ANALYZING VERBAL SUBCATEGORIZATION AIMED AT ITS COMPUTATIONAL APPLICATION

Izaskun Aldezabal and Patxi Goenaga

Abstract

The verb is one of the most important lexical components: it includes information regarding the necessary components that make up sentences and their features. This is precisely the domain of the analysis of subcategorization. However, specifying the subcategorization of each verb is a difficult task, mainly because of the following reasons: first, because the distinction of the semantic values and the alternations in each verb is problematic; and second, because of the presence of certain phenomena such as ellipsis, unspecification (of general and specific elements), and dependencies between Cases.

This work presents the following: after having reviewed the complex phenomena that are involved in verbal subcategorization, and contextualized these in our research area (i.e., in computational linguistics), we explain the procedure adopted to analyze 100 selected verbs, where Levin (1993) has been taken as point of departure. Once the research has been completed, we have defined what we have considered as subcategorization, namely, all the semantic/syntactic value(s) that we have defined for each verb (ssv), the set of outstanding elements in each ssv, their semantic specifications, and their Case realizations. Thus, we have tried to provide a coherent proposal as a base for grouping verbs depending on the goal.

1. Introduction: the need for sucategorization

Research on lexical components has become increasingly relevant in current theoretical and computational analyses for two reasons: first, because lexical information is the basic information that feeds other levels such as morphosyntax, syntax, semantics, etc., and second, because lexical information imposes conditions that determine the grammaticality and intelligibility of sentences. The verb is one of the most important lexical components. In fact, the verb includes information regarding the necessary components that make up sentences and their features. This is precisely the domain of the analysis of subcategorization.

The fact that, since the advent of generative grammar various proposals have arisen for defining the lexicon, suggests that this task involves many complications. As for us, Computational Linguists, we typically analyze real corpora, i.e., texts that are part of the everyday use of the Basque language. Thus, real corpora are the point of departure for all our analyses. In order to know and define the characteristic features of real

corpora, it is necessary to systematize a big number of phenomena. However, sources offered by general linguistics for this task have proven to be too scarce. The fact that corpora are the starting locus in computational linguistics implies two issues: on the one hand, the sentences under analysis are real sentences, and hence, we will encounter all types of sentences: long, short, grammatical and non-grammatical. On the other hand, sentences in real corpora are set in specific contexts.

All this suggests that we are dealing with components that still need analyzing in theoretical research. In other words, the tools that are available in theoretical linguistics are not sufficient to respond to the demands of automatic resources. One clear example is verbal subcategorization. Thus, computational linguistics adapts its resources by using the information that is available at the time, and it considers other ways in order to continue the research. The latter suggests that computational linguistics sets its own line of research largely.

Along these lines, the computer considers the corpus as a mere string of characters, and thus, the first step usually involves the analysis of the composition of words. Yet, the strings of characters that make up the corpus do not appear in isolation. Rather, they are set in specific contexts, and hence, it is necessary to predict the possible interpretations of words in connection with other surrounding words. Consider the following example: the word *iritziak* ('opinions') may appear in sentences like *Alkatearen iritziak herritarrak harritu ditu* ('The mayor's opinion has surprised the citizens') or in *Egunkariek herritarren iritziak plazaratu dituzte* ('Newspapers have published the opinions of the citizens'). Specifically, the word *iritziak* may appear in Ergative Singular or Absolutive plural. Moreover, *iritzi* has an ambiguous categorical status, and it may be a noun or a verb. To make matters worse, it may appear in a string like *iritzi dio* ('he/she believes'), where *iritzi* surfaces in the participial perfective form. All this implies that, were we to analyze such forms in isolation, they would be ambiguous, i.e., they would have various interpretations. Nevertheless, in order to advance into syntax, we need to cut such ambiguities by disambiguating processes. Among the possible analyses of the word, this process selects a single analysis (i.e., the correct one that corresponds to the context under consideration).

Here are the steps that we have taken to analyze sentences in real corpora:

a) The basis is a database, which includes the necessary information to morphologically isolate and analyze all the words in a sentence: the Basque Lexical Database (i.e., Euskararen Datu-Base Lexikala (henceforth EDBL)) (Aldezabal et al., 2001a). Thus, each item is classified in accordance with its lexical or morphosyntactic category and subcategory. The database is organized to carry out the so-called morphotactic relation (Alegria 1995, Urkia 1997), along the lines of the two-level morphology in Koskenniemi (1983). This means that the combinations between morphemes are included in the database itself. This provides as a result the morphological and morphosyntactic composition of words.
b) In order to reduce ambiguity, we have employed a disambiguating tool for Basque (Aduriz et al. 1997) that was created based on the Constraint Grammar (henceforth CG) formalism in Karlsson et al. (1995). This tool reduces the possible interpretations of words through definitions of rules that are based on

context. This disambiguating tool cuts categorical ambiguity almost entirely. However, ambiguity persists in cases where other factors such as Case or function are considered, which suggests that further information is necessary. One such type of information is verbal subcategorization, namely, specification of elements that are selected by verbs.

c) Yet, computational research has continued into syntax in two directions despite the persistence of ambiguity, but acknowledging the necessity for lexical information. One line of research has created a finite state system by extending the CG formalism (Tapanainen 1996); another line has created the PATR II formalism based on unification (Shieber 1986). The former creates new tags to form phrases based on the function of morphemes. This provides as a result a syntactically tagged sentence (Aduriz 2000, Arriola 2000). The later defines the unification-rules by employing the lexical information of morphemes. These rules meet the relations existing between the word level and phrase level by using the unification-equations (Gojenola 2000, Aldezabal et al. 2003).

d) However, the results obtained by the application of these formalisms are not very successful considering the following facts: first, some interpretations remain ambiguous in the morphological disambiguation process, and second, grammars suggest many combinations among the elements of the sentence, i.e., they create structural ambiguity. In order to minimize this problem, we have applied a Finite State technique based on automata and transducers. As a result, since the verbal context under consideration is reduced, ambiguity percentages are also significantly reduced (Aldezabal et al. 1999b, Aldezabal et al. 2001b). Thus, we are able to get a phrasal analysis of sentences in a corpus, and to use all the morphosyntactic information included in the phrase. In addition, we will often find that we get several interpretations for one sentence.

Let us consider an example of how this process is applied to a particular sentence (excluding ambiguity).

The above results show that there is no relation between the elements surrounding the verb; in other words, we assume that, in principle, the elements surrounding the verb somehow belong to the verb; there is no explicit distinction as to whether elements belong to the verb (the arguments of a verb) or to the sentence (adjuncts).

Things get even more complicated when sentences contain more than one verb, since, in principle the surrounding elements may be related to either predicate. In such cases, apart from not showing the argument/adjunct distinction stated above, there is no way of knowing to which verb phrases relate. This increases ambiguity, since all combinations are considered as legitimate options. Moreover, another arising problem is that clause boundaries within sentences cannot be delimited.

For all these reasons, at this point it is clear that, as is the case in theoretical linguistics, the computational treatment of language requires considering verbal subcategorization.[1] Thus, in this article we will show how subcategorization can be

[1] However, we need to mention that we have taken important steps in retrieving information pertaining to subcategorization by automatic or semi-automatic means (Arriola 2000; Aldezabal et al., 2001b, Aldezabal et al., 2001c).

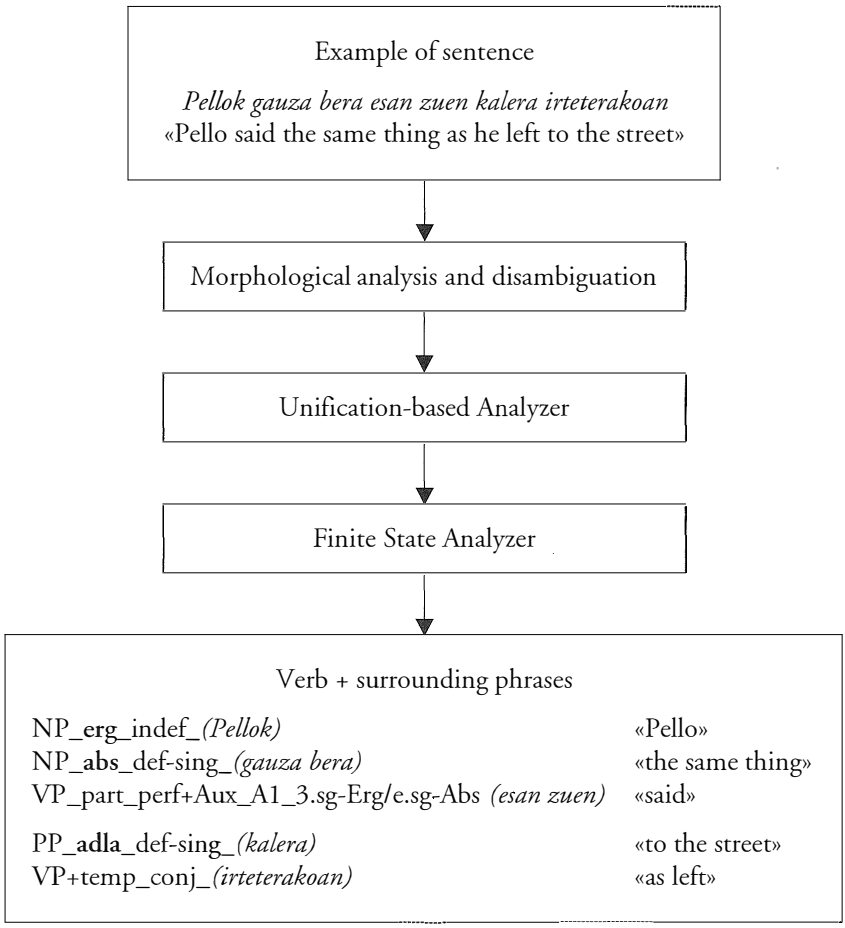

Figure 1. The general picture of the system with an example

analyzed in relation to the perspective of computational linguistics within the IXA group. The presentation of this research is organized as follows. Section 2 includes a presentation of the concept of subcategorization as well as a brief description of the proposals concerning the organization of the lexicon. In section 3, we present an overview of the procedure that we have selected for our work. First, and considering the above stated facts, we will take as a point of departure a proposal that bridges best the theoretical and computational approaches, namely the *English verb classes and alternations* by Levin (1993). Specifically, we will show the viewpoint and methodology included in Levin's work, as well as the gaps that we have detected in them. Next, we will explain the choice we have made for our work. Section 4 includes the overall conclusions drawn from the application of our procedure, the problems we encountered in doing so, the decisions we have made, and the specification of the pheno-

mena that were detected. Finally, section 5 is a summary of the article, and it includes the general conclusion drawn from the research.

2. On subcategorization

So far, in this article, we have suggested that the information in subcategorization pertains to the lexicon. However, in the literature we find various features and expressions that describe and designate this term. The term subcategorization arises parallel to the discussion on the autonomy of the lexicon within generative syntax,[2] which started when Chomsky published his second book *Aspects of the Theory of Syntax* (Chomsky 1965). In Chomsky (1965), the lexicon will become increasingly independent; lexical items include phonological and categorial information, and in the case of verbs, apart from phonological and categorial information, we will find information on subcategorization, selectional restrictions on arguments, and features pertaining to context. Subcategorization information includes the phrasal category (NP, TP, etc.) of the elements that are required by the verb, in other words, the specification of the syntactic realization of arguments. This was precisely what was considered to be in the so-called strict subcategorization.

Additionally, verbs were classified according to one of the two subcategorization structures that were suggested. On the one hand, predicates that had the subcategorization structure 'NP + V + NP' were transitive predicates, and those that displayed the structure 'NP + V' were classified as intransitives. In other words, when predicates contained an object they were called transitive predicates, and otherwise intransitives. Syntactic rules that made up sentences were defined in terms of this parameter. Yet, this first attempt in generative grammar was considered both redundant and too dependent on certain languages. Additionally, contrary to the above prediction, they realized on the existence of predicates that included a transitive auxiliary and a subject but no object (*irakin* ('to boil'), *iraun* ('to last'), *dimititu* ('to resign') and the like). This suggested that the terms transitive and intransitive were not clearly defined. As a solution, Chomsky in his *Lectures on Government and Binding* (Chomsky 1981) presented the influencing framework called Government and Binding Theory (henceforth GB). In this framework, grammars are viewed as computational systems composed of modules that include some universal principles and some parametric variations.

According to this proposal, predicates have the ability to assign a semantic feature called thematic role to each of its arguments (namely, to each of the participants that are necessarily involved in the action denoted by the verb). Additionally, verbs are capable of assigning the Case that will allow the realization of thematic roles in the syntax (the Case Filter). Moreover, thematic roles are hierarchically organized, which defines the function that arguments have in sentences. It is further assumed that thematic roles are invariably realized in specific phrasal categories. Thus, by the principle of Canonical Structural Realization (CSR), each thematic role is assigned the

[2] Before this date, we find the term *government*, which expresses the task of selection of pre/postpositions by the verb, which is a similar concept to current analyses involving verbal selection. Yet, the term subcategorization arises with Chomsky, and we have set our research after the term was suggested.

corresponding grammatical category. Hence, each predicate contains an argument structure in the lexicon, and the hierarchy and the CSR will determine the role and the syntactic realization of arguments.[3]

The new classification of verbs includes the following: unaccusative predicates (those that involve a purely intransitive auxiliary and a single argument, as in *etorri* 'to come'), unergative predicates (which involve a transitive auxiliary and a single argument, as in *iraun* ('to last')), and finally, transitive predicates (which involve a transitive auxiliary and two arguments, as in *eraman* ('to take')). Several theories have arisen attempting to explain single argument predicates.

Some authors started to claim that the structure of the lexicon is more complex than was standardly assumed, and they defended the existence of regularities in it. This attracted the attention of researchers towards the lexicon. It was claimed that such regularities arose from the interaction between semantics and syntax. The first to claim such a relation were Hale and Keyser (1987). Later, the proposal in Jackendoff (1990) has been the most successful one and the one to receive most attention. Jackendoff suggested a more abstract structure to represent the lexicon, namely the Lexico-Conceptual Structure (LCS). This structure is composed of various semantic primitives (among others, GO, STAY, CAUSE, TO, FROM, TOWARD, AWAY-FROM, VIA), an in turn, these primitives correspond to more general conceptual categories (Thing —or Object—, Event, State, Action, Place, Path, Property and Amount). For instance, primitives GO, STAY and CAUSE correspond to the conceptual category Event. In addition, the syntactic correspondence is defined also at this level. Thus, each lexical entry is defined in terms of the conceptual categories, primitives and their corresponding surface syntactic structures.

Other proposals were also suggested. For instance, the generative lexicon in Pustejovsky (1995), who suggests a complex structure for each lexical item (which includes argument structure, event-structure, and qualia structure), and obtains the surface structure through composition of all features that take part in the complex structure; Levin in her *English verbs classes and alternations* (1993) analyzes English verbs. Levin does not specify the entry corresponding to each lexical item (as she herself acknowledges). Rather, she suggests ways of organizing the entries. She notices that verbs that are similar in semantic nature accept the same syntactic structures. Thus, the fact that the ability of language is considered to be innate explains how speakers are capable of knowing what syntactic structures are allowed with predicates. This suggests that the first task is to figure out which syntactic structures we are facing in order to group predicates and to analyze their semantic components. It is clear that the internal composition of lexical items is still being debated and analyzed. Yet, there is a commonality underlying all the theories proposed: lexical items contain various types of features, and the existing relations of such features condition the correct syntactic realization of lexical items.

---

[3] Two clarifications are necessary at this point. First, there are those who support the view that it is necessary to define subcategorization (Grimshaw 1979; Rothstein (1992): among others). Second, the discussion on thematic roles is far from reaching consensus. Some suggest that roles may be distinguished in contrastive pairs (for example, the pairs *agent/ cause* and *goal/receiver* through the feature [±animate]). Others have suggested other proposals, among others, we find Dowty (1991) and Van Valin (1993), who compose roles by means of role hierarchies called 'protoroles' —or general roles— and by binary +/- features.

3. The procedure

As for the Basque lexicon, and more specifically, as for the verb, the EDBL defines the category, the subcategory and the word combinatorial options that are accepted within the verb (namely the morphotactic relation). Thus, it is clear that we are far from the complex composition of the lexicon proposed in the previous section, and that there is no reference to the components that are selected. This implies that, if in the near future we engage in completing the lexicon of the verbs for their application in automatic use, we will need to start by positing modest goals. Thus, although the ideal facts about the lexicon may be contained in theoretical proposals, practicality restricts our goals. To start with, our interest is to determine the surface realization of the components at the level of the sentence. This suggests a clear approximation to strict subcategorization.

Second, we need to take into account that the steps that we have taken so far in our group provide us with interesting available information, which includes phrases that compose sentences, including all the information contained in them. This will let us proceed to further analyses or to confirmation-processes.

All this suggests taking into account the work developed by Levin. In our view, the line proposed by Levin is roughly adequate, mainly for two reasons: first, from a computational linguistics perspective, because it engages in analyzing surface structures. Second, because it is aimed at organizing the lexicon of verbs. Thus, we have analyzed her proposal in detail and we have measured the advantages and disadvantages that it offers. In addition, Levin's work has served to analyze verbs in various languages such as Spanish and Catalan (Taulé 1995), French (Saint Dizier 1995), German, Korean and Bangla (Jones et al. 1994). The research on Spanish and Catalan deserves special mention. Because of the cooperative relation that we maintain with them, we have had the chance to get to know their work in detail; moreover, we hope that their experience will serve to guide us in our research (Vázquez et al. 2000).

3.1. Levin as point of departure

Levin claims that native speakers are capable of noticing many phenomena that appear in their language. One of them is the ability to notice among the various syntactic realizations of a particular verb. In other words, speakers are able to establish relations among the various structures —some of which imply semantic differences— that verbs display. They are also able to determine which structure(s) each predicate may accept, and which not. Levin employs the term diathesis alternations to name the different structures or, in other words, to name the pairs of structures of verbs that are related. Quoting:

> Verbs, as argument-taking elements, show especially complex sets of properties. As shown in B. Levin (1985b, in prep.) and other works, native speakers can make extremely subtle judgments concerning the occurrence of verbs with a range of possible combinations of arguments and adjuncts in various syntactic expressions. For instance, speakers of English know which *diathesis alternations* —alternations in the expressions of arguments, sometimes accompanied by changes of meaning— verbs may participate (Levin 1993: 2).

According to her, there is at least one common semantic feature in the syntactic variants of the alternations that verbs admit. This is precisely the reason why it is possible to classify verbs into groups:

> If the distinctive behavior of verb classes with respect to the diathesis alternations arises from their meaning, any class of verbs whose members pattern together with respect to diathesis alternations should be a semantically coherent class: its members should share at least some aspect of meaning (Levin 1993: 2).

Thus, after explaining the theory based on Lexical Knowledge in depth, Levin divides the content of her results into two parts: in the first part, she shows the alternations that she found in English, she provides the list of the verbs that take those alternations, and for each alternation, she describes their syntactic, semantic and (when applicable) morphological features. In total, she presents 80 alternations, and she divides them into 8 groups, which are, in turn, divided into further subgroups. In the second part of her work, and based on these alternations, she suggests 191 semantic subgroups in total, which are organized into 49 larger sections. Yet, we have detected several incoherencies in her procedure of analyzing alternations and grouping verbs. Here is the list of the incoherencies that we found:

— She does not always group verbs according to the alternations that verbs share.

- For instance, verbs of groups 9.1 *(Put verbs)* and 10.1 *(Remove verbs)* admit the same alternations, and yet, she classifies them into distinct groups.
- Another occasion when she turns verbs into distinct subgroups, is when they contain a semantic component introduced into the verb via suffixation. For example, this is the case of verbs in group 9 *(Verbs of Putting),* namely subgroup 9.9 *(Butter verbs)* and 9.10 *(Pocket verbs).* E.g.:

  9.9: *Lora buttered the toast*
  9.10: *Lydia pocketed the change*

  It is obvious that the basic structure of these derived verbs and that of the non-derived form (namely, the remaining subgroups in section 9) is different, and that syntactic structures are unable to relate the derived and non-derived forms. However, verbs of groups 9.9 and 9.10 admit/reject the same alternations, and thus, they should not be considered as 'syntactically' distinct; however, they are distinguished in Levin's system.
- Certain semantic groups do not display any alternations (for instance, group 52 *Avoid verbs*, and the subgroup 54.2 *Cost verbs* in section 54). This, according to Levin's methodology, would imply that verbs that display such structures do not accept any alternations, and hence, we would have to conclude that they do not form any group.

— She uses the term alternation in various senses. As was mentioned above, Levin considers alternations the pair of structures that certain verbs admit and that share certain common semantic property. Nevertheless, this is not always so. In fact, there are several alternations where only one structure is described (namely

those in 7.4.,[4] 7.5 and 8.4), and others that admit two structures, but where one of them is illegitimate (for instance 7.6.1, 7.6.2, 7.7, 8.1, 8.2, 8.3, 8.5 and 8.6). Moreover, we also find differences in those alternations that admit two legitimate structures: sometimes, one syntactic component drops in one structure; others, one component is added, and finally, sometimes, there is no component that is dropped, but the syntactic realization of such components changes.

— To finish up, for each semantic group, she does not specify the source of the structure that is considered as basic within alternations. It seems that the basic structure is already delimited (or it looks that she considers it to be so), and that based on this, she then lists the various alternations which are accepted in each case. Thus, there seems to be a gap in the methodology or theory that she proposes.

3.2. Our choice

Considering the problems in the previous section, rather than taking into account the semantic groups that she suggests for English, we decided to analyze 100 verbs in Basque by employing certain syntactic resources and by making use of the Corpus[5] that is available to us. When specifying our resources, we have taken into account how useful the selected resources may be for our computational tools. On the one hand, it is from these resources that we will retrieve useful data for our manual analyses, and on the other, those resources constitute the onset for our future research. However, we found it interesting to consider the alternations that we may find in Basque compared to those found in Levin (1993). The fact that we may find parallel alternations in Basque and English provides generality to the structures, and moreover, it may be relevant from a comparative perspective. Thus, we have considered the research that was developed within the IXA group (Aldezabal et al. 2002) as a basis, which includes a comparison of the alternations proposed by Levin with Basque.

As for the computational analyses in our research group, we have mentioned that the current computational tools analyze the phrases in sentences that appear in a corpus. These analyses provide as a result morphosyntactic information of phrases —namely information on number, definiteness and Case—. Additionally, our tools can easily provide us with the correct auxiliary that corresponds to the verb in each instance. These are ample resources that are available in the research. Next, we will describe the details of our line of research.

3.2.1. *Features considered in verbal analyses*

In order to complete the information pertaining to verbs, we have made use of two particular surface syntactic features when analising the sentences in the corpus.

— The type of auxiliary, by using the following typical means of expressing types of verbs: DA (purely intransitive), DU (transitive), DIO (ditransitive) and ZAIO (involving two arguments, one in dative and one in absolutive).

---

[4] The numbers in the alternations in the text strictly follow the ones in Levin (1993).

[5] The available corpora refer to the electronic samples of the daily *Euskaldunon Egunkaria* between January 1999 and May 2000.

— Case: we determine which cases verbs accept. However, we have only considered those Case markings that display a meaningful degree of presence in the corpus, specifically, only eight:

absolutive (41,79%), completive *-ela* (2,70%),
ergative (36,37%), instrumental (1,77%),
inessive (6,38%), sociative (1,52%),
dative (3,61%), ablative (1,34%)
adlative (1,28%).[6]

The remaining Cases have a percentage of presence lower than 1. This is the way we have analyzed Cases:

— On the one hand, Cases that outstand in frequency, namely those that are semantically closely related to the verb (or more specifically, those that we consider to be related to the verb), will be marked exceptionally. Thus, we will call these Cases 'outstanding Cases'.
— On the other hand, and in order to help distinguish between alternations and non-alternations pertaining to verbs, we have attempted to consider the constraints on the simultaneous appearances of Cases. In other words, we have analyzed Cases in terms of the restrictions that they impose on the realization of other Cases.

3.2.2. *Verbal values: syntactic/semantic values (ssv)*

The features described in the previous section will be assigned based on the different values that correspond to verbs. This is, indeed, the most complicated task. As it was mentioned above, the theory proposed by Levin suggests that, by virtue of their innate ability, speakers are able to determine the existing (and non-exiting) alternations pertaining to a verb. The underlying idea is that alternations share some semantic component. Hence, the crucial task is to determine which is/are the component(s) that alternations share. In fact, the semantic nature of such components (their semantic relation with verbs) determines how outstanding Cases are. Thus, our goal has been to determine those semantic components by analyzing 100 verbs in depth, and moreover, we have intended to identify the syntactic structures that are involved in alternations, i.e., to identify alternations, and those which are not. Thus, we have described several values for each verb, which are specified by their meaningful semantic components and by their syntactic Case realization. As a result, we have considered the values of verbs as semantic/syntactic values (*ssv*).[7]

---

[6] Note two important facts regarding Case: first, we have employed the term Case in a very general sense by including all declension Cases, both simple and complex (the later involving various words) as well as subordinating conjunctions (also simple and complex ones); second, works involving automatic retrieval of information regarding subcategorization consider all partitives as absolutives. This is the reason why, in manual analyses, we do not distinguish between appearances of partitive and absolutive.

[7] Note that the *ssv*-s that we have defined do not necessarily correspond to the verbal entries that are defined in dictionaries.

In addition, we have not distinguished between the two variants that belong to the alternation(s) of a verb (namely, the *ssv*-s that are related by some semantic component), nor the *ssv*-s that are not related to each other (namely, those that do not take part in alternations). Thus, various *ssv*-s are suitable for each verb (regardless of the existence of alternations among them). Note that we have not described *ssv*-s that are not in the corpus, although we acknowledge that there may be some.

We have designed a database in order to keep the information pertaining to *ssv*-s in a structures manner, and we have selected a marking-system to codify the information. Nevertheless, we have disregarded several topics to avoid the analysis from becoming too complex. For present purposes, we will only present the basics of this subject, and for more information, see Aldezabal et al. 2001 (forthcoming). First, we will present the topics that we have excluded from the research, and next, we will describe the marking-system that we have employed.

3.2.3. *Excluded topics*

3.2.3.1. *Impersonal, passive and antipassives*

Along the lines of Levin, in the task of marking different *ssv-s,* we have tried to solely resort to lexical values. From this perspective, it is well known that impersonal, passive and antipassive constructions are structures that are derived in the sense that they emerge as a result of applying some lexical operation to lexical structures. We have accepted this claim, and thus, when we have come across verbs that involve such constructions, we have not marked them as distinct in terms of *ssv*-s. Thus, when verbs appear in such constructions in the corpus, we have merely marked them as involving the values that they would have in non-impersonal and active sentences.

3.2.3.2. *Phrases without case*

Certain phrases are not formed by Case. These are adverbial phrases. We have not considered them because they do not display any Case.[8]

3.2.3.3. *The same case only once in each:* ssv, *except absolutive*

In the *ssv*-s, we will not mark the same Case more than once. If necessary, the Case will specify the possible semantic values that we have determined for the *ssv* in each instance. In other words, rather than distinguishing Cases we will distinguish semantic values. For example, we will not mark the two well-attested values of the ablative (source and path —or prosecutive value, Azkarate and Altuna 2001: 128) with two ablative markings, but rather, we will consider them as two legitimate values of the ablative in the same *ssv.* Nevertheless, we will make an exception; specifically we will accept two absolutives in the same *ssv.* Arguably, adjectives and nouns can form nominal predications that are formed with the absolutive (mostly

[8] This implies that certain legitimate values of verbs will be left out. Notice that adverbial phrases are sometimes necessary in the *ssv* definition of a verb.

with indefinite absolutive forms).[9] Yet, we will consider them as if they were first level nominal predicates, namely, only when the component in the *ssv* is most relevant.

3.2.3.4. *Lexicalized units*

The fact that we have considered Case does not imply that we have considered every phrase that contains some Case. It is well known that many of the Cases that we have selected display tendencies for lexicalization when they appear attached to other lemmas, either in the form of single words or in the shape of various forms. (e.g., *orduan* ('then'), *sekulan* ('ever'), *patxadan* ('relaxed'), *marmarrean* ('muttering'), *azken batean* ('after all'), *hitz batean* ('in a nutshell'), *gogotik* ('willingly'), *horratik* ('nevertheless'), *aspalditik* ('for a long time'), *inondik ere* ('absolutely not'), *gora* ('upwards'), *ahoz behera* ('face down'), *hankaz gora* ('upside down'), *adibidez* ('for example'), *negarrez* ('crying'), *beldurrez* ('in fear of'), etc.).

They also participate in various compounds (e.g. atez ate ('from door to door'), mendiz mendi ('form mountain to mountain'), etc.), and in complex declension Cases (e.g. -i buruz ('about something'), -tik at ('out of'), -n zehar ('through'), etc.). It may also be part of units that are composed of several phrases (e.g. *bostetik bi* ('two out of five'), *zazpitik lau* ('four out of seven'), *lurretik bost metrora* (five meters form the ground'), *egunetik egunera* ('day by day'), *goitik behera* ('thoroughly').

However, we need to mention that it is not easy to decide on the degree of lexicalization of such items. In fact, in our view, the fact that many such forms are in the process of lexicalization is related to the growing loss of the values that Cases have with respect to verbs in general. For example, based on what we have seen in our analyses, ablative Case involves values related to departure location, path and static setting of the entity, and adlative Case involves values related to goal. However, occasionally, ablative and adlatives receive other values too. For instance 'manner': *gogotik* ('willingly'); *hauturra* ('at someone´s discretion'). When this phenomenon happens lexicalization appears. Regarding units composed by more that one phrase, one of the reasons for considering them as units is that phrases in isolation do not make sense with respect to a particular *ssv* of a verb. In other words, what gets the value is the element resulting from the union of two phrases in the *ssv*. For example, in *Goitik behera busti zuten* ('they soaked him all over'), the unit shows 'manner'; in *Leihotik behera bota zuten* ('they threw him/her out of the window') the unit refers to the direction (and not to the departure and target points). In our view, this is precisely the reason why these strings should be considered as a complex declension Cases.[10] Thus, all the forms described above should be considered as lexicalized forms or units, and we

[9] Zabala analyzes predication relations in depth in her 1993 thesis entitled *Predikazioaren teoriak Gramatika Sortzailean (Euskararen kasua)*, where she includes several proposals for the elements that realize such predication relations.

[10] Let us mention that we have taken steps in analyzing units that contain various words (what we call Multi-word Lexical Units (MWLU)) (Aduriz et al., 1996). Moreover, there is current doctoral research on this topic in our group (Urizar, R.: *Kolokazioak euskaraz*). In addition, some research has been done in analyzing structures that contain various phrases from a semantic and pragmatic perspective (Garai & Ibarretxe 2002).

should analyze the syntactic and semantic values that they take as a whole. Since, automatically we have only analyzed units as postpositions (and since these forms are not among postpositions), we have decided not to analyze them.

3.2.3.5. *Cases that may have temporal reference*

It is well known that verbs may usually take phrases that contain temporal reference, and that temporal reference may be expressed by various types of Cases, such as inessives, ablatives, adlatives, instrumentals, sociatives, and also, absolutives —of course, only if one considers such forms as absolutives— *(gauean* ('at night'), *igandetik* ('since Monday'), *igandera* ('til Monday'), *arratsaldez* ('in the afternoon'), *igandearekin* ('with Sunday'), *bi egun* ('two days'), etc.). We know that, apart form setting the action denoted by the verb in temporal reference, these temporal references do not usually provide special information about the verb, and that most verbs accept such Cases.

Thus, when marking Cases, we have decided not to consider instances that contain temporal reference.

3.2.4. *The database and the marking-system*

3.2.4.1. *The database*[11]

The database contains five charts. There is one main chart, where we mark the type of auxiliary that corresponds to each *ssv* of the verb. Each of the remaining four charts corresponds to types of auxiliaries, and they contain a specification of Cases that will be analyzed in each chart. A small square beside the Case signals whether the case is accepted or not, and the Cases that we have determined as outstanding contain an additional domain that specifies their semantic value. The charts that correspond to auxiliaries have room for explanations, examples, and comments. Thus, after marking the type of auxiliary in the main chart, we fill the chart that corresponds to the auxiliary that we have marked.

3.2.4.2. *The marking-system*

We have employed three specific symbols in the marking-system, namely √, – and +. We have marked '√' the auxiliary and the outstanding Cases that are used in each *ssv.* Concerning cases, this symbol signals the following: 'it may appear, and it is outstanding'. In other terms, regardless of its presence/absence in the corpus, we consider that the Case has the ability to surface in the *ssv* under consideration, and it is typically outstanding. Assigning '√' to auxiliaries means that the verb under consideration takes the auxiliary in that *ssv*, although it may not appear conjugated. We employ the symbol '–' to express that a Case is unacceptable in a combination. Finally, we may find that, although a given Case is accepted, it is not closely related to the verb, namely, it is not an outstanding Case. Such Cases are marked with symbol +.

---

[11] For the moment being, the content and shape of this database is not available to the public. However, we are planning to include it in our webpage so that anyone can consult it.

Concerning Cases, it is well known that absolutive, ergative and datives are exceptional in displaying agreement with the auxiliary. As such, symbols '√', '–', and '+' on them contain a more specific meaning:

Absolutive and Ergative

Symbol '√' on absolutive and ergatives automatically implies that the verb requires them, but that the Cases may be absent due to ellipsis. In other words, these Cases 'may' appear, as is well known: when they are absent, it means that they are absent for elliptical reasons (the phenomenon of pro-drop). In contrast, instances where these Cases are absent for other reasons will be marked with '–' (namely as a distinct *ssv*) to signal they must be absent.

As for auxiliaries, we will mark the auxiliary type that the verb takes in the SSV.

Dative

We will mark the dative with '√':

— If a verb accepts the dative, where the dative is not a mere addition in instances that involve no dative. E.g.:

Pello adiskideen izenez ***ahaztu*** *da* → **Pello adiskideen izenez* ***ahaztu*** *zaizkio Anderri*

(Lit: Pello friends-of names-post forget is → *Pello friends-of names-post forget Aux(ABS-**DAT**) Ander-DAT)
Meaning: 'Pello forgot the name of his friends' → *'Pello forgot the names of his friends to Ander'.

(Correct structure: *Pellori adiskideen izenak* ***ahaztu*** *zaizkio,*

Lit.: Pello-DAT friends-post. names forget Aux (ABS-DAT)
Meaning: 'Pello forgot the names of his friends.'

— When solely the dative is accepted. E.g.:

Ekin *genion lasterrari* (e.g. from Sarasola 1996) → ****ekin*** *genuen lasterra*

(Lit.: Start Aux (ERG-DAT) run-DAT → *start Aux (ERG-ABS))
Meaning: 'We engaged in the task of running (i.e., we started running)'

— Finally, where the dative is mere addition, but appears very frequently. E.g.:

Lehen saria ***eman*** *zioten* (from Sarasola 1996)

Lit.: first prize give (ABS-DAT-ERG)
Meaning: 'They gave him/her the first prize.'

For these later instances, we will check whether the dative is very frequent in the corpus, and if so, we will mark it as outstanding. Where the dative is a mere addition

and is not frequent will be marked '+'. Of course, when the dative is not accepted we will mark it with '–'. Marking the dative does not imply that it will be reflected in the Auxiliary. Specifically, although the dative is marked with '+' —namely, when it is a mere attachment that is not frequent—, the auxiliary will be marked as either DA (ABS) or DU (ABS-ERG) (we will do the same, or course, when the dative is not accepted). Otherwise, the auxiliary will be marked as ZAIO (ABS-DAT) or DIO (ABS-DAT-ERG).

To summarize, these are the marking options that arise in the auxiliaries and the agreements.

	ABS	ERG	DAT
DA	√/–	–	√/+/–
ZAIO	√/–	–	√
DU	√/–	√/–	√/+/–
DIO	√/–	√	√

This means that, only instances that involve dropping of ergative and absolutive Cases will be considered as variants of an alternation, i.e., as separate *ssv-s* (the remaining Cases have the 'may appear' value signaled by '√').

3.2.5. *Alternations attested in both English and Basque*

As it was mentioned above, Aldezabal et al. (2002) analyze which alternations that have been proposed for English appear in Basque and which are absent. For present purposes, and without entering into details, among the ones that are accepted in Basque, we have selected instances that involve the Cases which were mentioned above as well as those forms that we have considered as lexical. Below is the list of the attested alternations illustrated by examples in English and Basque. The types of alternations are numbered according to the numbers in Levin's work. All these alternations have been marked according to the marking-system that we have suggested above. Since Levin considers the components that take part in the alternations as arguments (and she explicitly signals the ones that are not), we have marked the Case that such components show with '√'.

Here is the list:

Causative/Inchoative alternation; Levin's 1.1.2.1.

Eng. Janet broke the cup/The cup broke

Basq. *Janetek katilua puskatu zuen/Katilua puskatu egin zen*

Substance/Source alternation; Levin's 1.1.3.

Eng. Heat radiates from the sun/The sun radiates heat

Basq. *Beroa eguzkitik irradiatzen da/Eguzkiak beroa irradiatzen du*

Unspecified Object alternation; Levin's 1.2.1.
Eng. Mike ate the cake/Mike ate
Basq. *Mikek opila jan zuen/Mikek jan zuen*
Understood Reciprocal Object alternation; Levin's 1.2.4.
Eng. Anne met Cathy/Anne and Cathy met
Basq. *Annek Cathy topatu zuen/Anne eta Cathy topatu ziren*
Characteristic Property of Agent alternation; Levin's 1.2.6.1.
Eng. That dog bites people/That dog bites
Basq. *Zakur horrek jendeari hozka egiten dio/Zakur horrek hozka egiten du*
Characteristic Property of Instrument alternation; Levin's 1.2.6.2.
Eng. This knife cut the bread/This knife doesn't cut
Basq. *?Labana honek ogia mozten du/Labana honek ez du mozten*
Conative alternations; Levin's 1.3.
Eng. Paula hit at the fence/Paula hit the fence
Basq. *Paulak hesian/-ren kontra jo zuen/Paulak hesia jo zuen*
Locative Preposition drop alternation; Levin's 1.4.1.
Eng. Martha climbed up the mountain/Martha climbed the mountain
Basq. *Paula mendira igo zen/Paulak mendia igo zuen*
With preposition drop alternation; Levin's 1.4.2.
Eng. Jill met with Sarah/Jill met Sarah
Basq. *Jill Sarahekin topatu zen/Jillek Sarah topatu zuen*
Spray/load alternation; Levin's 2.3.1.
Eng. Jack sprayed paint on the wall/Jack sprayed the wall of paint
Basq. **Jackek horman pintura ihinztatu zuen/Jackek horma pinturaz ihinztatu zuen*
Simple Reciprocal alternation (Transitive); Levin's 2.5.1.
Eng. I separated the yolk from the white/I separated the yolk and the white
Basq. *Gorringoa zuringotik bereizi nuen/Gorringoa eta zuringoa bereizi nituen*
Simple Reciprocal alternation (Intransitive);[12] Levin's 2.5.4.
Eng. The oil separated from the vinegar/The oil and vinegar separated
Basq. *Olioa ozpinetik banandu zen/Olioa eta ozpina banandu egin ziren*
Body-Part possessor Ascension alternation; Levin's 2.12.
Eng. Selina touched the horse on the back/Selina touched the horse's back
Basq. (Lit.) *Selinak zaldia ukitu zuen bizkarrean*; (Meaning) *Selinak zaldiari bizkarra ukitu zion/Selinak zaldiaren bizkarra ukitu zuen*
Possessor object; Levin's 2.5.5.
Eng. I admired his courage/I admired him for his courage
Basq. *Bere kemena miresten nuen/Bere kemenagatik miresten nuen*

---

[12] We need to mention that we are unable to distinguish some alternations according to our marking-system, and hence, we have not marked them as distinct *ssv*-s. This applies to Simple Reciprocal alternation transitive' and 'Simple Reciprocal alternation intransitive'. Thus, we have listed them as accepted alternations, but keeping in mind that one variant of the alternation will not be considered as a separate *ssv*.

Attribute Object; Levin's 2.13.1.

Eng. I admired his honesty/I admired the honesty in him

Basq. *Bere zintzotasuna miresten nuen/Beregan zintzotasuna miresten nuen*

Possessor and Attribute alternation; Levin's 2.13.3.

Eng. I admired him for his honesty/I admired the honesty in him

Basq. *Bere zintzotasunagatik miresten nuen/Beregan zintzotasuna miresten nuen/Bere zintzotasuna miresten nuen*

Possessor subject (transitive); Levin's 2.13.4.

Eng. The clown amused the children with his antics/The clown's antics amused the children

Basq. *Pailazoak bere bihurrikeriekin haurrak entretenitu zituen/Pailazoaren bihurrikeriek haurrak entretenitu zituzten*

Time Subject alternation; Levin's 3.1.

Eng. The world saw the beginning of a new era in 1492/1492 saw the beginning of a new era

Basq. *Munduak aro berri baten hasiera ikusi zuen 1492an/1492k aro berri baten hasiera ikusi zuen*

Abstract Cause Subject alternation; Levin's 3.4.

Eng. He established his innocence with the letter/The letter established his innocence

Basq. *Bere inozentzia gutunaren bidez frogatu zuen/Gutunak bere inozentzia frogatu zuen*

Cognate Object construction; Levin's 7.1.

Eng. Sarah sang/Sarah sang a ballad/Sarah sang a song

Basq. *Sarah-k abestu egin zuen/Sarah-k balada bat abestu zuen/Sarah-k abesti bat abestu zuen*

3.2.6. *Selecting verbs*

The first task in analyzing verbs involves a selection of a set of verbs. For this purpose, we have made use of the Statistical Corpus of the XX. Century (i.e., *XX. mendeko euskararen corpus estatistikoa*). After selecting a sample of 22.000 words from the corpus, we have listed verbs according to degree of frequency in which they appear (overall 622 verbs), and, from this list, we have finally selected 100 verbs. We first present the criteria that we have followed for excluding verbs.

Excluding verbs that involve a clear derivational process

The list of selected verbs includes no verb involving clear and productive derivational processes. In section 3.1 of this article, where we described the proposal by Levin, we have argued that there are syntactic structural differences between a derived verb and its non-derived counterpart, where both contain parallel semantics. It is clear that they are syntactically distinct, and hence, along the lines of Levin's methodology, they are not syntactically comparable. We also mentioned that, in our view, Levin is not consistent in using her own methodology (among others, in cases where derivations are involved). However, this does not imply that we have initially discarded her methodology. Thus, we have excluded verbs that involve derivational processes, albeit

acknowledging the systematic process in them. We have preferred to analyze the general structure of verbs that involve no derivational process, and we leave the analysis of derived verbs based on general structures for future research. In fact, although these derivational processes are systematic, we believe that there is underlying complexity in the system (for instance, considering predicates such as *sartu* ('to put in') and *poltsikoratu* ('to pocket'), predicate *sartu* accepts the ablative Case —specially when expressing path— in addition to the adlative; in contrast, predicate *poltsikoratu* hardly accepts the ablative case). Thus, these are topics that require deeper research.[13]

Moreover, derivational processes are not sometimes very explicit; often, it is difficult to detect the components that take part in the composition of the verb, probably, because their birth is long back in history. For this reason, we have decided to exclude the following from our research. On the one hand, the clear and systematic derivational cases that we found in Basque in analyzing the verbal classes suggested by Levin, namely the forms composed of the following morphemes: *-etsi* (as in *onetsi* ('to accept'), *handietsi* ('to praise'), *-ztatu* (as in *ureztatu* ('to water'), *irineztatu* ('to flour'), *-ratu (*as in *poltsikoratu* ('to pocket'), *botilaratu* ('to bottle'), *-katu* (as in *mailukatu* ('to nail'), *kolpekatu* ('to hit'), and *-gabetu (*as in *hezurgabetu* ('to unbone'*), gazgabetu* ('to unsalt'). On the other hand, we have left out most of the derived semantic values that are attributed to suffix *-tu* (some of them also attested in the above analysis) in Gràcia et al. (2000). Specifically, these authors propose 6 interpretations for this suffix:

— Change in state/quality (-tu1, -tu2, -tu3, -tu8, -tu9): *gizondu* ('to become a man'), *izoztu* ('to ice'), *beldurtu* ('to (be) *frighten(ed)'), lotsatu* ('to (be) *embarrass* (ed)'), *zatitu* ('to divide'), *puskatu* ('to break'), *lasaitu* ('to calm'), *garbitu* ('to clean'), *mailakatu* ('to classify'), *lerrokatu* ('to align'), etc.
— Removal (-tu4): *larrutu* ('to skin'), *lumatu* ('to pluck feathers')
— Transmission (-tu7): *babestu* ('to protect'), *zigortu* ('to punish'), *aholkatu* ('to give advice'), etc.
— Change of Location (-tu6, -tu11, -tu12): *baztertu* ('to put aside'), *saihestu* ('to move sideways'), *alboratu* ('to approach'), *kaiolaratu* ('to cage'), *beruneztatu* ('to cover with lead'), *ureztatu* ('to water'), etc.
— Repetition (with some instrument) (-tu10): *mailukatu* ('to nail'), *mokokatu* ('to peck'), etc.
— Location (involving realization of the locus) (-tu5): *lumatu* ('to grow feather'), *hostatu* ('to become covered by leaves'), *loratu* ('to flower')

For our purposes, we have decided to only taken the first values into account.

Excluding verbs that are composed of more than one component

In our process of selection, we have excluded verbs that contain more than one component (e.g., *lo egin* ('to sleep'), *zain egon* ('to wait'), *axola izan* ('to matter'), *ari izan* ('be doing'), *barre egin* ('to laugh'), *bat egin* ('to unite'), *gogora ekarri* '(to remind'), *merezi izan* ('to be worth', etc.). In these cases, the component that appears together

[13] In this book, Odriozola (2003) makes a proposal on the regularities regarding verb derivation in Basque.

with the verb displays a close relation with it, which suggests that the verb and the accompanying component form a semantic unit. However, with respect to our project, the fact that they behave as a single unit produces syntactic structures that usually do not surface when the verb appears in isolation (for example, unlike the verb *ekarri* ('to bring'), the phrase *gogora ekarri* ('remind') accepts subordinate clauses of the *-ela* type. In addition, the element that accompanies the verb is not often the type of element that the verb would take in isolation. For example, in the phrase *hegaz egin* (literally 'wing-with do', meaning 'to fly'), the Case in the accompanying element is instrumental Case. However, as noted by Rodríguez and García Murga (2003), predicate *egin* in isolation does not include the instrumental Case as one of its outstanding Cases. These are some of the reasons that we have taken into account when determining whether a phrase should be considered as a unit or not. Nevertheless, there are units that involve several components where the accompanying element displays a syntactic structure that is compatible with the structure that the verb would take in isolation. In such cases, we have considered such complements as valuable elements of the verb, and the semantic value resulting from the composition must be expressed elsewhere (namely, by considering it as a single unit in the dictionary; this is parallel to the instances of lexicalized units that were described in section 3.2.3.4). However, there is much research that needs to be done on these complex units. It is a hard task to decide what elements belong to the verb itself or to the unit as a whole. We hope that our results serve for future research on this topic.[14]

After applying the above criteria for excluding verbs, let us next present the criteria that we have followed for selecting verbs.

— Frequency. We have selected verbs that display more than 1% frequency in the corpus: *izan*[15] ('to be', 'to have') (20,72%), *egin* ('to do') (6,98%), *egon* ('to be/stay') (4,44%), *esan* ('to say') (2,40%), *ikusi ('to see')* (1,75%), *eman* ('to give') (1,61%), *joan* ('to go') (1,49%), *jarri* ('to place/sit') (1,29%), *aritu* ('to be doing') (1,16%), *hartu* ('to take') (1,12%).

— Verbs that are interesting for our procedure: Among the verbs that display frequency rates lower than %1, we have selected verbs that are interesting for their subcategorization properties as well as for the Cases that they accept. Considering the criteria listed above, we have selected the following 100 verbs as our object of study.

— *abestu* ('to sing')
— *adierazi* ('to express')
— *afaldu* ('to have diner')
— *agertu* ('to appear')
— *ahaztu* ('to forget')
— *aldatu* ('to change')
— *amaitu* ('to finish')
— *argitu* ('to clarify')
— *aritu* ('to be doing')
— *asmatu* ('to figure out')
— *atera* ('to take out')
— *aurkitu* ('to find')
— *baieztatu* ('to confirm')
— *banandu* ('to separate')
— *barkatu* ('to forgive')
— *bazkaldu* ('to lunch')
— *besarkatu* (to embrace')
— *bete* ('to fill')

[14] Zabala (2002) has studied complex predicates. Her claims will be a good point of departure to work on this phenomenon.

[15] *Ukan* ('to have') also displays high frequency *(ukan* 6,34%), but we have subsumed it under *izan* ('to be'). Thus, we have added the frequency rate of *ukan* to the frequency of *izan.*

— *bilakatu* ('to become')
— *bisitatu* ('to visit')
— *dedikatu* ('to dedicate')
— *deitu* ('to call')
— *edan* ('to drink')
— *egin* ('to do')
— *egokitu* ('to adapt')
— *egon* ('to stay')
— *ehizatu* ('to hunt')
— *ekarri* ('to bring')
— *elkartu* ('to unite')
— *eman* ('to give')
— *entzun* ('to listen')
— *erabili* ('to use')
— *eragin* ('to cause')
— *eraman* ('to take')
— *erantzun*('to answer')
— *erre* ('to burn/smoke')
— *erreparatu* ('to notice')
— *esan* ('to say')
— *eskaini* ('to offer')
— *eskatu* ('to ask for')
— *etorri* ('to come')
— *eutsi* ('to hold')
— *existitu* ('to exist')
— *ezkondu* ('to marry')
— *flotatu* ('to float')
— *gainditu* ('to overcome')
— *gertatu* ('to happen')
— *gosaldu* ('to have breakfast')
— *grabatu* ('to tape')
— *hartu* ('to take')
— *haserretu* ('to get angry')
— *hasi* ('to start')
— *hautatu* ('to choose')
— *hautsi* ('to break')
— *hazi* ('to grow')
— *hil* ('to die')
— *hornitu* ('to supply')
— *hustu* ('to empty')
— *igo* ('to raise')
— *ikasi* ('to learn')
— *ikusi* ('to see')
— *irakin* ('to boil')
— *irakurri* ('to read')
— *iraun* ('to last')
— *iritsi* ('to arrive')
— *isildu* ('to quiet')
— *isuri* ('to pour')
— *izan* ('to be')
— *jaitsi* ('to descend')
— *jan* ('to eat')
— *jarri* ('to put')
— *jasan* ('to endure')
— *jaso* ('to raise')
— *jo* ('to hit')
— *joan* ('to go')
— *jokatu* ('to bet')
— *jolastu* ('to play')
— *kezkatu* ('to worry')
— *kokatu* ('to place')
— *konparatu*('to compare')
— *konturatu* ('to realize')
— *landatu* ('to plant')
— *landu* ('to elaborate')
— *laztandu* ('to caress')
— *loratu* ('to flower')
— *lortu* ('to achieve')
— *mintzatu* ('to speak')
— *moztu* ('to cut')
— *mugitu* ('to move')
— *nahastu* ('to mess')
— *onartu* ('to accept')
— *oroitu* ('to remember')
— *otu* ('to occur')
— *pasatu* ('to pass')
— *sartu* ('to enter')
— *topatu* ('to meet')
— *ukitu* ('to touch')
— *ulertu* ('to understand')
— *zeharkatu* ('to cross')
— *zintzilikatu* ('to hang')

4. Conclusions drawn from the analysis of verbs

We have drawn many conclusions after analyzing the 100 verbs in detail. In fact, because the different nature of the verbs —some are semantically heavy, and other are lighter— we have found various relevant phenomena.[16] For present purposes, we will mention three relevant phenomena: first, we will present the difficulties that we encountered in determining which are syntactic variants in a given alternation among the existing *ssv*-s of each verb, and which are not. We will further explain the decisions that we made in such instances. Next, we will briefly present and explain the semantic components that we have employed for distinguishing the *ssv*-s. Finally, we will clarify what we understand by subcategorization, and we will explain the difficulties and phenomena related to the realization of subcategorized elements in sentences.

[16] Further details on the results of the analysis are included in the dissertation research that will be available shortly (Aldezabal, forthcoming).

4.1. Distinguishing between syntactic variants and non-variants in an alternation

Our analysis reveals that some verbs are semantically heavier than others. Typically, semantically loaded verbs tend to have few semantic values, and the *ssv*-s that we have marked involve alternations of the same semantic value. In addition, most of the times they do not allow for alternations. We have found 21 verbs that lack alternations and involve a single semantic value, and 44 verbs that have been assigned more than one *ssv* and contain a single semantic value. Thus, out of 100 verbs, 65 involve a single semantic value. The remaining predicates have the ability to express more than one semantic value, and sometimes we find alternations within those semantic values.

It has not been an easy task to decide on the above facts. In fact, we have been forced to make certain decisions when we have encountered such problems.

This section describes the general problems that we have encountered.

— In the general meaning of some predicates (or better, the meaning that is most frequently attested in the corpus) certain Cases that do not appear to be relevant —usually the inessive— refer to the element in the absolutive, where the later specifies the particular location (versus the location of the event denoted by the verb). Sometimes, this phenomenon becomes relevant to the extent that it seems to induce a new different semantic value. Moreover, the element in the absolutive is different from the usual value of the verb (more specifically, for example, in the usual value of the verb the absolutive element is usually animate, and yet, in the new arising value of the verb, it involves a definite or abstract entity). We have considered these two phenomena (the fact that an element may take force and the fact that the absolutive has different value from the usual verb value) for marking a separate *ssv*. E.g. *etorri-3* ('to include'):

Bigarren liburu honetan ***badatoz****, gainera, aurrekoaren zuzenketak*

Lit.: 'Second book this-in come-they in addition, previous-det-gen corrections'

Meaning: 'This second book includes the corrections of the previous one'.

Elsewhere, in cases where the absolutive is not different from the usual value of verb we have not distinguished a separate *ssv*. For example, *erabili-0*:

Ez nuen aspaldian argazkirik poltsikoan ***erabiltzen***

Lit.: 'not did-I for a long time pictures-partitive pocket-in use-Nominalization-Inn

Meaning: 'I had not used pictures in my pocket for a long time.'

— Sometimes, the presence of certain Cases depends on the object or absolutive element that the verb takes. In such instances, some Case that, for a given verb has previously been considered as unacceptable becomes acceptable. Conversely, a Case that has been acceptable may become unacceptable. E.g.:

Egin ('to do'): adlative and ablative

... eta Artikotik Tropikora bidaia ***egin*** *zuen*

Lit.: 'and Artic-ADL Tropic-ABL trip made did'
Meaning: '...and he made the trip from the Artic to the Tropic.'

Here, the dative and adlative Cases, which are not commonly accepted by this verb are acceptable. Moreover, the dative Case, which is commonly accepted by this verb (with the value goal) is not acceptable. Thus, in such we have not accepted these adlative and ablative Cases, because, they arise as a result of some constraint on the element that is selected by the verb rather than by some constraint on the general value of the verb.

— We found that the semantic value may also be altered by the noun heading the phrase, but without altering other Cases. E.g., in the two examples with the verb *topatu (meet/encounter)* below:

Eskolan gazteleraz irakurtzean hitz arrotz asko ***topatzen*** *genituen.* Value: ENCOUNTER

Meaning: 'At school, we used to come across many unknown words when we were reading.'

Festibalak ***topatu*** *ditu estatu batuar aitabitxiak.* Value: INTENTIONALLY LOOK FOR AND FIND

Meaning: 'The godfather in the USA has found festivals.'

The following may also happen: the semantics of a verb may change according to context —often due to pragmatics— even in cases involving the same item.

Arazoen gainetik irtenbidea ***asmatzeko*** *eskatzen dizue, hala ere, gizarteak, urratsak egitea alegia.*

Meaning: 'However, despite the problems, society demands that a solution be sought.'

Ez da ikerketa sakonik egin eta horrelakoetan beti gertatzen da gauza bera, jendeak ***asmatu*** *egiten dituela gauzak.*

Meaning: 'No serious research has been done, and in such cases, people typically make things up.'

In the above two instance, we know that *irtenbidea* ('solution') and *gauzak* ('things') are usually sought/made up. However, these meanings are provided by context; without context, they would have merely meant 'figure out'. Such differences cannot be expressed by the resources that we have selected. Moreover, they are often determined by pragmatic factors. Thus, they involve further semantic specifications, and hence, we have not considered them as distinct *ssv*-s.

— We have mentioned that some verbs do not have much semantic load, i.e., they contain very little or general semantic information. In such instances, their sem-

antic value in each sentence is provided by the nature of the elements that they take in syntax. When faced with such cases, we have had to make certain decisions. First, we will present casuistry, and next we will specify what we have decided in each instance.

- Various semantic values may sometimes be realized with the same combination of Cases, and the differences are set in the head of the phrase, i.e. in *aldatu-3* ('change'):

  Oñatiko ur-hoditeria Urretxuko ur-biltegitik saihesbidera ***aldatzeko*** *proiektua eta lehendabiziko fasearen egite-lanak enkante bidez kontratatzeko baldintza.* VALUE OF CHANGE OF LOCATION

  Meaning: 'the project to change Oñati's water-pipes from Urretxu's water tanks to the by-pass and the condition to contract the first phase of the works through auction.'

  Izan ere, autonomi edo probintzia-mailara ***aldatu*** *nahi baditugu, zatikaturiko inkestak ez dira lehen bezain adierazgarriak.* VALUE OF CHANGE OF STATE

  Meaning: 'In fact, if we wish to change them into autonomy or a province, the divided surveys are not as meaningful as they were before.'

- Sometimes, the nature of the head of the phrase requires the Case combination to be fixed and syntactically explicit. For instance, *joan-2*:

  Urdailetik irteerara ***doan*** *zentimetroko hodia*

  Meaning: 'The one-centimeter duct that goes from the stomach to the exit.'

In this example, the phrase *zentimetroko hodia* expresses the path, and hence, rather than involving some meaning of movement it refers to its location. For this, it seems that the presence of the ablative or the adlative is necessary.

- Other times, different semantic values are expressed by various Case/value combinations. E.g.: *izan-1, izan-2, izan-4:*

  izan-1: *- Leopoldo, zu idazlea zara, baina zure familian idazle ugari* ***izan*** *dira, horrek zuregan eraginik izan du?*

  Meaning: 'Leopoldo, you are a writer, but there have been several writers in your family, did this have any influence on you?

  izan-2 Hitzarmena ***da*** *bidea*

  Meaning: 'A treaty is the (only) way/solution.'

  izan-4: *Ezer ez dute erraza* ***izan*** *ezta* ***izanen*** *ere*

  Meaning: 'Nothing was easy for them, nor it will be.'

- What changes (or specifies) the semantics of certain verbs is not the noun head of a phrase, but the presence of the phrase itself. For example, *bilakatu-1* and *egokitu-4:*

  Lianak suge ***bilakatu*** *ziren*

  Meaning: 'Whyps became snakes.'

  Gizartearen baloreak ***bilakatuz*** *doaz gizarte horren kontzeptuekin batera*

  Meaning: 'The values of society are developing parallel to the concept of society.'

  Lehen gazteek beraiek ***egokitzen*** *zituzten euren arauak unean uneko egoerara*

  Meaning: 'In the past, young people would determine their rules according to the situations.

  Betaurrekoak ***egokitu*** *zituen*

  Meaning: 'He/she adjusted his/her glasses.'

Considering the casuistry described above, we have decided the following: Those that display the same combination of Cases but change the semantic value according to the head will be included in the same *ssv*. Those that display the same fixed and syntactically explicit combination of Cases will be treated as different *ssv*-s. Those that show different values through different Case-combinations will be considered as different *ssv*-s. Finally, when the presence of a phrase changes/specifies the semantics, the case(s) that belong to the same *ssv* will be marked as optional and outstanding. However, the optionality will be specified in the explanations that will be provided for verbs, not in the marking-system.

This is a generalization of the phenomena that we have found. Yet, in most cases, the problems must be dealt separately in each verb.

4.2. The semantic specifications we have employed in defining the components of the *ssv*-s

We have made use of certain semantic specifications in order to define the most relevant features of each *ssv*. In fact, one of our goals in the onset was to determine such specifications. We may view such semantic specifications as thematic roles, since, in our view, thematic roles are semantic features of verbs, and therefore, they refer to the semantics of verbs rather than to positions and functions of arguments as is usually suggested. Moreover, in Basque, we need to consider that positions are not stable and that they are usually determined by the so-called *Topic*-structure. In addition, the specification of thematic roles has typically been decided in reference to typical or general values of verbs. However, we suggest that a thorough analysis of verbs requires defining various values of verbs, and in order to distinguish between different values, we need to consider additional features. Thus, in view of the procedure that is typically employed in defining thematic roles, we have preferred the term 'semantic specifications' rather than thematic roles.

We have noted that certain semantic specifications are only understood in relation to other semantic specifications. In other words, there is some dependency between certain semantic specifications. For example, if one component of a verb is an *affected_theme* or a *displaced_theme*, the remaining component (of course, in cases where the verb accept the latter) must be *cause*; when one component is *created_theme*, the other will be *producer*; when one is a *container* the other will be *content*. Where there is a *point of departure* there will be a *goal* —or at least it may appear—, and conversely. In contrast, other specifications such as the *experiencer,* the *theme,* and the *activity* do not show any implications.

Thus, it may happen that one element, say the *producer*, may additionally behave as *point of departure* because the sentence may contain some *goal* (when the set of its relevant specifications does not include *point of departure*). Alternatively, it may behave as a *goal* when the sentence includes a *point of departure* (when the set of its relevant specifications does not include any *goal*). After all, depending on the element of the sentence that we choose as target relation, we accept the fact that one component may have more than one semantic specification (the relation with *goal* is *point of departure*, and the relation to *created_theme*, instead, the *producer*).

However, note that these semantic specifications are not directly related to the so-called selectional restrictions. Thus, the semantic specification *cause* does not invariably refer to inanimate entities (in contrast to the definition given for thematic roles, where agents must be animate), or the specification *experiencer* does not imply *affected_object*. The semantic specifications that we have defined are related to the type of event denoted by the verb. Thus, when there is a change of state, we suggest that there is at least a *cause* and an *affected_object* regardless of their animacy. In general, when a predicate is an activity, we have taken the entity involved in the event as being an *experiencer*; it turns out that, in such cases, the entity involved in the event is not only animate but also human. Hence, the specification and assignment of semantic features depends on the way we view the semantics of the verb. Of course, we may view the semantics of verbs in various ways. As for our position, we have considered various viewpoints, and we have created a list of specifications that best fit the resources we have been considering. Only after we have analyzed the 100 verbs have we been able to define the set of specifications, and we have achieved it by basing on the semantics of the 100 verbs —and sometimes the alternations contained in them.[17]

The list of semantic specifications is provided below. However, note that we do not consider the list to be closed, in the sense that other demands may arise when we analyze other verbs in the future. We believe that we have provided an account of the overall casuistry of verbs. At present, the list contains 24 semantic specifications:

—created theme	—target location	—agent	—container
—displaced theme	—target state	—cause	—content
—affected theme	—departure location	—producer	—feature
—theme	—path	—experiencer	—activity
—state	—point of departure	—cause/experiencer	—measure
—location	—goal	—duration	—attitude

[17] For further details see Aldezabal (forthcoming).

We have also been able to specify certain selectional restrictions in some cases, because, in principle, semantic specifications do not have any implications with regards to selectional restrictions. Here is the list we have defined:

— [±biz] (+/-animate) [+giz] (+human)
— [±konkr] (+/-definite) [+lek] (+location)

Actually, we have selected further semantic specifications for defining entities when analyzing the 100 verbs. However, when defining the *ssv*-s in an abstract way, we have restricted to the list provided above.

Here is the list of the types of verbs that we have created based on those semantic specifications:

Verbs of change of state
Verbs of change of location
Verbs that indicate some change
Verbs that involve movement
Verbs that indicate change of psychological state
Verbs that indicate reaction
Verbs that indicate activity
Verbs that involve creation processes
Verbs involving interchange
Existentials, verbs of happening
Verbs that involve a stative location
Verbs that involve description
Verbs that indicate the passing of the time
Verbs that indicate possession
Verbs that indicate attitude
Verbs that indicate assignment of a feature
Opinion verbs

In the above list, certain verbs contain a richer, and hence, more specific information than others (for example, *verbs that indicate some change* vs. *verbs of change of state*; *verbs of change of state* vs. *verbs that indicate change of psychological state*). In fact, verbs that contain a general sense may obtain more specific values. For this task, we need to determine the relation existing between all the elements of the sentence. This is the reason why it is hard to define semantic sets coherently solely based on syntactic structure. In addition, alternations that are general provide a means of grouping verbs coherently and more abstractly (i.e. causative/inchoative alternation: change verbs). However, there are some verbs that contain the semantics carried out by sharing alternations, and nevertheless, do not display such alternation. Finally, there are some semantically similar verbs that do not display any alternations. Thus, there are various ways or parameters for grouping verb: those that share the semantics, those that contain the same number of relevant components, those that employ the same syntactic realization of such components, or those that share the same alternations. These parameters are not exclusive from each other.

We do not consider that Levin's proposal for classifying verbs may provide us with a coherent classification of verbs. Hence, the study of alternations is not enough to develop the decomposition or the internal composition of verbal items.

4.3. Conclusion: subcategorization from our viewpoint

This content of the article thus far shows that there are many difficulties in binding the internal semantic of verbs and their final meaning in sentences. By now, it is obvious that, in order to analyze the semantic value of verbs in sentences, we need to analyze in depth the internal structure of the verbs as well as the interrelation of the elements that make up sentences. This is even more obvious in verbs that are considered as primitives, such as *izan* ('to be'), *egon* ('to be/stay'), *mugitu* ('to move'), *bilakatu* ('to become'), *aldatu* ('to change'), etc.

Hence, we have not proposed specific groups of verbs. Instead, what we have done is to present the *ssv*-s of the 100 verbs we have analyzed (Case/value-combinations, including alternations), and determine the components that are outstanding in our view as well as the semantic specifications of such components. Thus, we will consider that, verbal subcategorization includes all those *ssv*-s, as well as the outstanding Cases of each *ssv*. In fact, Case specifications of components suggest what the syntactic realization of those components will be. However, this does not imply that all the elements that are included in the subcategorization must have a realization in the sentence. Hence, the fact that some element is semantically necessary and the fact that it may not appear syntactically are reflections of distinct phenomena. The next section presents such cases in detail.

4.3.1. *The presence of semantically categorized components in the sentence: unspecification and ellipsis. Dependency between cases*

It is clear that, apart from the Cases that show agreement in the Auxiliary, other elements (inessives, adlatives, ablatives, sociatives, instrumentals and those containing the suffix *-ela*) have also been taken as part of subcategorization in accordance with the semantics of verbs. However, the later, in contrast to the former, do not display agreement in the auxiliary. This hardens the task of determining their presence in the sentence. The next sections describe phenomena related to this issue.

4.3.1.1. *Unspecification and ellipsis*

Sometimes, the reason why a component is not present in the sentence is ellipsis. This is related to the phenomenon of pro-drop, whereby ergative, absolutive and dative elements may be absent in the sentence. However, even if these phrases may be absent, coreference with a previous argument rescues the interpretation that we need.

In contrast, sometimes we face the problem of unspecification. In other words, it is impossible to recover the element that is absent through ellipsis. More specifically, an element that is typical (in Levin's terms) or general (in terms of Vázquez et al.) in a verb, is not syntactically present with the purpose of reinforcing the event denoted by the verb. Sometimes, this object is attached to the lexical item and appears as a *cognate*. This is, in fact, what we find in the Unspecified Object Alternation and in Cognate Object Constructions.

This phenomenon has been widely analyzed in cases where the element is the semantic and syntactic object of the verb (mostly because, despite the presence of agreement suffixes in the auxiliary, there is no phrase in the sentence that may corefer

with such agreement). However, some authors (among others, Vázquez et al.) analyze unspecified cases that express *target location* or *departure location* in verbs such as those expressing displacement (or change of location). In addition, they also analyze cases of ellipsis involving *target* or *departure location* that are recoverable through coreference or some other devices (like deixis). After all, they pose cases parallel to the ones involving semantic and syntactic objects.[18]

The careful analysis of verbs has also revealed that, apart form the unspecification related to typical elements of a verb, there is unspecification that is based on pragmatic knowledge. In such instances, rather than a typical component, what is being unspecified is a specific element that we take as obvious based on our knowledge about the world, and yet, it does appear in the context. E.g.:

> *Lanestosako Herri Eskolan ere ikasle gehienek D ereduan* ***ikasten*** *dute [batxillergoa], izan ere 15 ikasleetatik 11 eredu honetan daude*
>
> Meaning: 'Similarly, in the town school of Lanestosa most students study in the D model [their secondary studies], in fact, out of 15 students, 11 belong to that model.'
>
> *Urduritasunik gabe* ***erre*** *zuen ordea [tabakoa]*
>
> Meaning: 'He/she smoked [cigarettes] with no nervousness.'

Thus, we may assume that we are facing such instances when the elements that we have considered as part of semantic subcategorization are not overtly realized. This is not easy to determine, however, since most of the times we do not know whether we have general unspecification, unspecification due to pragmatic factors, or whether unspecification results from the fact that the unspecified object is attached to the lexical element.

For example, in the case of the verb *konparatu* ('compare'), if the absolutive shows plural number, and if there is no sociative element in the sentence, it seems that, by default, we understand that the action of comparing involves reciprocity; hence, it seems that the lexical item includes this information, and that the sociative has the ability to specify it. The *ssv*-s of *bete-1-3* that belong to the verb *bete* ('to fill'), we know that something becomes full by filling something into it. However, the object that is used for filling may be absent, probably because the information is understood (for instance, a *sack* will be filled by some element that appears in the context, and similarly with objects such as *bottles; questionnaires* will be filled by answers, etc.). In *jarri-1* and *jarri 3 ssv*-s of the verb *jarri* ('to put'), although the outstanding Case is the inessive, sometimes, it is not explicit in the sentence because of the presence of a dative. However, in such cases, we understand that there is an element that is not specified and makes reference to some part of the body, and that the part belongs to the entity in the dative Case. For example, in the example *txapela jarri zion* (literally 'he put the beret'), we understand *buruan* ('in his head') as the locus of where he put the beret, since it is

[18] In addition, note that these authors consider unspecification of elements that denote *departure location* and *target location* as major alternations in what they call 'Trajectory verbs'. They locate verbs that express displacement (or change of location) within this concept of 'trajectory' or 'path'.

customary to put the beret into one's head. Similarly, in the *ssv*-s of a*ldatu-1-2-3-4* ('to change'), in the absence of ablative and adlative Cases, we assume that the change involves some change of state, unless the context forces some other reading. The *ssv jo-4* ('to keep on') typically implies an ablative and adlative, but the later does not usually surface, and where it does, it must be *aurrera* ('on/forward'). In this case, it looks like the item *aurrera* is sometimes included in the verb itself, and in others, it may surface syntactically.

To summarize, in all these cases we need to assume that the understood information is somehow included in the verb, and hence, it should be included and coded in the lexicon.

4.3.1.2. *Dependencies between cases*

In contrast to the examples in the previous section, not all elements that have been considered as involving outstanding Cases can appear in the text as freely. In other words, sometimes it seems that the presence of some Cases depends on the existence of other Cases. For instance, in the *ssv*-s, *pasatu-1-5* ('to pass'), when the ablative expresses the departure location or state of the source, the presence of the target location or state of the source must be explicit. E.g.:

> *...bata, lehen esan bezala, gaztelaniadunen ghettotik gure gizarte katalanera* ***pasatuko*** *direla pertsona batzuk, gazteak bereziki*
>
> Meaning: '...one, as was mentioned before, that several people, specially the young ones, will **pass** from Spanish-speaking ghettoes to our Catalan society.'

The converse does not hold, however. E.g.:

> *Erran diot juristak errandakoa, eta berak oso argi utzi nahi izan dit ni 3. gradura* ***pasatzeko*** *fax-a heltzeko denbora materialik ez dela izan*
>
> Meaning: 'I told him what the jurist said, and he wanted to make it clear to me that there has not been time for the fax that would allow my **passing** to the 3rd grade.

Similarly, in the case of *joan-2,* when the ablative expresses the departure location, the target location must be present, but here, the presence of the adlative forces the presence of the ablative. Consider the following example:

> *Urdailetik irteerara* ***doan*** *zentimetroko hodia*
>
> Meaning: 'The one-centimetre duct that goes from the stomach to the exit.'

However, in these instances of *joan,* we already mentioned that the head of the absolutive phrase has influence on the appearance of the ablative and the adlative.[19]

---

[19] This kind of dependency phenomena is analized in Boons (1987), within the "dependent point of departure" concept.

We need to conclude that much research needs to be done in the domains of contextual ellipsis, pragmatic ellipsis, and unspecification. In turn, this confirms that we need to take into account many complex phenomena when linking the internal structure of lexical items and their syntactic realization.

5. Summary and general conclusion

This work has presented the following. First, it has shown the complex phenomena that are involved in verbal subcategorization. Second, it has presented the line of work that we have developed in our field, i.e., in computational linguistics. It is clear that specifying the subcategorization of each verb is a difficult task due to the following reasons: first, because distinguishing the semantic values and the alternations in each verb is problematic, and second, because of the presence of phenomena such as ellipsis, unspecification (of general and specific elements), and dependencies between Cases.

After the research has been completed, we have defined what we have considered as subcategorization, namely, all the semantic/syntactic value(s) that we have defined for each verb *(ssv),* the set of outstanding elements in each *ssv*, their semantic specifications, and their Case realizations. We have employed various resources in order to define the components that make up subcategorization, and we have tried to provide a coherent proposal based on our resources.

In addition, considering all the phenomena that we have encountered, and along the lines of semantic decomposition, it is clear that we need to consider many features in order to determine the semantic value of predicates in specific contexts as well as to account for the different alternations. In order to complete this task, we would have to look at complex lexicons such as the one suggested by Pustejovsky (1995), and, apart from decomposition, we would have to specify the rules and features that serve in the composition of elements that make up verbs.

We need to point out that there is a big gap between what the current computational approach offers and the demands required by the conclusions of manual analyses. In other words, there is still much work left if we want the computational analyses to achieve the specifications achieved by manual analyses. However, the automatic resources will serve enormously in confirming the conclusions that we have obtained in the areas of combination of Cases, in the nature of the head of the phrase that bears Case, and with regards to outstanding Cases that are not present in the text.

To conclude, our main task has been to explore all these difficulties and to suggest subcategorizations for the initially selected 100 verbs. As we mentioned above, future research will include the confirmation by automatic tools, and at the same time, the analysis of more verbs based on the data we have provided; all these, by applying semi-automatic methods.

References

Aduriz, I., 2000, *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz.* Doctoral Dissertation, Basque Philology Section, University of the Basque Country.

—, I. Aldezabal, X. Artola, N. Ezeiza & R. Urizar, 1996, "Multiword Lexical Units in EUSLEM: A Lemmatiser/Tagger for Basque". In *COMPLEX'96 Forth Conference on Computational Lexicography and Text Research.* Budapest, Hungary.

—, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, K. Gojenola, M. Maritxalar, 1997, "Morphosyntactic disambiguation for Basque based on the Constraint Grammar formalism". *Proceedings of RANLP*.

Aldezabal, I., (forthcoming), *Aditzaren azpikategorizazioa aplikazio konputazionalari begira*. PhD dissertation. University of the Basque Country.

—, O. Ansa, X., Artola, A. Ezeiza, K. Gojenola, J.M. Insausti & Lersundi, M., 1999, (= Aldezabal et al., 1999a). *Euskararen Datu-Base Lexikala (EDBL): eskema berriaren proposamena.* Internal report. Department of Computer Sciences. University of the Basque Country.

—, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández & M. Lersundi, 2001, (= Aldezabal et al. 2001a), "EDBL: a General Lexical Basis for the Automatic Processing of Basque". *IRCS Workshop on Linguistic Databases*. Philadelphia.

—, K. Gojenola & M. Oronoz, 1999, (= Aldezabal et al., 1999b), "Combining Chart-Parsing and Finite State Parsing". *Proceedings of the European Summer School in Logic, Language and Information (ESSLLI)*. Student Session. Utrecht, the Netherlands.

—, K. Gojenola & K. Sarasola, 2000, "A Bootstrapping approach to parser development". *International Workshop on Parsing Technologies (IWPT 2000)*. Trento, Italy.

—, M. Aranzabe, A. Atutxa, K. Gojenola, M. Oronoz & K. Sarasola, 2001, (= Aldezabal et al. 2001b), "Applications of Finite State Transducers to the Acquisition of Verb Subcategorization Information". *Finite State Methods in Natural Language Processing*. ESSLLI Workshop. Helsinki.

—, P. Goenaga, K. Gojenola & K. Sarasola, 2001, (= Aldezabal et al., 2001c). "Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus". *SEPLN'2001*. September 12-14. Jaén.

—, K. Gojenola & K. Sarasola, 2003, "Baterakuntzan oinarritutako euskararen analizatzailea". *Euskalgintza XXI. Mendeari buruz. Euskaltzaindiaren XV. Biltzarra,* September 17-21. Bilbo and Baiona. (forthcoming).

—, M. Aranzabe, A. Atutxa, K. Gojenola & K. Sarasola, 2002, "Learning Argument/Adjunct Distinction for Basque". *ACL'2002 SigLex Workshop on Unsupervised Lexical Acquisition.* July 8-13. Philadelphia.

Alegria, I., 1995, *Euskal morfologiaren tratamendu automatikorako tresnak*. Doctoral Dissertation. Language and Computational Systems Section, University of the Basque Country.

Arriola, J.M., 2000, Hauta-Lanerako Euskal Hiztegi-*ko informazio lexikalaren erauzketa erdi-automatikoa eta bere integrazioa sistema konputazional batean*. Doctoral Dissertation, Basque Philology Section, University of the Basque Country.

Azkarate, M. & P. Altuna, 2001, *Euskal morfologiaren historia*. Elkarlanean.

Boons, J.P., 1987, "La notion sémantique de déplacement dans une classification syntaxique des verbs locatifs", *Langue Française* 76: 5-40.

Chomsky, N., 1965, *Aspect of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

—, 1981, *Lectures on Government and Binding. The Pisa Lectures*. Berlin, New York: Mouton de Gruyter.

Dowty, D., 1991, "Thematic protoroles and argument selection", *Language* 67: 547-619.

Euskaltzaindia, 1993, *Euskal gramatika laburra. Perpaus bakuna*. Bilbo: Euskaltzaindia.

Garai, K., I. Ibarretxe, 2002, "'...*-tik* ...*-ra*' egituraren azterketa". Euskararen Semantika eta Pragmatika Jardunaldiak (ILCLI).

Gojenola, K., 2000, *Euskararen sintaxi konputazionalerantz*. Doctoral Dissertation. Language and Computational Systems Section, University of the Basque Country.

Gràcia, M.T.; Azkárate, M.; Cabré, M.T.; Varela, S. *et al.,* 2000, *Configuración morfológica y estructura argumental: léxico y diccionario*. Resultados del proyecto de investigación DGICYT, PB93-0546-C04. University of the Basque Country.

Grimshaw, J., 1990, *Argument Structure*. Cambridge, Massachusetts: MIT Press.

Gruber, J., 1965, *Studies in Lexical Relations*. Bloomington: Indiana University Linguistics Club.
Hale, K.L. & S.J. Keyser, 1987, *"A view form the middle"*, *Lexicon Project Working Papers 10*. Center for Cognitive Science. Cambridge, Massachusetts: MIT Press.
Jackendoff, R.S., 1972, *Semantic Interpretation in Generative Grammar*. Cambridge, Massachusetts: MIT Press.
—, 1990, *Semantic Structures*. Cambridge, Massachusetts: MIT Press.
Jones D., R. Berwick, F. Cho, Z. Khan, K. Kohl, N. Nomura, A. Radhakrishnan, U. Sauerland & B. Ulicny, 1994, *Verb Classes and Alternations in Bangla, German, English, and Korean*. Massachusetts Institute of Technology center for Biological and Computational Learning and the Artificial Intelligence Laboratory.
Karlsson, F., A. Voutilainen, J. Heikkila & A. Anttila, 1995, *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
Koskenniemi K., 1983, *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. Doctoral Dissertation. University of Helsinki.
Levin, B., 1993, *English Verb Classes and Alternations. A preliminary Investigation*. Chicago and London. The University of Chicago Press.
Odriozola, J.C., 2003, "Verb-deriving processes in Basque", this volume, 185-222.
Pustejovsky, J., 1995, *The Generative Lexicon*. Cambridge, London: MIT Press.
Rodríguez, S. & F. García Murga, 2003, *"Predicados 'sustantivo + egin' en euskara"*. In J.M. Makazaga & B. Oyharçabal (eds.), *Euskal gramatikari eta literaturari buruzko ikerketak XXI. mendearen atarian. Gramatika gaiak,* Iker-14 (I), Euskaltzaindia, Bilbao, 417-436.
Rothstein, S., 1992, "Case and NP Licensing", *Natural Language and Linguistic Theory* 10, 119-139.
Saint Dizier, P., 1995, "A semantic classification of French verbs based on B. Levin's approach". Research report. IRIT.
Sarasola, I., 1996, *Euskal hiztegia*. Kutxa Fundazioa. Donostia.
Shieber, S.M., 1986, *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes 4, Stanford.
Speas, M., 1990, *Phrase Structure in Natural Language*. Kruwer, Dordrescht.
Talmy, L., 1985, *"Lexicalization patterns: semantic structure in lexical forms"*. In Shopen T., ed., *Language Typology and Syntactic Description*. Cambridge University Press: 57-149.
Tapanainen, P., 1996, *The Constraint Grammar Parser CG-2*. Publications of the Department of General Linguistics, 27. University of Helsinki.
Taulé, M., 1995, *Representación verbal en una base de conocimiento léxico*. Doctoral Dissertation, Barcelona.
Urkia, M., 1997, *Euskal morfologiaren tratamendu automatikorantz*. Doctoral Dissertation, Basque Philology Section, University of the Basque Country.
Van Valin, R.D., 1993, *Advances in Role and Reference Grammar*. Amsterdam: John Benjamins Publishers.
Vázquez, G., A. Fernández & M.A. Martí, 2000, *Clasificación verbal. Alternancias de diátesis*. Quaderns de Sintagma 3. Edicions de la Universitat de Lleida.
Zabala, I., 1993, *Predikazioaren teoriak Gramatika Sortzailean (Euskararen kasua)*. Doctoral Dissertation. Basque Philology Section, University of the Basque Country.
—, 2002, "Predicados complejos vascos". Manuscript.