INTERNATIONAL CONFERENCE

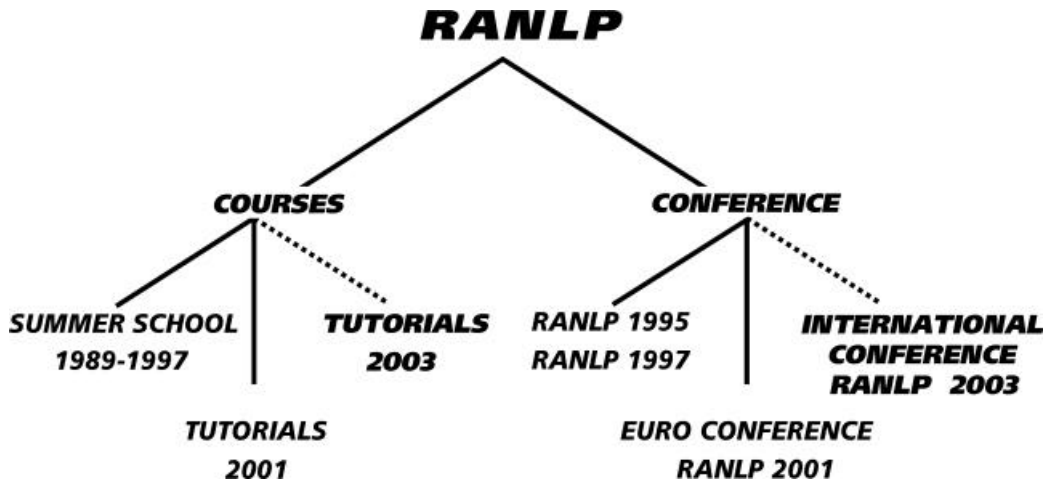
RECENT ADVANCES IN

NATURAL LANGUAGE PROCESSING

Supported by the European Commission (IST conference grant)

P R O C E E D I N G S

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov



Borovets, Bulgaria

10-12 September 2003

INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING'2003

PROCEEDINGS

Borovets, Bulgaria
10-12 September 2003

ISBN 954-90906-6-3

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

ORGANISERS AND SPONSORS

The International Conference RANLP - 2003 is organised by

Linguistic Modelling Department,
Central Laboratory for Parallel Processing (CLPP),
Bulgarian Academy of Sciences (BAS)
and
The Bulgarian Association for Computational Linguistics

The International Conference RANLP - 2003 is supported by

The European Commission (IST conference grant)

CLPP-BAS (BIS-21 Centre of Excellence)

The Bulgarian Association for Computational Linguistics

OntoText Lab., Sirma AI Ltd., Sofia, Bulgaria

Prosyst Ltd., Sofia, Bulgaria

The team behind RANLP - 2003

Galia Angelova (BAS, Sofia)
Kalina Bontcheva (Sheffield University)
Ruslan Mitkov (University of Wolverhampton)
Nicolas Nicolov (IBM, T.J. Watson Research Center)
Nikolai Nikolov (The Bulgarian Association for Computational Linguistics)

PROGRAMME CHAIR

Ruslan Mitkov (University of Wolverhampton)

PROGRAMME COMMITTEE

Elisabeth Andre (University of Augsburg)
Galia Angelova (BAS, Sofia)
Amit Bagga (Avaya Labs Research)
Lamia Belguith (LARIS-FSEG, University of Sfax)
Branimir Boguraev (IBM, T.J. Watson Res. Center)
Kalina Bontcheva (Sheffield University)
António Branco (University of Lisbon)
Sylviane Cardey (University of FrancheComte)
Nicoletta Calzolari (University of Pisa)
Eugene Charniak (Brown University, Providence)
Dan Cristea (University of Iasi)
Walter Daelemans (University of Antwerp)
Ido Dagan (Bar Ilan Univ. & FocusEngine, Tel Aviv)
Robert Dale (Macquarie University)
Hercules Dalianis (Royal Inst. of Techn., Stockholm)
Alexander Gelbukh (Nat. Polytechnic Inst., Mexico)
Ralph Grishman (New York University)
Walther von Hahn (University of Hamburg)
Jan Hajic (Charles University, Prague)
Graeme Hirst (University of Toronto)
Eduard Hovy (ISI, University of Southern California)
Aravind Joshi (University of Pennsylvania)
Martin Kay (Stanford University)
Alma Kharrat (Microsoft Natural Language Group)
Manfred Kudlek (University of Hamburg)
Shalom Lappin (King's College, London)
Anke Luedeling (Humboldt University, Berlin)
Nuno Mamede (INESC, Lisbon)
Carlos Martin-Vide (Univ. Rovira i Virgili, Tarragona)
Tony McEnery (Lancaster University)
Beáta Megyesi (Royal Institute of Technology, Stockholm)
Rada Mihalcea (University of North Texas)
John Nerbonne (University of Groningen)
Nicolas Nicolov (IBM, T.J. Watson Research Center)
Kemal Oflazer (Sabanci University, Istanbul)
Constantin Orasan (University of Wolverhampton)
Chris Paice (Lancaster University)
Manuel Palomar (University of Alicante)
Gerarld Penn (University of Toronto)
Fabio Pianesi (IRST, Trento)
Stelios Piperidis (ILSP, Athens)
Gábor Prószéky (MorphoLogic, Budapest)
Stephen Pulman (Oxford University)
James Pustejovsky (Brandeis University)
Jose Quesada (University of Seville)
Dragomir Radev (University of Michigan)
Allan Ramsay (UMIST, Manchester)
Lucia Rino (Sao Carlos Federal University)
Anne de Roeck (Open University)
Laurent Romary (INRIA, Lorraine)
Harold Somers (UMIST, Manchester)
Richard Sproat (AT&T Labs Research)
Keh-Yih Su (Behavior Design Corporation)
Kristina Toutanova (Stanford University)
Isabelle Trancoso (INEC, Lisbon)
Jun'ichi Tsujii (University of Tokyo)
Hans Uszkoreit (University of Saarland)
Piek Vossen (Irion Technologies, Delft)
Yorick Wilks (Sheffield University)
Michael Zock (CNRS, Tokyo Institute of Technology)

REVIEWERS

In addition to the members of the Programme Committee, the following colleagues were involved in the reviewing process:

Le Ha An (University of Wolverhampton)
Rie Ando (IBM)
Victoria Arranz (Univ. Politècnica de Catalunya)
Patricio Martínez-Barco (University of Alicante)
Catalina Barbu (University of Wolverhampton / University of Brighton)
Svetla Boytcheva (Sofia University)
Sabine Buchholz (Toshiba Research Europe Ltd.)
Grace Chung (Corporation for National Research Initiatives)
Borja Navarro Colorado (Universidad de Alicante)
Pernilla Danielsson (University of Birmingham)
Richard Evans (University of Wolverhampton)
Kerstin Fischer (University of Hamburg)
Laura Hasler (University of Wolverhampton)
Diana Maynard (University of Sheffield)
Rafael Muñoz-Guillena (University of Alicante)
Preslav Nakov (Berkeley University)
Ani Nenkova (Columbia University)
Maximiliano Saiz-Noeda (University of Alicante)
Viktor Pekar (University of Wolverhampton)
Jesús Peral (University of Alicante)
Krasimira Petrova (Sofia University)
Kiril Simov (BAS, Sofia)
Mark Stevenson (University of Sheffield)
Jana Sukkarieh (Oxford University)
Valentin Tablan (Sheffield University)
Hristo Tanev (IRST, Trento)
Doina Tatar (Babes-Bolyai University)
Isabel Verdaguer (Universitat de Barcelona)
Karin Verspoor (Los Alamos)
José Luis Vicedo (University of Alicante)
Zhu Zhang (University of Michigan)

PROGRAMME COMMITTEE COORDINATORS

Albena Strupchanska (BAS, Sofia)
Milena Yankova (BAS, Sofia)

TABLE OF CONTENTS

Description and Exploitation of BDS: a Syntactic Database about Verb Government in Spanish

Eva Mª. Muñiz* and **Marta Rebolledo** and **Guillermo Rojo** and **Mª. Paula Santalla** and **Susana Sotelo**
Dptm. of Spanish Language, University of Santiago de Compostela
Avda. Burgo das Nacions, s/n, E-15782, Santiago de Compostela, Spain
{emuniza, mrlemus, grojo, fempsr, fesdocio}@usc.es

Abstract

In this document, a description is given of a syntactic database about verb government in Spanish. We first describe the framework that inspired the development of the database (Section 1.1), and then present the analysed corpus (Section 1.2) and the database itself (Section 1.3). Section 2 contains a description of the structure of the database itself, as well as of the annotation system used in it. Section 3 presents some results of the work done for the exploitation of the data contained in BDS, as well as the current stage of the main exploitation that we intended to do of it when we decided to start the work: the elaboration of a Spanish Dictionary of Verb Structure and Government. Section 4, finally, indicates the new directions that the project has recently taken or is likely to take in the near future.

1 BDS description

1.1 Framework

In the last fifteen years approximately, the research group Sintaxis del español (Spanish Syntax[1]) has been mainly interested in the structure, especially in their functional constituents, of clauses, these being conceived of as the grammatical units organised around a verb that plays the function of predicate. To go further in this field, the group, in 1988, decided to start a research project, which should account for three main ideas that came out from our previous work:

- Syntactic schemes, that is, the overall organisation of the syntactic functions at the level of the clause, are more important than syntactic functions by themselves.
- As verbs play the function of predicates within clauses, the study of clause schemes can be undertaken by looking at the syntactic structures in which verbs are actually found.
- Verbs are actually found within the syntactic structures in which they can be found, that is, verbs determine the syntactic structures, or clause schemes, in which they appear. As a result of this, the approach to the study of clause schemes must consist of the study of the syntactic schemes that each verb can determine, the kind of information collected in a Dictionary of Verb Structure and Government.

On these grounds, it was clear what the project should consist of, i.e., the collection of the analyses of a great deal of clauses, in order to deliver a reliable and representative, from the quantitative point of view, amount of data about the syntactic behaviour of verbs. Since we decided to analyse the object corpus selected by hand, only after some years of work[2], we finally obtained the syntactic database, henceforth BDS[3], that is the object of this presentation, and consists of the analysis of the syntactic context of the approximately 160,000 verbs that appear in the contemporary part of the Hispanic Texts Archive of the University of Santiago, henceforth ARTHUS[4].

1.2 ARTHUS

The ARTHUS corpus is constituted by million and a half words of texts taken from all the Hispanic countries. It includes oral samples as well as novels, press and theatre, all of them published

*The research described in this document has been partially supported by the Spanish Government under projects PB90-0376 and BFF2000-0381, and the Autonomous Government of Galicia (Xunta de Galicia) under projects XUGA82710088 and PGIDT00PX120410.

[1]URL: http://www.sintx.usc.es.

[2]In the more than ten years already in which BDS has been being developed, a great number of researchers have collaborated, in different degrees and with different tasks, in this work. At the moment (April 2003) are still working on it the following members of the initial team, all of them from the University of Santiago de Compostela, except one from the University of Vigo: Francisco García Gondar, José María García-Miguel (University of Vigo), Belén López Meirama, Inmaculada Mas Álvarez, María José Rodríguez Espiñeira, Guillermo Rojo and Victoria Vázquez Rozas. Apart from these, other researchers are currently collaborating in the extension of BDS and its application to new grammatical studies: Fernando Castro Paredes, Eva Mª. Muñiz Álvarez, Marta Rebolledo Lemus, María Paula Santalla del Río and Susana Sotelo Docío.

[3]Stands for *Base de datos sintácticos*, see (Rojo 95).

[4]Stands for *Archivo de textos hispánicos de la Universidad de Santiago*.

between 1980 and 1990. The distribution of this corpus appears in Table 1:

Gender	Spain	Amer.	Total	%
Fiction	385,661	153,245	538,906	37.19
Essay	168,511	89,207	257,718	17.78
Theatre	212,507	0	212,507	14.66
Press	166,804	0	166,804	11.51
Oral	207,948	65,122	273,070	18.85
Total	1,141,431	307,574	1,449,005	
%	78.77%	21.23%		

Table 1: ARTHUS design.

1.3 BDS

BDS is a database each record of which contains the syntactic analysis of one clause of the ARTHUS corpus. The type of information collected for each clause in the various fields of each record concerns the following aspects:

- Data about the verb that plays the function of predicate: lemma and location in the corpus (text, page-line-column).
- Data about the clause as a whole: a) type of clause, b) function of the clause, c) voice, d) mood, e) polarity, f) periphrasis, g) mood and tense of the main verb form, h) mood and tense of the verb form of a possible embedding clause, i) number and person of the main verb form, j) number of arguments, k) order of constituents.
- Data about each of the syntactic functions identified within the clause: a) data that are specific for each function (things like type of impersonality structure in the case of the subject, or prepositions introducing them in the case of prepositional complements), and b) data, such as type of syntactic category, animation, countability, determination and number, that are specified for all syntactic functions.
- Additional information: syntactic properties that, although they are not strictly related with the arguments selected by the predicate of a clause, are, for different reasons, considered interesting.

By way of example, we show, in Figure 1, the record[5] of BDS that contains the analysis of the clause whose verb is found in line 25 of page 42 of the text El Sur[6].
In the top of Figure 1, we can see the verb (*abandonar*, "to leave") and the location of the example, together with the value for voice, in this case, active. In the bottom of the figure, the text of the example is showed with various lines of context. In the central part of the figure, the first two columns on the left indicate that the clause in question, *de que tú pudieras abandonarme*[7], has the following characteristics: It is a that-clause, it functions as a prepositional modifier, it has affirmative polarity, it has declarative mood, it includes periphrasis *poder* + infinitive, the verb is in imperfect tense and subjunctive mood, tense and mood of the main clause are not considered relevant, the verb is in second person singular, the main verb is not a multiword one, the clause includes two syntactic arguments and these are found in the order S(ubject)V(erb)[8]. The three last columns on the right, on the other hand, particularly describe the two syntactic arguments identified in the example in question, a subject and a direct object. The description of these syntactic arguments, here expressed by means of numerical keys[9] is verbalized in Figure 2: there is a subject in this clause, the subject is a second person personal pronoun, it is animate and countable, and it has values definite for determination and singular for number. There is also a direct object, expressed by means of a clitic pronoun *me*, that is, first person singular, whose referent is animate and countable.

2 BDS structure and annotation system

The final result of the work done by manually analysing the ARTHUS corpus, is a database that contains 160,000 records approximately, each with 61 fields that store, together with the identification of the example in question (verb, text and location of the verb and clause of the example), the information that concerns the multiple aspects of the syntactic analysis of clauses –especially those related with verb government– mentioned in Section 1.3.

More revealing than its –quite simple– structure as a database, is, however, the annotation system used to encode the information within the relevant fields of the records of the database, that is, the kind of tags used to refer to the syntactic information encoded for each clause. It is in this respect, precisely, where the originality of the

[5] Figure 1 presents the information as it is showed by an interface developed for internal use within the research group: we have only translated the epigraphs.

[6] A. García Morales. 1985. *El Sur (seguido de Bene)*. Anagrama, Barcelona.

[7] "That you could leave me".

[8] In this case we do not include the direct object in the description of the order of arguments because, as it is a clitic pronoun, it has a fixed position, immediately preceding the verb, in the clause.

[9] See Section 2. As in the case of Figure 1, in Figure 2 we have only translated the epigraphs.

```
ABANDONAR [SUR: 42, 25]          Act.

General characteristics.                        Subj. D. Obj.

Type              That-clause     Character       1       1
Function          Prep. Mod.      Type
Polarity          Affirmative     1st clitic pr.          11
Mood              Declarative     2nd clitic pr.
Periphrass        poder + Inf     Prep.
Mood/Tense        llegaras        Unit            22
Mood/Tense                        Animat.         11      11
Person            2nd.  sing.     Determ.          1
Multiword.                        Number           1
Number arg.       2               Refnt.
Order             SV
Add. data.

42,22 tus cartas, tu proposición de volver con ella,
42,23 abandonándonos a nosotras. >Me equivoco? En mis ca-vilaciones
42,24 de niña sobre lo que yo consideraba tu secreto
42,25 nunca apareció la posibilidad de que tú pudieras aban-donarme.
42,26 Yo sabía tan poco de ti... Mi mirada era tan
42,27 corta.
42,28 Decidí visitar a aquella mujer. Ahora sabía que vivía
```

Figure 1: Example of BDS (I).

```
                ABANDONAR [SUR: 42, 25] Act.

                                    Subject         Direct Object

Character                           Explicit        Found
Type
1st clitic pronoun                                  me
2nd clitic pronoun
Preposition
Unit                                Pers. Pron 2nd.
Animation                           Animate count.  Animate not count.
Determination                       Definite
Number                              Singular
Referent of predicative complement

42,22 tus cartas, tu proposición de volver con ella,
42,23 abandonándonos a nosotras. >Me equivoco? En mis ca-vilaciones
42,24 de niña sobre lo que yo consideraba tu secreto
42,25 nunca apareció la posibilidad de que tú pudieras aban-donarme.
42,26 Yo sabía tan poco de ti... Mi mirada era tan
42,27 corta.
42,28 Decidí visitar a aquella mujer. Ahora sabía que vivía
```

Figure 2: Example of BDS (II).

BDS more strongly stands out. The syntactic information contained in BDS is, in fact, encoded by means of a large set of hierarchically organised numerical keys, in a way so that, within each tag, each number in its position refers to a syntactically relevant feature on the basis of which all of the items differentiated by the previous number on the right within the tag can be grouped. This means that, if necessary, the information can be very easily retrieved with different degrees of detail. Let us consider, by way of example, the 55 different tags that are available to the annotator when filling in fields 17, 24, 31, 38, 45, 51 and 56 of the form, the fields that contain the information about the type of unit that plays the function of, respectively, subject, direct object, indirect object, first prepositional complement, second prepositional complement, agent and predicative complement.

Figure 3 shows an important part of these 55 different tags. As can be observed, the degree of detail is very high, and, for instance, each different relative, interrogative-exclamative or personal pronoun has a different tag. However, if we do not take into account the last number of the tag, all the personal pronouns are grouped together, the feature of person being lost. The same happens in the case of relative and interrogative pronouns, the last number of whose tags refers to each of the pronoun items of these types functional in Spanish.

Figure 4 shows another example of a set of tags: those that may appear in field 4, which contains the information about the type of clause constituted by the example in question. Tags starting with number 2 refer to all direct object clauses, within them numbers 1, 2 and 3 to the right distinguish *that*-clauses, infinitive clauses and other types of direct object clauses.

In total, 256 tags of the form described are used by the whole annotation system[10].

With all the previous explanations, the description of BDS that can be given within the limits of this brief exposition has been completed. It is, then, at this point of the paper that a discussion about what the BDS resource is can and must be faced. BDS should not be considered as

```
10    Nominal Phrase
11    Demonstrative pronouns
12    Indefinite and quantitative pronouns
13    Nominalised relative clause
14    Nominalisation of adj., prepositional
      phrase, etc.
15    Possessive pronouns
21    Personal pronoun, first person
22    Personal pronoun, second person
23    Personal pronoun, third person
34    Personal pronoun, polite form usted
311   Relative que
312   Relative quien
313   Relative cual
314   Relative cuanto
315   Relative donde adonde
316   Relative cuando
317   Relative como
318   Relative el que/la que/los que/las
      que/lo que
319   Relative el cual/la cual/los cuales/
      las cuales/lo cual
321   Interrogative-exclamative qué/por qué
322   Interrogative-exclamative quién
323   Interrogative-exclamative cuál
324   Interrogative-exclamative cuánto
325   Interrogative-exclamative dónde/adónde
326   Interrogative-exclamative cuándo
327   Interrogative-exclamative cómo
41    That-clause + indicative
42    That-clause + subjunctive
...
```

Figure 3: Types and subtypes of units.

a syntactically annotated corpus or treebank in the style of the Penn treebank (Marcus *et al.* 93), the Susanne corpus (Sampson 95), the Nijmegen corpus (van Halteren & Oostdijk 93) and similar resources, in which syntactic tags are inserted in, or aligned with, the text. Instead of this, BDS is a pure database, in the more strict sense of the word, that contains the syntactic data that correspond to the analysis, done by hand, of (almost) all the clauses that appear in the ARTHUS corpus, the connection with the text being possible only by means of the reference (text, page-line-column, contained in fields 2 and 3 of the database) of the example in question. Such an organisation should not be contemplated as better or worse in general: at the moment, we simply think that it has proven to be adequate for the objective aimed at when we decided to start it, i.e., to collect information about verb government in Spanish, avoiding the execution of previous stages of analysis, as well as allowing the retrieval of the information in an easy way. The drawbacks of a resource like BDS appear, on the one hand, when we try to obtain information about linguistic as-

[10] This is, however, a somewhat deceptive amount, because information like the number and order of arguments (contained in fields 14 and 15) of the database has not been categorised as a tag. The same goes for all the keys that may appear in field 61 of *additional information*.

```
0   Independent clause
12  Coordinated clause
21  That-clause
22  Infinitive clause
23  Other direct-object clauses
31  Relative clause
32  Nominalised relative clause
4   Adverbial clauses
5   Gerund clause
6   Participle clause
7   Conditional, reason, result, concessive,
    comparative and similar clauses
81  Non direct-object clauses introduced by
    certain conjunctions
82  Finite clause introduced by a multiword
    conjunction
83  Non-finite clause introduced by a multiword
    conjunction
9   Other
```

Figure 4: Types of clauses.

pects not connected with verb government (data about, for instance, phrases or circumstances), and, on the other hand, when we come across the fact that the frequencies of the verbs documented in BDS, and, consequently, of the linguistic phenomena related with them, are sometimes very low (in the 160,000 clauses analysed, we have documented 3.550 verbs, 748 of these appear only once in the corpus, 1.679 appear less than five times). Obviously, this drawback arises from the fact that BDS has been collected by hand, but a further extension of the resource in the same way does not seem feasible, so that we try now to develop an automatic system of analysis, for which the data of BDS are being very helpful, see (Álvarez *et al.* 98) and, especially, (Santalla 02), in which the process of elaboration of a lexical database of verbs from BDS is described.

3 Exploitation

Since 1992 approximately, BDS has been exploited as a syntactic dictionary, as well as a basis for the development (still in course) of a *Dictionary of Verb Structure and Government*, which, as observed above, was, since the very beginning, our main goal. We present here a brief description of the form and the results of the exploitation of BDS for the study of different aspects of verbs, mainly focusing on those that concern the generation of the type of information that should be contained in a Dictionary of Verb Structure and Government.

3.1 Verbs

As the result of the analysis of the 160,000 simple clauses in the corpus, BDS offers the best data on verbal frequencies in current Spanish. With respect to verb lemmas, for instance, it must be noted that, differently to what has been the practice in other studies on textual data, in BDS the distinction between main and auxiliary uses of verbs is carefully respected and, as a consequence of that, the data on the frequency of, for example, the verb *haber* "to have" reflects only its use as a main verb and not as the auxiliary form of compound tenses and some other periphrastic constructions.

With this in mind, the frequency distribution in BDS is in accordance with the usual profile in a textual corpus: some verbs present a very high frequency and many other verbs have a low or very low frequency (more than 20% are *hapax legomena* in ARTHUS). In the high part of the spectrum, it is interesting to emphasize the fact that the 32 more frequent verbs in the corpus (less than the 1%) sum up a percentage higher than the 50% of the total verbal occurrences in the corpus.

Much more interest, mainly because of the lack of relevant data until now, has the picture of functional schemes frequencies in current Spanish[11]. Taking into account our concept of syntactic scheme (see Footnote 14), it is relevant to verify that only 158 functional schemes are documented in a corpus composed of some 160,000 clauses. Only 36 of them reach a frequency equivalent to the 0.1% or more, but the sum of these 36 syntactic schemes raise the 98.36% of all the analysed clauses. As showed in Table 2, the active construction composed of subject and direct object is the more frequent construction in Spanish: it has been documented in the 40% of the clauses in our corpus and in the 70% of the verbs contained in it.

Table 2[12] contains two different aspects in the frequency of syntactic schemes: the general frequency and the number of verbs presenting the scheme. The importance of the distinction is clear in the scheme Active voice: Subject-Predicative of Subject: it supposes the 6.34% of all the clauses in the corpus, but only the 1.83% of the verbs in it

[11] See (Rojo 01) for more details.

[12] Keys for the table: S, Subject; D, Direct object; PS, Predicative of Subject; I, Indirect object; ADV, adverbial complement; PC, Prepositional Complement; PD, Predicative of Direct object.

Scheme	Freq.	% Clauses	Verbs	%
Active S D	62,022	39.06	2,421	70.44
Active S	19,462	12.26	1,176	34.22
Active S PS	10,069	6.34	63	1.83
Active S D I	9,249	5.83	624	18.16
Active S ADV	6,732	4.24	179	5.21
Active S	6,416	4.04	816	23.74
Active S PC	5,084	3.20	321	9.34
Active S I	5,046	3.18	222	6.46
Active S PC	4,222	2.66	370	10.77
Active S D PD	4,115	2.59	95	2.76

Table 2: The more frequent syntactic schemes in BDS.

have been documented with this construction[13].

3.2 Towards a Dictionary of Verb Structure and Government

However, although we consider these results about the frequencies of verbs, on the one hand, and schemes, on the other one, very revealing, as we explained in Section 1, our main interest in research is the structure of the clause and its functional constituents, focusing in the verb as the element that determines the appearance of other syntactic functions, such as subject, direct object, etc., in the clause. For this reason, the central issue in the BDS project are not verbs and schemes independently of each other, but verbs and schemes to the extent that they are associated with each other.

With this in mind, we have designed around BDS a system that permits us (and other members of the research community, see below) to look at the data of BDS in the more adequate form to show the information about the Spanish verbs in context, focusing on the fact that these are used with certain schemes and subschemes[14]. As we explained in Section 2, BDS is a classical database that contains all the information encoded for each example analysed of the ARTHUS corpus (Figure 1 and Figure 2). All this information constitutes a very rich degree of detail that, although it may be pertinent for the study of certain phenomena, happens to be an inconvenience for other types of searches (a clear example in this respect are the 55 different keys for the identification of the type of unit underlying syntactic functions, a degree of detail that is not relevant when you are interested in, for instance, the general structure of the clause, see Figure 3).

In order to account for this wealth of information in a reasonable way, a set of computer programs takes care of handling all such information without modifying the original data: profiting from the hierarchical structure of the annotation system, it groups certain degrees of detail found in BDS and automatically adds the new information that, in two additional fields, identifies the scheme and subscheme documented by each example. This constitutes a new, clustered, form of the information that is used to elaborate a secondary version of BDS, a collection of derived files that primarily contains:

- One record per each verb in each syntactic scheme (example: *abandonar* in the scheme Active voice: Subject-Direct object; *abandonar* in the scheme Active voice: Subject-Direct object-ADVverbial complement).
- One record per each verb in each syntactic subscheme (example: *abandonar* in the subscheme Active voice: animate Subject-inanimate Direct object).

In all cases, this information is enriched with statistical data about the frequency of the verb, of the scheme and of the subscheme, which means that these secondary files contain all and only the information necessary to account for the association of verbs, schemes and subschemes. These secondary files, in fact, serve as basis for a web-based application that enables interested researchers to access this second version of the database[15]. The type of queries that can be posed to this application, and the results obtained from them, perfectly illustrate what we considered since the very beginning of the project that would be its main exploitation.

Queries can be posed to the system from different points of view. Among other less relevant, for our purposes here, possibilities of search, we can, on the one hand, ask the system for syntactic schemes or subschemes in order to obtain verbs that have been documented in BDS in such schemes or subschemes (see Table 3).

[13] In fact, the occurrences of verbs *ser* and *estar* presenting this construction suppose the 4.72% of the data in the whole BDS and, indeed, the 75% of the clauses with this scheme.

[14] With respect to schemes, it must be noticed that in our approach, a syntactic scheme of a verb is conceived of not only as a series of functional elements, as usual, but as this series of functional elements in combination with voice construction (periphrastic passive, active or middle). Subschemes, on the other hand, consist of the addition of relevant syntactic and semantic characteristics to each functional element found in the scheme (characteristics such as animate/inanimate, infinitive clause, adverbial phrase, etc.).

[15] URL: http://www.bds.usc.es

Verb	Frequency	% Verb
ABOLLAR	1	100.00
ABONAR	5	35.71
ABRASAR	2	20.00
ABRIR	43	6.51
ABROCHAR	3	20.00

Table 3: Verbs with a scheme Active voice: Subject-Direct object-Indirect object (partial result).

```
ABANDONAR [197 examples; 10 schemes; 16 subschemes]

Active S (F=2; 1.02% of verb)
  San (F=2; 100.00% of sch.)
Active SD (F=171; 86.80% of verb)
  San Dan (F=32; 18.71% of sch.)
  San Dinan (F=122; 71.35% of sch.)
  SinanDan (F=11; 6.43% of sch.)
  SinanDinan (F=6; 3.51% of sch.)
Active SD ADV (F=5; 2.54% of verb)
  San Dan ADVinan(en) (F=1; 20.00% of sch.)
  San Dinan ADVinan(en) (F=3; 60.00% of sch.)
  San Dinan ADVinan(sobre) (F=1; 20.00% of sch.)
Active SD PC (F=1; 0.51% of verb)
  San Dinan PCinan(a) (F=1; 100.00% of sch.)
```

Figure 5: Active schemes and subschemes of verb *abandonar*.

On the other hand, we can look up a verb, and the engine will give us back information about the schemes and subschemes documented in BDS for this verb (see Figure 5): and note that, if we specify some real examples for each entry of this list of schemes and subschemes associated with a verb, the result will already be very close to what is considered a syntactic dictionary of verbs.

According to Figure 5, the context in which verb *abandonar* has been found in the ARTHUS corpus, documents its association with four different schemes: a) Active voice: Subject, 2 examples, which means 1.02% of the frequency of the verb, b) Active voice: Subject-Direct object, 171 examples, which means 86.80% of the frequency of the verb, c) Active voice: Subject-Direct object-ADVerbial complement, 5 examples, which means 2.54% of the frequency of the verb, and d) Active voice: Subject-Direct object-Prepositional Complement, 1 example, which means 0.51% of the frequency of the verb. For each of these schemes, the subschemes documented for verb *abandonar* in the ARTHUS corpus are listed immediately below in Figure 5 (*an* means "animate noun phrase or similar", *inan* means "inanimate noun phrase or similar").

4 Future developments

Apart from the extension of BDS, already mentioned in Section 1.3, by means of the automatic addition of new examples of those verbs less frequently occurring in the ARTHUS corpus, we are at the moment enriching the information already collected for each example of BDS with the specification of the particular meaning, among all those associated with the verb in question, functional in the example in question, in order to finally obtain the Spanish Dictionary of Verb Structure and Government that was our objective when we decided to start the project. Simultaneously, we are adding information about the semantic class of the verb meaning identified in each example, as well as of the words that function as the nucleus of its arguments.

In addition to this, by other research groups of our centre of philological studies, a diachronic extension of BDS is already in course, as well as the application of the same principles for the development of a similar resource for the (modern) Galician language. BDS, finally, is currently being used as the main source for the Spanish part of a Spanish-German Valency Dictionary.

References

(Álvarez *et al.* 98) Concepción Álvarez, Pilar Alvariño, Adelaida Gil, Teresa Romero, Mª. Paula Santalla, and Susana Sotelo. AVALON, una gramática formal basada en corpus. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, (23):132–139, 1998.

(Marcus *et al.* 93) Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 2(19):313–329, 1993.

(Rojo 95) Guillermo Rojo. La base de datos sintácticos del español actual. *Español Actual*, (50):15–20, 1995.

(Rojo 01) Guillermo Rojo. La explotación de la base de datos sintácticos del español actual (bds). In Josse De Kock, editor, *Lingüística con corpus. Catorce aplicaciones sobre el español*, pages 255–286. University of Salamanca, Salamanca, 2001. Also available in http://www.bds.usc.es/?url=masinfo.html.

(Sampson 95) Geoffrey Sampson. *English for the Computer. The SUSANNE Corpus and Analytic Scheme.* Clarendon Press, Oxford, 1995.

(Santalla 02) Mª. Paula Santalla. *A Formal Grammar for Phrase Level Analysis Applied to Information Retrieval.* University of Santiago de Compostela, Santiago de Compostela, 2002.

(van Halteren & Oostdijk 93) Hans van Halteren and Nelleke Oostdijk. Towards a syntactic database. the tosca analysis system. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English language corpora: Design, analysis, exploitation. Papers from the 13th ICAME conference*, pages 145–161. Rodopi, Amsterdam, 1993.