Illarraza, A., Insausti, J.M., Sarasola, K. Different issues in the design of a general-purpose Lexical Database for Basque. NLDB´95, 299-313, Versailles, France, 1995.

Alegria, I., Artola, X., Sarasola, K., Urkia, M. Automatic morphological analysis for Basque. Literary & Linguistic Computing, Vol. 11, N. 4, 193-203, Oxford University Press. Oxford, 1996.

Arriola, J.M. and Soroa, A. Lexical Information Extraction for Basque. CLIM´96, Montreal, 1996.

Brent, M. Automatic acquisition of subcategorization frames from untagged text. In Proceedings. Of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, C.A., pp.193-200, 1991.

Briscoe, T., and Carroll, J. Generalized probabilistic LR parsing for unification-based grammars. Computational Linguistics, 19:25-60, 1993.

Briscoe, T., and Carroll, J. Automatic extraction of subcategorization from corpora. In Proceedings of ACL, SIGDAT Workshop on very Large Corpora. Copenhagen, August 1997.

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

Manning, C. Automatic acquisition of a large subcategorisation dictionary from corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp.235-242, 1993.

Pustejovsky, J and Boguraev, B. Lexical knowledge representation and natural language processing. In Pereira, F. and Gross, B. (eds.). Natural Language Processing Massachusetts: The MIT Press, pp.193-223, 1994.

Sarasola I. *Euskal hiztegia.* Kutxa Fundazioa: Donostia, 1996.

Sperberg-McQueen, C.M., Burnard, L. Guidelines for Electronic Text Encoding and Interchange. Chicago & Oxford, 1994.

Tapanainen, P. The Constraint Grammar Parser CG-2 University of Helsinki. Publications n. 27, 1996.