AUTOMATIC EXTRACTION OF VERB PATERNS FROM *HAUTA-LANERAKO EUSKAL HIZTEGIA*

Jose Mari Arriola, Xabier Artola, Aitor Soroa

Abstract

This paper presents some of the results obtained by means of the method we developed for the study of verb usage examples, emphasizing as we do so that the primary aim was the development of a method rather than the results per se, and dwelling on the importante of shallow syntactic patterns in obtaining the patterns of the verbs studied. We are concerned with the extraction of verb patterns from the verb entries examples of an ordinary dictionary in machine readable version. The corpus of verb usage examples that we have analysed is composed of 13.089 examples. A shallow analysis allowed us to detect the verb chains and phrasal units that appear with the verb under study. The use of an SGML (Standard Generalized Mark-up Language) data structure to represent the analysed verb entry examples facilitates the extraction of the information contained in this data structure. We present an evaluation of the basic subcategorization patterns found and the principal problems encountered in the automatic extraction of them.

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, Aldezabal *et al.* 2001); and (2) the acquisition of a basic subcategorization information of verbs to support our parsing tools. The practical goal of our work is to enrich the information in verb entries with their corresponding basic subcategorization 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 subcategorization 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 subcategorization information causing half of the parse failures (Briscoe and Carroll 1993), attempts have been made to construct, from empirical data, lexicons that encode

[ASJU Geh 46, 2003, 127-146]
http://www.ehu.es/ojs/index.php/asju

information about predicate subcategorization 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 can and should be combined in the acquisition of this kind of information. The main reasons for deciding to use the verb examples in particular were these:

— More controlled analyses: the dictionary contains, together with other information about each verb, a statement of what type of auxiliary it takes, as well as certitude that the verb will be there.
— Comparison with the main corpus: as we said above, the examples may be considered a kind of specialized corpus because they have been taken from the general corpus. We can thus study low-frequency verbs by obtaining basic information about them from the examples, without needing to resort to much larger corpora.

In view of these reasons, the initial assumption, as stated earlier, is that the examples in the dictionary will be of use in determining the basic subcategorization of verbs.

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 one entry 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. All this information is given implicitly or explicitly in the hierarchical structures of dictionary articles, which are quite complex. The structural complexity presents some problems that must be treated in the analysis and interpretation of the articles. 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 type 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 carried out: (1) the parsing of the internal structure of the articles; (2) the definition of a grammar of entries that covered the general structure of the dictionary (as a Definite Clause Grammar (DCG) in Prolog) and (3) the conversion of the labelled structure which was encoded automatically following the Text Encoding Initiative (TEI) guidelines (Sperberg-McQueen *et al.* 1994). The TEI guidelines have been applied to the dictionary with considerable ease.

As a result of this conversion process we recognised the structure of the 98,49% of the entries with all the information contained in them, being the error rate of 3% (evaluation based on a sample). 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.

Through the work of adaptation we have taken a first step to facilitate the study of dictionary examples. It also provides an opportunity to take note of the problems and weaknesses of the lexicographer's approach for building the dictionary. The work of preparation for subsequent automatic analysis makes manifest the dictionary's structure; this is seen particularly in the parsing grammar. This is the grammar that the lexicographer had in mind when producing the dictionary.

3. Corpus of vers 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 actual 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 expected this made them appropriate for the subcategorization extraction process. However, sometimes we had to reject some examples as material for automatic subcategorization, when these consist of incomplete sentences containing syntactic structures that are not pertinent to the verb under consideration. Consider for example *Zaldiak alhatzen diren soroa* 'The field where horses graze'. Here a relative clause is used as an example to indicate the usage of the verb *alhatu* 'to graze'. A shallow parse would correctly detect the absolutive subject, *zaldiak* 'horses', but the other noun phrase, *soroa* 'field', has no argument function vis-à-vis the verb *alhatu*. There is no criterion for deciding between a subject or object function for *soroa*, without specifying another verb outside the relative clause, which is not provided in the example. Since only the relative part of the sentence is given, no choice is possible. Information extracted from such examples will therefore show a higher proportion of error.

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 (Arriola *et al.* 1999). The main bases 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. CG is based not on context-free grammars but on rules encoded in finite state automata. The fact that is morphology-based makes it attractive in our case because of Basque's morphological complexity. Moreover, the fact that it is aimed to process real texts and implemented through automata makes it a robust and efficient tool. For these reasons a decision was made in favour of CG for the writing of a general Basque parser (Aduriz *et al.* 2000). We also believe it to be an adequate solution for the purpose of analysing the verb examples in EH. As Abney (1997) points out, shallow parsers have been used, among other things, for extracting subcategorization patterns. Therefore we developed a shallow syntax, a constraint grammar for Basque or EUSMG, following CG formalism.

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

Example 1. Example before the analysis process: *Basurdeek ausikiko gaituzte gutxien ustean* 'The wild boars will bite us when we least expect it'

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 morphologically and syntactically ambiguous. 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.

In order to improve 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. 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 agree 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 by 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[2] 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 that have been developed apart from the general grammar. These subgrammars are CG-style grammars that contain mapping rules.

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

---

[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.

[2] The syntactic function tags designed for Basque are based on the Constraint Grammar formalism. The set of categories, syntactic functions and abbreviations used in the article are explained in Appendix A.

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 that 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 to mark up the non-continuous verb chains are:

— %NCVCHI: this tag is attached to the initial element of a non-continuous verb chain.
— %NCVCHC: this tag is attached to the second element of a non-continuous verb chain.
— %NCVCHE: this tag is attached to the final element of a non-continuous verb chain.

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

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

Example 2. A non-continuous verb chain and its corresponding syntagmatic units: *Euriak ez du ia kalea busti* 'The rain has scarcely wetted the street'

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 mark up 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 recognition of phrase units. In Example 3 some examples of the analyses got after applying the above mentioned subgrammars are shown:

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

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

4.4. 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 designed 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.

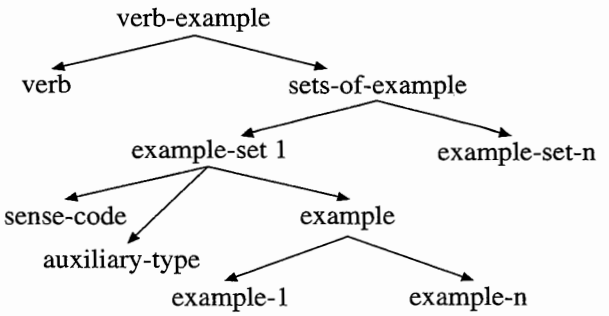

Figure 1. Outline of the organisation of examples

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 this 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 verb usage example we shown in 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>
               <PER_ABS>3rdPER_ABSS</PER_ABS>
               <PER_ERG>3rdPER_ERGS</PER_ERG>
            </Person>
            <Mood-Time>Present_of_Indicative</Mood-Time>
            <Main-Verb>
               <Chain>jaso</Chain>
               <Syntactic-Function>@-FMAINVERB</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>
```

Example 4. The verb usage example seen in example 3 represented in SGML

5. Evaluation of the analysis

The results of the analysis are referred to the above mentioned subgrammars applied to the output of the disambiguation grammar.

5.1. Evaluation of the verb chains and the phrasal units stablished

After marking verb chains and phrasal units, a random sample of 400 examples was taken out of the total of 13.089 examples. We checked this sample manually, looking at two points in particular:

1) Whether the chain labels were assigned correctly.
2) Whether any elements that should have had a label lacked one. Elements that should have a chain label are those forming part of phrasal units and verb chains discussed in the preceding section.

With regard to the first point, 84 of the examples contained a phrasal unit or verb chain that escaped correct detection. Thus 79% were labelled properly. Wrong labelling occurred chiefly for the following reasons:

— Ambiguity remaining in the examples. Since the chunk marking strategy is based on syntactic functions, ambiguity of syntactic function is a source of problems. But not all ambiguities affect the chunk marking phase. There will be problematic ambiguity when a single word contains both a major syntactic function and a minor one. This kind of ambiguity is of low frequency; it does not reach 2%.
— Disambiguation errors. In this section we include the consequences of incorrect assignments of syntactic function, which affect the identification of chunks.
— Unknown words. These are words for which there is no entry in the Lexical Database for Basque. The words also get analysed by lexicon-independent lemmatisation, but in such cases it is more difficult to get a correct analysis.
— Coordinate phrases. The rules for such structures need to be refined and improved.
— Postpositional structures. We have incorporated some postpositions, but the coverage is incomplete and many are not recognised; these are important for studying verb behaviour.
— Unpredicted structures in parsing label chains. For instance, modifications are necessary in the label set used for parsing structures such as *-ik ena*, as in *Arbolarik ederrena* (English gloss: 'the prettiest tree').
— Other errors. This category includes, *inter alia*, errors inherited from previous phases, such as one case in which a verb's category had been wrongly read as an example due to a mistake occurring in dictionary preparation.

Concerning the second point, elements that should have a chain label are those forming part of phrasal units or verb chains discussed in section 4.3. Therefore we do not take into account for this evaluation certain elements lacking labels, where we have not given rules for them to be labelled as parts of a chunk so they cannot be evaluated. Elements falling outside the labelling rules given include, among others, linkers,

conjunctions, relative clauses, multiple-word lexical units, etc. The chains recognised, with the exception of discontinuous verb chains, are all continuous.

5.2. Evaluation of the assignment of syntactic functions to phrasal units

To measure the accuracy of assignment of syntactic functions to the phrasal units detected, we created a random sample mirroring the characteristics of the whole set of examples, and performed a manual assignment of functions to each phrase. After the manual analysis, we compared this with that obtained automatically. This sample contained 1.211 examples, of which we only checked those containing a single verb, numbering 646.

The following criteria were used:

— We checked for the following functions: subject, object, indirect object and adverbial.
— We checked whether the functions assigned by manual and automatic means agreed. Disagreement, or error, might consist of incorrect marking or failure to mark.

The following table shows the results of the evaluation:

PHRASES	TOTAL	CORRECT	WRONG
MARKED AS SUBJECT	177	126	51
MARKED AS OBJECT	358	251	107
MARKED AS INDIRECT OBJECT	21	20	1
MARKED AS ADVERBIAL	220	213	7

Table 1. Results of the evaluation of the assignment of functions of phrases

As the table shows, indirect object and adverbial function assignment was successful. The weak point is assignment of subject and object functions. Nevertheless we consider the results obtained quite good, since %70 were correctly labelled and our syntactic disambiguation grammar is still under development.

With regard to subject and object assignment, some errors resulted from the difficulty of assigning these functions to arguments of verbs in non-finite form. In such cases, although there is only one verb, we lack the help given by finite auxiliaries whose agreement with subjects and objects facilitates the assignment of syntactic function. There are further difficulties with verbs for which the auxiliary-type specification in the dictionary is not helpful, as with the specification DA-DU (which indicates that the verb may be either intransitive or transitive). Even though such sentences may look simple, with the available resources there is no way to determine, in such examples, the function of every phrase associated with a non-finite verb. To do this, the lexicon needs to contain subcategorization information. For example: *Lana banatu* 'Distribute work'. To determine that *lana* 'work' is the object, the lexicon would have to specify what kind of objects the

verb *banatu* can take. Here there would be a specification of the thematic role of the object. We could then differentiate object from subject: the lexicon would need to state that this verb's agent is animate, whereas its object is inanimate. Thus it is very important for the thematic roles of verbs to be specified, to know what features make it possible for such an element to be either the subject or the object, where it might potentially be either.

Apart from the results shown in the table, the number of phrasal units recognised in the automatic analysis disagrees with that obtained manually (see 5.1, and remember that 79% were correctly detected), and consequently, the number of phrases marked for a given function may be larger of smaller in the automatically marked sample. The automatically marked sample shows 40 more phrasal units than the manually analysed one. On detecting the phrases belonging to a verb and their syntactic function and case, the shallow pattern that emerges is therefore distorted. For example, in *Meza azkendu zen arte* ('Until the mass was finished'), two 'subjects' are found: *meza* (a noun) and *arte* (a subordinating conjunction that happens to be homonymous with a noun), and the result would be to classify this as a verb taking two subjects.

6. Criteria for verb classification

As mentioned earlier, we obtained the analysis of each example through shallow parsing, and proceeded to extract from that analysis features that might be relevant for work on subcategorization. Given the wealth of data, examples may be classified in numerous ways, but in the present case we chose to focus on case and syntactic function. We based our classification of the syntactic structures obtained on the syntactic functions/cases @SUBJ_ERG, @SUBJ_ABS, @OBJ_ABS and @ZOBJ_DAT. With a classification based upon these functions and cases, we examined the lexically realized items that carried these markers in the dictionary examples. Given that it is extremely common in Basque that items related by agreement to the verb are not overtly realized, we should remark that such elided items are not included in our classification.

Of the examples of finite verbs studied, in 500 out of 2.700 there is neither an ergative subject, an absolutive subject, an absolutive object nor a dative indirect object. It is also common in other cases for one or another of these functions to undergo elision; the type of argument most commonly elided is the ergative subject. This fact is significant, and suggests that other cases appearing in shallow structure, cases not included in our shallow patterns, ought to be considered when studying subcategorization. Probably some cases/functions falling outside our analysis of syntactic structure should be included for consideration when determining whether or not they participate in argument structure. Thus for example local cases participate in the argument structure of certain verbs. Here are a few verbs that appeared in classes lacking any ergative subject, absolutive object or indirect object (ZERO-@SUBJ_ERG-@OBJ_ABS-@ZOBJ_DAT) and the cases that occur with each:

atera 'go/take out': 8 examples with local cases: ABL and INE (out of 32 total)
igo 'go up': 4 times ALA and 1 INE (22 total)
iritsi 'arrive, reach': 2 ALA, 2 INS, 1 INE, 1 ABL (17 total)
itzuli 'return': 5 ALA and 1 ABL (32 total)
hurbildu 'approach': 2 ALA and 1 INE (14 total)
dudatu 'doubt': 3 INS (6 total)

In these verbs, which are mainly verbs of motion, the cases that chiefly appear overtly are local cases. With some other verbs the instrumental occurs, such as, in our examples, *aldatu* 'change', *baliatu* 'use', *begiratu* 'look after', and *burlatu* 'make fun (of)'. The cases mentioned are frequently excluded from studies of argument structure, but as we have shown, they probably ought to be considered.

Our reason for not having taken these into account is that they are not the most common cases or functions to participate in argument structure. Since, overall, they rarely appear in a verb's specification for argument structure, they were not made a criterion for establishing the classes. However, more directed analyses can be carried out using the query system,[3] in order to look at examples of verbs taking local cases/functions, for instance. We have extracted the complete analysis of such examples and consequently dispose of information about the cases and functions of phrasal units associated with a given verb. We know what examples are given for each verb, with examples classified according to the sense of the verb and subcategory. This information is preceded by an indication of the verb's participle, the verb's sense, its subcategory and an example number; in this way examples are uniquely indexed. Each index is followed by a shallow parse, first showing the auxiliary type pertaining to the verb according to the dictionary entry, and then pairs of syntactic function and case.[4] If any other verb complexes occur in the same example, this is indicated by the sign MP (for 'subordinate clause') accompanied by + for subordinate or - for non-subordinate.

Thus for example the following patterns are listed for the verb *bultzatu* 'push, press':

```
bultzatu, bultza, bultzatzen.

bultzatu-A0.-DU-1    DU.@SUBJ_ERG-@OBJ_ABS.
bultzatu-A0.-DU-2    DU.@OBJ_ABS.MP+
bultzatu-A0.-DU-3    DU.@ADLG.
bultzatu-A0.-DU-4    DU.@SUBJ_ABS-@OBJ_ABS @PRED_ABS.MP-
bultzatu-A0.-DU-5    DU.@OBJ_ABS-@OBJ_ABS-@ADLG_ABZ-@OBJ_ABS-@OBJ_ABS-@ADLG.MP+
bultzatu-N1.-DU-1    DU.@SUBJ_ERG.MP+
bultzatu-N1.-DU-2    DU.@OBJ_ABS.
bultzatu-N1.-DU-3    DU.@SUBJ_ERG-@ADLG_ALA.
bultzatu-N1.-DU-4    DU.@SUBJ_ERG-@OBJ_ABS.MP-MP+
bultzatu-N1.-DU-5    DU.@ADLG_ABZ-@OBJ_ABS.
bultzatu-N1.-DU-6    DU.@OBJ_ABS.
```

Example 5. Basic verb patterns for the verb *bultzatu* 'push, press'

The shallow pattern class of each verb was obtained automatically and we defined a code identifying the verb examples occurring in each of those patterns. An example will

[3] 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 (Arriola *et al.* 1999).

[4] Syntactic function and case are linked by an underline character. A hyphen separates function/case pairs.

serve to show what kind of information the code contains. The example is *bultzatu-A0.-DU-2*:

— the participle (used as the verb's citation form), in this case *bultzatu*.
— sense index: specifies the sense, subsense or nuance of the verb in this example, e.g. *A0*.
— auxiliary type: the type of auxiliary indicated in the dictionary (DA, DU, DIO, ZAIO, or DA-DU). In this case, *DU*.
— example number: the examples for each verb are numbered, e.g. *2*.

The appendix of the thesis (Arriola 2000) lists all the verb examples classified by verb, in such a way as to show what shallow syntactic structures show up with what verbs. However, when classifying verbs in the next section, we shall only take function and case into consideration. The appendix shows all examples, but below we will select a few for illustrative purposes, following the above-mentioned criterion.

It needs to be noted too that the set of syntactic functions (Arriola 2000) that were defined affects the range of structures that can be recognised. The shallow structures that are detected correspond, of course, to those defined in our set of syntactic functions. Now these functions are adequate from the point of view of the parser, but when applied to the examples some of the functional distinctions turn out to be undesirable. The distinctions in question are very difficult to decide upon automatically, and consequently incorrect syntactic structures will sometimes be assigned. For example, distinguishing the nominal predicate function @PRED usually led to incorrect identification of structures. In principle we consider it necessary for subcategorization to distinguish the @PRED function; the trouble is that accurate detection of this function is hard to achieve, precisely because the lexicon lacks information about subcategorization. Therefore, it was thought advisable to proceed in our initial analysis without distinction of the function in question.

False recognition of patterns was also caused by the specification, where a subordinate clause was involved, of its function within the main clause. Even though inclusion of such distinctions in the set of syntactic functions is justified on linguistic grounds, this is not appropriate for the purpose of the method we developed. If for example, a verb has associated with it a non-finite subordinate clause, we may detect the subordinate clause but be unable to determine what the non-finite clause's role is vis-à-vis the main clause. To do this requires assistance from subcategorization information. In practice, then, more detailed syntactic functions hinder the disambiguation process and make it more likely for errors to occur in the information that is extracted.

Thus with regard to the set of syntactic tags, it may be concluded from our experiment that specification of the function of subordinate clauses in relation to a main clause, as part of the set of syntactic functions, ought to wait until subcategorization has been described. Likewise, the function of nominal predicate, @PRED, should be specified once there is a working subcategorization. At that point we would have the option of specifying what kind of subordinate clause each verb can take and the functions of the subordinate clauses.

7. The set of shallow patterns detected

In this section we present the shallow patterns that were extracted. The following diagramme shows what patterns were found:

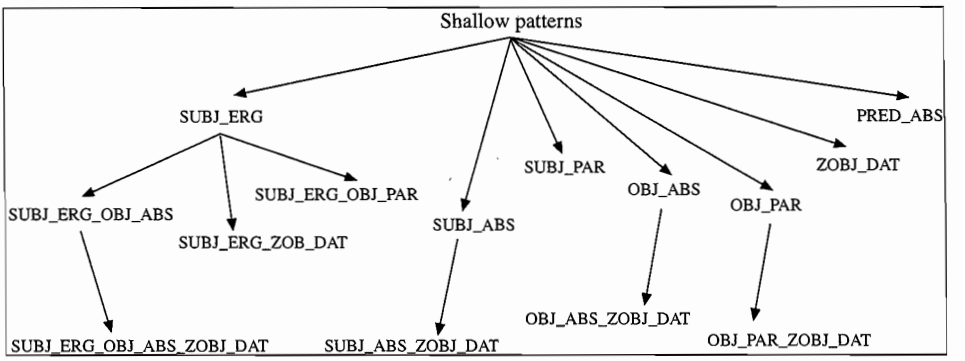

Figure 2. Surface patterns[5] in the examples

As we said before, we consider syntactic functions and cases when classifying examples. In this way, different verbs will be grouped together according to the shallow syntactic functions and cases with which they occur. Although verbs coincide in taking those functions and cases, criteria clearly need to be developed for a finer classification. The present classification is merely a modest first step. Work could begin on thematic roles on the basis of this material, among other sources.

These patterns merely show what structures each verb accepts. As we have pointed out, it takes a deeper analysis to determine what the obligatory arguments of these verbs are. Some authors argue that semantics should come under consideration here, in addition to other factors; Levin (1993) claims that the semantics of a verb determines its syntactic behaviour. In order to facilitate such analyses, we have decided to include information about which sense a verb is used in for each example. However, this task, among others, is for the future.

8. Automatically derived shallow patterns: difficulties and evaluation

In this section we will discuss the main difficulties encountered for classifying verbs on the basis of the methods developed and the reliability of the resulting classification. With regard to the difficulties, we will talk about the limitations of shallow syntax, the limited usefulness of position, and certain features of these verb examples. Following this we evaluate the classification, using measures of reliability for each pattern on the basis of an analysis of a sample.

---

[5] The shallow patterns that are detected correspond, of course, to those defined in our set of syntactic functions.

8.1. Limitations of shallow syntax

In developing the shallow syntax section we took an important step towards verb classification, labelling explicitly the phrasal units and verb complexes associated with a given verb with chunk marker tags (4.3). Thus we must take into account what we are able and unable to detect, i.e. what kinds of phrase (4.3). We furthermore evaluated the phase of phrase detection at the end of the section 5, noting the kinds of problem or error occurring with those phrases that could be detected. We find that of the phrases recognised, 79% were tagged correctly, that is, 79% of the chunks are correctly parsed. It is also necessary to consider the reliability of function and case identification in correctly marked chunks (5.2).

Considering what was said in the sections mentioned, it should be noted that the shallow syntax also fails to specify the relations between main and subordinate clauses. Thus we cannot use data from examples containing more than one verb for classification purposes. For example:

> *Liburu askoz **baliatu** dira idazlan hori prestatzeko.* 'They have **used** a lot of books to prepare that study.'
>
> The lexicographer is illustrating the use of *baliatu* 'use'. But our method is incapable of distinguishing whether *idazlan hori* 'that study' is the direct object of *baliatu* 'use' or of *prestatu* 'prepare'. Thus we cannot be sure of getting a correct analysis, which would be as follows:
>
> *Liburu askoz **baliatu** dira* [*idazlan hori prestatzeko.*] 'They have **used** a lot of books [to prepare that study].'

For a deeper analysis of such sentences, subcategorization data would need to be specified in the lexicon. But of course that information was not available when we started developing the parser.

With our resources it is very difficult to use the parser we developed to determine automatically which verb each argument (or potential argument) belongs to in multiple verb sentences. The information extracted would contain more mistakes if these were included, since the parser has no way of dealing with this problem. Such results would then require much manual work to determine whether automatically produced patterns were right. We preferred for the information extracted automatically to be more reliable and require less manual checking. This led us to study one-verb sentences, but we used some multiple-verb sentences to study the usefulness of position.

8.2. The use of position

We used position to help determine, in examples with more than one verb, which phrases (or subordinate clauses) go with which verb. We attached a number to each phrasal unit and verb complex detected, to indicate the order in which they occur. The order does not determine what function arguments have, except for focalisation, focused elements being placed immediately before the verb. But our hypothesis is that potential arguments and verb complexes do not appear just anywhere, but will normally occur in the vicinity of the verb in whose subcategorization they are included. On this ass-

umption, examples containing more than one verb were truncated according to the following criteria:

— When the verb under primary consideration precedes another verb complex, items following the second verb complex are ignored.
— Conversely, if another verb complex precedes the verb complex we are interested in, items preceding the first verb are ignored.

In the former case, where a second verb complex occurs later than the verb under consideration, then, it was decided not to count phrasal units occurring after the second verb. The example is truncated at that point; however, the second verb complex itself *is* counted, since it is possible that this might be part of the subcategorization of the verb we are considering. For example:

— Original example (the first two verbs in the example are underlined; the verb whose subcategorization is being analysed is in bold): *Zure okerrak **tapatu** nahirik egin dituzu pausuak, zer enganio egin didazun jakitun daude auzoak.*
— The same example after applying the criterion of position, i.e. truncated: *Zure okerrak **tapatu** nahirik egin dituzu...*

What we have done is to truncate the example appearing in the dictionary in order to limit our analysis to the part that remains after truncation. The rationale for this is that pertinent information about the verb being considered is located in the part of the example remaining after truncation, whereas the part of the original example that has been removed does not contain information relevant to the verb under consideration. However, this truncation criterion can give erroneous results, as for example when the two verbs are related by coordination. In such cases the two verbs may share the same arguments, but these will fail to get included in the analysis. For example:

— Original example: *Edanak eragiten ditu eta **erasaten** gauza lotsagarriak* 'Drink brings about, and **causes to be said**, shameful things'
—Truncated example: *...eragiten ditu eta **erasaten** gauza lotsagarriak* '...brings about, and causes to be said, shameful things'

Here our criterion leads us to exclude *edanak* 'drink' from the analysis, even though this is in fact the subject of *erasaten* 'causes to be said'.

Despite our awareness of the complexity of these issues, in our development of a shallow syntax we considered position a useful criterion and applied the truncation principle. To enhance the usefulness of this approach, it would be preferable to be able to take into account conjunctions, linkers and punctuation, assigning position to these and referring to them in the course of the truncation process. But recourse to these elements fell outside the scope of this study.

8.3. Evaluation of the patterns

It is important to evaluate the shallow patterns yielded by the verb classification in order to measure the patterns' reliability. We did this on the basis of section 5.2, checking for each pattern, on the basis of the criteria presented there, how often right or wrong syntactic functions and cases have been assigned. The evaluation was done

over a sample, which contains 1.211 examples of which 646 have a single verb. The 406 examples with more than one verb and the 159 examples in which none of the syntactic functions and cases that we have considered for verb classification occur are omitted.

The evaluation results represent comparisons between automatic and manual classifications. For each pattern, the functions and cases taken into account to classify verbs were checked. As we have said, we looked at whether or not the right functions and cases were assigned. We also remark on functions not appearing in the manual analysis of the sample but marked in the automatic analysis. The results show that when there is only an absolutive subject or object in a pattern, accuracy is lower than when these co-occur with other functions. For instance, the results for pattern OBJ_ABS are not as good as those for patterns OBJ_ABS-ZOBJ_DAT and SUBJ_ERG-OBJ_ABS. Indeed, labelling these functions correctly is the biggest problem. Nonetheless the results for pattern SUBJ_ERG are fairly good. Patterns SUBJ_ABS-ZOBJ_DAT and OBJ_PAR are not very reliable, while the most reliable are OBJ_ABS-ZOBJ_DAT and ZOBJ_DAT.

9. Conclusions

Despite the difficulties we encountered in the preceding section, and although the information obtained is *shallow*, we believe that the information may be useful not only as progress in syntactic analysis but also for methodological development. This requires integrating the information obtained into the lexicon for application in parsers. It will take deeper analysis to decide how to incorporate the extracted subcategorization data into the lexicon or parser in such a way as to be useful for parsing.

We also claim to have helped in the aim of facilitating the study of subcategorization in Basque. In that sense we think that the classification achieved provides valuable material for further analysis.

We initially expected the dictionary examples to provide a good source of material for the study of verb behaviour, and as a consequence of the work we have performed on them, that expectation is now even stronger, since the examples have been tagged syntactically and the basic chunks identified. Moreover, the materials have now been converted from plain text to a richer format using SGML, so that all this information will be the more accessible. Use of this encoding also facilitates the development of a query system; new methods and opportunities for research have thus been created (Arriola *et al.* 1999). Through the identification of numerous features, the material can now be employed to study various aspects of verb behaviour. In our own study we have used case and syntactic function, as was seen in section 7, to classify verbs.

We have developed a shallow syntax, with recognition of verb complexes and associated phrasal units, in order to extract a verb classification. If, however, we wish to go beyond the parsing of those units, deeper parsing is required. Specification of the subcategorization of verbs makes it possible to move forward from the analysis of phrases and verb complexes to the analysis of more complex sentences. To develop deeper parsing, of course, we will need to have information on subcategorization that should be specified in the lexicon. In our case, however, we set out with no such information, our goal being to discover which phrases and verb complexes occur in

association with individual verbs, inasmuch as that was possible. There is something of a vicious circle here. On the one hand we perceive the need to strengthen the syntax component in order to obtain information about subcategorization, and on the other, subcategorization information is essential for parser improvement. Notwithstanding, we believe the shallow analysis achieved is a valuable aid for further work on Basque subcategorization.

10. References

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

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

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.

Alegria, I., Artola, X., Sarasola, K., Urkia, M., 1996, "Automatic morphological analysis for Basque", *Literary & Linguistic Computing* 11.4, 193-203, Oxford University Press. Oxford.

Arriola, J.M., 2000, *Euskal hiztegiaren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen Analisia Murriztapen-gramatika Baliatuz Azpikategorizazioaren Bidean*, Ph. D. dissertation, Gasteiz.

— and Soroa, A., 1996, "Lexical Information Extraction for Basque". In *Proc. of CLIM'96*, Montreal.

—, 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.

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

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

— and —, 1997, "Automatic extraction of subcategorization from corpora", *Proceedings of ACL, SIGDAT Workshop on very Large Corpora.* Copenhagen.

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.

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

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

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

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

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

Appendix A

@+FAUXVERB: finite auxiliary verb.
@+FMAINVERB: finite main verb.
@<NCOMP: postposed adjectival.
@ADVERBIAL: adverbial.
@CASE_MARKER_MOD>: modifier of case bearing item.
@-FAUXVERB: non-finite auxiliary verb.
@-FMAINVERB: non-finite main verb.
@LOK: linker.
@NCOMP>: preposed adjectival.
@OBJ: object.
@PRED: predicative.
@SUBJ: subject.
@SUBJ_ERG: ergative subject (in this pattern we find transitive verbs with no object).
@SUBJ_ERG-@OBJ_ABS: ergative subject and absolutive object (transitive verbs with an object).
@SUBJ_ABS: absolutive subject (this pattern occurs with intransitive verbs).
@SUBJ_ABS-@ZOBJ_DAT: absolutive subject and dative indirect object.
@OBJ_ABS: absolutive object.
@OBJ_PAR: partitive object.
@ZOBJ_DAT: dative indirect object.
@OBJ_ABS-@ZOBJ_DAT: absolutive object and dative indirect object.
@ZOBJ: indirect object.
1stPER_PL: first person of plural.
3rdPER_ABS: third person of singular (absolutive).
3rdPER_ERG: third person of singular (ergative).
3rdPER_PL: third person of plural.
ABS: absolutive on nominals.
ABZ: ablative of direction.
ADJ: adjective.
ADVERB: adverb.
ALA: alative.
AUXV: auxiliary verb.
CERTAINTY: certainty.
COMMON: common.
DA: intransitive auxiliary.
DAT: dative.
DEFINITE: definite.
DET: determiner.
DIO: transitive auxiliary (with dative object).
DU: transitive auxiliary.
ERG: ergative.
GEL: genitive of location.
GEN: genitive of possesion.
INS: instrumental.
IWLP: interpretation with less probabilities.
LOK: link particle.
LOT: link particle.
MP: subordinative clause.
NOUN: noun.
PART: participle.
PL: plural.
S: singular.
SIMPLE: simple.
SUPERLATIVE: superlative.
SYNTHETICV: synthetic verb.
TRANSITIVE: transitive.
UNDEFINITE: undefinite.
V: verb.
ZAIO: intransitive auxiliary (with dative object).
ZERO-@SUBJ_ERG-@OBJ_ABS-@ZOBJ_DAT: verbs that appeared in classes lacking any ergative subject, absolutive object or dative indirect object.