Acquiring Verb Subcategorization from Spanish Corpora

Grzegorz Chrupała

grchrupc7@docd4.ub.edu
Universitat de Barcelona
Department of General Linguistics
PhD Program "Cognitive Science and Language"

Supervised by Dr. Irene Castellón Masalles

September 2003

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The principal goal of the present study is to review the state-of-the-art in both theoretical and applied research on the phenomenon of verb subcategorization and such related issues as diathesis alternations and verb classification systems. Furthermore we set out to assess the progress in, and the perspectives of, the effort to automatically acquire verbal subcategorization frames from linguistic corpora. We review existing research on methods of acquisition developed for English and propose to evaluate how well similar methods can be applied in the context of Spanish. To this end we implement a small-scale experimental system for extraction of subcategorization frames from Spanish partially parsed corpora and experimentally assess its performance.

In chapter 2 we discuss the approaches to verb subcategorization in some major linguistic theories. We briefly sketch the principles behind each of the theories discussed and their major contributions to the understanding of the combinatorial properties of verbs.

The theories we cover are Government and Binding, Categorial Grammar, Lexical-Functional Grammar, Generalized Phrase Structure Grammar and Head-Driven Phrase Structure Grammar. We finally make some observations on the differences in how these various theories account for verb subcategorization, with special emphasis on the treatment of subjects.

In chapter 3 we focus on a specific aspect of verbal subcategorization: diathesis alternations. We explain what is meant by diathesis and what diathesis alternations are in general, and then we proceed to describe in some detail this phenomenon in Spanish and the account given of it by Vázquez et al. (2000). We discuss how these authors explain diathesis alternations in terms of underlying changes in the conceptualization of the event being described.

In chapter 4 we present different approaches to verb classification. We mention the relevance of semantic decomposition and the we proceed to discuss systems such as Levin classes, intersective Levin classes and the classification proposed by Vázquez et al. We then describe lexicographical databases such as WordNet, VerbNet and FrameNet.

In chapter 5 we tackle the applied issues central to the present investigation, i.e. verb subcategorization acquisition. We describe motivations for this effort as well as the problems involved in acquisition of information from linguistic corpora. We then discuss the different methods used for evaluating the performance of acquisition systems, and finally describe research that has been done in this area to date. We describe the progress in the field since its beginnings and notice the relative maturity of this line of research and the related technology for the English language.

Finally in chapter 6 we describe our own exploration of the possibilities and problems involved in acquiring subcategorization from Spanish corpora. We describe the rationale and methodology of the project and explain the issues behind some design choices. We go on to present some implementation details and the functioning of the system in general. We then describe the experimental evaluation of the system and assess the perspectives for subcategorization acquisition in Spanish, indicating some potentially interesting directions of further research.

Chapter 2

Verb Subcategorization in Linguistic Theory

2.1 Introduction

Subcategorization is the word traditionally used to refer to the subdivision of major syntactic categories, particularly verbs, according to what other constituents they co-occur with. Thus, the category of verbs can be split into subcategories such as transitive, intransitive, ditransitive and other kinds of verbs based on the number and type of syntactic arguments these verbs require. What we normally think of as a single verb may belong to more than one subcategory, that is it may appear in different syntactic pattern. This pattern is called the **subcategorization frame** (SF) and can be described as the order and category of the constituents co-occurring with the verb in question. Thus, in English, a verb such as *give* occurs in the slots in one of the following subcategorization frames: `NP _ NP NP or NP _ NP PP`.

The subcategorization of a lexical item is one of the most important pieces of information associated with it. It is vital for both theoretical linguistics and in practical applications. It is indispensable in computational lexicons if they are to be useful for natural language processing. Parsing can be greatly enhanced by providing the parser with lexical entries of verbs containing detailed information on their combinatorial properties, i.e. their subcategorization frames.

The treatment of subcategorization varies across linguistic theories. In the following sections we will offer an overview of the different approaches and compare their relevance for subcategorization acquisition.

2.2 Government-Binding and related approaches

In this section we will briefly outline the treatment of verb subcategorization within the Government-Binding (GB) framework. Government-Binding

Theory was developed by Chomsky and others in 1980's and built on the previous work within Transformational Grammar. It has possibly been the most influential brand of generative theory in theoretical linguistics.

One of the central parts of GB is the **X-bar theory**. This theory is an attempt to factor out similarities among different phrase structures. Looking at the makeup of different phrases in a language, common patterns are discernible: for example the relative ordering of verbs and their objects, and prepositions and their objects, tend to be the same within a particular language. X-bar theory generalizes these commonalities and proposes a scheme that accommodates all or most of structures in a language.

The version of the template specified by the X-bar theory consists of three levels, corresponding to nodes designated as X", X' and X, where X stands for a lexical head. The X' node has as its daughters the lexical head and its arguments: the constituents that the head subcategorizes for. The X" node is mother to constituents that act as specifiers (such as determiners) or modifiers (such as adjectives or non-argumental prepositional phrases).

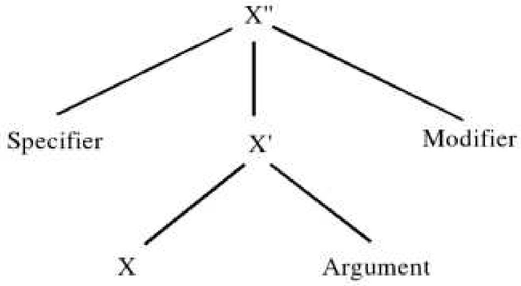

Figure 2.1: X-bar structure for English

The X" node (also known as XP) is called the **maximal projection** of the lexical head. For example VP is the maximal projection of V. The relative ordering of constituents is not actually specified in the X-bar theory, but rather is accounted for by independent principles of grammar. What is important is the hierarchy of projections.

This same scheme is applied to sentences, although the naming conventions are violated in this case. The maximal projection is often referred to as S', the intermediate projection is S and the head is, depending on the version of the theory, one of abstract constituents such as INFL (for inflection).

The four basic categories N, V, A and P are defined in terms of pairs of binary features V (verbal) and N (nominal), which arguably accounts for certain generalizations that apply across categories.

Table 2.1: Binary features of basic categories

	[+N]	[-N]
[+V]	A	V
[-V]	N	P

In GB the role of phrase structure rules is assumed by the combination of X-bar templates and subcategorization frames of heads. In principle, X-bar theory allows arguments to be any maximal projection. It is the subcategorization frames of heads that act as a filter to rule out ungrammatical sentences such as **John gave Mary.*

An important feature of GB is the fact that *subjects are not subcategorized for* by the verbal head. The domain of subcategorization is limited to the maximal projection containing the head. In GB subjects are typically outside of VP, i.e. they are not sisters to the verbal head. This leads to GB predicting a number of subject/object asymmetries in syntax (Sells, 1985).

2.3 Categorial Grammar

The group of grammar formalisms collectively know as Categorial Grammar descends from a tradition different from that of phrase-structure grammars. Its roots are in philosophy of language and formal logic. Work by theorists such as Ajdukiewicz, Montague and Ben-Hillel laid the foundations of this theory.

According to Bach, (after Wood (1993)) there are three basic principles underlying the apparent diversity of theories within the CG paradigm. Firstly, language is analyzed as consisting of functions and arguments rather than phrase structures. Unlike phrase-structure grammars, which are configurational, CG is a functional-type formalism.

Secondly, CG insist on a close correspondence between syntax and semantics: a syntactic description of a linguistic unit also carries its compositional semantics.

The third characteristic feature of CG is its monotonic nature. It is averse to posit abstract devices such as movement or transformations common in GB-type theories.

In CG the concept of rules of grammar, conceived of as separate from lexical items, is superfluous. The combinatory properties of words are encoded directly in lexical entries; thus CG is a radically lexicalist theory.

There are many different notations even for basic, unextended CG in current use. Here we will present the principles of CG in the system used by Steedman. In the following account, based on (Wood, 1993), we present

CG at its simplest. This version only postulates two atomic categories derived from the two central concepts in the philosophy of language. These are names of entities (**e**) and propositions carrying truth values (**t**). These two concepts are represented in more linguistic approaches as **N** for *name* and **S** for *sentence* respectively. The above atomic categories are 'complete' or 'saturated'. Other categories need other expressions to complete them. These incomplete categories can be seen as functions from the missing expressions, i.e. their arguments, to the categories resulting from the combination with the arguments. For example, an intransitive verb such as *walks* needs one argument, a name of an entity, such as *Mary* to form a complete expression (sentence) *Mary walks.* On this view the category of intransitive verbs in the notation used here would be S\N, with the first symbol, S denoting the result, the symbol N denoting the argument needed and the direction of the slash indicating relative word order: the argument must appear to the left of functor. A phrase such as *likes ice-cream*, whose combinatory properties are the same as those of *walks* would have the same category S\N. The transitive verb *likes* is then that category which, when completed by an NP to the right, and then completed by another NP to the left of the resulting expressions, forms a sentence; in CG notation it comes out as (S\N)/N.

There is no category corresponding to the notion of verb; different verb and VP's belong to different complex categories. This results from the radical lexicalism of CG, and also has the undesired effect that it becomes difficult to express generalizations about inflectional patterns and the like.

Another part of CG is the set of rules that make it possible to decide on the grammaticality of a sentence and derive a semantic interpretation for it. The most basic operation is the application of a functor to its arguments – i.e. combining an non-saturated category with a preceding or following category to form the 'result' category. For example:

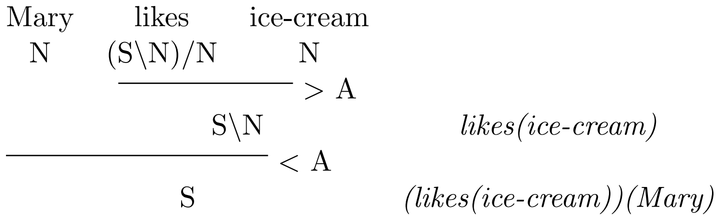

Figure 2.2: Derivation of a sentence in CG

Each operation of function application is underlined and annotated with the rule used and its direction. Semantics is likewise built by function application (A for Application in figure 2.3). This is how core CG works: it is not entirely adequate for human language and so a variety of extensions have been proposed. The minimal set of two atomic categories is commonly augmented: for example for more sophisticated treatment of nouns and determiners, a distinction is made between common nouns (CN or N) and proper nouns/noun phrases (PN or NP). Extensions are also made to the set of rules used in deriving a sentence. Lambek calculus is a classical set of such rules. Some rules are binary, i.e. they combine categories, others are unary, and permit to convert one category into another. Apart from (1) function application, exemplified above, Lambek introduced (2) associativity, (3) composition and (4) raising.

Another extension to core CG regards the use of Attribute-Value Matrices (AVMs) and **unification** for representing complex feature bundles associated with categories. AVMs and unifications are discussed in more detail in the following sections on GPSG and HPSG.

Ideas from CG have influenced developments in other theories, for example in GPSG and HPSG. GPSG uses a feature called SLASH in its account of unbound dependencies such as topicalization and WH-constructions. A category with this feature, C[SLASH C'], also written as C/C', is a constituent of type C, from which a subconstituent C' is missing, which is analogous to how non-atomic categories work in CG.

In HPSG the mechanism of arguments being 'canceled off' non-saturated categories is analogous to the way in which arguments are removed from the SUBCAT (or COMPS) list of the head in the process of building a headed phrase (except that HPSG allows non-binary branching). In CG the idea that the ways in which linguistic units can combine with each other is totally specified in the categories associated with lexical items is taken to its logical conclusion. Other approaches have made use of this fundamental insight in their own treatment of subcategorization.

2.4 Lexical-Functional Grammar

The Lexical-Functional Grammar was developed by Ron Kaplan and Joan Bresnan. As its name indicates, is espouses lexicalism. Phenomena treated in GB by means of Move-α, such as passivization, are dealt with by lexical rules which specify the relation between the active and passive forms of verbs. LFG, unlike GB and like all the other approaches discussed in this chapter, is a monostratal, transformation-free theory.

The LFG model of syntax consists of two parts, the **c-structure** and the **f-structure**. The first encodes such interlinguistically variable properties as word order and phrase structure.

F-structure, on the other hand is meant to be fairly stable across languages and to express the relations between the functional constituents of a phrase. Those constituents are **grammatical functions** such as SUBJ (subject), OBJ (object), or XCOMP (open complement). Thus LFG accords theoretical, primitive status to the notion of grammatical function, which GB treats as reducible to phrase structures. Although c-structures, together with the lexicon, determine the f-structures, there is no direct mapping from c-structures to f-structures, and each obey their own specific constraints.

F-structures are built based on information from two sources. One are **functional annotations** associated with c-structures. For example:

1. $$\begin{array}{cccc} \text{S} & \rightarrow & \text{NP} & \text{VP} \\ & & (\uparrow \text{SUBJ}){=}\downarrow & \uparrow{=}\downarrow \end{array}$$

The arrows in the annotation refer to the function of the annotated constituent. The up-arrow means that the function refers to the mother of the node while the down-arrow indicates the node itself. So the first NP annotated as (↑ SUBJ) means that this NP is the SUBJ of its mother, i.e. the S, or more precisely, that the f-structure carried by the NP goes to the S's SUBJ attribute. Similarly, the VP's annotation (↑=↓) indicates that the VP's f-structure is also S's f-structure – which can be paraphrased as VP being the functional head (Sells, 1985).

The other source of information is the lexicon. A simplified lexical entry of a verb would look as the following:

2. $$\begin{array}{lllcc} \textit{paint} & \text{V } (\uparrow \text{PRED}) = \text{'paint} & < (\uparrow \text{SUBJ}) & (\uparrow \text{OBJ}) > \text{'} \\ & & | & | \\ & & \text{Agent} & \text{Theme} \end{array}$$

The category of the lexical item is indicated (V) as well as its semantics and subcategorization information. After the lexical entry combines with inflectional morphemes, information about tense, person, etc. is added. Lexical forms subcategorize for forms rather than categories. This allows for non-standard categories to realize functions in a sentence (e.g. non-NP subjects, cf. Sells (1985, ch.4)). Functions are also linked to arguments of the **Predicate-Argument Structure**. In (2) above, the SUBJ function is linked to the Agent role and the OBJ to Theme. In contrast to GB, in LFG subject forms part of the verb's subcategorization frame.

2.5 Generalized Phrase-Structure Grammar

The Generalized Phrase-Structure Grammar was developed by Gerald Gazdar and others in the 1970s and 1980s. More recently it mutated into Head-Driven Phrase-Structure Grammar, which will be discussed in the following section. In the present section we will take a closer look at subcategorization in the original GPSG (Gazdar et al., 1985).

One of the motivations for the development of GPSG was reaction to Chomsky's claim that adequate treatment of human language could not be achieved with phrase structure grammars. This claim justified the use of transformations and multistratal theories of grammar. GPSG is an attempt to extend traditional phrase structure grammars so they can handle the phenomena that only transformations were supposed to be able to explain. This theory also emphasized the necessity of formalization. Thanks to its simple monostratal architecture and the formal notation it introduced, it was much easier to implement computationally than theories such as GB.

Even though GPSG started out as an augmented phrase-structure grammar, in its mature version it does not have phrase-structure **rewrite** rules. Instead these are replaced by **immediate dominance rules**, or ID-rules, that indicate the tree hierarchy of constituents but not their relative order. The ordering is described by **linear precedence statements**. This is more economical and flexible than traditional rewrite rules, which collapse both sorts of information, in that it factors out redundancy and allows for languages with freer word-order than English.

In GPSG a category is a set of feature-value pairs. For example the category traditionally represented as NP corresponds to the following set:

3. {<N,+>,<V,->,<BAR,2>}

For the category N, the feature-value set would be similar but the BAR feature would be 0.

Features in GPSG can have either atomic values or values that are themselves feature-value sets. One such feature is AGR (agreement).

4. {<AGR,{<N,+>,<V,->,<BAR,2>,<NUM,3>,<GEND,FEM>,<PLU,->}>}

The above notation indicates agreement with a 3rd person feminine singular NP.

The BAR feature corresponds to the bar-level concept in X-bar Theory, which GPSG adopts. One important difference between the basic X-bar scheme as found in GB and the one used by GPSG is the fact that in the former the S is the projection of V rather than of an abstract category such as INFL. Abstract categories are unavailable and undesirable as a consequence of GPSG being a monostratal system.

In GPSG subcategorization frames of verbs are implemented by means of the feature SUBCAT whose value is an integer corresponding to an IP-rule describing the structure in which they are inserted. This feature is encoded in lexical entries: multiple frames mean multiple entries in the lexicon. As an example consider the lexical entries in 5.

5. (a) <*weep*,[[-N],[+V],[BAR 0],[SUBCAT 1]],{*slept*},**sleep'**>
   (b) <*devour*,[[-N],[+V],[BAR 0],[SUBCAT 1]],{},**devour'**>

6. (a) VP → H[1]
   (b) VP → H[2], NP

The value of the SUBCAT feature in these entries references the ID rules in 6. As a consequence, the verb *sleep* can only appear in trees where it is the only daughter of VP. On the other hand, *devour* must have an NP as a sister, thus assuring its correct behavior as transitive verb. The rules in 6 are denominated **lexical** ID-rules. They are characterized by the fact that they introduce a lexical head – this is apparent in the category H being annotated with an integer corresponding to the value of SUBCAT in verb lexical entries.

There are also other rules, which do not provide arguments for lexical heads. For example:

7. (a) S → X^2, H[-SUBJ]
   (b) NP → Det, N^1

The first rule states that an S can consist of a [BAR 2] phrase and a VP. The second one says an NP is made up of a Det and a [BAR 1] N category. This kind of rules that do not refer to the value of SUBCAT in lexical entries are called *non-lexical* ID-rules.

Another important notion in GSPG are *metarules*. As the name indicates, these are rules that take rules as their input and produce other rules as their output. They extend the basic phrase structure grammar. Metarules in GPSG are used, for example, to derive rules licensing passive sentences from those that describe active ones. Their use permits to factor out redundancy that would otherwise be present in the grammar, and also provides a principled treatment of regular correspondences apparent between active and passive constructions.

As a consequence of the fact that SUBCAT indexes verbs into immediate dominance rules, heads only subcategorize for their sisters. This in turn means that subjects are *not* subcategorized for, as they are not immediately dominated by VPs; rather, as can be seen in 7a, subjects appear in non-lexical rules. The verb, however, still plays a pretty central role in GSPG: sentences are 'maximal projections' of verbs in terms of X-bar theory.

2.6 Head-Driven Phrase-Structure Grammar

HPSG is an eclectic theory of grammar combining insights from a variety of sources, most notably GPSG, CG and GB. Like GPSG it stresses the importance of precise formal specification. The theory uses typed feature structures in order to represent integrated linguistic signs. The types are described by means of a multiple inheritance hierarchy, which helps avoid redundancies.

HPSG is more lexicalist than most other theories. Most linguistic information is contained in the lexical entries. The remainder is specified in principles and rules of a very general kind. Syntax and semantics are not largely independent, as in the approaches described above, but rather are tightly integrated in the same framework. The semantic component of HPSG is based on situation grammar (Barwise and Perry, 1983).

In HPSG subcategorization, information is specified in lexical entries. In the standard version of the theory, as exposed in Pollard and Sag (1987) and Pollard and Sag (1994), the subject is treated in a way similar to other arguments. Verbs have a SUBCAT feature whose value is a list of **synsem** objects corresponding to values of the SYNSEM features of arguments subcategorized for by the head. The order of these objects corresponds to the relative **obliqueness** of the arguments, with the subject coming first, followed by the direct object, then the indirect object, then PPs and other arguments.

In Chapter 9 of Pollard and Sag (1994) the authors present a revised version of the theory, where subject and non-subject arguments are treated differently. This revision was motivated by a series of technical arguments put forward by Borsley, who argues that a singled-out subject accounts for various data (mainly from English and Welsh) in a more parsimonious way. The phenomena he discusses include simplifying the notion of possible non-head, subcategorization of non-predicative prepositions and blocking subject traces, among others.

In their revision the authors propose three different features to replace SUBCAT, namely SPR (SPECIFIERS), SUBJ (SUBJECT) and COMPS (COMPLEMENTS). Below we present the treatment of verbal subcategorization in Sag and Wasow (1999), which is simpler in that only two out of these three features are used: SPR and COMPS.

Non-subject arguments (**complements**) are specified by the COMPS feature. Its value is an ordered list of feature-structure descriptions corresponding to the complements taken by the verb. So, for example, for an intransitive use of a verb such as *bajar* (as in *Los precios de la fruta han bajado*), the value of the COMPS feature would be an empty list. On the other hand, for the transitive meaning of this same verb (as in *La frutería ha bajado los precios*) it would be a one-element list, its sole item specifying an NP argument. One of the generic rules, the Head-Complement Rule (or Schema) [1] assures that when a head combines with its complements, only complements specified in the head's COMPS list will be licensed. One can think of the complements as being removed from the COMPS list in the process of building a headed phrase. After a head has been 'saturated' (i.e. it has combined with all the complements that it subcategorizes for), its

[1] The notion of **rule** in HPSG is not really a separate language construct. Words, phrases and rules are all represented by signs:

> A grammar rule is just a very partially specified sign which constitutes one of the options offered by the language for making big signs from little ones. (Pollard and Sag, 1987)

mother's COMPS list is empty (Sag and Wasow, 1999).

Subject arguments are dealt with in a manner analogous to complements. Subjects are treated as a kind of specifier. Verbs have a SPR (SPECIFIERS) feature, whose value is also a list of feature-structure descriptions. There is a constraint which makes sure that unless otherwise specified by a rule, the SPR and COMPS of the mother are identical to those of the head daughter. Thanks to this principle (known as the Valence Principle) the SPR list in a lexical entry of a verb gets 'propagated up the tree' up to the point when the verb has combined with all its arguments from the COMPS list and is ready to combine with the subject argument. This combination is licensed by the Head-Specifier Rule, similar to the Head-Complement Rule.

As noted above, HPSG uses feature structures to represent linguistics signs. A linguistic sign in HPSG can be loosely thought of as based on the notion of sign proposed by Ferdinand de Saussure (1959), i.e. as a pairing of sound and meaning. In HPSG, each sign has two basic features: a PHON feature, which represents the sound, or phonology of the sign, and the SYNSEM feature which combines syntactic, semantic and pragmatic information. Above we have seen briefly the treatment of syntactic arguments of a verb. These need to be linked in some way to semantic arguments, i.e. the participants of the event denoted by the phrase. In HPSG this is achieved by unifying the feature structure descriptions on the COMPS and SPR lists with feature structure descriptions representing semantic arguments in the set of **predications** that is the value of the feature RESTR (RESTRICTION). Most of the above is brought together in a simplified feature-structure illustrating the transitive meaning of *bajar*.

$$\begin{bmatrix} \mathit{word} & \\ \text{PHON} & \langle \mathit{bajar} \rangle \\ \text{SYNSEM} & \begin{bmatrix} \mathit{synsem - struct} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \mathit{verb} \\ \text{SPR} & \langle \boxed{1}\text{NP}_i \rangle \\ \text{COMPS} & \langle \boxed{2}\text{NP}_j \rangle \end{bmatrix} \\ \text{ARG-ST} & \langle \boxed{1}, \boxed{2} \rangle \\ \text{SEM} & \begin{bmatrix} \text{MODE} & \mathit{prop} \\ \text{INDEX} & s \\ \text{RESTR} & \left\langle \begin{bmatrix} \text{RELN} & \mathit{lower} \\ \text{SIT} & s \\ \text{LOWERER} & i \\ \text{LOWERED} & j \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 2.3: Representation of sign in HPSG

The indices *i* and *j* link the semantic arguments to the syntactic ones. NP_i is shorthand for a feature-structure description of an NP whose SYNSEM | SEM | INDEX has value *i*. The ARG-ST (ARGUMENT-STRUCTURE) feature in Figure 2.3 is present in lexical heads, and its value is the concatenation of the values of SRP and COMPS. This list is used in HPSG's Binding Theory to provide a rank order for all arguments of a head.

2.7 Discussion

We have reviewed the treatment verb subcategorization receives in some major linguistic theories. Notwithstanding their important theoretical differences and technical details, all provide some mechanism whereby verbal lexical items can specify what syntactic arguments they can combine with and how these syntactic arguments are linked to the semantic arguments, i.e. thematic roles.

One important dimension of difference between the approaches discussed is the degree to which the treatment of different sort of arguments subcategorized for is unified or differentiated. One extreme point in this continuum is occupied by LFG, where each of the grammatical functions receives a separate 'attribute' in the f-structure.

Another position is to treat all the arguments in a unified manner, except subjects. This is how most versions of GB and GPSG work. In most versions of GB, verbs don't subcategorize for subjects at all, as subjects are external to VPs, and subject-verb agreement is dealt with by abstract categories such as INFL. In GPSG subjects are also excluded from verbal subcategorization frames (though the AGR feature does link verbs with subjects). In both theories this is motivated by 'a principle of locality', which means that "subcategorization must be satisfied in some local structural domain" (Sells, 1985, p. 88).

Early versions of HPSG attempted to fully unify the treatment given to subjects and other types of complements. They all appear on the SUBCAT list of the verbal head, and the differences in syntactic behavior of different complements are accounted for in terms of their relative position on that list, which reflects their rank order along the dimension of 'obliqueness'.

Such an approach allows for simple and consistent treatment of the verb-subject agreement and the semantic restrictions imposed on subjects by verbs. However, in a number of constraints specific reference was required to the first member of the SUBCAT list, and singling subjects out provides a number of technical advantages for the theory (Pollard and Sag, 1994, ch. 9). So in later versions of HPSG an intermediate solution is adopted, where *all* verb's arguments are gathered in a single ARG-ST list, but subjects and other complements are differentiated by the use of separate SUBJ and COMPS features. It seems that a position similar to the one used in later versions of HPSG integrates conflicting arguments from syntax and semantics and allows a theory to predict an asymmetry in syntactic behavior between subjects and non-subjects, while at the same time taking into account the dependency of subjects on verbal heads.

Chapter 3

Diathesis Alternations

3.1 Introduction

Verbs typically occur in more than one subcategorization pattern and the linking between the syntactic and semantic arguments can vary. Such variations in verb syntax and semantics are often referred to as **diathesis alternations**. Verbs tend to cluster in groups according to the alternations they participate in, and they often share some meaning components. In the following sections we briefly review some of the research done on diathesis alternations, concentrating on Spanish data.The phenomena discussed below are also relevant to verb classification in general, which we will review in the following chapter.

3.2 Diathesis

The concept of **diathesis**, although frequently used in the expression **diathesis alternation**, does not have a universally agreed-upon definition. Sometimes it is treated as the synonym of **voice**. Mel'čuk and Xolodovič (1970) may well have been the first to distinguish between the two terms, using diathesis to mean a more general phenomenon than voice: syntactic realization of verbs' argument structure. Voice is then used to mean specifically the kind of diathesis that affects the morphological form of verbs.

Basically the same definition is adopted by Kharkovsky and Tesnière, but they differ as to their definition of voice. For Tesnière, voice is synonymous with valence, i.e. the number of syntactic arguments a verb adopts. For Kharkovsky and colleagues voice is a specific verb form by which diathesis is realized (Vázquez et al., 2000); we will use here voice and diathesis with these meanings.

3.2.1 Alternations

A verb displays a diathesis alternation if, for the same basic verb meaning, there are alternative ways of realizing the semantic arguments in syntax, or if some of these arguments are not realized. Some typical examples of such alternations follow:

8. (a) Encarni cargó el carro de latas de cerveza.
   (b) Encarni cargó latas de cerveza en el carro.

9. (a) Mabel threw the ball to Damian.
   (b) Mabel threw Damian the ball.

10. (a) Asunción rompió el ordenador.
    (b) El ordenador se rompió.

11. (a) El govierno ha bajado el IVA.
    (b) El IVA ha bajado.

These alternations receive names such as *'load/spray'* or **locative** alternation in 8, **dative** alternation in 9, and **transitive/unaccusative** or **causative/anticausative** alternation in 10 and 11. In English, especially the dative and *'load/spray'* alternations have been studied extensively. It has been noticed that, in these two alternations, the participants in the situation described are the same, and the basic meaning expressed stays the same. There are, however, differences in the details of the semantics. Thus in 8a the trolley would normally be understood to be full of beer cans as the result of the action of loading, whereas in 8b no such entailment or implicature is involved.

In the case of the dative alternation, two subcategorization frames are involved, Propositional Object (PO) in 9a and Double Object (DO) in 9b. In many cases, no clear semantic difference between the two alternatives is detectable. However, there are restrictions on the kind of verbs that accept one or the other frame, as well as restrictions on what kind of entities can appear in the NP slots of the frames, which has led researches to posit different semantic representations for the alternatives of this alternation. As an example, Pinker (1989) proposes the following semantics for the two frames:

13. **Prepositional Object**
    NP_0 CAUSES NP_2 to GO TO NP_1

    Double Object
    NP_0 CAUSES NP_1 to HAVE NP_2

This difference in the representation of meaning between the PO and DO frames is used to explain some of the restrictions observed in their distribution with verbs: for example from 13 it follows that in order for the PO construction to be grammatical, NP_2 must undergo movement. Similarly, in DO the NP_1 must be selectionally consistent with possession. So the meaning representations in 13 account for the following contrasts in grammaticality:

14. (a) The nasty smell gave Egbert nausea.
    (b) *The nasty smell gave nausea to Egbert.

15. (a) Helga sent her daughter to Greece.
    (b) *Helga sent Greece her daughter.

Other researchers have refined Pinker's analysis or proposed alternative explanations (for one such account see Krifka (2000)). Providing an elegant and economic representation in the lexicon of alternative linkings between verb syntax and semantics is a major goal of the research on diathesis alternations.

3.3 Diathesis alternations in Spanish

Naturally, the phenomena of diathesis alternations with all its apparently confusing complexity call for a reductionist account. The question is whether it is possible to explain or coherently classify all the different alternations with the accompanying shifts in meanings by appealing to some more basic feature or set of features that interact to produce the observed alternations. Below we review the proposal presented in Vázquez et al. (2000). These authors argue that diathesis alternations are reflection of changes in the conceptualization of the event or state that is denoted by the verb and its arguments.

In the specification of the semantics of diathesis alternations a hierarchy of 'meaning components' is used. The authors tier those components on three levels, along the dimension of diminishing generality. Thus the ones on the first level are shared by all verbs, while those lower down in the hierarchy are progressively less universal.

Level 1 time, space, entity

Level 2 property, initiator, manner

Level 3 change, trajectory ...

Entity is that element which the predication is about. The **initiator** collapses the more familiar notions of 'agent', 'experiencer' and similar roles.

The **property** component is that which is being asserted of the **entity** in stative constructions. The components on the third level result form the semantic decomposition of specific groups of lexical items.

The authors discuss two basic types of oppositions involved in diathesis alternations. The first one is a change in the way the described event is conceived of. The other one involves cases where a given verb can describe either events or states.Within the first of these oppositions, **change of focus** and **underspecification** are further distinguished. The second group of oppositions (aspectual) comprises the **resultative** and **middle**, and the **personal temporally unmarked stative** constructions.

3.3.1 Change of focus

The authors consider sentence-initial elements to be focalized. This is somewhat surprising, as under standard assumptions the more typical position for material under focus is sentence-final (e.g. Jackendoff (2002, sect. 12.5)). However, the exact definition of what constitutes focus does not seem to affect the general argument, which is: Changes to the syntax-semantics mappings that alter the position of the constituents linked to specific participants in the event will affect those participants' saliency in the discourse, and thus have a direct effect on the **information structure** of the sentence. The speaker stresses the increased relevance of some aspect of the event at the cost of others.

Three diathesis alternations are identified as being due to focus change: **causative/anticausative**, **holistic** alternation and **inversion**.

Causative/anticausative

In this alternation the 'focus change' affects the initiator. The two alternatives in the alternation involve: (1) expressing the initiator ('cause') in the subject position and (2) omitting the initiator from the overt syntactic form or expressing it by means of a prepositional phrase. For example:

16. (a) La retirada del activista conservador Gary Bauer redujo el pelotón de aspirantes presidenciales republicanos a cuatro.
    (b) El pelotón de aspirantes presidenciales republicanos se redujo a cuatro debido a la retirada del activista conservador Gary Bauer.

The two poles of the alternation can be instantiated in several Spanish-specific constructions. These are briefly presented below.

The prototypical causative construction involves a causal or agentive initiator (those two differ according to the degree of intentionality they display). The 'causativeness' can be expressed synthetically (a) or periphrastically (b):

17. (a) El fuego arrasó 50 hectáreas de bosque.

(b) El calor hizo sudar a Nina.

The anticausative group of constructions is characterized by the initiator being either absent or in a 'non-prominent' position in the structure of the sentence, where by 'non-prominent' the authors mean a non-subject, non-topical position, such as a sentence final prepositional phrase. Only constructions that alternate with a causative equivalent are considered to belong to this category. The various types of anticausatives differ as to the following two features:

1. Type of initiator
   - (a) Cause: **prototypical anticausative**, **anticausative of process**
   - (b) Agent: **passive**
2. Telicity
   - (a) Process: **prototypical anticausative**, **anticausative of process**, **passive**
   - (b) State: **prototypical anticausative**, **passive**

The **prototypical anticausative**[1] subtype involves those constructions where the affected entity is expressed in the subject position. Spanish examples typically involve either intransitive or pronominal constructions.

18. (a) El escándalo ha hecho bajar las cotizaciones de Telefónica estrepitosamente.

    (b) Las cotizaciones de Telefónica han bajado estrepitosamente.

19. (a) El incidente desató la rabia de nuevo en El Ejido.

    (b) La rabia se desató de nuevo en El Ejido.

It will be noted in 18 that anticausatives can alternate either with periphrastic (this is more frequent for the intransitive subtype) or with synthetic causatives (typically in the case of the pronominal subtype).

Another type of anticausative construction discussed is the **anticausative of process**, which is characterized by the occurrence of a *non-affected* entity in the subject position. The distinguishing test consists in the fact that action realized on the entity does not produce a result.

20. (a) El alcohol ha hecho soñar a María cosas terribles está noche.

    (b) Esta noche María ha soñado cosas terribles.

[1] This construction also receives other denominations, such as **inchoative** or **inaccusative**.

(c) *María está soñada.

The last type of anticausative proposed is the passive, where the initiator component is agentive in the transitive construction that the passive alternates with. In Spanish the passive can be syntactically expressed by means of a periphrastic construction with forms of the verb *ser*, or by means of a pronominal construction with *se*.

21. (a) Las autoridades han cerrado las fronteras.
    (b) Las fronteras han sido cerradas (por las autoridades).
    (c) Se cerraron las fronteras *(por las autoridades).

While subjects can be expressed in the *ser* passive by means of a prepositional phrase headed by *por*, this is not possible with *se* passives. It should also be noted that the pronominal passive is syntactically identical to the prototypical anticausative, but semantically they can be distinguished according to the initiator: for passives it has to be agentive, while for prototypical anticausatives it is causal. With verbs that admit both types of initiators, these constructions are semantically ambiguous.

22. (a) El niño ha mezcaldo las pinturas.
    (b) Las pinturas se mezclaron.
    (c) Las pinturas fueron mezcladas.

23. (a) Se han roto los platos.
    (b) Se han roto los acuerdos.

Notice how 22b can have two readings, whereas in 22c only the agentive meaning is available. Sentences in 23 illustrate how the preferred interpretation of an ambiguous *se* construction depends on the whether the entity is typically affected by non-volitional causes (a) or voluntary agents (b).

Another type of construction that should be distinguished from both the prototypical anticausative and passive is the **impersonal** construction. Impersonals lack an explicit or elided syntactic subject and they have a generic interpretation. Similarly to pronominal passives, they are formed with *se*, but unlike passives there is no subject-verb agreement between the verb and the constituent which expresses the entity. Compare:

24. (a) Se señalaron las características que deberá tener el proyecto.
    (b) Se señaló las características que deberá tener el proyecto.

The sentence in 24a is a pronominal passive with the verb in plural to agree with the subject *las características*. In 24b the verb is in singular, and there is no subject (unless we reinterpret *se* as a ‘dummy’ subject). Impersonals such as 24b are uncommon, and ungrammatical in some dialects. More typical uses involve human entities (25a), cases where there is no explicit constituent expressing the entity (25b), or cases where the verb governs a preposition (25).

25. (a) A los detenidos se les acusa de prevaricación.
    (b) Se vive bien en España.
    (c) Se ha experimentado con animales.

The alternations Vázquez and colleagues consider to be of the **causative/anticausative** type have traditionally been treated separately. Their account unifies many phenomena that, notwithstanding their diversity, share a common core of changes undergone by the information-structure of the sentence. This provides an analogous analysis of constructions that intuitively seem similar, e.g. *Se discutieron muchas cuestiones* and *Se habló de muchas cuestiones* or the English *This bed has been slept in* and the Spanish *Se ha dormido en esta cama.*

Holistic

By the holistic alternation the authors understand a construction pair where a semantic argument denoting a complex entity may be either expressed by a single syntactic constituent, or else be decomposed in two different constituents. The associated change in focus would then consist in emphasizing the entity described as a whole, as opposed to focusing on some specific aspect or property of this entity.

26. (a) Raimundo me irrita con su impuntualidad.
    (b) Me irrita la impuntualidad de Raimundo.

27. (a) He mezcaldo la harina con el azúcar.
    (b) He mezcaldo la harina y el azúcar.

In sentences (a) above the complex argument is expressed by two different constituents, whereas in sentences (b) it combined in a single syntactic constituent. In 26b the mechanism of combination is a prepositional phrase while in 27b the phenomenon involved is that of coordination.

Inversion

Under this rubric the authors include a variety of alternations where two semantic arguments exchange their position in the order of syntactic constituents of the sentence, and thus exchange their position under focus. Here are included some of the most studied alternations, discussed in section 3.2, such as the locative and dative alternations (Spanish lacks the latter). Other constructions exemplifying this category involve alternations affecting the subject, such as:

28. (a) El sol irradia calor.
    (b) El calor irradia del sol.

As can be seen, this particular alternation exchanges the relative positions of the two arguments involved. Formal changes occur as well: in 28b one of the NPs becomes a PP, the indefinite *calor* becomes definite *el calor*. Apart from modifications to the information structure, the element in the subject position acquires initiator-like properties. As the authors observe, with this type of alternations it is frequent for the opposition to be lexicalized and expressed by two different verbs (e.g. *dar/recibir, comprar/vender*).

3.3.2 Underspecification

Vázquez et al. include under this category the alternations that involve the expression vs non-expression of one of the verb's semantic arguments. Thus one of the alternatives in the alternation has more information specified than the other. In other words, one is more *specific* while the other is more *general*.

Unlike in anticausative constructions, the elision of one argument does not cause the other to change its position in the syntactic frame of the sentence.

29. (a) Trini está comiendo sopa de cebolla.
    (b) Trini está comiendo.

Sentence 29a simply provides more information than sentence 29b, without a shift in the information structure.

The underspecification alternations are closely related with the notion of **transitivity**. Some traditionally transitive verbs such as *comer* above can be used without a direct object. On the other hand, some other verbs, usually classified as intransitive, allow objects:

30. (a) La debutante cantó.
    (b) La debutante cantó un aria.

31. (a) Mi abuelo ha dormido.
    (b) Mi abuelo ha dormido la siesta.

Yet another group of verbs incorporate an implicit object, (also known as **cognate** object). This can be normally expressed if it is additionally specified. In Spanish examples are harder to come by than in English, but some exist:

32. (a) El dueño anterior de este bar aguaba el vino *(con agua).
    (b) El dueño anterior de este bar aguaba el vino con agua de grifo.

33. (a) Llueve.
    (b) Llueve una lluvia muy fina.

These alternations should be distinguished from the phenomenon of ellipsis, where the elided material can be recovered from context. Ellipsis does not entail an opposition in the quantity of information provided, since the elided element has already appeared in the discourse and forms part of the information available. Thus no semantic opposition of underspecification is involved and ellipsis is not considered to be a diathesis alternation.

3.3.3 Resultative construction

We shall now consider the members of the subdivision of alternations that are due to an **aspectual opposition**, where one member of the alternation has an eventive interpretation and the other a stative one. The stative constructions differ further in the prominence of the 'stativeness'.

The **resultative construction** has a clear stative reading. It is formed periphrastically with *estar* + participle, and expresses the result of the action undergone by the entity. The process which leads to the result is not expressed in this construction: the result is conceived of as separate from the action which produces it. The initiatior in resultatives can be either agentive or causal and is typically not expressed:

34. (a) Los propietarios han cerrado la fábrica.
    (b) La fábrica está cerrada.

35. (a) La lluvia ha mojado las calles.
    (b) Las calles están mojadas.

There are also variants of this construction where the participle is replaced by an adjective (36) and others where the verb *quedar* takes place of the more usual *estar* (37).

36. (a) El camarero ha llenado los vasos.
    (b) Los vasos están llenos (*llenados).

37. (a) Laura ha manifestado su opinion.
    (b) Su opinión ha quedado manifestada.

3.3.4 Middle construction

In alternations involving the **middle construction** an opposition is expressed between an event situated in a specific time and space on the one hand and an atemporal, non-situated states. Middle constructions typically contain elements reinforcing the stative interpretation, such as adverbials (38), modal verbs (39) or negation (40).

38. (a) Evaristo fundió el plomo.
    (b) El plomo se funde facilmente.

39. Estas setas se pueden comer.

40. Esta fruta no se come.

Similar to anticausatives, Vázquez et al. propose a sub-classification of middle constructions based on what type of eventive construction they alternate with. **Passive middle constructions** (38b) alternate with agentive causatives (38 above). **Prototypical anticausative middle constructions** display a causal agent in their alternating pair (41), while **middle constructions of process** alternate with constructions with a *non-affected* entity (42).

41. (a) La humedad ha estropeado la madera.
    (b) La madera se estropea (con la humedad).

42. (a) Las vitaminas hicieron crecer a los niños.
    (b) Los niños crecen rápidamente.

It will be noticed that the middle constructions are formally similar to anticausatives discussed in section 3.3.1. The difference between the two groups is semantic, namely the lack of specification of time and place in the case of middle constructions. Unlike anticausatives, middles are stative. They describe a property of the entity, and that accounts for the lack of spatiotemporal specification that characterizes them.

Personal temporally unmarked stative

This type of construction, similarly to the previous one, is not specified for time. Unlike with middle constructions, however, in **personal temporally unmarked statives** the argument corresponding to the initiator of the corresponding causative alternative stays in the subject position.

43. (a) Purificación ha leído mucho.
    (b) Purificación lee mucho.

44. Fumar durante el embarazo perjudica la salud de su hijo.

45. En este país no dejan nunca propina.

The sentence in 43a is stative inasmuch as it is a predication about a property of the entity expressed as subject. As exemplified in 44 the entity can also be derived from a causal initiator. Finally, 45 illustrates that the entity can occasionally be expressed by constituents other than the subject, such as a locative prepositional phrase.

3.3.5 Conclusions

The classification of diathesis alternations is Spanish proposed by Vázquez, Fernández and Martí organizes the diversity of constructions according to unifying semantic criteria that are meant to group together diathesis changes that affect the meaning of sentences in similar ways. Within those larger semantically-based categories further subdivisions are proposed based on details of eventive structure and syntactic subcategorization.

Chapter 4

Verb classification

4.1 Introduction

Verb (and other part of speech) classification efforts have a variety of goals. Some of them aim mainly at providing a comprehensive lexical database for use in lexicography, natural language processing and other applications. WordNet, VerbNet and FrameNet, discussed in section 4.5, can be included in this category. Other schemes, such as Levin classes, purport to provide mechanisms that allow to derive a verb's syntactic behaviour, i.e. its subcategorization frames and the diathesis alternations it participates in, from semantic principles. This is usually done by decomposing the verb meaning into more primitive elements that account for that verb's specific syntactic combinatorial properties. We have seen a simple example with Pinker's account of the restrictions on the dative alternation (3.2.1). In this chapter we discuss such issues involved in verb classification, and present some verb-class related projects and resources for English and Spanish.

4.2 Semantic decomposition

One way to simplify the task of providing a coherent and explanatory classification of verbs that would account for their syntactic and semantic properties is to try to find the basic atoms of meaning that lexical items are composed of. It is hoped that the ways in which these semantic primitives interact will help explain, for example, verbal subcategorization frames and the diathesis alternations a verb participates in. Although finding a set of psychologically plausible primitives that could be used to exhaustively compose the meaning of any verb has proved difficult, there has been some progress. Research on this topic includes Miller and Johnson-Laird (1976), Wierzbicka (1985), and Jackendoff (1990).

This last author proposes that verb meanings are composed of functions. These are derived form a non-standard, augmented version of type-logic, where the usual primitive types e (entities) and t (truth values) are replaced by a much richer set of ontological objects such as Object, Event, Path and others. In common type-logic notation a function from semantic objects of type a into semantic objects of type b is written as $\langle a, b \rangle$. Jackendoff's most basic function BE is thus:

46. BE: <(X,Y), State>, X and Y are an ordered pair, where the types of X and Y depend on semantic field

The **semantic field** referred to is an additional feature associated with the function, which determines the character of its arguments and the sort of inferences that can be drawn. Thus if this feature is *Spatial*, then the X argument is an object and Y is its location. If the semantic field is *Possession*, then X is an object and Y the person possessing it. With this feature equal to *Scheduling*, X is an event and Y is a period of time. The BE(X,Y) function is a conceptualization of states. A similar function which underlies event verbs is STAY(X,Y).

The *go* family of verbs have a function GO(X,Y), which conceptualizes the event of X (object) traversing Y, which is a Path (or Trajectory). Paths can be built by providing a start point and an end point, or simply specifying a direction. These and other functions are used to construct situations (States and Events). Other families of functions are aspectual functions such as INCH(X) and PERF(X) (for inchoative and perfective, respectively), which are involved in encoding aspect, and the causative functions such as CAUSE, LET and HELP.

The functions are used to build up skeletons of verb meanings and to explain some facts about verb valency. The lexical entry of the verb *enter*, with its meaning decomposed into primitives, is as follows:

47. $/\varepsilon\text{ntr}/_i\ \text{V}_i\ [_{\text{Event}}\ \text{GO}([_{\text{Object}}\ \text{X}], [_{\text{Path}}\ \text{TO}([_{\text{Place}}\ \text{IN}([_{\text{Object}}\ \text{Y}])])])]_i$

There are two free variables X and Y, which need to be satisfied by NP arguments, so *enter* is a transitive verb. On the other hand, in a verb such as *fall*, the second argument to the GO function is incorporated, i.e. contains no free variables; it is $[_{\text{Path}}\ \text{DOWNWARD}]$. So *fall* only accepts one argument, which fills the X variable. A similar analysis applies to the verbs *put, butter* and *pocket. Put*, the one with most free variables, requires the agent, the patient and the location to be provided. In the case of *butter* the patient is incorporated, so this verb requires the agent and location as its arguments. The verb *pocket* (as in *Paul pocketed the penny*) the reverse is true: the location is incorporated, and agent plus patient are required. This method, with all its generally acknowledged limitations, provides a basis for a principled and meaningful verb classification.

4.3 Levin classes

4.3.1 Beth Levin's classification

In her influential *English Verb Classes and Alternations* Beth Levin (1993) has proposed a comprehensive classification of over 3000 English verbs, using syntactic criteria to achieve coherent semantic classes. Her hypothesis is that verbs that participate in the same diathesis alternations will also share basic meaning components, and thus will tend to cluster in semantically delimited groups. This should be so because the underlying semantic components in a verb constrain its possible arguments, as already illustrated in the previous section and in section 3.2.1. Levin's approach is somehow the reverse of what we have shown in the previous section. Rather then deriving syntactic frames and possible diathesis alternations from semantic primitives identified in a verb, it proceeds in the other direction. By looking at a verb and what alternations it allows, as well as contrasting it with similar verbs, Levin tries to isolate the right combination of semantic components that would result in the observed behavior.

For example, verbs such as *cut* and *break* are similar in that both participate in the transitive and middle constructions:

48. (a) John broke the window.
    (b) Glass breaks easily.

    (a) John cut the bread.
    (b) This loaf cuts easily.

However, only *break* verbs can also occur in the simple intransitive (i.e. anticausative):

49. (a) The window broke.
    (b) *The bread cut.

Another contrast is the ability of *cut* to appear in the **conative** construction. The semantic distinction expressed in the conative is that the action is being directed at the object, but may not succeed, i.e. it does not necessarily *affect* the object. Compare:

50. (a) John valiantly cut at the frozen loaf, but his knife was too dull to make a dent in it.
    (b) *John broke at the window.

The explanation of these fact is given in terms of the meaning components specified for both verbs. *Cut* describes a series of actions directed at the goal of separating something into pieces. These actions are characterized by a specific manner of performing them, recognizable as cutting. The end-result of these actions is not specified, only the attempted goal, and so it is possible to perform them without achieving the goal. Thus 50a makes sense. In the case of *break*, the only thing specified is the end-result, that is the object separating into pieces. So if this end-result is not achieved, there is no breaking at all, which accounts for the incongruity of 50b. In this way the differing alternations allowed by *cut* verbs and *break* verbs serve to identify semantic distinctions between these two groups.

This approach works fine in many cases. But Levin has classified a large amount of data and her method does not always scale well. Some classes contain verbs which are not closely related in meaning, e.g. the *braid* class, which include: *bob, braid, brush, clip, coldcream, comb, condition, crimp, crop, curl,* etc. (Dang et al., 1998). Others have complained that for few of the classes the meaning components are explicitly stated, and that in most of the groups not all the verbs share the alternations stipulated (Vázquez et al., 2000).

Yet another shortcoming has been identified by Baker and Ruppenhofer (2002). They compare Levin classes with the classification developed by the FrameNet project (see section 4.5.3). In FrameNet, classes are based on empirical data extracted from linguistic corpora. They notice that in many Levin classes there are some members that are not attested in some of the constructions associated with their class. The verb *telephone* (which belongs to Verbs of Instrument of Communication), based on its class membership, should occur in the following frames:

51. (a) ?Mom telephoned me the good news.
    (b) ?Mom telephoned me that she was ill.
    (c) ?My brother, mom telephoned me, was now in the hospital

None of these uses, however, is attested among the 1200 examples of the verb *telephone* in the British National Corpus. Of itself, it does not necessarily mean that telephone does not allow these frames, but it does strongly suggest so.

From these issues it seems that Levin's classification, as the subtitle of her work indicates, is indeed preliminary. Others have tried to build on her data and elaborate on and modify her approach.

4.3.2 Intersective Levin Classes

One important refinement of standard Levin classes are **intersective** Levin classes proposed by Dang et al. (1998). One of their main goals is to make Levin's classification less ambiguous, so that it can be interfaced with WordNet (see 4.5.1). The ambiguity arises because of the fact that many verbs are listed in more than one class, and it is not clear how to interpret it. It might indicate that for each listing there is a separate sense involved, or it might be that one of the senses is primary and the syntactic behavior specified for that sense takes precedence over other senses.

Dang et al. created additional classes, which augment Levin's categorization. These intersective classes are formed by isolating set intersections between existing Levin classes and removing the members of these intersections from the original class. The resulting intersective classes are subject to the condition that they contain at least three members; this allows to filter out spurious intersections were overlap between classes is due to homophony.

The authors then show how the intersective classes improve on isolating semantic components shared by class members. The semantically heterogeneous Levin class of *split* verbs includes *cut, draw, kick, knock, push, rip, roll, shove, slip, split* etc. They are grouped together because they manifest an extended sense 'separate by V-ing'. Verbs such as *draw, pull, push, shove, tug, yank* belong here because of the meaning component of exerting 'force' they have. The 'separate' interpretation is only available for these verbs in specific frames such a 52a and 52b but not 52c.

52. (a) I pulled the twig and the branch apart.
    (b) I pulled the twig off the branch.
    (c) *I pulled the twig and the branch.

The adverb *apart* adds the meaning component of 'change of state', which in combination with 'force' produces the 'separation' interpretation.

These marginal *split* verbs are also listed in the *carry* and *push/pull* classes, and so they form an new intersective class. This class is characterized by having the 'force' semantic component. Depending on the particular frame they are used in, they display behavior characteristic of any one of the intersecting Levin classes that list them.

53. (a) Nora pushed at the package.
    (b) Nora pushed the package to Pamela.
    (c) Nora pushed the branches apart.
    (d) *Nora pushed at the package to Pamela.

In 53a *pushed* acts as a verb of exerting force (no separation or caused motion implied). In b it behaves like a *carry* verb, implying caused motion, while in c it acts as a *split* verb, with separation implied. When we try to combine the component of 'exerting force' with other mutually incompatible components, as in 53d, an ungrammatical sentence results. Based on such data, the authors convincingly argue that intersective Levin classes group verbs according to more coherent subsets of semantic features than Levin's original classification.

4.4 Spanish: verbs of change and verbs of path

In this section we review in some detail a proposal of verb classification for Spanish, presented by Vázquez et al. (2000) and based on their account of diathesis alternations, which we have already discussed in section 3.3. Their project aims at establishing a classification that would explain verbal behavior in terms of a theoretical model permitting to form generalizations valid for a large number of verbs. The proposal is based on combined syntactic and semantic criteria, with especial emphasis on interface phenomena.

The authors include insights from prototype-based approaches to classification, where classes have central members who possess all or most of the features characteristic of the class, whereas other members only have each a subset of these features. For the classes they postulate, they define a set of central properties which are shared by all members, and other more marginal features which are common only to a subset of members. They have studied approximately 1000 verbs, divided into two large groups.

4.4.1 Verbs of change

This group includes those predicates where an object is affected by an action realized by a causal initiator. The 'change' consists in the object passing from one (initial) state to another (resulting) state. This change can be either physical (for verbs such as *romper, borrar, congelar*) or mental (*abatir, maravillar, sorprender*).

Meaning components

The basic meaning components for this class are the initiator, the entity and the change (this last is class-specific). The initiator corresponds to the cause of the event, while the entity is the object affected by the action predicated in the verb. The change is, logically, the transition from the initial to the resulting state. The voluntariness of the initiator is, in general, not a distinguishing feature for this class, and most verbs admit both voluntary and involuntary interpretations. Those verbs that only admit a voluntary initiator as a subject, such as *decorar*, are not members of this class.

As for the entity, the resulting state in which it is put by the action of the initiator, can be either permanent (54a and e), temporary (54b and d) or gradual (54c), depending on the verb and the nature of the entity.

54. (a) Se ha desintegrado.
    (b) Sa ha aburrido mucho.
    (c) Las temperaturas descienden.
    (d) Edgardo se ha roto la pierna.
    (e) El cristal se ha roto.

The authors define **affectation** to exclude entities that change location, or that come into being as a result of the action they undergo. Also excluded are those entities that are caused by the action, as with verbs such as *provocar*. Some psychological verbs, such as *amar*, do not belong to the *verbs of change* class either. This motivated by the fact that the entity is not clearly affected, and that these verbs do not occur in the prototypical anticausative construction.

Event structure

The event structure of verbs of change is complex: it combines a process and the resulting state. Thus this class of verbs prototypically participates in the causative/anticausative alternation. In the causative construction, according to Vázquez et al., both 'subevents' are equally emphasized. The anticausative frame emphasizes more the resulting state than the process. In anticausative sentences, due to the fact that they are mainly about the resulting state, that is a property of the entity, the entity is always present, while the initiator can be omitted.

Alternations

Like in the case of Levin classes, the participation in a shared set of diathesis alternations it the main criterion for class membership. Three groups of alternations have been distinguished:

- those that are decisive in the semantic characterization of the class and as such are common to all members
- those that are of secondary importance
- those that class members do not participate in.

Main alternations There are two principal alternations characterizing the *verbs of change* class:

- the *prototypical anticausative*
- the *resultative.*

In the former, all the main meaning components and all elements of the event structure are present. The latter opposes an event and a state, and accordingly neither the initiatior nor the change appear directly in this construction.

Secondary alternations

- A considerable number of verbs participates in the *middle* alternation. In many cases there are restrictions as to what kind of entities can occur in these constructions.
- The subgroup of verbs that admit agents in the subject position participate in the *passive* alternation.
- Some verbs appear in a construction denominated the *middle passive*, e.g. *Este cristal se rompe fácilmente.*
- Psychological verbs belonging to this class also participate in the *holistic* alternation.

Disallowed alternations All the verbs belonging to the *verbs of change* class systematically fail to participate in the following alternations:

- *inversion*
- *underspecification*

4.4.2 Verbs of path

This class includes verbs expressing the change in location of an object. A **path** (or **trajectory**) is covered by the object between two points, the origin and the destination. The concept of path adopted here is a broad one, as it includes both changes in physical location and more abstract, extended meanings of change of place, such as changes in possession (*comprar, dar*) or communicative exchanges (*decir, responder*).

A distinction should be made between those verbs that express a change in location, and those that simply indicate movement, where change in location is secondary. *Correr* belongs to the former group, while *bailar* is member of the latter; only the first type is included in *verbs of path.* The authors exclude most perception verbs, such as *oler, ver* and *mirar* from the class. They do include, however, *escuchar* and *oír* arguing that these, being the reverse of communication verbs such as *decir*, share enough features common to the class to be included.

With some verbs it is not clear whether they should belong to verbs of change or verbs of path. Two examples include predicates with incorporated objects: *ensuciar* and *embotellar.* They both entail that some entity is affected: something is dirty, and something is bottled, respectively. They also both entail that some object changes location: dirt is transferred onto the affected object, and the affected object is put in a bottle, respectively. The authors decide to classify the *ensuciar* as a verb of change, because it emphasizes the affectation aspect, and *embotellar* as a verb of path, since the aspect of change of location is more prominent in this verb.

Meaning components

The basic semantic components are the initiator, the entity and the path. The entity component is typically expressed by an NP as in 55a.

55. (a) El cartero lleva las cartas a sus destinatarios.
    (b) El profesor habla de la historia de Grecia a sus alumnos.
    (c) El niño dijo que no lo volvería a hacer.
    (d) Marisol confesó ante los presentes: "No soy culpable"

With verbs of communications, the entity can also be expressed as a PP (55b), a subordinate clause (55c) or a quotation (55d).

Also note that the initiator and entity can in some cases be combined in the same object, as in the case of verbs of autonomous movement (i.e. *Los estudiantes van a la manifestación*). It is also possible for the initiator to coincide with either the origin or the destination of the path, with verbs such as *obtener* (where initiator = destination) or *vender* (with initiator = origin).

Path is a complex component. In addition to origin and destination it includes **route** (or **via**), which is typically expressed with *por* PPs (i.e. *Pilarín ha ido de Logroño a Huesca por Sabiñánigo*). Another subcomponent of path is the **direction** (cf. Jackendoff's TOWARDS), normally expressed with a PP headed by *hacia* or *en dirección a.*

Event structure

The verbs belonging to this group are not quite uniform as to their telicity. The relevant contrast can be observed by comparing *llegar*, which is telic, with *correr*, which is not. The members of the class also differ as to whether they emphasize the origin (*marcharse*), destination (*aterrizar*) or the route (*errar*). This process of emphasizing one subcomponent can also be achieved by means of an adjunct, as in *Nadó hacia la orilla.*

Based on these points, the authors posit eventive structures consisting of two 'subevents': either origin and process or process and destination, with one or the other or both emphasized by specific verbs, or by verbs in combination the adjunct. This approach adapts Pustejovsky (1995)'s analysis, developed for verbs of change, to this class: the event structure is complex, consisting of a process and a preceding or following telic subprocess.

Alternations

Main alternations There is only one alternation shared by all the members of the *verbs of path* class: *underspecification*: one or more of the path subcomponents is omitted, as in *Los gaurdias arrastraron al preso (hacia la celda).* There are restrictions on what subcomponents can be non-expressed:

often, if the destination is omitted, so must be the origin. In *Los operarios han carreteado la mercancía del camión al almacén* either both subcomponents are omitted or both must be expressed; **Los operarios han carreteado la mercancía del camión* is impossible.

Verbs like *venir* do allow the omission of destination while expressing origin (*Los peregrinos han venido de todas las partes del mundo*). It should be observed, however, that this is a case of ellipsis rather than underspecification, since this verb has an incorporated deictic referent for the destination (namely the place where the speaker is at the moment of utterance). The route subcomponent is, in general, admitted by members of the class, as is its omission, although there seem to be certain weak restrictions on its omission for certain verbs (*pasar, deambular*).

Secondary alternations None of the following alternations is common with class members:

- *Passive* alternation. Verbs one of whose arguments is an NP can participate in this alternation. When the NP expresses the path component, the pronominal form is more readily accepted: *Se caminaron muchos kilómetros* vs **Muchos kilómetros fueron caminados.*

- *Passive middle*, such as in *El Danubio se cruza difícilmente.*

- *Underspecification* involving the affected entity is extremely uncommon. Some communication verbs, such as *hablar*, participate in it.

- *Holistic.* Participating verbs are those where initiator = destination, e.g. *Le compré el coche a Sebastián* vs *Compré el coche de Sebastián.*

- *Inversion* in verbs such as *cargar*

Disallowed alternations Members of this class systematically fail to participate in the following alternations:

- *Prototypical anticausative*

- *Anticausative of process*

- *Anticausative middle*

4.4.3 Discussion

In contrast to other schemes, the classification sketched above is unusually globalizing. It groups together many verb types that have been traditionally treated separately. This responds to the authors' unifying perspective. They try to isolate underlying primitives and base their classification on those rather than get distracted by the superficial diversity of observed behavior.

They make some inevitable trade-offs, losing in granularity while they gain in generality. Ultimately the correct resolution depends on the intended application of the classification.

4.5 Lexicographical databases

Below we review three lexicographical projects relevant to issues of verb classification. Their goal is to provide semantically and/or syntactically structured online lexical databases for use in computational linguistics, natural language processing, lexicography, language teaching and other applications.

4.5.1 WordNet

WordNet is the oldest and the best-known online lexical database for the English language (Miller et al., 1990). It arose as an attempt to leverage the computational resources of computers in the service of lexicography and to create a database of English words that would reflect the organization of the mental lexicon. Instead of listing words alphabetically as in printed dictionaries, WordNet organizes them according to semantic criteria. Relationships in Wordnet are between word-meanings and between word-forms. Word meanings are represented by sets of synonyms, or **synsets**. The relations of **hyponymy** (**ISA** relations) and **meronymy** (**HASA** relations) obtain between synsets. **Antonymy** relates word-forms; even though the synsets {*rise, ascend*} and {*fall, descend*} are conceptual 'opposites', *descend* is not the antonym of *rise*, but *fall* is. WordNet treats separately words belonging to different grammatical categories. Four major categories are represented: nouns, verbs, adjectives and adverbs. Below we discuss how WordNet organizes verbs (Fellbaum, 1998).

Verbs in WordNet

The organization of WordNet is based on semantic criteria, and verb classification follows this design decision. Verbs are organized into synsets and those are related among themselves by a number of relations. The most important of those is **troponymy**, which includes several manner relations, and can be paraphrased as *X is a particular way to Y*, e.g. *Limp is a particular way to walk*. Another way to associate verbs in WordNet is by the **causal relation**, which relates two verb concepts, one causative and the other resultative. This relation holds, for example, between *give* and *have*. Only lexicalized pairs are included. WordNet adopts the relational approach to verb categorization: the smallest units of analysis are lexicalized meanings, and they are not further decomposed into semantic primitives. On occasion aspects of semantic decomposition are implicit in the relations between synsets. For example, members of synsets that are associated with others by means of the causative relation contain the semantic primitive CAUSE.

WordNet is designed to mimic lexical memory rather than the bulk of lexical knowledge, so the elements of verbal knowledge that do not contribute significantly to the organization of the mental lexicon have been mainly disregarded. Only the most basic syntactic information is provided: each verb synset has one or more sentence frames associated with it. The frames are very rudimentary. For the synset {*give, (transfer possession of something concrete or abstract to somebody)*} the frames provided look like this:

56. (a) `Somebody ----s somebody something`
    (b) `Somebody ----s something to somebody`

The semantic roles of the constituents represented by `somebody` and `someone` are not indicated. The diathesis alternations that verbs can participate in are not systematically indicated, either. Often the alternatives in a diathesis alternation are assigned to different senses. This causes multiplication of senses which are, in principle, regular meaning extensions derivable from the core semantics of a verb.

A parallel project, called EuroWordNet, has elaborated databases for various European languages, including Spanish, along the same lines as the original English WordNet (Vossen, 1998).

4.5.2 VerbNet

VerbNet (Kipper et al., 2000a; Kipper et al., 2000b) is a project which aims at remedying some of the shortcomings in WordNet's treatment of the syntax-semantics interface. It consists of a static part, made up of verb entries, and a dynamic part, which captures the syntax associated with verb classes. The latter is implemented in a Lexicalized Tree-Adjoining Grammar and constrains the entries to allow a compositional interpretation in derivation trees. The verb classification adopted in VerbNet is based on intersective Levin classes, discussed in section 4.3.2. A verb entry contains links to the classes which correspond to the different senses of the verb. For each of the senses there is also a link to the set of WordNet synsets that best reflects this sense.

Verb classes bring together information that is generally applicable to class members. It is intended to be detailed enough for most computational applications. Class-specific thematic roles are listed. Syntactic frames in which these roles are expressed are specified together with the selectional restrictions that semantically constrain the frame arguments.

Each frame also includes predicates that describe participants at various stages of the event described by the frame. The event structure is also represented. This is done by decomposing the event into a tripartite structure. For each predicate, the time functions *during(E), end(E)* and

Thematic Roles		Agent(A), Patient(P), Instrument(I)
Basic Transitive	A V P	manner(during(E),directedmotion,A) ∧ manner(end(E),forceful,A) ∧ contact(end(E),A,P)
Conative	A V at P	manner(during(E),directedmotion,A)

Table 4.1: Two VerbNet frames for *hit* verbs

result(E) specify whether a predicate is true at the moment associated with the function. This allows to capture the complex semantic composition of many verbs (such as for example verbs of change), where the distinction between a process and a result of this process is important if we want to be able to predict the behavior of a verb.

The semantics of each syntactic frame is captured by a conjunction of predicates. The table 4.1 illustrates two of the frames for the *hit* class of verbs. The static frames described above are mapped onto elementary TAG trees (Joshi, 1985), and semantic predicates are associated with each tree. There are also predicates associated with auxiliary trees: for example PPs headed by *across* specify the path of the object via some other object, which is expressed by the complement of the PP. The compositional semantics of the sentence can thus be built up from the trees used to derive it: it is the conjunction of the predicates associated with them.

VerbNet in a way complements WordNet. The project pays special attention to issues WordNet largely ignores, namely the syntax-semantic interface. This aspect of verbal behavior is represented in considerable detail, building on previous research by Levin and her successors.

4.5.3 FrameNet

FrameNet (Baker et al., 1998; Johnson et al., 2002) is a project dedicated to developing an online lexical database for English predicating words. It seeks to provide empirically sound data by making extensive use of linguistic corpora, mainly the British National Corpus (`http://www.natcorp.ox.ac.uk/`). The central goal is to provide for each sense of each word the range of semantic and syntactic combinatory possibilities.

It consists of three major integrated components: a lexicon, a frame database and annotated example sentences. The lexicon entries contain links to the frame database, which captures generalizations on syntactic and semantic behavior of words. They are also linked to the annotated sentences illustrating the use of the lexical item described in the entry. At the end of 2002 the database contained approximately 6000 lexical items (verbs, nouns and adjectives).

The theory underlying the semantic and syntactic architecture of the frame database is Frame Semantics, developed by Charles Fillmore. A frame is a script-like construct representing a situation. Its components are **Frame Elements** (**FE**s), i.e. the participants, props and other roles involved in the situation. These are in a way similar to the familiar notion of thematic roles. In Frame Semantics, however, the FEs are only meaningful in a given frame, i.e. they have no global validity. Frame Elements are the semantic arguments of the frame predicate. Syntactic and semantic properties of predicating words, such as verbs, can be characterized by relating each of their senses to a particular frame. Frames are organized as an inheritance hierarchy, with more specific frames including information from more general parent frames (Johnson and Fillmore, 2000).

Frame Elements can be realized in actual syntactic patterns: they are of specific **Phrase Type** such as NP, PP etc., and have a specific **Grammatical Function** such as External Argument (i.e. subject), Object etc. These realizations are documented in FrameNet by analyzing and annotating relevant corpus data. For each lexical entry, a complete set of its syntactico-semantic combinatorial properties is constructed.

FrameNet provides very comprehensive and empirically validated data on the syntactic behavior of predicating words, but how it presents these data is different in an important respect from Levin's and VerbNet's approach. FrameNet classification is based on shared semantics; it makes no assumptions as to whether shared syntactic behavior is caused by similar semantic composition. So verbs belonging to the same frame do not necessarily all participate in the same diathesis alternations. As long as they express the same underlying situation, as represented by the frame, the details of their syntactic behavior do not determine their class membership. For example, Levin distinguishes a class of *butter* verbs. These are like her *load/spray* verbs, except for the fact that they only explicitly realize the object that changes location if it is more specific than their internal incorporated object. Thus *butter* verbs, due to this detail of their syntactic behavior, have to be in a class of their own in Levin's approach. For FrameNet, the deciding factor is the shared semantics, which is the similar to *load/spray* verbs, so all these verbs are placed together in the Filling frame (Baker and Ruppenhofer, 2002).

There is a Spanish FrameNet project (`http://gemini.uab.es/SFN`) underway, working in collaboration with the Berkeley FrameNet team.

Chapter 5

Subcategorization Acquisition

Verb subcategorization is one of the most important pieces of information in a lexicon. According to the radical lexicalist position, together with the combinatorial properties of other grammatical categories, it accounts for all the syntax there is in a language. Even if we assume a more moderate position, verb subcategorization still determines a major part of syntax. It is especially important in computational applications. Extensive information on syntactic frames that verbs occur in greatly improves accuracy of natural language parsing.

Given the above arguments it is obvious that computational lexicons with accurate subcategorization frames are a very valuable resource. Unfortunately, they are tedious to produce manually, and rapidly fall out of date. Extraction of subcategorization frames from text corpora, if it works, is a cheap and robust alternative to manual compilation.

The purpose of a verb subcategorization acquisition systems is to utilize implicit information present in text such as machine-readable linguistic corpora. They try to infer the combinatorial properties of verbs by analyzing their distribution and patterns of usage in large amounts of textual data. By integrating information from a big enough sample of naturally occurring text it is possible to extract generalizations about syntactic behavior of verbs and record those findings in a lexical databases.

The exact method of acquisition of subcategorization patterns varies (we discuss several methods in section 5.2). In very general terms, however, it involves performing some sort of pattern recognition on the corpus data. This step, the detection of putative SF, uses SF **cues**, i.e patterns that indicate with a certain probability the presence of a specific subcategorization frame. In general, not all these finds are accepted, but rather some sort of selection is performed, often involving statistically estimating the probability that a certain proportion of cues does in fact correspond to a real SF for a given verb. The performance of many acquisition systems is then experimentally evaluated, using standard evaluation measures and methods. In the following sections we describe first some basic concepts used in evaluation and then we review a number of SF acquisition systems.

5.1 Evaluation measures

The ultimate demonstration of success for an NLP application is its performance in some language-processing task such as understanding, dialog, summarization, spell-checking etc. However, in the process of development of a specific system dedicated to some specific subtask, such as extraction of verb subcategorization frames from corpora, it is useful to have agreed-on evaluation measures. This makes it possible to test objectively if modifications to the system have yielded better results, and compare the performance of different systems.

5.1.1 Precision, recall and the F-measure

Two frequently used measures that were adapted to NLP from information retrieval research are **precision** and **recall**, the latter also called coverage (Manning and Schütze, 1999). Precision is the percentage of the results returned by the system that are correct. In the context of information retrieval, it would be the number of documents selected that are relevant, divided by the total number of documents selected. In SF acquisition, it could be the proportion of verb-SF pairings which are correct, to the total number of pairings generated.

Recall is the proportion of the correct results returned to the total number of results that should have been returned. In information retrieval, this would normally be the number of relevant documents selected by the system to the total number of relevant documents. In SF acquisition it might be the number of correct verb-SF assignations made by the system to the number of such pairings for the same verbs listed in some standard that we use for comparison.

There is usually a trade-off between precision and recall: one can maximize one at the cost to the other. By returning all possible SF frames for all verbs one can achieve 100% recall, but the precision will be abysmal. Thus it can be convenient to combine both in a single measure to evaluate overall performance of the system. Such a metric is called the **F measure** and is defined as follows;

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

where P is precision, R is recall and α is the weight given to precision (between 0 and 1). Often precision and recall are weighted equally. With

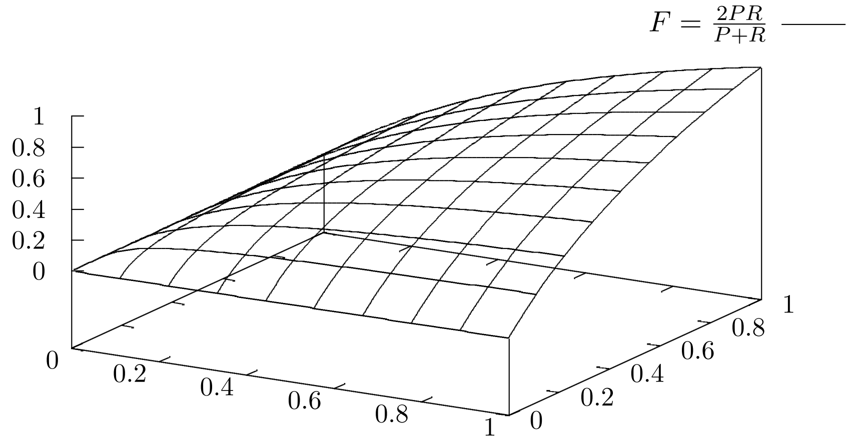

Figure 5.1: F as function of precision and recall.

$\alpha = 0.5$, the above equation reduces to

$$F = \frac{2PR}{P + R}$$

Figure 5.1 shows how F depends on precision and recall for $\alpha = 0$. If either P or R are equal to 0, F-measure is 0 as well; if both P and R are equal to 1, so is F. Also if both P and R are equal to some value β, $F = \beta$ as well.

5.1.2 Types and tokens

The performance of an NLP tool such as a subcategorization acquisition system can be evaluated in terms of **types** and **tokens**. The meaning of the technical term *type* roughly corresponds to its meaning in general use: there is one type per each set of distinct but identical items in some collection of items. Every one of these individual items is, in turn, a token. A string such as *a rose is a rose is a rose* contains eight tokens, i.e. individual words, and three different types, i.e. 'a', 'is' and 'rose'.

In the SF acquisition, a type would usually correspond to a pairing of a specific verb with a specific subcategorization frame. **Type recall** is evaluated by dividing the number of correct pairings generated by the system, by the total, expected number of such associations. This last figure is taken from a source often named a **gold standard**. Such a standard can be obtained in at least two different ways:

- Large dictionary. We can take the verb-SF pairs found in some preexisting source, such as a comprehensive manually compiled dictionary. One disadvantage of this method is that the set of SFs used by the system and the dictionary may be different and it may be difficult to map one to the other. Another problem is that the dictionary may contain SFs not attested for a given verb in the corpus we use. Or, conversely, the dictionary may be lacking in coverage and not include SFs present in the corpus.

- Manual construction. We can manually analyze the same data as the system and thus determine the correct set of SF-verb associations. Here the main issue is that it is time-consuming: at least the same portion of corpus must be analyzed that is used to test the performance of the system.

Tokens, in subcategorization acquisition, are the actual occurrences of SF in analyzed corpus data, i.e. the SFs that verbs in the corpus sentences are assigned. Although it is more usual to report recall and precision over types, **token recall** is also sometimes used. In order to caculate it, for each verb token in running text sample data, we check if its categorization frame is listed in the acquired subcategorization lexicon. Token recall is the number of tokens for which the correct frame appears in the lexicon, divided by the number of all tokens in the sample. This measure may be used to estimate, for example, the performance a parser equipped with the acquired frames would have.

5.2 SF acquisition systems

Not surprisingly, the first SF extraction methods were developed for English. A major part of subsequent research also focussed on this language. In this section we present an overview of these systems.

5.2.1 Raw text

The first system for extraction of subcategorization frames from corpus data was presented by Brent (1991; 1993). It worked with raw, untagged corpus of the *Wall Street Journal*. It was capable of assigning six predetermined, simple subcategorization frames. As in English subjects are always subcategorized for, they were ignored. The frames were:

- V NP (direct object)
- V NP S (direct object + clause)
- V NP INF (direct object + infinitive)

- V S (clause)
- V INF (infinitive)

Several steps were involved in the extraction of frames. First, verbs had to be identified. This was done by identifying in the corpus words that occurred both with and without the *-ing* suffix. These were treated as potential verbs. An occurrence of a potential verb was treated as verb unless it was preceded by a determiner or a preposition other than *to*.

The next step was to identify SFs for words categorized as verbs. Brent chose to use only those patterns in the data that identified a particular SF with fairly high certainty. This was achieved by the relying on the appearance of closed-class lexical items such as pronouns in the patterns (as in 58) so as to avoid misguiding cues, such as those in 57.

57. (a) I expected the man who smoked to eat ice-cream.
    (b) I doubted the man who liked to eat ice-cream.

58. (a) I expected him to eat ice-cream.
    (b) *I doubted him to eat ice-cream.

Even though Brent used only such reliable cues, the SF detection still produced some spurious matches. For example the verb *refer* is wrongly classified as taking an infinitival complement based on sentences such as *I referred to changes made under military occupation*. With growing corpus size such erroneous judgments tend to degrade the performance of the system. The remedy applied by Brent was to submit the judgments made (which can be called hypothesis) to a statistical test which decides whether there is a sufficient number of cases to warrant the permanent inclusion of an SF in the lexical entry for a given verb. This is called **hypothesis testing** and involves the formulation of the **null hypothesis** (i.e. that a given SF is *not* correct for a given verb) and then either accepting or rejecting it, based on the probability of it being false. The specific test used by Brent was the **binomial hypothesis test**. To perform it we record the number of SF cues (n) found for a verb, and the number of cues (m) that indicate SF f_i. Also needed is an estimate of the probability that a cue for f_i occurs with a verb that does not have frame f_i, i.e. error probability p^e.

Each occurrence of the verb is an independent trial, with probability p of the cue being wrong. For such events, the probability that they will happen exactly m times in n trials is given by the following equation:

$$P(m,n,p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

Thus in order to calculate the probability of the event occurring *at least* m times out of n trials we need to sum all the probabilities for values between

m and n:

$$P(m+, n, p) = \sum_{k=m}^{n} P(k, n, p)$$

We reject the null hypothesis if this probability is less than some threshold we have decided on, typically 0.02 or 0.05 (Korhonen, 2002; Manning and Schütze, 1999).

Brent's system showed very high precision with low coverage. This is due to his use of only reliable cues, which are infrequent. Additionally, for some SFs there are no reliable cues. For example many verbs subcategorize for an NP and a PP (e.g. *They assist the police in the investigation*). However, most NP PP occurrences are cases where the PP modifies the NP (Korhonen, 2002). So this frame and other similar ones cannot be reliably acquired in Brent's system.

5.2.2 Tagged and chunked text

Manning (1993) addressed some of the problems inherent in Brent's approach. He used a tagged corpus as input to the SF detector. This increases the number of cues available. Some errors that appeared in Brent's systems due to misclassification of words are less likely to occur since POS classification is performed by a dedicated tagger. SF detection is done by means of a finite state parser.

Manning used all sorts of cues, even those that are relatively unreliable: the hypothesis being that high unreliability does not matter as long as adequate hypothesis testing is performed. For example if a cue has error-rate $= 0.25$ and occurs 11 out of 80 times, it is still a good indicator of the putative SF being correct, since the probability of the null hypothesis is $p^e \approx 0.011 < 0.02$. An SF detector that uses dubious cues, such Manning's, returns nothing or the wrong result in most cases — so here the hypothesis testing phase gains more importance and, for it to work adequately, accurate error probability estimates are necessary. Manning adjusts these estimates empirically. Based on the system's performance he sets them considerably higher than Brent, who achieved the best performance with values of the order of 2^{-8}. Manning, in contrast, got best results with error estimates ranging from 0.02 to 0.25.

Manning's system can learn a larger set of subcategorization frames, even those that have no reliable cues. Still, the results show a pattern familiar from Brent's system: for the set of 40 verbs tested, precision was 0.9 while type recall was 0.43, giving an F-measure of 0.58.

SF detector's performance can be also improved by providing a more structured input that POS tagged text: that is by chunking it. **Chunks** are segments of text corresponding to non-recursive cores of major phrases (Abney, 1991). A chunk, in English, includes the part of the constituent to the left and including the head, but no material to the right of the head.

Thus, a verbal chunk normally ends with the verb, excluding any possible arguments. For Spanish the heuristic has to be slightly different, due to different word order patterns within constituents. Compare a and b in 59 below:

59. (a) ($_{NP}$ We) ($_{VP}$ lack) ($_{NP}$ the means) ($_{VP}$ to do) ($_{NP}$ that)

    (b) ($_{NP}$ Las nuevas generaciones) ($_{VP}$ se preocupan) ($_{NP}$ cada vez) ($_{ADV}$ menos) ($_{PP}$ por la música clásica)

Chunking (unlike full non-statistic parsing) has the advantage that it can be performed without already knowing subcategorization patterns of verbs that appear in text. Even though it structures text in a pretty elementary way from the syntactic point of view, it nevertheless permits the SF detector to work on a rather higher level than individual tagged words. This can significantly improve the accuracy of the output from detection.

Gahl (1998) presents an SF extraction tool for the British National Corpus (BNC). It can be used to create subcorpora with different SFs for verbs, nouns and adjectives. The SF detection makes use of regular expressions over POS and morphosyntactic tags, lemmas and sentence boundaries, thus being equivalent to a chunker. No hypothesis testing is performed and no evaluation measures are reported.

Lapata (1999)'s system also uses the BNC (POS tagged and lemmatized) as input. She develops a chunk grammar to recognize verb groups and NPs, using a corpus query tool called GSearch (Keller et al., 1999). Lapata's aim was to extract corpus segments containing the SF-patterns: V NP NP, V NP *to* NP and V NP *for* NP. SF detection produced output with a high level of noise. In order to deal with this issue, Lapata postprocessed the data using linguistic heuristics. The next step involved hypothesis testing. She reports trying both the binomial hypothesis test and a relative frequency cutoff (empirically established).

In the systems reviewed above the SF detection module outputs a relatively large amount of noise. This is due to various factors, but mostly caused by the limitations of heuristics used in chunking (such as the longest match heuristics) and it's inability to deal with non-trivial syntactic phenomena such as ambiguity of attachment. They mostly rely on the statistical filter to remove the noise from data. This strategy works in many cases but has its own limitations. Brent, Manning and Lapata observe that the hypothesis testing is ineffective for low-frequency subcategorization frames.

5.2.3 Intermediately parsed text

As a next step in the direction of providing SF detectors with higher-level data, *intermediately parsed* text has been used. Intermediate parsers are more sophisticated than chunkers, even though they typically continue to rely on POS tagged data, without prior knowledge of subcategorization. Data are parsed into trees such as the following (Korhonen, 2002):

60. ($_S$ ($_{NP}$ He) ($_{VP}$ ($_{VP}$ has remained) ($_{AP}$ very sick)))

In the majority of cases intermediate parsers are probabilistic. They can be trained to assign weightings to alternative parse possibilities thus producing less errors than chunk parsers using simple-minded heuristics.

One of the systems which has made use of such highly structured data is described by Ersan and Charniak (1996). Their primary goal was to perform syntactic disambiguation, but their system is also capable of SF detection. It worked by statistically analyzing word usage in a corpus and then refining a probabilistic context-free grammar based on these statistics. This new grammar is used to parse data again. SF detection is done by mapping the 1209 VP rewrite rules in the grammar to 16 SFs, and for each verb checking which of the rules has been used to parse it. This allows to assign to this verb the SF that corresponds to the rule. They then filter the data using Brent's binomial hypothesis test. Error probabilities are established empirically. They report a precision of 0.87 and type recall of 0.58. The F-measure for these values is 0.7.

Briscoe and Carroll (1997) describe a system capable of acquiring a set of 161 subcategorization frames. Raw text is tagged, lemmatized and statistically parsed. The parser uses a unification-based grammar. Frames associated with verbs are then extracted and the features on heads and arguments are examined in order to determine if the pattern should be included as an SF for the verbal head. The parser is not provided with lexical information, so it inevitably generates some erroneous output, failing in cases of argument/adjunct ambiguity or in cases where semantic knowledge is needed to resolve a particular argument pattern.

The final step involves using the binomial hypothesis filter to weed out erroneously assigned SFs. They evaluate their results against two different gold standards:

- Manual analysis of data. This gave a precision of 0.77 and type recall of 0.43 (F-measure 0.55). They also report token recall, at 0.81.

- Dictionary (ANLT + COMLEX). The precision obtained was 0.66 and type recall 0.36 (F-measure 0.47)

Carroll and Rooth (1998) also use a robust statistical parser to process text used to acquire subcategorization frames. They implement an unsupervised learning algorithm to iteratively refine a probabilistic context-free grammar. Their system extracts 15 SFs and their relative frequencies. The final output contains errors from sources similar to those described above, Carroll and Rooth choose not to use hypothesis testing, so the putative SFs are stored directly in the generated lexicon. They evaluated their results against a dictionary-based gold standard (*Oxford Advanced Learner's Dictionary, OALD*) The precision obtained was 0.79, with type recall at 0.75 (F-measure 0.77).

Although parsers used in systems described in this section generate less errors that simple chunkers, some error-types are difficult to avoid, as they spring from lack of lexical information (which cannot be straightforwardly provided as it is exactly what the system is trying to acquire) or semantic knowledge, which is also unavailable. Given that errors cannot be expected to disappear completely, statistical filtering gains importance. The most frequently used test is still the BHT introduced by Brent, but it fails to discriminate correct results from errors for low-frequency data.

Korhonen (2000; 2002) considers the hypothesis-testing to be the major bottleneck in automatic subcategorization frame acquisition. She goes on to propose some refinements to this test to remedy the problem. The system she uses for hypothesis generation (i.e. generating putative SFs) is the one developed by Briscoe and Carroll, described above. She claims that the major source of filtering inefficiencies are caused by the fact that the conditional SF distribution (i.e. specific to a verb) does not correlate sufficiently with unconditional SF distribution (regardless of verb). Instead of assuming unconditional distribution, Korhonen proposes to base distribution estimates on the way verbs are grouped into semantic classes. Her system semantically classifies verbs according to their predominant senses and this classification is used to guide subcategorization acquisition, by providing class-specific probability estimates. She reports that this method improves the results with low-frequency verb-SF associations, demonstrating that semantic generalizations can be successfully used to guide otherwise syntax-driven SF acquisition. She performs a comprehensive evaluation of her results, comparing several methods of smoothing distributions based on verb semantics with a manually produced gold standard. The best set of results, obtained for semantically classified verbs only (LEX-D), achieved a precision of 0.87 and type recall of 0.71, for an F-measure of 0.78.

5.3 Discussion

In table 5.1 we provide a summary of the evaluation results reported by the systems discussed in this section. It should not be taken as a direct comparison of their relative performance, since the systems started with different objectives, used different corpora as their input, set out to extract varying numbers of subcategorization frames and analyzed varying numbers of verbs.

It is obvious that subcategorization acquisition methods have progressed significantly and have grown in sophistication since the first serious system presented by Brent. In more recent approaches SF detection works on highly

System	**No. SFs**	**No. verbs**	**Data size**	**Standard**	**Precision**	**Recall**	**F**
Brent (1993)	6	33	1.2M	manual	0.96	0.76	0.85
Manning (1993)	19	40	4.1M	OALD	0.9	0.43	0.58
Ersan and Charniak (1996)	16	30	36M	OALD	0.87	0.58	0.7
Briscoe and Carroll (1997)	161	7	1.2M	manual	0.77	0.43	0.55
Briscoe and Carroll (1997)	161	14	1.2M	ANLT + COMLEX	0.66	0.36	0.43
Carroll and Rooth (1998)	15	100	30M	OALD	0.79	0.75	0.77
Korhonen (2002)	45	75	20M	manual	0.87	0.71	0.78

Table 5.1: Results obtained in SF acquisition by different systems

structured parsed input. Both the generation of presumable frames and the filtering of errors have been perfected. Automatic SF acquisition in English has matured considerably since its inception.

Unfortunately the same cannot be said of most other languages. For Spanish, research on large scale automatic subcategorization acquisition has not, to our knowledge, been reported. In the following chapter we describe our preliminary exploration of the possibilities and issues that arise in adapting some insights, gleaned from English-specific research, to SF acquisition in Spanish.

Chapter 6

Subcategorization acquisition in Spanish: a preliminary study

In the project described in this chapter we set out to determine how well subcategorization acquisition techniques such as those described in the preceding sections work for Spanish. We adopted and adapted an existing scheme of classification of subcategorization frames form the SENSEM database project (Fernández et al., 2002). We implement a tool which searches partially parsed corpora and detects potential verbal SFs: in the study described below we have used this system to acquire subcategorization patterns for ten Spanish verbs. The detection is based on trying to find matches for "templates", which are typical syntactic patterns associated with specific subcategorization frames.

We have implemented a relatively simple system which allows us to explore the issues involved in extracting subcategorization information from chunked corpora. Simplicity was one of the objectives: we have tried to keep the project easy to understand and easy to modify. This gives the flexibility necessary in a pilot study such as the present one, where most time is spent experimenting with different settings and adjustments and observing how the system's performance reacts. After describing the project's rationale, resources used and giving some details on implementation we describe the experimental evaluation of the acquisition task.

6.1 Design and Resources

In general lines, the approach to verb subcategorization adopted for the purposes of developing our acquisition system is based on work by Vázquez et al. (Vázquez et al., 2000) presented in sections 3.3 and 4.4. One important assumption shared with their approach, and one that influences to some degree the design of the system is to consider subjects as subcategorized for by verbs. Vázquez and colleagues' stance on this issue seems to be mainly motivated by their largely semantic approach to verb classification. They try to avoid splitting verbs entries into many verbs closely related in meaning and differing mainly in the conceptualization of the situation the predicate expresses. Rather they choose to treat all those different meanings as regular extensions of the same core semantics of the verb.

In order to identify the meaning components underlying the syntactic behavior of verbs, it is necessary to consider all participants in the situation described by the verb, including the one expressed by the subject, as semantically they are all equally relevant. Including the subject in subcategorization is important if lexical entries are to contain such refinements as appropriate linking between thematic roles and syntactic arguments, or selectional restrictions on the verb's arguments. We also chose to adopt this position, even though at this stage our system extracts purely syntactic information, and ignoring subjects would likely work fine as well.

6.1.1 SENSEM database

One of our goals has been to develop a tool that ultimately could be used as a basis for a system that would automatically acquire subcategorization frames for the SENSEM verb database, so compatibility with that project was an important point to keep in mind.

Subcategorization frames

SENSEM is based on Vázquez and colleagues' work and contains a variety of information associated with verbs, such as definitions, patterns of syntactic arguments and associated thematic roles, prepositions governed, selectional restrictions and examples of use. For specifying subcategorization frames a number of classes are used. The classes' names combine information on the diathesis alternation and syntactic arguments of the verb. So for example `caus-2np` corresponds to the frame associated with a causative diathesis, with two arguments, a subject and a direct object, while `anti-pr-np` refers to a pronominal verb in an anticausative diathesis with a single argument, a subject. Thus, the first frame is exemplified in *Charly García presentó su disco Influencia y desató una fiesta ante una sala repleta*, while the second can be seen in *Vino la lluvia, bajaron las corrientes, se desataron los vientos y dieron contra aquella casa.* A complete list of those classes is given in appendix A. As can be seen, most of them encode some semantic information (which is not dealt with at this stage of our study). Many frames are not distinguishable, or difficult to distinguish from each other, by purely syntactic means: for example `caus-np` and `anti-np`, or `pas-se-np-pp` and `anti-pr-np-pp`. The frames in the first pair are always syntactically equal.

caus-2np	caus-compl-np
caus-compl-np-pp	caus-2np-pp
caus-2np-pp	caus-np-pp/anti-np-pp
caus-np-2pp	caus-np/anti-np
pas-se-np/anti-pr-np	anti-pr-2np/caus-pr-2np
pas-se-np-pp/anti-pr-np-pp	pas-se-2pp
imp-se	result-estar-part-np
result-estar-part-np-pp	result-estar-adj-np
anti-dejar-part-np	anti-dejar-part-np-pp
anti-dejar-adj-np	caus-hacer-inf-2np
caus-hacer-inf-2np-pp	
caus-hacer-compl-2np/caus-hacer-compl-2np-pp	

Table 6.1: SF classes

The second pair differs in that the passive diathesis only occurs with *se*, while in the pronominal anticausative first and second person clitics such as *te* and *nos* are possible. But then, in most uses of this frame in text the pattern is exactly the same. So we have decided to prune and compact the list, combining identical syntactic frames in one and discarding some that could not be correctly treated due to limitations of the system. We named the resulting frames by combining the most common two of the corresponding SENSEM classes. We have ended up with the classes listed in 6.1.

As can be seen we have additionally adopted the classes `caus-compl-np` and `caus-compl-np-pp` for frames with sentential complements, as they are easy to distinguish from NP complements and provide extra information that can be useful. If needed, they can be mapped to the corresponding frames with NP complements. A more detailed discussion of the SFs that our system acquires and what exactly they stand for can be found in section 6.2.

Choice of verbs

Given the preliminary nature of this project we have decided to choose a small amount of common Spanish verbs, which are comprehensively covered in the SENSEM database. There are ten such verbs, belonging to the two groups studied by Vázquez et al. (2000). They are:

Verbs of change *bajar convertir dejar desatar deshacer llenar preocupar reducir sorprender*

Verbs of path *bajar decir dejar*

Note that the verb-forms *bajar* and *dejar* are listed in both groups, since these verbs have at least one meaning in each of the classes. For example compare 61a and b:

61. (a) ¡Baja la música!

    (b) ¡Baja la persiana!

The first use should be classified as a verb of changes – the entity is affected, i.e. the volume of music becomes lower. In the second use, the entity is moved downwards along some path.

The ten verb forms listed above are those that the system acquires subcategorization for. The different meanings are not distinguished: the system produces verb-form-SF pairings as its output.

In the class of verbs of change, we have also included representatives of the subclass of verbs of psychological change such as *preocupar* and *sorprender*. As will be seen in section 6.4.3, these two verbs show a curious pattern of behavior that caused some interesting problems both for the acquisition system and for the linguist evaluating the system's performance.

This varied group of verbs was chosen with the goal of permitting to test the system with verbs participating in the major alternations studied by Vázquez et al., across a wide range of syntactic constructions.

6.1.2 Chunked corpus

Our most important source of data is the corpus from which SFs are acquired. The corpus is a sample of approximately 370000 words of newswire from the *El Periódico* daily. This sample has been obtained by extracting from a larger corpus all sentences with occurrences of one of the ten verbs studied.

decir	5582
dejar	1729
convertir	1259
reducir	481
bajar	340
sorprender	242
preocupar	179
llenar	165
desatar	80
deshacer	64

Table 6.2: Exemplars of each of 10 verbs in corpus sample

It is POS tagged and processed by a partial parser. The system used to tag and chunk the corpus consists of a suite of tools know as *Integrated Morpho-Syntactic Analyzer* (MS), developed by a group of researchers at the Universitat Politècnica de Catalunya and the Universitat de Barcelona (Atserias et al., 1998). A sentence processed by MS is shown in table 6.3. As can

```
[ word=" ] [ pos=Fe ] [ lema=" ]
[ word=Me ] [ pos=patons ] [ lema=yo ] [ num=s ] [ pers=1 ]
[ word=hace ] [ pos=grup-verb ] [ lema=hacer ] [ num=s ] [ pers=3 ]
[ word=mucha|ilusion ] [ pos=sn ] [ lema=ilusion ] [ num=s ] [ gen=f ]
[ word=debutar ] [ pos=infinitiu ] [ lema=debutar ]
[ word=en|este|espacio ] [ pos=grup-sp ] [ anchor=en ] [ lema=espacio ]
[ word=" ] [ pos=Fe ] [ lema=" ]
[ word=, ] [ pos=Fc ] [ lema=, ]
[ word=dijo ] [ pos=grup-verb ] [ lema=decir ] [ num=s ] [ pers=3 ]
[ word=el|musico ] [ pos=sn ] [ lema=musico ] [ num=s ] [ num=m ]
[ word=punt ] [ pos=Fp ] [ lema=punt ]
EOP
```

Table 6.3: A sentence chunked by MS

be seen each chunk of the sentence is laid out on a separate line. The chunk is described by means of a simple attribute=value scheme. Features are either atoms (`pos=grup-verb`) or pipe-separated lists (`word=mucha|ilusion`). The meaning of the attributes is the following:

word The word-form or series of word-forms making up the chunk.

lema Lemma: the canonical form of the semantic head of the chunk. Compare `word=hace` with `lema=hacer` or `word=mucha|ilusion` with `lema=ilusion`.

pos Part of Speech. One of the grammatical categories the chunk is assigned to.

num Number, s (singular) or p (plural).

gen Grammatical gender, either m (masculine) or f (feminine).

pers Person, either 1, 2 or 3

anchor This feature is used with chunks assigned to `grup-sp` (prepositional phrase), `conj-subord` (subordinating conjunction) etc. to indicate the syntactic head, as opposed to the semantic head, which is indicated by `lema`. For example: `[ word=en|este|espacio ] [ pos=grup-sp ] [ anchor=en ] [ lema=espacio ]`

The most common categories used as values of the pos feature are listed below:

sn Noun Phrase. When a noun phrase is modified by one or more PPs, they are normally not included in the NP chunk.

grup-verb Finite Verb. This chunk includes auxiliaries and orthografically fused clitics.

patons Clitic pronoun.

grup-sp Prepositional Phrase. Nested PP are normally treated as separate chunks.

coord Coordinating cunjunction.

conj-subord Subordinating conjunction.

morfema-verbal Verb clitic (*se*).

sa Adjectival Phrase.

s-a-ms Verbal adjective/participle.

Fe Punctuation: quotation marks.

Fc Punctuation: comma.

Fd Punctuation: colon.

The MS chunking parser does a lot of useful low-level processing on the text. In addition to the sort of information exemplified above, it also attempts to recognize named entities and when it detects one, it puts an appropriate label, such as PERSON, ORGANIZATION or MISC as the value of the `lema` attribute. Informal observations while developing our system seem to indicate that it is quite accurate in assigning the PERSON label but is rather non-discriminating with ORGANIZATION. The relevance of this issue will become apparent in section 6.2. Some other chunker errors that are inherited by our acquisition system will also be commented on in section 6.3.

6.1.3 Spanish WordNet

Initially we intended to use exclusively syntactic information for SF detection, on the argument that it would make the system less dependent on external resources and its design would be overall simpler. Such an approach has certainly been largely successful in SF acquisition for English.

Soon it became obvious that purely syntactic information was inadequate, and to insist on this initial design decision would actually make the implementation more rather than less complicated. The reason for this is the Spanish-specific phenomenon of marking a class of direct-object NPs with the preposition *a*. Although in actual use it is more complex, we can assume that to the first approximation this marking is obligatory for NPs that are human: this includes people, groups of people and metonimic references to those. So information about this feature of semantic heads is essential if we are to be able to distinguish, for example, between direct objects and locatives, or between direct objects and subjects that happen to follow verbs, as is on occasion the case in Spanish.

These motives made us decide to use the Spanish WordNet lexical database as the source of information to use when deciding whether a noun is human or not. The general architecture of the Spanish WordNet is the same as that of WordNet, described in section 4.5.1. The data we have used come from Spanish WordNet 1.0, which is the result of the combined efforts of the following Spanish research groups:

- UPC NLP group at TALP Research Center (`http://www.lsi.upc.es/~nlp/`)
- UB CL group at CLIC. (`http://clic.fil.ub.es/`)
- UNED NLP group. (`http://sensei.lsi.uned.es/NLP/`)

In this study we have used the noun subsystem of the Spanish WordNet. The database itself consists of various files, jointly defining relations between synsets. One file contains the 'variants', that is the actual word-forms. Each record contains the word's POS tag, the ID number of the synset it belongs to, the sense number and a confidence score.

Another file records various properties of synsets and a third one indicates the relations that hold between the synsets. It records the type of relation – the one we were interested in our study is the `has_hyponym` relation – as well as the source and target synsets' IDs and POS tags. With these data it is possible to track noun hyponymy hierarchies, and we have done so in order to determine whether a given noun has meanings with the [+ HUMAN] feature (see the next section for details).

6.2 Implementation

The acquisition system used in the present study is basically a collection of small computer programs that perform different subtasks involved in SF detection, extraction and evaluation. Those tools are developed in Scheme, a programming language from the Lisp family that offers facilities for both functional and imperative programming. The implementation of Scheme used is Gauche (`http://www.shiro.dreamhost.com/scheme/gauche`). Scheme in general and Gauche in particular are well-suited to rapid development of prototype-level tools and to natural language processing.

The steps that the system performs up to and including the stage of SF detection are the following:

```
(((lema . ") (pos . Fe) (word "))
 ((pers . 1) (num . s) (lema . yo) (pos . patons) (word Me))
 ((pers . 3) (num . s) (lema . hacer) (pos . grup-verb) (word hace))
 ((num . s) (gen . f) (lema . ilusion) (pos . sn) (word mucha ilusion))
 ((lema . debutar) (pos . infinitiu) (word debutar))
 ((lema . espacio) (anchor . en) (pos . grup-sp) (word en este espacio))
 ((lema . ") (pos . Fe) (word "))
 ((lema . ,) (pos . Fc) (word ,))
 ((pers . 3) (num . s) (lema . decir) (pos . grup-verb) (word dijo))
 ((num . s) (gen . m) (lema . musico) (pos . sn) (word el musico))
 ((lema . punt) (pos . Fp) (word punt)))
```

Figure 6.1: Chunks converted to S-expressions

Conversion to S-expressions The whole corpus has been converted to Lisp-style S-expressions, which allows the leveraging of facilities that Scheme provides for searching and manipulating lists. Figure 6.1 shows a corpus segment in this notation.

'Humanizing' Prior to any further processing, the NPs and PPs in the chunked corpus were augmented with feature `human` (`yes` or `no`). This information was extracted from the Spanish WordNet.

SF detection Now the actual SF detection takes place. Each SF label has a number of associated patterns defined over attribute-value sets. The system searches the corpus for these patterns for the chosen 10 verbs and records any matches found.

6.2.1 Checking HUMANness

We use a pretty straightforward heuristic to determine whether a particular noun is or not human. The corpus is searched for candidate nouns. To those chunks whose `lema` is `PERSON`, the feature `human=yes` is added directly. We have decided not to apply the same rule to `ORGANIZATION` as it gave too many false positives. Other nouns – the ones that appear as the value of the `lema` feature in chunks whose `pos` is either `sn` (NP) or `grup-sp` (PP) – have to be checked against WordNet. This consists in, for each meaning of the noun, recursively checking if its hypernyms include one of the following synsets:

- 4865:
  {*alma, humano, individuo, mortal, persona, ser_humano*}

- 1778647:
  {*homínido*}

caus-2np →

$$\begin{bmatrix}\text{pos} & \text{sn}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{grup-verb}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{sn}\\ \text{human} & \text{no}\end{bmatrix}$$

caus-compl-np →

$$\begin{bmatrix}\text{pos} & \text{sn}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{grup-verb}\end{bmatrix}, \begin{bmatrix}\text{lema} & \text{que}\\ \text{pos} & \text{conj-subord}\end{bmatrix}$$

pas-se-np-pp/anti-pr-np-pp →

$$\begin{bmatrix}\text{pos} & \text{sn}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{morfema-verbal}\\ \text{lema} & \text{se}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{grup-verb}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{grup-sp}\end{bmatrix}$$

Figure 6.2: Canonical templates for three SFs

- 17008:
  {*grupo, colectivo, agrupación*}

These three synsets have been determined empirically by checking the hypernyms of a number of human and non-human nouns and observing the presence of which synsets discriminated between the two categories. This heuristic seems to work well most of the time. The deficiencies we have observed in WordNet coverage were most pronounced in the absence of many common feminine nouns such as *abogada* or *profesora*: they are meant to be derivable from the masculine forms. At this stage we have chosen to ignore this gap in coverage, so this is an area of possible improvement.

6.2.2 SF templates

The SF detection is performed by taking a list of patterns, which we will call **templates** and scanning the corpus for segments that match those templates. As each template is associated to a SF class, finding a match to the template is equivalent to the detection of a cue for this SF. The templates themselves are roughly equivalent to regular expressions over attribute-value sets. In order to avoid redundancies, each SF has a basic, canonical template associated with it. Additional templates are then derived by means of a set of *metarules.* These are transformers that can add, delete, swap or otherwise alter elements of a template. Figure 6.2 shows the canonical templates for three SFs. For readability they have been represented as sequences of AVMs – they are actually implemented as Scheme lists, and are listed in full in appendix B.

6.2.3 Template expansion

The canonical templates on their own would obviously be not much use. They would only be matched very infrequently and so most opportunities of using a cue would be missed. This is why there is a mechanism that takes templates as input and performs various operations on them, outputting new, transformed templates.

This additional mechanism is stored in a template expansion table, which consists of what could be named metarules. Conceptually, each metarule specifies a left-hand-side pattern, which has to be matched by (a portion) of a template if the metarule is to be applied, and a right hand side, which specifies the result of the application to the input (portion of) template. A template which was output by a metarule can be inputted to another metarule. In fact, it usually is, as the rules in the template expansion table are applied in order to a list of templates. The first metarule is tried on all templates in the initial list, and successfully produced derived templates are appended to the initial list. Now the second metarule is tried on this updated list of templates, as so on. The effect is a greatly expanded list of templates, some of which have been created by applying various metarules. With the current version of the table of metarules the list expands from the initial 22 canonical patterns to 364.

Unfortunately, in the current implementation metarules are not particularly readable: the left-hand-side of a metarule is the same sort of attribute-value set as in the case of templates, but the right-hand-side is actually an anonymous Scheme function which is called with the list representing the portion of the input template that was matched by the left-hand-side, and whose return value is substituted for that matched portion in the template output by the metarule. The complete list of metarules can be found in Appendix C.

As an example of what metarules can be used to achieve, consider the following:

62.
```
(((((pos . grup-verb))
   ((pos . sn) (human . no))) .
   (lambda (m) `(((pos . patons)) ,(list-ref m 0))))
```

This could be represented in a more transparent manner using the familiar AVM notation:

63. $\boxed{1}\begin{bmatrix}\text{pos} & \text{grup-verb}\end{bmatrix}, \begin{bmatrix}\text{pos} & \text{sn}\\ \text{human} & \text{no}\end{bmatrix} \rightarrow \begin{bmatrix}\text{pos} & \text{patons}\end{bmatrix}, \boxed{1}$

This rule takes a template that has a subsequence matching `grup-verb`$_i$ `sn[-human]`, and creates a new template, where this subsequence is replaced with `patons grup-verb`$_i$, where any additional features (such as `lema` or `num`) that `grup-verb`$_i$ had in the original template are preserved in the derived template. Incidentally, this example also shows how the `human` feature is used. The left-hand-side of the metarule specifies a template segment that would match a verb group followed by a non-human NP. This means this metarule will not be applied to a template that explicitly specifies the NP as human. Such a template which would be meant to capture a verb - subject sequence and it would be wrong to covert this sequence to a clitic-pronoun - verb sequence, as the right-hand-side of the metarule does, since subjects cannot be clitic pronouns in Spanish.

Many of the different variations on the basic, canonical pattern associated with each SF can be captured by deriving additional templates with metarules. The phenomena treated in this way in our program include:

Negation The chunker does not include the negative *no* in `grup-verb`, so extra templates are made which have it in the appropriate place (i.e. before any clitic pronouns). Ordering of negation relative to clitics is achieved by ordering the metarules: we can ensure that the pronominalization metarule will be tried on templates produced by the negation metarule by placing the former after the latter on the list of metarules.

Pronominalization Templates with clitic pronouns are derived from templates that specify full non-human NPs (direct objects) or human PPs headed by *a* (direct or indirect objects).

Subject Templates for subject omissions and inversions are also produced, but no templates that specify only the bare `grup-verb` are derived.

Direct speech Templates for direct speech are derived form those that specify standard NPs. Here punctuation is of considerable help.

The actual steps involved in searching sentences for matches to templates can be summarized as follows:

64. (a) Order the expanded list of templates, first by length and then by specificity. The comparison function passed to the sort routine takes two templates and returns the longer one, or if they are of equal length, the one that is specified for more features.

    (b) Repeat for each sentence in the corpus:

      i. Split the sentence into overlapping segments, each of which only contains one finite verb (i.e. one `grup-verb`).

      ii. Repeat for each sentence segment:

        A. Try to match the segment to templates in the sorted list, in order, until the first match is found

        B. If a match is found to one of the templates, record the match data (frame matched, the portion of the segment that matched, the containing sentence) in a table under the verb and the frame name. If no match is found, record the segment under the null frame (labelled `none`).

The steps 64a and 64(b)iiA in combination ensure that the longest and the most specific match will be chosen. Some such rule is necessary to resolve cases when a segment portion matches more than one template, if we want the system to return at most one match. The alternative would be to collect all matches and weight them in some way. In the present study we have opted for simplicity and adopted the first solution. There is no profound motivation for the rule of the longest and most specific match. In fact, the longest match is in many cases wrong, most commonly with non-argumental PPs (we have tried to account for this fact by limiting which prepositions can appear is certain templates involving PPs). Still, this simple heuristic proves adequate most of the time.

The actual matching of a sentence segment to a pattern represented by a template is achieved by an operation of straightforward non-recursive unification. Chunks are represented by sequences of attribute-value sets, and values are always atomic. Thus, for two sequences (the template and some subsequence of a segment) to match, attribute-value sets at the same position in both sequences must have no conflicting values.

The algorithm described above produces a table which can then be printed or displayed in various formats, inspected for hints of possible improvements, and finally used to produce a list of SF-verb pairs hypothesized. Once we have obtained such a list it can be statistically analyzed in order to reject some putative pairings and accept others. The optimal settings of the parameters used in hypothesis testing can be found by manipulating them and observing the system's performance, as measured by the standard evaluation measures described in section 5.1. The following section describes how we measured the performance of our system and presents the results obtained.

6.3 Evaluation

In order to evaluate the performance of the system, the whole corpus was analyzed and the data thus obtained recorded. Also a random sample of 20 occurrences of each of the ten verbs was extracted: 200 sentences in total. The following data was extracted:

- the verb,
- the subcategorization frame detected,
- the text segment that matched it,
- the containing sentence

One copy of this sample was stored as returned by the system. Another copy was manually analyzed and correct SF frames were assigned: these were the reference data. These two datasets were the basis of various evaluation tests. The simplest test we performed simply tested the performance of the system in raw SF detection. The SFs assigned to the 200 verb tokens were compared one by one to those recorded in the manually analyzed reference data: agreement between the two datasets is counted as a 'hit', disagreement a 'miss'. The proportion of hits to all choices made indicates the system's precision in SF detection. The number of tokens assigned correct SFs was 112, so the detection precision was 122/200 = 0.56. This measure can be used during the development of the system in order to check how well enhancements to the detection algorithm translate to improved SF assignment.

Another simple measurement we performed was the following: from both the reference dataset and the system output we extract the verb-SF pairings. Each pairing is counted once, i.e. we count types rather than tokens. From both sets of pairings those involving the null frame `none` are removed. Now by comparing the two sets we can check the standard metrics of type recall and type precision. Note that at this stage we do not do any filtering: we simply accept the pairings that appear in the sample of 200 tokens as valid, as statistical filtering on such a small number of tokens would not make much sense.

The set obtained from the system's data is the *selected* set, while the one extracted from the manually analyzed data is the *target* set. The number of true positives was (40), false positives (21), and false negatives (8). These numbers give a type recall of 0.83, type precision of 0.65 and the F-measure of 0.73. These results are quite good considering that no filtering was performed to remove the false positives, but this is not really surprising as we have used exactly the same data (200 tokens) to hypothesize SF-verb pairs in the selected set as were used in the target set. If we use the SF-verb pairs hypothesized on the basis of the whole corpus, and we accept all of them as valid, precision will suffer, and will tend to get worse with corpus size due to a large number of false positives. For comparison, if the previous test is performed on a selected set based on the whole corpus, precision falls to 0.57, i.e. just about half of the SF-verb types are actually valid.

So if we want to use all the data the system hypothesizes, we need to filter them. For our system we have tried two methods: a relative-frequency cutoff and the binomial hypothesis filter. The relative frequency method only accepts a putative SF if the percentage of times the system detects it for a given verb is higher than some threshold. The optimal level of this figure is normally established empirically, and can vary for different SFs. In our case we have found that the system performed best with values of the minimum relative frequency between 0.03 and 0.07. For the value of 0.07 the precision was at 0.77, recall at 0.7 and the F-measure at 0.74. This can be considered a relatively good result, and the relative frequency method has the advantage of being very simple to apply.

Cutoff	**Precision**	**Recall**	**F**
0.03	0.65	0.85	0.74
0.05	0.72	0.75	0.73
0.07	0.77	0.70	0.74

Table 6.4: Evaluation for filtering with relative-frequency cutoff

We have decided to check if this performance level could be improved on by using a more sophisticated filtering technique, i.e. Brent's binomial hypothesis test. This test is computationally much more expensive as can be appreciated from the number of factorials in the equation in section 5.2.1. It also works best with a good estimate of error probability, which we did not have. The results of filtering with this test for a few values of the error estimate are given in table 6.5. We rejected the null hypothesis for likelihoods below 0.02.

p^e	**Precision**	**Recall**	**F**
0.005	0.62	0.69	0.65
0.010	0.68	0.68	0.68
0.015	0.71	0.66	0.69
0.025	0.76	0.60	0.67
0.050	0.79	0.47	0.59

Table 6.5: Evaluation for filtering with BHT

So from these tests it is not apparent that BHT is an improvement on relative-frequency cutoff. The best way of driving up the performance of the filter would presumably be to use frame-specific values for cutoff and/or error probability. There is certainly potential for better filtering but given the preliminary nature of this study it remains a task for further investigation.

Detection precision	0.56
Token recall	0.87

Table 6.6: Detection precision and token recall

One final measure we have obtained is the token recall. We have created a SF-verb list (lexicon) using data from the whole corpus and applying the frequency-cutoff filter of 0.7. Now for each verb token in the reference data we checked if its manually determined SF was listed in the lexicon. This was the case in 175 cases out of 200, which gives a figure of token recall of 0.87.

The results obtained in these various measures seem encouraging. Although it is not directly comparable to some large-scale SF acquisition systems described in 5.2, the evaluation of our system seems to indicate that even a simple, prototype tool can effectively extract useful information from corpora. It also seems to be the case that with small amount of adaptation a methodology similar to that used for English SF acquisition works reasonably well for Spanish.

6.4 Sources of errors

In this section we briefly review some ways in which our system fails, and as can be seen from the raw detection score, it does in almost half the cases. These failures can be divided into two major groups: either they are inherited from other components the system relies one, or they spring from the limitations of the system itself. There is also a minor third group of issues, where it is not clear if the behavior in question is an error or rather some problem with our assumptions about what the correct behavior is.

6.4.1 Inherited errors

Some of these are caused by the chunking parser that our system relies so heavily on. As an example consider the following sentence:

65. *Fue una rueda de prensa larguísima (50 minutos) llena de explicaciones tácticas, todas ellas encaminadas a ensalzar la validez del sistema, a justificar todas sus decisiones y a liberarse de la responsabilidad de lo que ocurre en el campo.*

The chunker interprets the adjective *llena* as a verb. There are also some cases of confusion between the relative *que* and the subordinating *que*, and a variety of other mistakes. We have not yet quantified the contribution of different types of errors to performance degradation, but from informal observation it seems that parser errors are not the major factor.

The other external resource we rely on is WordNet. There are words such as “abogada” or “nadie” that are not covered by WordNet and thus erroneously treated as non-human by our system. In other cases it is not clear that the NP in question denotes human entities, and yet they are marked with the direct object *a*, for example: *El Parlament deja sin sala de muestras al museo.* The first issue would be easy to fix by augmenting WordNet’s coverage, but this second seems to be rather more demanding.

6.4.2 System’s own errors

There are of course a lot of these, as is to be expected from an experimental tool. One particular example results from the lack of sophistication of the unification operation used in pattern matching. It does not allow variables: an attribute is either specified, and it unifies with an equal value, or is unspecified and unifies with any value. Currently there is no way to constrain two values to be equal without fully specifying them, which prevents the correct treatment of subject-verb agreement. So this additional source of information is unavailable to the system with the consequent drop in performance. Another more fundamental limitation is the fact that the system uses very little semantic information, and wherever semantics is necessary to resolve some ambiguity, there is a possibility of failure. Also lacking are other sources of linguistic knowledge: for examples idioms are not recognized and are treated just as regular usage of verb complementation patterns. It is not clear that *Su gestión dejó mucho que desear* should be used as evidence in determining the subcategorization frame of *dejar*.

6.4.3 Bug or feature?

On occasion the system disagrees with human judges on issues where it is not clearly evident who is right. In the manual analysis of uses of *desatar* as in *Sabe usted qué debe hacer si se desata un incendio en la sala?* we have systematically assumed the `pas-se-np/anti-pr-np` SF, whereas the system always indicated `anti-pr-np-pp/caus-pr-np-pp`, treating the locative PP as argumental. Given a considerable number of examples with this pattern of complementation, we should probably consider whether place is indeed an essential meaning component in a verb which means that something 'violently comes to being'.

Another difficult case is the behavior displayed by two verbs of psychological change *sorprender* and *preocupar*. Although these are often traditionally classified as taking an accusative direct object in sentences such as *A Mercedes la preocupa la salud de su madre* , the preferred form is to use the dative pronoun *A Mercedes le preocupa la salud de su madre*. Our corpus is too small to provide evidence either way, but a Google search for the exact string "a ella la preocupa" returns three hits, while "a ella le preocupa" shows approx. 175 matches. Given this, in the manual analysis of corpus data we assigned such uses to the `caus-np-pp/anti-np-pp` SF. The systems chooses `caus-2np`, in agreement with the traditional analysis. This is not because of any special evidence but precisely because of lack of evidence that would make it possible to distinguish between these two options. In the majority of cases, dative cannot be distinguished from accusative in Spanish clitics, the only clear exception being *la* and *las*. So the system has to make an arbitrary decision, which is probably wrong in the case of *preocupar*, but would be right for *ayudar*. This issue indicates that in order to make some distinctions, an acquisition system should optimally be provided with some linguistic generalizations. In deciding on the subcategorization pattern of *preocupar* it would help to know how each alternative correlates with the occurrence of other SF. Explicit knowledge on the general patterns in diathesis alternations would facilitate decisions on the validity of particular SF-verb pairings. For example, it is known that the absence of the passive alternation such as **Mercedes es preocupada por la salud de su madre* is a strong indication that in the active version *Mercedes* is not the direct object. Such information could be used to guide the learning of an acquisition system.

6.5 Conclusions and further research

We have implemented an experimental tool for the extraction of verb subcategorization information from partially parsed Spanish corpora. We have obtained promising preliminary results at roughly the same level of basic performance as analogous systems developed for English. Even though we have dealt with a reduced number of SF classes and with only ten verbs, nothing prevents us, in principle, from using a system based on the one presented in this study to perform subcategorization acquisition on larger scale.

We have also argued that using more explicit linguistic knowledge, especially about relations between different diathesis alternations, might lead to improved performance of an acquisition system.

A further obvious area of improvement is the hypothesis testing. Here a more informed and individualized treatment of SFs based on empirical tests has a potential to enhance acquisition results. Another idea to explore is the approach adopted by Korhonen (2002). Her semantically-assisted filtering step, based on grouping verbs into syntactico-semantic classes, could be used in Spanish as well, leveraging the research on verb classes done by Vázquez et al. (2000). In order to determine which of these potential directions of further research proves most worthwhile and perhaps to uncover other unsuspected challenges, it is necessary to submit the approach presented in this study to further tests, using larger and more varied amounts of data and quantifying the contribution of different factors and error sources to the final performance. It would also be interesting to check how well the standard metrics applied in this study reflect the system's general usefulness as measured in task-based testing.

Bibliography

Abney, S. (1991). Parsing by chunks. In Abney, S. and Tenny, S., editors, *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, Dordrecht.

Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Màrquez, L., Martí, M. A., Padró, L., Placer, R., Rodríguez, H., Taulé, M., and Turmo, J. (1998). Morphosyntactic analysis and parsing of unrestricted Spanish text. In *First International Conference on Language Resources & Evaluation (LREC'98)*, volume 2, pages 1267–1272, Granada.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montréal.

Baker, C. F. and Ruppenhofer, J. (2002). FrameNet's frames vs. Levin's verb classes. In Larson, J. and Paster, M., editors, *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, pages 27–38.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes.* MIT Press, Cambridge, Mass.

Brent, M. R. (1991). Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 209–214, Berkeley.

Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262.

Briscoe, E. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Carroll, G. and Rooth, M. (1998). Valence induction with a head-lexicalized pcfg. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, pages 36–45, Granada.

Dang, H. T., Kipper, K., Palmer, M., and Rosenzweig, J. (1998). Investigating regular sense extensions based on intersective Levin classes. In Boitet, C. and Whitelock, P., editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 293–299, San Francisco, California. Morgan Kaufmann Publishers.

Ersan, M. and Charniak, E. (1996). A statistical syntactic disambiguator program and what it learns. In *Connectionist, Statistical and Symbolic*

Approaches in Learning for Natural Language Processing, pages 146–157. Springer-Verlag, Berlin.

Fellbaum, C. (1998). A semantic network of English verbs. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, chapter 3. MIT Press, Cambridge, Mass.

Fernández, A., Saint-Dizier, P., Vázquez, G., Benamara, F., and Kamel, M. (2002). The VOLEM project: a framework for the construction of advanced multilingual lexicons. In *Proceedings of the Language Engineering Conference*, Hyderabad.

Gahl, S. (1998). Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 428–432, Montreal.

Gazdar, G., Klein, E., Pullum, G. K., and A., S. I. (1985). *Generalized Phrase Structure Grammar.* Blackwell, Oxford.

Jackendoff, R. (1990). *Semantic Structures.* MIT Press, Cambridge, Mass.

Jackendoff, R. (2002). *Foundations of Language. Brain, Meaning, Grammar, Evolution.* Oxford University Press, Oxford and New York.

Johnson, C. R. and Fillmore, C. J. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 56–62, Seattle.

Johnson, C. R., Fillmore, C. J., Petruck, M. R. L., Baker, C. F., Ellsworth, M., Ruppenhofer, J., and Wood, E. J. (2002). *FrameNet: Theory and Practice.* Word Wide Web, `http://www.icsi.berkeley.edu/~framenet/book/book.html`.

Joshi, A. K. (1985). How much context sensitivity is necessary for characterizing structural descriptions: Tree Adjoining Grammars. In Dowty, L. K. and Zwicky, A., editors, *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, pages 206–250. Cambridge University Press, UK.

Keller, F., Corley, M. W., and S., T. (1999). GSearch: A tool for syntactic investigation of unparsed corpora. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, pages 56–63, Bergen.

Kipper, K., Dang, H. T., and Palmer, M. (2000a). Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

Kipper, K., Trang, D. H., Schuler, W., and Palmer, M. (2000b). Building a class-based verb lexicon using TAGs. In *Fifth TAG+ Workshop.*

Korhonen, A. (2000). Using semantically motivated estimates to help subcategorization acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 216–223, Hong Kong.

Korhonen, A. (2002). Subcategorization acquisition. Technical Report

UCAM-CL-TR-530, University of Cambridge Computer Laboratory, `http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-530.pdf`.

Krifka, M. (2000). Mannner in dative alternation. In *WCCFL 18: Proceedings of the Eighteenth West Coast Conference on Formal Linguistics*, Sommerville/Medford. Cascadilla Press.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 397–404, Maryland.

Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation.* University of Chicago Press, Chicago and London.

Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, pages 235–242, Columbus.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Mass.

Mel'čuk, I. A. and Xolodovič, A. A. (1970). K teorii gramatičeskogo zaloga. 'Towards a theory of grammatical voice'. *Narody Azii i Afriki*, 4:111–124.

Miller, G. (1976). *Language and Perception.* Harvard University Press, Cambridge, Mass.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Pinker, S. (1989). *Learnability and Cognition. The Acquisition of Argument Structure.* MIT Press, Cambridge, Mass.

Pollard, C. and Sag, I. (1987). *Information-Based Syntax and Semantics, Volume 1: Fundamentals.* University of Chicago Press, Stanford.

Pollard, C. and Sag, I. (1994). *Head-Driven Phrase-Structure Grammar.* University of Chicago Press, Chicago.

Pustejovsky, J. (1995). *The Generative Lexicon.* MIT Press, Cambridge, Mass.

Sag, I. and Wasow, T. (1999). *Syntactic Theory. A Formal Introduction.* University of Chicago Press, Chicago.

Saussure, F. ((1919) 1959). *Course in General Linguistics.* McGraw-Hill, New York.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theories.* Center for the Study of Language and Information, Stanford.

Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht.

Vázquez, G., Fernández, A., and Martí, M. A. (2000). *Clasificación verbal. Alternancias de diátesis.* Edicions de la Universitat de Lleida, Lleida.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis.* Karoma, Ann Harbor.

Wood, M. (1993). *Categorial Grammars.* Routledge, London and New York.

Appendix A

Subcategorization classes in the SENSEM database

caus-2np
caus-pr-2np
proc-pr-2np
caus-2np-pp
caus-pr-2np-pp
proc-pr-2np-pp
caus-np-pp
caus-pr-np-pp
proc-pr-np-pp
caus-np-2pp
caus-pr-np-2pp
proc-pr-np-2pp
caus-np
caus-pr-np
proc-pr-np
pas-se-np
pas-se-np-pp
pas-se-pp
pas-se-2pp
pas-se
pas-ser-part-np
pas-ser-part-np-pp
caus-hacer-inf-2np-pp
caus-hacer-inf-2np
caus-hacer-compl-2np
caus-hacer-compl-2np-pp

anti-pr-np
anti-pr-np-pp
anti-np
anti-np-pp
anti-dejar-part-np
anti-dejar-part-np-pp
anti-dejar-adj-np
result-estar-part-np
result-estar-part-np-pp
result-estar-adj-np
refl-pr-np
refl-pr-2np
rcpr-pr-np
rcpr-pr-2np
refl-pr-np-pp

Appendix B

Canonical templates for SFs

```
(((template . caus-2np))
 ((pos . sn))
 ((pos . grup-verb))
 ((human . no) (pos . sn)))
(((template . caus-compl-np))
 ((pos . sn))
 ((pos . grup-verb))
 ((lema . que) (pos . conj-subord)))
(((template . caus-compl-np-pp))
 ((pos . sn))
 ((pos . grup-verb))
 ((anchor . a) (pos . grup-sp))
 ((lema . que) (pos . conj-subord)))
(((template . caus-2np-pp))
 ((pos . sn))
 ((pos . grup-verb))
 ((human . no) (pos . sn))
 ((pos . grup-sp)))
(((template . caus-2np-pp))
 ((pos . sn))
 ((pos . grup-verb))
 ((human . no) (pos . grup-sp))
 ((human . no) (pos . sn)))
(((template . caus-np-pp/anti-np-pp))
 ((pos . sn))
 ((pos . grup-verb))
 ((pos . grup-sp)))
(((template . caus-np-2pp))
 ((pos . sn))
 ((pos . grup-verb))
 ((pos . grup-sp))
 ((xanchor . no) (pos . grup-sp)))
(((template . caus-np/anti-np))
 ((pos . sn))
 ((pos . grup-verb)))
(((template . pas-se-np/anti-pr-np))
 ((pos . sn))
 ((lema . se) (pos . morfema-verbal))
 ((pos . grup-verb)))
(((template . anti-pr-2np/caus-pr-2np))
 ((pos . sn))
 ((lema . se) (pos . morfema-verbal))
 ((pos . grup-verb))
 ((human . no) (pos . sn)))
(((template . pas-se-np-pp/anti-pr-np-pp))
 ((pos . sn))
 ((lema . se) (pos . morfema-verbal))
 ((pos . grup-verb))
 ((pos . grup-sp)))
(((template . pas-se-2pp))
 ((pos . sn))
 ((lema . se) (pos . morfema-verbal))
 ((pos . grup-verb))
 ((pos . grup-sp))
 ((xanchor . no) (pos . grup-sp)))
(((template . imp-se))
 ((lema . se) (pos . morfema-verbal))
 ((pos . grup-verb))
 ((pos . sn)))
(((template . result-estar-part-np))
 ((pos . sn))
 ((lema . estar) (pos . grup-verb))
 ((pos . s-a-ms)))
(((template . result-estar-part-np-pp))
 ((pos . sn))
 ((lema . estar) (pos . grup-verb))
 ((pos . s-a-ms))
 ((pos . grup-sp)))
(((template . result-estar-adj-np))
 ((pos . sn))
 ((lema . estar) (pos . grup-verb))
 ((pos . sa)))
(((template . anti-dejar-part-np))
 ((pos . sn))
 ((lema . dejar) (pos . grup-verb))
 ((pos . sa)))
(((template . anti-dejar-part-np-pp))
 ((pos . sn))
 ((lema . dejar) (pos . grup-verb))
 ((pos . sa))
 ((pos . grup-sp)))
(((template . anti-dejar-adj-np))
 ((pos . sn))
 ((lema . dejar) (pos . grup-verb))
 ((pos . sa)))
(((template . caus-hacer-inf-2np))
 ((pos . sn))
 ((lema . hacer) (pos . grup-verb))
 ((pos . infinitiu))
 ((human . no) (pos . sn)))
(((template . caus-hacer-inf-2np-pp))
 ((pos . sn))
 ((lema . hacer) (pos . grup-verb))
 ((pos . infinitiu))
 ((human . no) (pos . sn))
 ((pos . grup-sp)))
(((template .
  caus-hacer-compl-2np/caus-hacer-compl-2np-pp))
 ((pos . sn))
 ((lema . hacer) (pos . grup-verb))
 ((pos . infinitiu))
 ((lema . que) (pos . conj-subord))
 ((pos . sn)))
```

Appendix C

Metarules

```
((((pos . sn))
 ((pos . grup-verb))) .
 ,(lambda (m) `(((pos . conj-subord) (lema . que)) ,(list-ref m 1))))

((((pos . sn))
 ((pos . grup-verb))) .
 ,(lambda (m) `(((pos . coord) (lema . y)) ,(list-ref m 1))))

((((pos . sn))
 ((pos . grup-verb))) .
 ,(lambda (m) `(,(list-ref m 1)
          ((pos . sn) (human . yes)))))

((((pos . sn))
 ((pos . grup-verb))
 ()
 ) .
 ,(lambda (m) `(,(list-ref m 1)
          ,(list-ref m 2)
          )))

((((pos . grup-verb))) .
 ,(lambda (m) `(((pos . neg) (lema . no)) ,(list-ref m 0))))

((((pos . grup-verb))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(((pos . patons)) ,(list-ref m 0))))


((((pos . grup-verb))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(,(list-ref m 0)
          ((pos . grup-sp) (anchor . a) (human . yes)))))


((((pos . infinitiu))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(,(list-ref m 0)
          ((pos . grup-sp) (anchor . a) (human . yes)))))

((((pos . grup-verb))
 ((pos . sn))
 ((pos . grup-sp) (anchor . a))) .
 ,(lambda (m) `(((pos . patons))
          ,(list-ref m 0)
          ,(list-ref m 1))))

((((pos . grup-verb))
 ((pos . grup-sp) (anchor . a))
 ((pos . sn))) .
 ,(lambda (m) `(((pos . patons))
          ,(list-ref m 0)
          ,(list-ref m 2))))

((((pos . grup-verb))
 ((pos . grup-sp) (anchor . a))
 ((pos . conj-subord) (anchor . que))) .
 ,(lambda (m) `(((pos . patons))
          ,(list-ref m 0)
          ,(list-ref m 2))))

((((pos . grup-verb))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(,(list-ref m 0)
          ((pos . Fd))
          ((pos . Fe)))))

((((pos . grup-verb))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(((pos . Fe))
          ((pos . Fc))
          ,(list-ref m 0))))

((((pos . grup-verb))
 ((pos . grup-sp) (anchor . a))
 ((pos . sn) (human . no))) .
 ,(lambda (m) `(,(list-ref m 0)
        ,(list-ref m 1)
        ((pos . Fd))
        ((pos . Fe)))))

((((pos . sn))
 ((pos . patons))
 ((pos . grup-verb)))
 .
 ,(lambda (m) `(,(list-ref m 1)
        ,(list-ref m 2)
        )))

((((pos . sn))
 ((pos . morfema-verbal))
 ((pos . grup-verb)))
 .
 ,(lambda (m) `(,(list-ref m 1)
        ,(list-ref m 2)
        )))
```