Noname manuscript No.
(will be inserted by the editor)

The Spanish DELPH-IN Grammar

Montserrat Marimon

Received: date / Accepted: date

Abstract In this article we present a Spanish grammar implemented in the *Linguistic Knowledge Builder* system and grounded in the theoretical framework of *Head-driven Phrase Structure Grammar*. The grammar is being developed in an international multilingual context, the DELPH-IN Initiative, contributing to an open-source repository of software and linguistic resources for various Natural Language Processing applications. We will show how we have refined and extended a core grammar, derived from the LinGO Grammar Matrix, to achieve a broad-coverage grammar. The Spanish DELPH-IN grammar is the most comprehensive grammar for Spanish deep processing, and it is being deployed in the construction of a treebank for Spanish of 60,000 sentences based in a technical corpus in the framework of the European project METANET4U (Enhancing the European Linguistic Infrastructure, GA 270893GA)[1] and a smaller treebank of about 15,000 sentences based in a corpus from the press.

Keywords Spanish · Grammar · Deep Processing · HPSG · LKB · DELPH-IN

1 Introduction

This article presents the development of a Spanish grammar as part of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN).[2] DELPH-IN is an international research initiative that, on the basis of contributions from

Universitat de Barcelona
Gran Via de les Corts Catalanes 585
08007-Barcelona
Tel.: +34 93 4034695
Fax: +34 93 3189822
E-mail: montserrat.marimon@ub.edu

[1] http://www.meta-net.eu/projects/METANET4U/.

[2] http://www.delph-in.net/.

its members, has created an open-source repository of software and linguistic resources for Natural Language Processing (NLP) applications.

The linguistic resources that are available in the DELPH-IN repository include precise grammars and treebanks for a wide variety of languages, including English (Flickinger, 2002; Oepen et al, 2002), French (Tseng, 2004), German (Crysmann, 2005), Japanese (Siegel and Bender, 2002; Hashimoto et al, 2007), Korean (Kim and Yangs, 2003), modern Greek (Kordoni and Neu, 2005), Norwegian (Hellan and Haugereid, 2004), Portuguese (Branco and Costa, 2008; Branco et al, 2010), and the Spanish grammar we present in this article.[3]

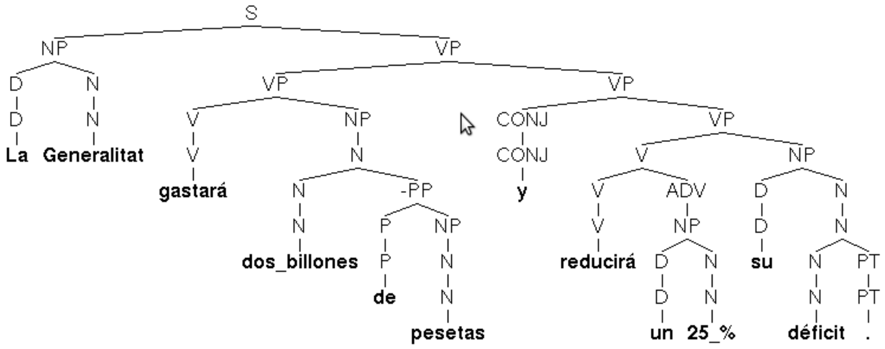
All DELPH-IN grammars are grounded in the theoretical framework of *Head-driven Phrase Structure Grammar* (HPSG) (Pollard and Sag, 1987, 1994), a constraint-based lexicalist approach to grammatical theory where all linguistic objects (i.e., words and phrases) are represented as typed feature structures, and they use the *Minimal Recursion Semantics* (MRS) semantic representation (Copestake et al, 2006). Using unification of typed feature structures, the MRS representation assigns a syntactically flat semantic representation to linguistic expressions which offers, by means of labeling of arguments (arg0, arg1, arg2, arg3) and their co-indexation, a list of semantic relations and a set of syntactic limitations on possible scope relations among them.

DELPH-IN grammars are implemented in the *Linguistic Knowledge Builder* (LKB) system, a grammar development system that includes a parser and a generator, visualization tools, and a set of debugging facilities (Copestake, 2002). The analysis produced by the LKB system for parsed sentences simultaneously displays a traditional syntactic phrase structure tree and an MRS representation. An example is shown in Fig. 1 with the sentence *La Generalitat gastará dos billones de pesetas y reducirá un 25% su déficit* ('The Generalitat will spend 2 billion pesetas and will reduce its deficit by 25%'). Further software tools for grammar development and deployment that are available in the DELPH-IN repository are the `[incr tsdb ()]` competence and performance profiling platform (Oepen and Carroll, 2000) and the PET parser for efficient processing (Callmeier, 2000).

The basis of the development of the Spanish DELPH-IN grammar was an early version of the LinGO Grammar Matrix, developed in 2004 (Bender and Flickinger, 2005). That early version was derived from the types and constraints defined in the English and Japanese DELPH-IN grammars that were cross-linguistically valid, and it included a set of basic grammar types that covered a small set of linguistic phenomena –basic word order, sentential negation, main-clause yes-no questions, and a small set of lexical properties–, thus constituting a 'core' grammar available to grammar developers starting new grammars.[4]

[3] An earlier version on the grammar was briefly presented in (Marimon, 2010).

[4] The current version of the LinGO Grammar Matrix is defined as a web-based interface accessible from: http://www.delph-in.net/matrix/customize/matrix.cgi. A description of it can be found in (Bender et al, 2010).



```
{e2:
 x6:_el_q[]
 e10:_gastar_v[ARG1 x6:named(string),ARG2 x11:part_of]
 x11:part_of[ARG1 x13:_peseta_n]
 x11:non_free_relative_q[]
 x13:udef_q[]
 e2:_y_c[L-INDEX e10:_gastar_v,R-INDEX e21:_reducir_v]
 e21:_reducir_v[ARG1 x6:named(string),ARG2 x25:_déficit_n]
 e26:unspec_loc[ARG1 e21:_reducir_v,ARG2 x27:quant_n]
 x27_art_indef_q[]
 x27:implicit_q[]
 u35:predsort[ARG1 x27:quant_n]
 x25:_su_q[]
 u41:poss[ARG1 x25:_déficit_n,ARG2 x40_pron]
 x40:pronoun_q[]
}
```

Fig. 1 Phrase structure tree and MRS representation for *La Generalitat gastará dos billones de pesetas y reducirá un 25% su déficit.*

In this article we describe how we have refined and extended the core LinGO Grammar Matrix to achieve a broad-coverage grammar, getting a representation close to deep understanding to assist in the solution of NLP tasks such as co-reference annotation, event identification, and normalized predicate-role representation crucial for NLP applications such as information extraction, text summarization, question answering, and machine translation. The Spanish DELPH-IN grammar is the most comprehensive grammar for Spanish deep processing, and it is being deployed in the construction of two treebanks: the IULA Treebank, a large treebank of 60,000 sentences based in a technical corpus, and the Tibidabo treebank, a smaller treebank of about 15,000 sentences based in a corpus from the press.[5] This article, thus, describes significant contributions not only to the verification of the LinGO Grammar Matrix project, but also to theoretical grammar approaches to Spanish and deep linguistic processing in NLP applications.

[5] (Pineda and Meza, 2003, 2005) describe a basic grammar for Spanish implemented in the LKB system that has 15 syntactic rules, 180 lexical entries, and 120 lexical rules.

As we have already mentioned, the Spanish DELPH-IN grammar is open-source and it can be downloaded from http://svn.emmtee.net/trunk.

After this introduction, this article is organized as follows: Section 2 describes the Spanish DELPH-IN grammar components. Section 3 briefly shows the linguistic coverage that we have achieved. Due to space limitations, we will not go into detail about the implementation of all the linguistic phenomena that the grammar deals with; instead, in Section 4, we present the implementation of cliticization phenomena for standard peninsular Spanish –including cliticization, clitic doubling, and clitic climbing– and the closely related phenomena of reflexive and reciprocal constructions, and the so-called impersonal and passive *se*-constructions. Note that these constructions are highly frequent in Spanish and their implementation is central to deep processing of Spanish, but they were not covered by the LinGO Grammar Matrix. Section 5 reports on the performance of the grammar when parsing unrestricted text. Finally, Section 6 presents the conclusions.

2 The Spanish DELPH-IN grammar components

A grammar implemented in the LKB system consists of three basic components: inflectional rules, a lexicon, and syntactic rules. This section describes these components of the Spanish DELPH-IN grammar.

2.1 Inflectional rules

The inflectional rules in the LKB system perform the morphological analysis of surface forms.

The Spanish DELPH-IN grammar, however, uses an external pre-processor, the FreeLing toolkit (Padró et al, 2010),[6] which receives a sentence, morphologically annotates each word by dictionary look-up, and performs Hidden Markov Mode (HMM) disambiguation. The morphological analysis step of FreeLing includes the application of a cascade of specialized processors that annotate punctuation symbols, multi-words, and Named Entities (NEs), including numerical expressions, date/time expressions, ratios, percentages, monetary amounts, and proper names.[7]

The advantages of using such an external morphological analyzer are several. First, FreeLing is an efficient morphological analyzer that provides broad-coverage. Reusing this resource provided faster broad-coverage grammar development than defining inflectional rules for a language such as Spanish that shows a substantial morphology. Note that Spanish verbs are highly inflected; there are 48 distinct simple forms distributed along 8 simple tenses for each of

[6] http://nlp.lsi.upc.edu/freeling.

[7] FreeLing also includes a guesser to deal with words which are not found in the lexicon by computing the probability of each possible PoS tag given the longest observed termination string for that word.

ncms :=
%suffix ()

$${}_{\textit{n\_masc-sg}}\begin{bmatrix} \text{SYNSEM} \mid \text{LOCAL} \begin{bmatrix} \text{CAT} \mid \text{HEAD} & \text{noun} \\ \text{AGR} \mid \text{PNG} & \begin{bmatrix} \text{PN} & \text{3sg} \\ \text{GEN} & \text{masc} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Fig. 2 Mapping rule for the FreeLing NCMS tag into a feature structure.

the three classes or conjugations. Besides, verbs present well-known morphophonemic alternations; (Bosque, 2010), for example, proposes a model that includes 69 classes of irregular verbs (excluding variants that obey some systematic spelling rules).

Second, reusing this external resource reduced the workload required for the development of a large-coverage lexicon. This is because all instances of a given NE class identified by FreeLing share the same syntactic and semantic properties. Then, instead of encoding a lexical entry for each NE instance, we can encode a unique lexical entry for each NE class, while providing a complete coverage for this type of expressions.

Finally, our approach also helps in controlling ambiguity in parsing, hence improving parsing efficiency. In this regard, our goal was, in fact, to provide the best balance between parsing efficiency and accuracy. To achieve this, we pass on to the parser the ambiguities that are not resolved with high reliability by the HMM tagger. Examples are the ambiguity pronoun-conjunction of the word *que* (that), proper names at sentence beginning, or multiword units. For those words and tags, the HMM tagger decisions are ignored (no analysis is discarded) when found at the specified position, passing all possibilities to the deep parsing to be resolved by the symbolic grammar. On the other hand, parsing failures due to discrepancies about the categories assumed in FreeLing and the grammar are avoided by simple substitutions of tags in the interfacing module. This is the case, for instance, of deictic adverbs like *here*, *there*, *today*, *tomorrow*, etc., which FreeLing tags as adverbs while the grammar lexicon encodes them as pronominal signs.

The integration of FreeLing is done using the LKB Simple PreProcessor Protocol (SPPP).[8] Then, we use the inflectional rule component of the LKB system to pass the morphological analysis from FreeLing on to the grammar, by simple mapping of PoS tags into feature structures. An example is given in Fig. 2, which shows the mapping rule for the FreeLing NCMS tag, for masculine singular common nouns, into a partial feature structure.

[8] SPPP assumes that a pre-processor runs as an external process to the LKB that communicates with its caller through its standard input and output channels. See http://wiki.delph-in.net/moin/LkbSppp.

ejemplo_n := n_pp_c_le &

$$\begin{bmatrix}\textsc{stem} & \text{ejemplo}\\ \textsc{synsem}\,|\,\textsc{lkey}\,|\,\textsc{keyrel}\,|\,\textsc{pred} & \text{“\_ejemplo\_n\_rel”}\end{bmatrix}$$

Fig. 3 Lexical entry for *“ejemplo”*.

2.2 The lexicon

The second component of an LKB grammar, i.e., the lexicon, contains the lexical entries of the grammar. Each lexical entry basically consists of a unique identifier, a lexical type, an orthography, and a semantic predicate. Fig. 3 shows, as an example, the lexical entry for the common noun *ejemplo* ('example').[9]

2.2.1 Lexical Types

Lexical types represent the classes of words that a lexicon in the LKB system contains and are defined on the basis of shared syntactic and semantic properties. Following well-established theoretical HPSG proposals, these lexical types are organized into a multiple inheritance type hierarchy (i.e., subtypes may inherit properties from more than one supertype higher in the hierarchy) allowing for lexical generalizations shared by several subtypes to be captured only once.

The LinGO Grammar Matrix provided basic definitions for both open word classes (e.g. intransitive, transitive, and ditransitive verbs, non-argumental nouns and adjectives) and closed word classes (e.g. definite and indefinite articles, propositions selecting NPs and VPs, and complementizers). To obtain a broad-coverage lexicon that would allow coverage for the wide range of syntactic constructions that are found in Spanish, we extended the small set of lexical types provided by the LinGO Grammar Matrix.

Extensions to the LinGO Grammar Matrix required both refining its lexical types and adding new types. On the one hand, we refined the original classification by adding finer-grained lexical types that provided with more detailed descriptions of the valence frames, to distinguish, for instance, clauses in indicative and subjunctive (we give some examples below). On the other hand, we created new lexical types to address a wide range of syntactic constructions that either fall beyond the scope of a 'core' grammar or are characteristic of Spanish, but not of English or Japanese (remember that the early LinGO Grammar Matrix was derived from these two DELPH-IN grammars). Thus, for example, we created a fine-grained classification of closed class items, in-

[9] Lexical type names consist of four fields separated by an underscore. The first three fields specify the part-of-speech, the complements that the type selects for (separated by a hyphen), and optional annotations to distinguish lexical types with the same part-of-speech and complement selection; the last field is always the suffix 'le' (lexical entry). Thus, the type name *n_pp_c_le* that we show in the example is for nouns selecting for a PP_{de} complement, '*c*' indicates that the noun is countable.

Table 1 Number of lexical types.

category	number of lexical types
verb	236
noun	27
adjective	36
adverb	43
preposition	34
determiner	43
pronoun	46
conjunction	38
complementizer	3
auxiliary verb	3
named entity	7
total	516

cluding Spanish clitic pronouns. The result is a comprehensive classification of Spanish words consisting of 516 lexical types.

Table 1 shows the distribution of the lexical types along the syntactic categories. Due to space limitations, we will not go into detail about the 516 lexical types. Instead, here, we will summarize the linguistic constraints that we have taken into account for identifying the different classes.

Lexical types for verbs. Verbal types in the Spanish DELPH-IN grammar are described by a set of lexical types that distinguish fine-grained classes of Spanish verbs that occur in intransitive, transitive, ditransitive, and predicative constructions. These lexical types are defined by a set of subtypes that specify argument relationships (arg0, arg1, arg2, arg3) and a valence frame describing the constituents associated to these arguments. Valence frames are central to identify verb classes, and they are organized into a hierarchy that, first, classifies verbs according to the number and the syntactic category of their arguments, which include NPs, PPs, ADJPs, ADVPs, and CPs. Then, valence frames are further constrained in terms of: optionality (of complements, as well as of marking preposition and of the complementizer introducing finite completive clauses), preposition classes for verbs of location and verbs of movement (constraints on the marking prepositions that are allowed to co-occur with verbs are set on the lexical items), control and raising relations, mood (indicative or subjunctive) of clausal subjects and complements, pronominal clitics, and, finally, those frame alternations that in the grammar are handled by means of lexical rules and that we will see below.

These constraints are organized into a multiple inheritance type hierarchy that we partially show in Fig. 4, which shows the different lexical classes we considered for transitive verbs taking a nominal subject and a propositional complement.

Briefly, as can be observed in the figure, all these verbs belong to the supertype *v_cp_prop-or-ques* for verbs taking verbal complements. This type is divided into four subtypes: *v_cp_ques* for verbs taking question, *v_cp_prop* for

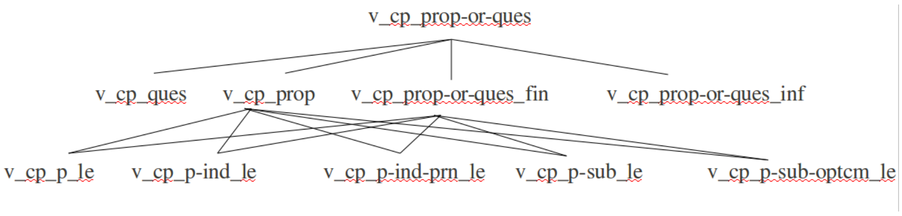


Fig. 4 Lexical type hierarchy for verbs taking nominal subjects and verbal complements.

verbs taking propositions, *v_cp_prop-or-ques_fin* for verbs taking finite verbal complements, and *v_cp_prop-or-ques_inf* for verbs taking nonfinite verbal complements. The second and the third subtypes are the basis of our cross-classification that includes five subtypes: *v_cp_p_le* for verbs taking finite clauses both in indicative and in subjunctive, e.g., *entender* ('to understand') as in (1.a); *v_cp_p-ind_le* for verbs taking finite clauses in indicative, e.g., *saber* ('to know') as in (1.b); *v_cp_p-ind-prn_le* for pronominal verbs taking finite clauses in indicative, e.g., *figurarse* ('to think') as in (1.c); *v_cp_p-sub_le* for verbs taking finite clauses in subjunctive, e.g., *querer* ('to want') as in (1.d); and, finally, *v_cp_p-sub-optcm_le* for verbs taking finite clauses in subjunctive where the complementizer is optional, e.g., *sentir* ('to be sorry') as in (1.e).

(1) a. *Entiende que no sea/será fácil alcanzar el objetivo.*
'He understands that it will not be easy to achieve the goal.'
b. *Este año sabía que podía/*pudiera.*
'This year I knew that I could.'
c. *Uno se figura que el otro se figura/*figure y así se montan la cosas.*
'One thinks that the other thinks and then things happen.'
d. *Quiere que los partidos paguen/*pagan por sus miembros corruptos.*
'He wants that the political parties pay for their corrupt members.'
e. *Siento (que) no haya tenido la oportunidad de competir.*
'I am sorry that he did not have the opportunity to compete.'

Lexical types for nouns. Nominal types in the Spanish DELPH-IN grammar are also described by a set of lexical types that, in identifying the classes of Spanish common nouns, associate arguments to complements, if any. Complements of nouns in Spanish are restricted to marked NPs (e.g., *los amigos de Peter* ('Peter's friends'), *la preocupación de María por sus hijos* ('Mary's concern for her children')), and marked finite and nonfinite verbal complements (e.g., *una ventaja de que no llueva es que no hay ocasión de perder el paraguas* ('an advantage that it does not rain is that there is no chance of losing the umbrella')). Our nominal types also identify several classes of quantifying nouns, including partitive, pseudo-partitive, group nouns, and temporal measure nouns (e.g., *mayoría de estudiantes* ('majority of students'), *grupo de personas* ('group of people'), *quilo de peras* ('kilo of pears'), *semanas de negociación* ('weeks of negotiation'). Further constraints classify common nouns into uncountable, countable and/or mass, whereas, lexical semantics (e.g., human, animate, semiotic,...) are specified in each lexical entry.

Lexical types for adjectives. Spanish adjectives in the Spanish DELPH-IN grammar are cross-classified on the basis of their valence frames, which describe the syntactic characteristics of their complements, and further constraints that indicate: control and raising relations, their position within the NP (i.e., whether they precede and/or follow the noun they modify; e.g., *mero hecho/*hecho mero* ('the mere fact'), *vaso lleno/*lleno vaso* ('a full glass'), *mirada dulce/dulce mirada* ('sweet look')); their grade (i.e., positive, comparative, superlative); whether they are gradable or not (e.g., *una revista muy cara/*muy mensual* ('a very expensive/*very monthly journal')); whether they are intersective or scopal; and whether they are predicative or not, and, if they are predicative, the copula verbs they co-occur in predicative constructions, which allows to distinguish, for example, the two reading of *atento*, as in *Juan está atento a las notícias* ('Juan pays attention to the news') and *Juan es atento con sus compañeros de clase* ('Juan is attentive to their classmates').

Lexical types for adverbs. Leaving apart types for closed classes of adverbs that identify various classes of deictic, relative, interrogative, and degree adverbs, lexical types for adverbs classify adverbs into scopal or intersective adverbs, which in turn have subtypes specifying whether they may co-occur with degree adverbs (e.g., **muy quizás/muy probablemente iremos al cine* ('*very maybe/very probably we will go to the cinema')), and their position with respect to the element they modify (e.g., *recién llegado/*llegados recién* ('reciently arrived')).

Lexical types for closed class items. The Spanish DELPH-IN grammar contains a fine-grained classification of closed class items that distinguishes the different types of Spanish determiners (definite and indefinite articles, universal and non-universal quantifiers, and demonstrative, possessive, comparative, relative, interrogative, and exclamative determiners), prepositions (marking prepositions and prepositions that subcategorize for NPs, VPs, APs, ADPs, and PPs, and head PPs that modify verbs, nouns, adjectives and PPs), pronouns (personal, definite and indefinite intransitive, partitive and pseudo-partitive, relative, and interrogative pronouns), conjunctions (coordinating and subordinating conjunctions), complementizers, and auxiliary verbs.

Lexical types for NEs. Finally, the Spanish DELPH-IN grammar has lexical types that distinguishes the different classes of NEs annotated by FreeLing; i.e., numerical expressions, date/time expressions, ratios, percentages, monetary amounts, and proper names (cf. Section 2.1).

Table 2 Number of words and lexical entries, and average numbers of entries per word.

category	number of words	number of entries	#entries/word
verb	4,314	7,973	1.85
noun	27,686	27,939	1.01
adjective	10,229	10,433	1.02
adverb	4,050	6,792	1.68
preposition	256	956	3.73
determiner	45	47	1.04
pronoun	66	70	1,06
conjunction	135	145	1.07
complementizer	4	5	1,25
auxiliary verb	1	3	3
named entity		9	
total	46,786	54,372	1,16

Table 3 Number of words distributed according to the number of lexical entries.

category	1e	2e	3e	4e	5e	6e	7e	8e	9e	10e	11e
verb	2,477	894	481	257	109	46	24	12	6	3	5
noun	27,436	247	3	-	-	-	-	-	-	-	-
adjective	10,292	53	73	15	-	-	-	-	-	-	-
adverb	4,071	2,700	21	-	-	-	-	-	-	-	-

2.2.2 Lexical Entries

In order to achieve broad-coverage, DELPH-IN grammars require broad- coverage lexicons; i.e., a large amount of lexical entries that instantiate the lexical types. The Spanish DELPH-IN grammar has a full coverage lexicon of closed word classes and it contains 53,137 lexical entries for open word class items.[10] The grammar also includes a set of generic lexical entry templates for open classes to deal with unknown words. These lexical entry templates are activated when the system cannot find a particular lexical entry to apply to provide robustness to the grammar.

Table 2 shows the number of words and lexical entries that we have for each syntactic category.[11] The average numbers of entries per word for open class categories is: 1.85 for verbs, 1.01 for nouns, 1.02 for adjectives, and 1.68 for adverbs; however, some verbs have as many as 11 lexical entries encoding their different valence frames and readings. Table 3 shows, for open word classes, the number of words distributed according to the number of lexical entries they have.

[10] For the sake of simplicity, the Spanish DELPH-IN grammar does not bind words to lexical types in the morphological lexicon of the FreeLing toolkit, which has more than 500,000 full-form entries. This approach also allows the two components, which have been developed independently, to be maintained independently of each other.

[11] This table also shows the number of lexical entries we have defined for NEs.

Table 4 Number of type alternations.

category	num. of type alternations
verb	700
noun	45
adjective	58
adverb	38

Finally, Table 4 gives the number of *type alternations* that we find in open class words. Type alternations are the alternations of two or more lexical types that are found in the words of the lexicon.

Type alternations show up different word classes as defined by the set of lexical types that depict the range of valence alternations. An example is the alternation *v_np_le/v_cp_p_sub_le* that we find in transitive verbs e.g., *querer* ('to want') to encode the different syntactic forms we find in their complement position: NP and sentential complements, as in the following example: *quiero silencio/quiero que te calles* ('I want silence'/'I want you to shut up').

However, a large number of type alternations are motivated by the implementation of polysemic words, since different lexical types are used to encode the different semantic readings of a given word. An example is the alternation *n_pp_psd-part_le/n_pp_c_le* that we find in words like *diente* and *vaso* to encode the pseudo-partitive readings ('clove (of garlic)', 'glass (of water)') and the body part readings ('tooth', 'vessel').

2.2.3 Lexical Rules

As we have just described, alternations between valence frames such as present in transitive verbs are expressed in the lexicon, by means of separate lexical entries. These alternations, however, primary refer to the distinction between NP and sentential complements, both propositions and questions. The alternation between finite and nonfinite clauses that is also found in their complement position (e.g., *quiero que vengas/quieres venir* ('I want you to come'/'you want to come')) is predicted by a lexical rule that changes the verb form of the complement and provides control constraints.

Lexical rules, thus, are unary rules that perform valence changing operations on lexical items (i.e., they apply on lexemes before the application of inflectional rules), generating new lexical items, and, in that, they reduce the number of lexical entries to be manually encoded in the lexicon.

Lexical rules in the Spanish DELPH-IN grammar are also used to account for the following verb alternations (or *diathesis alternations* cf. (Levin, 1993)):

- Active/passive alternation:
  Desalojaron los 11 pisos del inmueble. / Los 11 pisos del inmueble fueron desalojados.
  'They evacuated the 11 flats of the building.' / 'The 11 flats of the building were evacuated.'

- Causative/inchoative alternation:
  Aumentaron los precios industriales un 0,2% en abril. / Los precios industriales aumentaron un 0,2% en abril.
  'In April, they increased the industrial prices by 0,2%.'/ 'The industrial prices increased in April by 0,2%.'
- Personal/impersonal alternation:
  Contrataron a tres estudiantes para el proyecto. / Se contrató a tres estudiantes para el proyecto.
  'They hired three students for the project.' / 'Three student were hired for the project.'
- Active/passive (with *se*) alternation:
  Proyectarán imágenes de su participación en diversas películas. / Se proyectarán imágenes de su participación en diversas películas.
  'They will show images of his participation in several movies.' / 'Images of his participation in several movies will be shown.'

Finally, lexical rules are also used in the implementation of cliticization phenomena, as we will see below.

The Spanish DELPH-IN grammar has a total of 70 lexical rules.

2.3 Syntactic rules

The third component of a grammar implemented in the LKB system are the syntactic rules. Syntactic rules are phrase structure rules that combine words and phrases into larger constituents and compositionally build up their semantic representation.

The LinGO Grammar Matrix provided basic definitions of head-subject, head-complement, and head-adjunct HPSG schemata, together with definitions of head-initial and head-final phrase types, and head-only (or unary) phrase types dealing with optionality and extraction. The LinGO Grammar Matrix, therefore, already provided material to account for a description of a core Spanish grammar. However, in order to obtain a broad-coverage grammar, we extended these phrase structure type definitions and implemented fine-grained types that provided with more constrained definitions of HPSG schemata. So, for example, we distinguished five rules for inverted subjects dealing with: inverted subjects in declarative sentence with verbal heads that require inverted word order (2.a), inverted subjects in relative (2.b) and interrogative clauses (2.c), inverted subjects in imperatives (2.d), inverted subjects of infinitives (2.e), and inverted coordinated subjects, which, as we illustrate in (2.f), may show partial agreement.

(2) a. *A Guillem le gustan las espinacas.*
'Guillem likes spinach.'
b. *Este es el coche que compró mi hermano.*
'This is the car that my bother bought.'

Table 5 Number of phrase structure rules of the Spanish DELPH-IN Grammar.

HPSG schemata	num. of phrase structure rules
head-subject	6
head-complement	6
head-adjunct	48
head-marker	1
head-specifier	9
head-filler	39
non-headed phrase	85
unary phrase	13
other	23
total	230

c. *¿Qué coche compró tu hermano?.*
'Which car did your brother buy?'
d. *Vete tú, ya me quedo yo.*
'(You) leave, I stay.'
e. *La clase acabó tras irrumpir los manifestantes en el aula.*
'The lecture finished when the protesters burst into the room.'
f. *En la corte existía/existían el favoritismo y la corrupción.*
'Existed in the court patronage and corruption.'

The current version of the Spanish DELPH-IN grammar has 230 phrase structure rules. Table 5 distributes these rules along headed-phrase and non-headed-phrase (for coordination) HPSG schemata, unary phrase structure rules (for optionality and extractions), and "other", which includes unary rules –dealing, for instance, with bar NPs–, binary rules –for punctuations marks, for instance–, and quaternary rules for verbal ellipsis (covering gapping and conjunction reduction).

3 The linguistic coverage of the Spanish DELPH-IN grammar

The Spanish DELPH-IN grammar deals with a wide range of constructions in Spanish, including: main clauses with canonical surface word order and word order variations, valence alternations, determination, agreement, null-subject, compound tenses and periphrastic forms, raising and control, passives, (basic) comparatives and superlatives, all types of relative clauses, unbounded dependency constructions, cliticization phenomena, constructions with *se*, coordination, and nominal and verbal ellipsis.

Due to space limitations, only some of these phenomena can be described. We will focus on the cliticization phenomena in the standard peninsular Spanish –including cliticization, clitic doubling, and clitic climbing– and the closely related phenomena of reflexive and reciprocal constructions, and the so-called impersonal and passive *se*-constructions. As we have already pointed out, these constructions are highly frequent in Spanish and their implementation is central to deep processing of Spanish, but they were not covered by the LinGO

Grammar Matrix. We will present the main features of the implementation of these phenomena and we will illustrate the different semantic representation that the grammar produces for these constructions.

4 The Spanish clitic pronouns

4.1 Linguistic description

4.1.1 Cliticization

Spanish clitic pronouns are unstressed object pronouns that appear adjacent to a host verb, either attached to its right, the so-called *enclitics*, or as independent lexical units in front of it, known as *proclitics*. Infinitives, gerunds, and non-negated imperatives have enclitic pronouns (3.a-c), verbs in personal forms always require proclitics (3.d), and past participles cannot have clitics (3.e).

(3) a. *Quiero comprarlo.*
want to buy-clitic (acc)
'I want to buy it.'
b. *Estoy comprándolo.*
am buying-clitic (acc)
'I am buying it.'
c. *Cómprenlo. / No lo compren.*
buy-clitic (acc) / don't clitic (acc) buy
'buy it.' / 'don't buy it.'
d. *Lo compro/compraba/compré/compraré.*
clitic (acc) buy/bought/will buy
'I buy/bought/will buy it.'
e. **He comprádolo.*
have bought-clitic (acc)
'I have bought it.'
f. *Lo he comprado.*
clitic (acc) have bought
'I have bought it.'

As we show in (3.e-f), in compound tenses, Spanish clitics must "climb" in the syntactic structure and they must appear as proclitics in front of the auxiliary verb *haber* ('to have'). These phenomenon is referred to as *clitic climbing.*

Clitic climbing can also occur with modal and aspectual verbs (4.a-b), subject-control verbs (4.c), causative verbs (4.f), and perception verbs (4.g). Thus, if one of these verb classes appears, the clitic may attach to the main verb or it may stay within the embedded verb. But the clitics that belong to the embedded clause need to form a cluster; either they all attach to the main verb or they all stay within the embedded verb, and sentences like (4.d-e) are ungrammatical. Note that in (4.f-g) the accusative clitic is an argument of the embedded verb and the dative clitic is an argument of the causative verb and, therefore, the two clitics can be separated.

(4) a. *Puedo hacerlo. / Lo puedo hacer.*
can do-clitic (acc) / clitic (acc) can do
'I can do it.'
b. *Sigo haciéndolo. / Lo sigo haciendo.*
continue doing-clitic (acc) / clitic (acc) continue doing
'I countinue doing it.'
c. *Quiero hacerlo. / Lo quiero hacer.*
want to do-clitic (acc) / clitic (acc) want to do
'I want to do it.'
d. **Me quiere darlo.*
clitic (dat) wants to give-clitic (acc)
'S/he wants to give it to me.'
e. **Lo quiere darme.*
clitic (acc) wants to give-clitic (dat)
'S/he wants to give it to me.'
f. *Me permitieron hacerlo. / Me lo permitieron hacer.*
clitic (dat) allowed to do-clitic (acc) / clitic (dat) clitic (acc) allowed to do
'They allowed me to do it.'
g. *Me vieron hacerlo. / Me lo vieron hacer.*
clitic (dat) saw to do-clitic (acc) / clitic (dat) clitic (acc) saw to do
'They saw me doing it.'

Unlike French and Italian, where clitics and full phrases are considered to be in strict complementary distribution within the clause, Spanish clitic pronouns may also appear together with the complement they refer to, in what is known as *clitic doubling* constructions.

– IO-doubling is always possible, and it is obligatory when the complement is a strong pronoun (5.a), and in constructions that introduce a benefactive (5.b), an experiencer (5.c), or an inalienable possessor (5.d), among others.

(5) a. *Le dí el regalo a él.*
clitic (dat) gave the present to him
'I gave the present to him.'
b. *Le preparé la cena a Guillem.*
clitic (dat) prepared dinner for Guillem
'I prepared dinner for Guillem.'
c. *A Guillem le gustan las espinacas.*
Guillem clitic (dat) likes spinach
'Guillem likes spinach.'
d *A Guillem le duele la muela.*
Guillem clitic (dat) hearts his tooth
'Guillem has a toothache.'

– DO-doubling is also obligatory when the complement is a strong pronoun (6.a), and it is preferred when the complement refers to an human entity and it is realized by the pronoun *todo* ('everything') (6.b), a numeral preceded by an article (6.c), and by the indefinite pronoun *uno* ('one') when it refers to the speaker (6.d); otherwise DO-doubling is not allowed.

(6) a. *Me vieron a mí.*
clitic (acc) saw me
'They saw me.'
b. *Yo lo sé todo.*
I clitic (acc) know everything
'I know everything.'
c. *Los ví a los cuatro.*
clitic (acc) saw the four
'I saw the four of them.'
d. *Si la oyen a una hablando, se ponen furiosos.*
If clitic (acc) hear one talking, go mad
'If they hear one talking, they go mad.'

4.1.2 Reflexive and reciprocal constructions

The clitic pronouns *me*, *nos*, *te*, *os*, and *se* can also appear in *reflexive* and *reciprocal* constructions, both as enclitics and as proclitics (7).

(7) a. *Te peinas. / Peínate.*
clitic (reflex) comb / comb-clitic (reflex)
'you comb your hair.' / 'Comb your hair.'
b. *Te peinas el pelo. / Peínate el pelo.*
clitic (reflex) comb your hair / comb-clitic (reflex) your hair
'you comb your hair.' / 'Comb your hair.'
c. *Nos abrazamos llorando. / Abrazémonos.*
clitic (reflex) hugged crying. / hug-clitic (reflex)
'We hugged each other crying.' / 'Let's hug each other.'

In these constructions, pronouns substitute direct object and indirect object and are co-indexed with the subjects, which in reciprocal constructions are always plural or coordinated.

In addition, these clitic pronouns are also found with so-called *inherent reflexive* verbs (or *pronominal verbs*); i.e., verbs which require a clitic pronoun co-indexed with the subject and which lack the corresponding non-reflexive form (8).

(8) *Te resfriarás.*
clitic (reflex) will caught a cold
'You will caught a cold.'

4.1.3 Constructions with se

The form *se* can also appear in the so-called *passive* and *impersonal se* constructions, which we illustrate in (9.a) and (9.b), respectively.

(9) a. *Se proyectarán imágenes de su participación en diversas películas.*
passive-marker will show images of his participation in several movies
'Images of his participation in several movies will be shown.'
b. *Se contrató a tres estudiantes para el proyecto.*
impersonal-marker hired three students for the project
'Three students were hired for the project.'

In these constructions, the verb occurs with the clitic *se*, which is not a verbal argument but a grammatical marker.

In *passive constructions* the verb has a unique argument (arg2) which is the syntactic subject. This construction can only appear with transitive verbs.

Unlike passives, *impersonal constructions* do not have an overt subject; in this construction, the verb appears in third singular person and the complement is the arg2. Another difference is that this construction can appear not only with transitive verbs, but also with intransitive verbs (10.a), unaccusative verbs (10.b), and verbs taking sentential complements (10.c).[12]

(10) a. *Se cree en milagros.*
impersonal-marker believes in miracles
'One believes in miracles.'
b. *Aquí se vive bien.*
Here impersonal-marker lives well
'Here life is good.'
c. *Se ve cómo caen las gotas de lluvía.*
impersonal-marker sees how fall the raindrops
'One sees how the raindrops fall.'

4.2 Implementation

Based on a set of well-known criteria proposed by (Zwicky and Pullum, 1983) to distinguish between affixes and clitics, Spanish clitic pronouns are commonly considered in the literature as verbal affixes that have to be treated in the morphology (see, for instance, (Fernández Soriano, 1999)), similarly to the analysis proposed by (Miller and Sag, 1997) for French and (Monachesi, 1998) for Italian within the theoretical framework of HPSG. However, due to the Spanish orthographic conventions (in Spanish orthography the proclitics are written separately and the enclitics are written attached to the verbxs) this

[12] (Mendikoetxea, 1999), in addition, distinguishes *medio se*-constructions, where, like in *passive constructions*, the verb has a unique argument (arg2) which is the syntactic subject and which usually precedes the verb. In the Spanish DELPH-IN grammar we treat *medio constructions* as a sub-class of *passive constructions*.

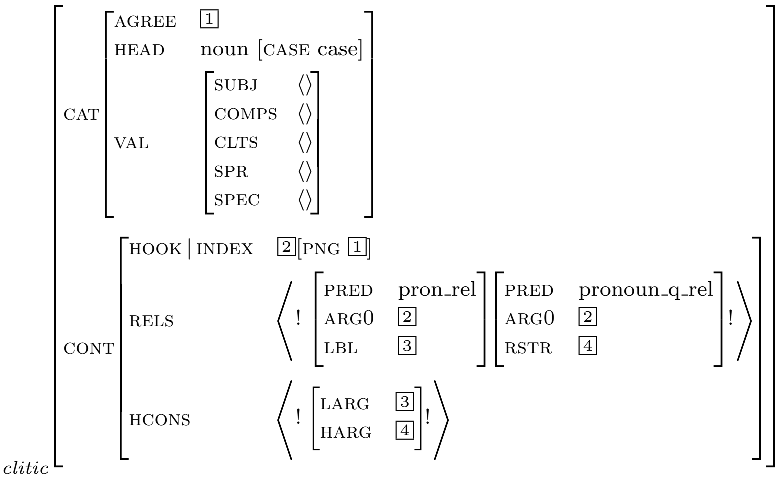

AGREE 1
HEAD noun [CASE case]
CAT
VAL
SUBJ ⟨⟩
COMPS ⟨⟩
CLTS ⟨⟩
SPR ⟨⟩
SPEC ⟨⟩
HOOK | INDEX 2[PNG 1]
RELS ⟨! [PRED pron_rel, ARG0 2, LBL 3] [PRED pronoun_q_rel, ARG0 2, RSTR 4] !⟩
CONT
HCONS ⟨! [LARG 3, HARG 4] !⟩
clitic

Fig. 5 Type for clitics.

approach is not adopted in the Spanish DELPH-IN grammar, where enclitics are treated in the inflectional rule component of the LKB system and proclitics are treated in the syntax, as we will see below.[13] Thus, in the Spanish DELPH-IN grammar, clitics are not considered featural information used in morphology and phonology for the realization of the cliticized verb form, as in (Miller and Sag, 1997) and (Monachesi, 1998),[14] but syntactically independent words, which are members of a CLTS list of their host verb.

Fig. 5 shows the basic definition of Spanish clitic pronouns, distinguished by the values of the features CASE and PNG (for person, number, and gender). Briefly, the MRS is defined in the feature CONT with features HOOK, RELS (for relations), and HCONS (for handle constraints). The attribute HOOK introduces the INDEX attribute, which denotes the index variable of the clitic itself and which is token identical with the ARG0 attribute of the pronoun relation (*pron_rel*) within the RELS list.[15] Note that pronouns in the DELPH-IN Spanish grammar lexically introduce a quantifier relation (*pronoun_q_rel*). Scopal constraints which hold between the pronoun and the quantifier relation is set in the HCONS feature.

4.2.1 Cliticization

To account for cliticization, we have implemented several Complement Cliticization Lexical Rules (CCLRs) that allow the realization of clitic pronouns

[13] In the implementation of modern Greek clitic doubling constructions in the modern Greek DELPH-IN grammar, proclitics are also treated in the syntax (Kordoni and Neu, 2005). (Pineda and Meza, 2005) also propose this dual approach to Spanish object clitics.

[14] In (Monachesi, 1998) clitics are members of the feature CLTS and in (Miller and Sag, 1997) are members of the ARG-ST (argument structure) of the verb.

[15] Boxed numbers indicate that two features are token-identical.

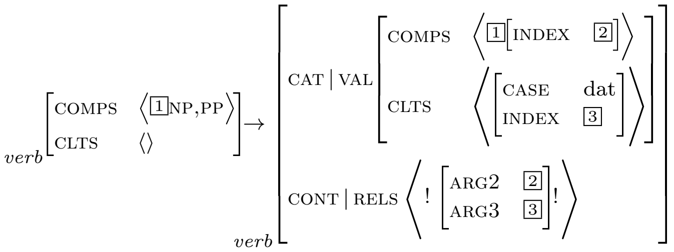

verb
COMPS ⟨1NP,PP⟩
CLTS ⟨⟩
→
verb
CAT | VAL
COMPS ⟨1[INDEX 2]⟩
CLTS ⟨[CASE dat, INDEX 3]⟩
CONT | RELS ⟨! [ARG2 2, ARG3 3] !⟩

Fig. 6 Dative Complement Cliticization Lexical Rule.

```
{e2:
 x4:pronoun_q[]
 e2:_comprar_v[ARG2 x8:_regalo_n,ARG3 x4:pron]
 x8:undef_q[]
}
```

Fig. 7 MRS representation of *les compré regalos.*

as arguments.[16] These rules remove one element in the COMPS list and add to a CLTS list a clitic pronoun whose INDEX is token-identical to the corresponding argument feature of the verb's relation.

Fig. 6 shows the Dative CCLR trigged by ditransitive verbs, and Fig. 7 illustrates the MRS representation that the grammar produces for cliticization with the sentence *les compré regalos* (clitic (dat) bought presents ('I bought presents for them.')), where the clitic pronoun instantiates the ARG3 of the verb's relation.

4.2.2 Clitic doubling

To allow clitic doubling constructions, we have implemented 9 Clitic Doubling Lexical Rules (CDLRs). These rules also have the effect of adding a clitic pronoun to the CLTS list, but in these rules verbal complements are maintained, and the INDEX of the clitic is token-identical to the value of a feature AFFIX in the verb's relation. These rules also restrict the AGREE(ment) features of the clitic to be identical with the AGREE features of the complement.

Fig. 8 shows the Dative CDLR trigged by ditransitive verbs, and Fig. 9 shows the MRS representations produced for clitic doubling, exemplified with the sentence *les compré regalos a los niños* (clitic (dat) bought presents for the children ('I bought present for the children')), where the the clitic instantiates the AFFIX of the verb's relation.

[16] The Spanish DELPH-IN grammar has 14 CCLRs. Diversification of the CCLR allows to control the order within the clitic cluster when more than one complement is cliticized (imposing the additional constraint that the "spurious *se*" is used instead of the dative clitic when it precedes third person accusative clitics) and when object clitic pronouns occur in reflexives and the impersonal constructions. Alternatively, to control the order within the clitic cluster, (Pineda and Meza, 2005) develop a clitic lexicon consisting a set of 100 clitic pronoun sequences.

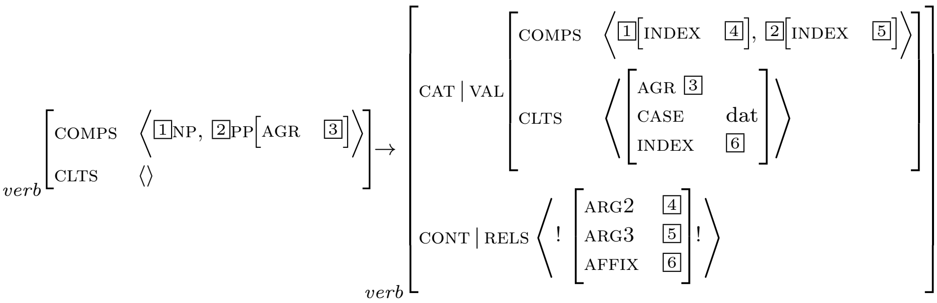

verb
COMPS ⟨1NP, 2PP[AGR 3]⟩
CLTS ⟨⟩
→
verb
CAT | VAL
COMPS ⟨1[INDEX 4], 2[INDEX 5]⟩
CLTS ⟨[AGR 3, CASE dat, INDEX 6]⟩
CONT | RELS ⟨! [ARG2 4, ARG3 5, AFFIX 6] !⟩

Fig. 8 Dative Clitic Doubling Lexical Rule.

```
{e2:
 x4:pronoun_q[]
 e2:_comprar_v[AFFIX x4:pron,ARG2 x8:_regalo_n,ARG3 x9:_niño_n]
 x8:undef_q[]
 x9:_el_q[]
}
```

Fig. 9 MRS representation of *les compré regalos a los niños.*

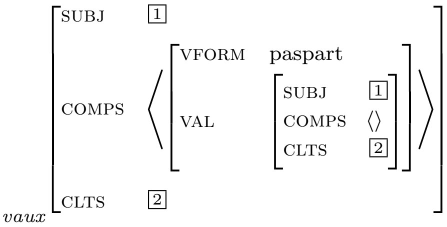


Fig. 10 Type for the auxiliary *haber* in the Spanish DELPH-IN grammar.

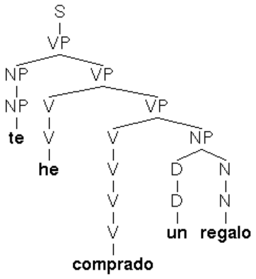
4.2.3 Clitic climbing

Our approach to clitic climbing follows from lexical constraints and local syntactic combination.

In compound tenses, the CLTS requirements of the participle lexically determine those of the auxiliary's lexeme; i.e., auxiliaries and participles they select for share their clitics, as described in Fig. 10. Because auxiliaries select for saturated participles, they first combine with VP complements and then with pronominal clitics, producing, for example, phrase structure trees and MRS representations like Fig. 11 for (11).

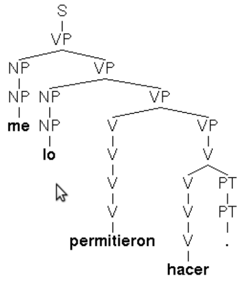
(11) *Te he comprado un regalo.*
clitic (dat) have bought a present
'I have bought a present for you.'

The same approach has been adopted in the analysis of periphrastic and subject-control verbs, whose value for CLTS are also token-identical to the value of the CLTS list of their verbal complement.



```
{e2:
 x4:pronoun_q[]
 e2:_comprar_v[ARG2 x9:_regalo_n,ARG3 x4:pron]
 x9:art_indef_q[]
}
```

Fig. 11 Phrase structure tree and MRS representation for *te he comprado un regalo*



```
{e2:
 x4:pronoun_q[]
 x9:pronoun_q[]
 e2:_permitir_v[ARG2 e16:_hacer_v,ARG3 x4:pron]
 e16:_hacer_v[ARG1 x4:pron,ARG2 x9:pron]
}
```

Fig. 12 Phrase structure tree and MRS representation for *Me lo permiten hacer*

In the case of causative and perception verbs that we illustrated in (4.f-g) and we repeat in (12.a-b), clitic climbing can be described in terms of 'clitic composition'; i.e., the two clitics represent arguments of different verbs.[17]

In this case, first, the clitic requirements of the complement are lexically passed up to the CLTS list of causative (or perception verb), then the Dative CCLR that applies to the causative verb adds its own clitic requirements to the CLTS list.

Fig. 12 shows the phrase structure tree and MRS representation the grammar displays for (12.a), where the clitic *lo* instantiate the ARG2 of the relation

[17] The same approach is described in (Pineda and Meza, 2005).

$$ {}_{verb}\left[\begin{array}{ll}\text{CAT}\mid\text{VAL} & \left[\begin{array}{ll}\text{SUBJ} & \langle\boxed{1}\rangle\\ \text{COMPS} & \langle\text{NP},\ldots\rangle\\ \text{CLTS} & \langle\rangle\end{array}\right]\end{array}\right] \rightarrow {}_{verb}\left[\begin{array}{ll}\text{CAT}\mid\text{VAL} & \left[\begin{array}{ll}\text{SUBJ} & \boxed{1}\text{NP}_i\\ \text{COMPS} & \langle\ldots\rangle\\ \text{CLTS} & \left\langle\text{NP}_i\ [\text{INDEX}\ \ \boxed{2}]\right\rangle\end{array}\right]\\ \text{CONT}\mid\text{RELS} & \left\langle !\ [\text{ARG2}\ \ \boxed{2}]!\right\rangle\end{array}\right] $$

Fig. 13 Reflexive Complement Cliticization Lexical Rule.

$$ {}_{\mathit{v\_-\_prn\_le}}\left[\begin{array}{ll}\text{CAT}\mid\text{VAL} & \left[\begin{array}{ll}\text{SUBJ} & \left\langle\text{NP}\ [\text{INDEX}\ \ \boxed{1}\,]\right\rangle\\ \text{COMPS} & \langle\rangle\\ \text{CLTS} & \left\langle\text{NP}\ [\text{INDEX}\ \ \boxed{2}\,]\right\rangle\end{array}\right]\\ \text{CONT}\mid\text{RELS} & \left\langle !\ \left[\begin{array}{ll}\text{ARG1} & \boxed{1}\\ \text{AFFIX} & \boxed{2}\end{array}\right]!\right\rangle\end{array}\right] $$

Fig. 14 Lexical type for intransitive pronominal verbs (*v_-_prn_le*).

of the embedded verb *hacer* ('to do') and the clitic *me* instantiates the ARG3 of the relation of the causative verb *permitir* ('to allow'), as well as the ARG1 of the embedded verb *hacer*.

(12) a. *Me lo permitieron hacer.*
clitic (dat) clitic (acc) allowed to do
'They allowed me to do it.'
b. *Me lo vieron hacer.*
clitic (dat) clitic (acc) saw to do
'They saw me doing it.'

4.2.4 Reflexive and reciprocal constructions

For the analysis of reflexive and reciprocal constructions we have adopted the same strategy as in cliticization. We have implemented two CCLRs –the Reflexive CCLR and the Reciprocal CCLR– that allow the realization of clitic pronouns as arguments of these verbs. These rules remove one element in the COMPS list and add to the CLTS list a clitic pronoun whose INDEX is token-identical to the corresponding argument feature of the verb's relation. For these constructions, these CCLRs, in addition, co-index the reflexive clitic with the subject. The Reciprocal CCLR impose the additional constraint that the subject and clitic must be plural. Fig. 13 shows the Reflexive CCLR.

As for pronominal verbs, which in (Miller and Sag, 1997) are treated as lexemes that require one or more argument to be of type *affix*, our approach also follows from lexical constraints. Pronominal verbs are defined by lexical types that require an element in the CLTS list whose INDEX is token-identical to an AFFIX argument of the verb's relation. Fig. 14 shows the lexical type for pronominal intransitive verbs (*v_-_prn_le*).

```
{e2:
 x4:pronoun_q[]
 e2:_peinar_v[ARG2 x4:pron]
}
```

Fig. 15 MRS representation of *te peinas*.

```
{e2:
 x4:pronoun_q[]
 e2:_resfriar_v[AFFIX x4:pron]
}
```

Fig. 16 MRS representation of *te resfriarás*.

$$\mathit{verb}\begin{bmatrix}\text{SUBJ} & \langle \text{NP} \rangle \\ \text{COMPS} & \langle \boxed{1}\text{NP},\ldots \rangle \\ \text{CLTS} & \langle\rangle\end{bmatrix} \rightarrow \mathit{verb}\begin{bmatrix}\text{SUBJ} & \boxed{1} \\ \text{COMPS} & \langle \ldots \rangle \\ \text{CLTS} & \langle [\text{AGR}\,|\,\text{PTYPE} \quad \text{impers}] \rangle\end{bmatrix}$$

Fig. 17 Lexical Rule for passive *se*-constructions.

$$\mathit{verb}\begin{bmatrix}\text{SUBJ} & \boxed{1} \\ \text{COMPS} & \boxed{2} \\ \text{CLTS} & \langle\rangle\end{bmatrix} \rightarrow \mathit{verb}\begin{bmatrix}\text{SUBJ} & \boxed{1}[\text{AGR} \quad \text{3sg}] \\ \text{COMPS} & \boxed{2} \\ \text{CLTS} & \langle [\text{AGR}\,|\,\text{PTYPE} \quad \text{impers}] \rangle\end{bmatrix}$$

Fig. 18 Lexical Rule for impersonal *se*-constructions.

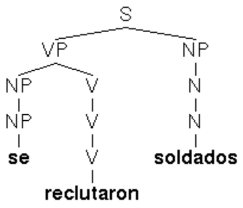
The distinct MRS representations that the grammar produces for reflexive constructions and pronominal verbs are illustrated in Fig. 15 and Fig. 16, respectively, with the sentences *te peinas* (clitic (refx) comb ('you comb your hair.')) and *te resfriarás* (clitic (refx) will caught a cold ('you will caught a cold.')). As can be observed, in the reflexive construction the clitic pronoun instantiates the ARG2 of the verb's relations, whereas in pronominal verbs, the clitic pronoun instantiates the AFFIX feature.

4.2.5 Constructions with se

In the Spanish DELPH-IN grammar, *se*-constructions are generated by means of two different lexical rules.

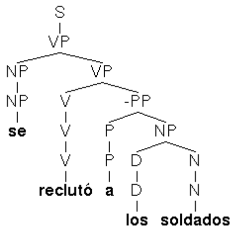
The lexical rule for passive se-constructions, shown in Fig. 17, removes the direct object from the COMPS list and places it as the subject, and adds to the CLTS list a clitic pronoun of type impersonal. The lexical rule for impersonal se-constructions, shown in Fig. 18, also adds to the CLTS list a clitic pronoun of type impersonal, but in these constructions the complement is maintained and the subject is restricted to be unexpressed. This rule impose the additional constraint that the unexpressed subject must be third person singular.

Fig. 19 and Fig. 20 show the output that the grammar produces for passive, respectively, impersonal constructions with the sentences *se reclutaron soldados* and *se reclutó a los soldados* ('soldiers were recruited'). As can be observed,



```
{e2:
 x4:pronoun_q[]
 e2:_reclutar_v[ARG2 x9:_soldado_n,AFFIX x4:pron]
 x9:undef_q[]
}
```

Fig. 19 Phrase structure tree and MRS representation for *se reclutaron soldados*



```
{e2:
 x4:pronoun_q[]
 e2:_reclutar_v[ARG2 x8:_soldado_n,AFFIX x4:pron]
 x8:_el_q[]
}
```

Fig. 20 Phrase structure tree and MRS representation for *se reclutó a los soldados*

the MRS represents the same argument structure for both sentences, where *soldados* instantiates the ARG2 of the verb's relation, and the clitic *se* instantiates its AFFIX feature. However, the grammar produces two distinct phrase structure trees: in the passive construction, *soldados* is the syntactic subject of the verb, and it combines with the VP node (after having combined the clitic with the verb); in the impersonal construction, (*a los*) *soldados* is the complement of the verb, and it is combined with the V node before combining the clitic with the verb.

4.2.6 Enclitics and proclitics

As we have already said, enclitics are treated in the inflectional rule component of the LKB system by means of a set of rules that are trigged by the PoS tag that FreeLing assigns to them. These rules apply on inflected items, and, like the morphological inflectional rules (cf. Section 2.1), they map FreeLing tags

Table 6 Grammar performance.

sentence length	# sent.	# parsed sent.	# grammar failures	# time-out	# annotated sent.
1-5	872	802 (92%)	70 (8%)	-	681 (78%)
6-10	1,420	1,260 (89%)	126 (9%)	34 (2%)	1,072 (76%)
11-15	1,877	1,409 (75%)	287 (15%)	181 (10%)	1,132 (60%)
total	4,169	3,471 (83%)	483 (11%)	215 (5%)	2,885 (69%)

into feature structures. The effect of these rules is that of removing the clitics from the CLTS list.

Proclitics are treated in the syntax by means of the *clitic-head* phrase structure rule. This rule allows a VP head and a clitic pronoun appearing on the left of the verb to combine. Like the set of rules dealing with enclitics, the effect of this rule is that of removing the clitics from the CLTS list. This rule applies recursively until the CLTS list is empty.

5 Evaluation

As we have already mentioned, the Spanish DELPH-IN grammar is being deployed in the construction of two treebanks: the IULA Treebank, a treebank of 60,000 sentences based in a technical corpus from the fields of Law, Economy, Genomics, Medinice, Computing Science, and Environment, and the Tibidabo treebank, a smaller treebank of about 15,000 sentences taken from newspaper articles.

Following (Oepen et al, 2002; Hashimoto et al, 2007; Branco et al, 2010), we are using the corpus annotation environment of the DELPH-IN framework to annotate the corpora. Using this framework, the annotation process is divided into two parts: first, the corpus is parsed using the Spanish DELPH-IN grammar; then, the best parse is manually selected. The DELPH-IN framework also provides a Maximum Entropy (ME) based parse ranker that ranks the parses generated by the grammar, allowing the annotator to focus on the n most likely trees, typically to less than 500 top readings (Toutanova et al, 2005), and thus reducing the required annotation effort. Statistics are gathered from disambiguated parses and can be updated as the number of annotated sentences increases.

Table 6 reports on the grammar performance when parsing a subset of the Tibidabo corpus, containing the sentences up to 15 words.

The second column shows the number of sentences up to 15 words that the target corpus has, distributed along sentence length. The third column shows the number of sentences for which the grammar produces an output. Parsing failures in the remaining sentences are basically due to two reasons. First, the processing components –as any other complex software in development stage– certainly show some deficiencies which are responsible for 11% of the parsing failures. Second, 5xs% of the input sentences reach time-out limit set in the parsing engine (which is set at 60 seconds per sentence), because they get a too large number of analyses. The fourth and fifth columns show the number of failures due to grammar deficiencies and time-out, respectively. In the sixth column we show the number of annotated sentences; i.e., the number of sentences for which we have selected an analysis.

6 Conclusions

We have presented the Spanish DELPH-IN grammar; a Spanish grammar implemented in the LKB system and grounded in the theoretical framework of HPSG that is being developed as part of the international multilingual DELPH-IN Initiative. We have described the grammar components, showing how, on the basis of a core grammar as defined by an early version of the LinGO Grammar Matrix, we have achieved a large-coverage grammar. We have also described some important aspects of deep processing of Spanish, illustrating the different analyses that the grammar produces for closely related constructions.

7 Acknowledgments

This work was funded by the *Ramón y Cajal* program of the Spanish *Ministerio de Ciencia e Innovación.* Part of this work was carried out during a three-month research visit at CSLI of the Stanford University funded by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* under the programe *Beques per a estades per a la recerca fora de Catalunya.* The author is grateful to the anonymous reviewers for their constructive and helpful comments on the earlier version of the paper. The author also thanks all DELPH-IN members, special thanks to Dan Flickinger for fruitful discussions and Stephan Oepen for answers to numerous question about the LKB system.

References

Bender EM, Flickinger D (2005) Rapid prototyping of scalable grammars: towards modularity in extensions to a language-independent core. In: Proceedings of IJCNLP'05 (Posters / Demos), Jeju Island, Korea, pp 203–208

Bender EM, Drellishak S, Fokkens A, Poulson L, Saleem S (2010) Grammar Customization. Research on Language & Computation 8(1):23–72

Bosque I (2010) Nueva gramática de la lengua española: Manual. Real Academia Española, Asociación de Academias de la lengua española, Espasa Calpe, Madrid

Branco A, Costa F (2008) A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version A. 4.1. TR-2008-17. Tech. rep., Universidade de Lisboa, Faculdade de Ciências, Departamento de Informatica

Branco A, Costa F, Silva J, Silveira S, Castro S, Avelãs M, Pinto C, Graca J (2010) Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In: Proceedings of LREC-2010, La Valletta, Malta

Callmeier U (2000) PET a platform for experimentation with efficient HPSG processing. In: Dan Flickinger, Stephan Oepen, Jun-Ichi Tsujii and Hans Uszkoreit (ed) Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation, Cambridge University Press, pp 99–108

Copestake A (2002) Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford

Copestake A, Flickinger D, Pollard C, Sag IA (2006) Minimal Recursion Semantics: an Introduction. Research on Language and Computation 3(4):281–332

Crysmann B (2005) Syncretism in German: a Unified Approach to Underspecification, Indeterminacy, and likeness of Case. In: Proceedings of HPSG'05, Lisbon, Portugal

Fernández Soriano O (1999) El pronombre personal. Formas y distribuciones. Pronombre átonos y tónicos. In: Ignacio Bosque and Violeta Demonte (ed) Gramática descriptiva de la lengua española, Madrid: Espasa, pp 1209–1273

Flickinger D (2002) On building a more efficient grammar by exploiting types. In: Dan Flickinger, Stephan Oepen, Jun-Ichi Tsujii and Hans Uszkoreit (ed) Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation, Cambridge University Press, pp 1–17

Hashimoto C, Bond F, Siegel M (2007) Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. Language Resources and Evaluation (Special issue on Asian language technology) 42(2):117–126

Hellan L, Haugereid P (2004) NorSource - an Excercise in the Matrix Grammar Building Design. In: Emily M Bender, Dan Flickinger, Frederik Fouvry and Melanie Siegel (ed) A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI, Vienna, Austria

Kim JB, Yangs J (2003) Korean Phrase Structure Grammar and Its Implementations into the LKB System, paper presented at the 17th Pacific Asia Conference on Language, Information, and Computation

Kordoni V, Neu J (2005) Deep Analysis of Modern Greek. In: Keh-Yih Su, Jun-Ichi Tsujii and Jong-Hyeok Lee (ed) Lecture Notes in Computer Science, Vol 3248, Springer-Verlag Berlin Heidelberg, pp 674–683

Levin B (1993) English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago

Marimon M (2010) The Spanish Resource Grammar. In: Proceedings of LREC-2010, La Valletta, Malta

Mendikoetxea A (1999) Construcciones con *se*: medias, pasivas e impersonales. In: Ignacio Bosque and Violeta Demonte (ed) Gramática descriptiva de la lengua española, Madrid: Espasa, pp 1631–1722

Miller PH, Sag IA (1997) French Clitic Movement without Clitics or Movement. Natural Language and Linguistic Theory 5(3):573–639

Monachesi P (1998) Decomposing Italian clitics. In: Sergi Balari and Luca Dini (ed) Romance in HPSG, CSLI publications. Stanford, pp 305–357

Oepen S, Carroll J (2000) Performance Profiling for Parser Engineering. In: Dan Flickinger, Stephan Oepen, Jun-Ichi Tsujii and Hans Uszkoreit (ed) Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation, Cambridge University Press, pp 81–97

Oepen S, Flickinger D, Toutanova K, Manning CD (2002) LinGo Redwoods. A Rich and Dynamic Treebank for HPSG. In: Proceedings of TLT 2002, Sozopol, Bulgaria

Padró L, Collado M, Reese S, Lloberes M, Castelón I (2010) FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In: Proceedings of LREC-2010, La Valletta, Malta

Pineda L, Meza I (2003) Una gramática básica del español en HPSG. Tech. rep., DCC-IIMAS, Universidad Nacional Autónoma de México

Pineda L, Meza I (2005) The Spanish pronominal Clitic System. Procesamiento del Lenguaje Natural 34:67–104

Pollard C, Sag IA (1987) Information-based Syntax and Semantics. Volume I: Fundamentals. CSLI Lecture Notes, Stanford

Pollard C, Sag IA (1994) Head-driven Phrase Structure Grammar. The University of Chicago Press and CSLI Publications, Chicago

Siegel M, Bender EM (2002) Efficient Deep Processing of Japanese. In: 3rd Workshop on Asian Language Resources and International Standardization, COLING-2002, Tapei, Taiwan

Toutanova K, Manning CD, Flickinger D, Oepen S (2005) Stochastic HPSG parse disambiguation using the Redwoods corpus. Research on Language & Computation 3(1):83–105

Tseng J (2004) LKB Grammar Implementation: French and beyond. In: Emily M Bender, Dan Flickinger, Frederik Fouvry and Melanie Siegel (ed) A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI, Vienna, Austria

Zwicky A, Pullum G (1983) Cliticization vs. Inflection: English n't. Language 59(3):502–513