María Paula SANTALLA DEL RÍO
(University of Santiago de Compostela)

FROM BDS TO THE AVALON LEXICON AND GRAMMAR: VERB SCHEME CLUSTERS GENERATION AND THEIR EXPLOITATION WITHIN THE GRAMMAR

0. INTRODUCTION

0.1. BDS

In the title of this paper two concepts are already contained whose knowledge is necessary for the understanding of the rest of the exposition. These are BDS and formal grammar and formalism. BDS, *Base de Datos Sintácticos*, on the one hand, is a database which resulted from a classical corpus-linguistics approach for the collection of the manual analysis – having in mind constitutive and functional analysis principles – of the syntactic context of 160.000 verb forms approximately, which belong to the contemporary part of the *Archivo de Textos Hispánicos de la Universidad de Santiago* (ARTHUS). In this database, the appearance of each verb form from the corpus has been individually registered with detailed information about the syntactic configuration of the linguistic sequence governed by the verb: we can find their general information about the clause that contains the verb, as well as detailed information about each of the arguments found in it. Comprehensive descriptions of BDS, which should be too space-consuming to be included here, can be found in Rojo (1995, 2001), Muñiz, Eva M[a] et al. (2003) and Mas/Rojo (2004), This Volume, pp. 101–108.

0.2. Formal grammar and formalisms. AGFL

A formal grammar, on the other hand, is a description of the structure of a language encoded in a certain formalism, i. e., in a restricted notation that is adequate for the expression of grammar rules. AGFL[1], *Affix Grammar over Finite Lattices*, is a formalism suitable for this task, we include here a toy-grammar written in this formalism in order to explain its basic features.

[1] Developed by professor C. H. A. Koster and his research group in the Department of Software Engineering of the Catholic University of Nijmegen, http::www.cs.kun.nl/agfl.


```
SENTENCE:
  SUBJECT ( person, gender, number ),
    PREDICATE ( mood, tense, person, number ).

SUBJECT ( person, gender, number ):                    GRAMMAR
  Pronoun ( person, gender, number ).
PREDICATE ( mood, tense, person, number ):
  Verb ( mood, tense, person, number ).

mood :: INDICATIVE; SUBJUNCTIVE.                       METAGRAMMAR
tense :: PRESENT; IMPERFECT; PERFECT; FUTURE.
person :: FIRST; SECOND; THIRD.
gender :: MASCULINE; FEMININE.
number :: SINGULAR; PLURAL.

Pronoun ( THIRD, MASCULINE, SINGULAR ): "alguien".     LEXICON
Verb ( INDICATIVE, PRESENT, THIRD, SINGULAR ): "viene".
```

Figure 1. Example AGFL Grammar.

AGFL is a formalism for the description of two-level grammars. The first level, which accounts for what is considered the *grammar* in a strict sense, consists of rules of the type of those showed on the top of the previous figure: something placed on the left-hand side of a colon is composed of (or: rewrites as) the series of elements, separated by commas, found on the right-hand side of the colon, until we find a period that closes the series and the rule. For instance, in the grammar of Figure 1, the first rule states that a sentence of the language described by the grammar rewrites as a subject (enriched with a series of parameters enclosed between brackets: person, gender and number) followed by a predicate (also enriched with a series of parameters enclosed between brackets: mood, tense, person and number). Next, by way of, respectively, the second and the third rules, the subject and the predicate, both with the same parameters referred above, rewrite as, respectively, a pronoun and a verb, each of which also enriched with the same parameters that extend the subject and the predicate.

In addition to this, we can observe that the third group of rules showed by Figure 1 corresponds also to the first level of description of the grammar. These rules have, in fact, the same form that we described for the previous ones (something placed on the left-hand side of a colon rewrites as something placed on the right-hand side of the colon, period), but they must be distinguished from them because these account for the definition of the terminal symbols of the grammar, i. e., the "words" of the language. In our example there are only two words: the pronoun *alguien* ("somebody"), with the values "third", "masculine" and "singular" for, respectively, the parameters person, gender and number, and the verb *viene* ("comes"), with the values "indica-

tive", "present", "third", "singular" for, respectively, the parameters mood, tense, person and number.

Finally, the question logically arises of where do we specify the possible values of the parameters (or: affixes). These are actually accounted for by means of the rules of the second level of the formalism, what we call the *meta-grammar*, which consists of rules of the type of the second group showed by Figure 1: something, an affix variable, placed on the left-hand side of a double colon can take one or more than one of the values that appear, separated by semicolons, on the right-hand side of the double colon. A period closes the list of possible values for the affix in question. For instance, parameter mood of our example-grammar might take either one of the values indicative or subjunctive, or both.

0.3. Parser generators and parsers

Apart from being a goal in its own right, if the formalism in which they are written is associated with a parser generator, formal grammars can also be automatically converted into parsers that can in turn be used to automatically analyse linguistic input according to the descriptions of the language as encoded within the grammars. The AGFL formalism has this possibility and, therefore, by processing the grammar of Figure 1, we can produce a parser that, when asked to analyse the sequence "alguien viene", produces the output of Figure 2, which is the analysis of this sequence according to the grammar of Figure 1.

```
alguien viene$EOS$
--------------
parsing     1     (00.001 s.)

SENTENCE
  SUBJECT(THIRD,MASCULINE,SINGULAR)
    Pronoun(THIRD,MASCULINE,SINGULAR)
      "alguien"
  PREDICATE(INDICATIVE,PRESENT,THIRD,SINGULAR)
    Verb(INDICATIVE,PRESENT,THIRD,SINGULAR)
      "viene"
```

Figure 2. Parsing of the sequence *alguien viene*.

0.4. The argument of this exposition

With these elements, it is our purpose to explain how BDS has contributed to the elaboration of an AGFL formal grammar, which we have called AVALON, for the generation of a parser that can be used to analyse spanish texts.

1. BDS AND THE LEXICON OF AVALON

1.1. Verb government

Let us consider again the lexicon definition of the verb *viene* in Figure 1. If, together with the values for mood, tense, person and number, we had wanted to store information about the syntactic behaviour of this verb – the arguments and voice associated with it –, we would have, instead of the one in Figure 1, a lexicon definition similar to the one in Figure 3, which includes two additional values, intransitive and active, that give indications about the syntactic behaviour of verb *venir*.

```
Verb ( INTRANSITIVE, ACTIVE, INDICATIVE, PRESENT, THIRD, SINGULAR ):
      "viene".
```

Figure 3. Lexicon definition with subcategorization information.

This is the kind of information that can be derived from a linguistic resource like BDS, which primarily collects information about verb government in Spanish – with the additional advantage that, in doing so, the information encoded in the lexicon would be *corpus-*, instead of *dictionary-, based*, that is to say, based on real data instead of on intuitions. Therefore, in a natural way, what BDS can provide to the AVALON grammar are the data about the kind of arguments and their organization required by the verbs contained in the AVALON lexicon. Or, in other words, the values of the affixes that hold this information in each of the verbal definitions found in the lexicon.

1.2. The AVALON affixes about verb government

Which are the affixes that hold this information in the AVALON grammar? Those appearing in Figure 4, for the purposes of this exposition largely simplified with respect to their real values and internal hierarchy:

```
mvtype :: S; SPC; SDO; SIO; SPR; SDOPC; SDOPR; SIOPC; SDOIO; SIOPR; SDOIOPC.
voice :: ACT; ACT_IMP; MIDD; MIDD_IMP.
preptype :: A; ANTE; ... TRAS
```

Figure 4. AVALON affixes for subcategorization information.

The affix *mvtype* specifies information about the number and type of the arguments that we can find in the syntactic sequence governed by a verb: since -S- stands for subject, -PC- for prepositional complement, -DO- for direct object, -IO- for indirect object and -PR- for predicative complement, the value S, for instance, indicates that a verb can only be combined with a subject, the value

SPC, indicates that a verb can be combined with a subject and a prepositional complement, and so forth.

The affix *voice* simultaneously encodes information about voice and impersonality, and can take one of the values ACT, if the verb can be combined with a subject and appears in active voice, ACT_IMP, if the verb cannot be combined with a subject and appears in active voice, MIDD, if the verb can be combined with a subject and appears in middle voice, and MIDD_IMP, if the verb cannot be combined with a subject and appears in middle voice.

The affix *preptype*, finally, specifies the preposition that introduces a prepositional complement in case the verb can be combined with one.

According to this, a valid lexicon definition in the AVALON lexicon is, for instance, the following:

Verb (SPC, MIDD, A, THIRD, SING, PRESENT, INDICATIVE): "dedica".

Figure 5. Lexicon definition of a verb in the AVALON lexicon.

Where the underlining indicates the contribution of BDS to this lexical entry, i. e., the values selected by this verb (*dedicar*, in contexts like *Se dedica a la construcción*, "He works in building") for the affixes *mvtype*, *voice* and *preptype*.

1.3. From BDS to the AVALON lexicon

In this Section we try to explain how do we obtain this information from BDS, i. e., how the values for the affixes *mvtype*, *voice* and *preptype* are derived from BDS and assigned within the lexicon definitions that account for the main verbs collected in the AVALON lexicon.

1.3.1. Configurations and schemes

To do this, we must start by distinguishing the two concepts configuration and scheme. Both concepts refer to a combination of the four elements that we consider that determine the syntactic behaviour of a verb: the **arguments** and their organization (or: **voice**) that it selects, its relation with the subject argument (or: **impersonality**) and the preposition required to introduce a prepositional complement if one is selected by the verb. Beyond this similarity, however, configuration refers to this combination as it appears in a real linguistic context, while scheme does as we associate it with a verb in our lexical competence. In computational terms, we can say that the configuration is the combination of argument subcategorization, voice, impersonality and prepositional requirements as it appears in the input sequence that has to be analysed, while scheme is the same combination as it appears in the lexicon that we use for the analysis.

The distinction just introduced between configurations and schemes has important consequences for the argument that we are addressing in this Section, i. e. how, from BDS, we can extract the information that, in the AVALON lexicon, concerns argument subcategorization, voice, impersonality and prepositional requirements for each verb contained in the lexicon. These consequences are those that derive from the fact that what BDS, which encodes the analysis of the great majority of the clauses of the ARTHUS corpus, documents are verbal configurations, while what we want to store in the AVALON lexicon is the information that forms the schemes that underlie these configurations. Obviously, the argument of this Section can now be redefined in the following way: how do we derive the AVALON lexicon schemes from the BDS configurations?

1.3.2. Operations of reduction

The instrument to do this are what we call the operations of reduction, i. e., certain operations that reduce configurations into schemes by means of the application of rules similar to the one in Figure 6.

The examples of BDS that show active voice impersonal configurations due to the presence of periphrasis *haber que* are reduced in the AVALON lexicon to active voice schemes whose verb type includes SUBJECT and all the arguments specified in the BDS analysis.

Figure 6. Example rule of reduction.

According to this rule, the analysis in BDS of, for instance, the sequence *Hay que salir*-"We must go out", causes the inclusion of the scheme *salir*-"to go out"/S/ACT in the AVALON lexicon. Therefore, all the entries corresponding in the lexicon to the different forms of verb *salir* must have each the adequate values for parameters such as mood, tense, person, gender or number, and, apart from these, all of them the same values **S** and **ACT** for, respectively, parameters *mvtype* and *voice*. Figure 7 shows one of these entries:

Verb (S, ACT, THIRD, SING, PRESENT, INDICATIVE): "sale".

Figure 7. Lexicon definition of a form of verb *salir* in the AVALON lexicon.

With respect to the operations of reduction, it is worth remarking, finally, that, as there are many possible and valid theoretical approaches to the relationships between configurations and schemes, there are also many ways, or more exactly degrees, of applying operations of reduction. We want to mention here, for their importance later both for the formalization and for the performance of the system, one especially relevant characteristic of our particular approach: we try to encode only those relationships which, being more sistem-

atic, can guarantee that the associated operations of reduction can be applied fully automatically, without any human intervention, in order to lay only correct schemes. On the grounds of automation, for instance, we do not encode the relationships that hold between what could be considered middle configurations based on active schemes that include a direct object (i. e. the relationship between *se lava*-"he wash himself" and *lavar algo o a alguien*-"to wash something or somebody"). And, what is more important, on the same grounds, we do not encode the relationships that could be identified between what could be considered configurations based on schemes that include optional arguments. This means that, for instance, we do not have in the AVALON lexicon SDO schemes associated with verbs that can also be added an optional indirect object; on the opposite, if a verb appears in BDS in two configurations of this type (with subject and direct object, on the one hand, and with subject, direct object and indirect object, on the other) we have two schemes in the AVALON lexicon (one SDO and another SDOIO). With these criteria, for developing the AVALON lexicon from BDS, we have been applying 11 rules of the type of the one in Figure 6.

Table 1 shows, for verb *abandonar*, "to leave, to give up", the data of BDS used by the operations of reduction and the result of them.

Relevant fields of BDS								Scheme
6	9	16	21	28	35	37	55	
1		2, 1						S/ACT (2 examples)
1, 3	Ø, 27, 10, 30, 15, 24, 33, 1, 23, 14, 15, 4	2, 1, 34, 30	1					SDO/ACT (175 examples)
1, 3		2, 1	1		1	a		SDOPC/ACT/A (3 exeamples)
1, 3	Ø, 30	1, 2	1		1	en		SDOPC/ACT/EN (5 examples)
1		2	1		1	sobre		SDOPC/ACT/SOBRE (1 example)
2		1			1			S/MIDD (3 examples)
2		2, 1, 30			1	a		SPC/MIDD/A (8 examples)

Table 1. Result of the operations of reduction for verb *abandonar*.

The BDS relevant fields for the operations of reduction (field 6 of voice, field 9 of type of periphrasis, fields 16, 21, 28, 35 and 55 of type of, respectively, subject, direct object, indirect object, prepositional complement and predicative complement, and field 37 of preposition introducing a prepositional complement in case there is one) are showed by the columns on the left, while the resulting schemes for the AVALON lexicon are showed by the column on the right. The first row shows a resulting S/ACT scheme – meaning that the verb

abandonar might be, in active voice, combined with a subject and no more arguments – derived from two examples of BDS in which, in field 6 of voice, we find the key 1 (all BDS keys are numerical keys), meaning that voice is active, and, in field 16 of type of subject, we find one of the values 1 or 2, respectively meaning that there are an explicit and an implicit subject. Next, although we will not explain them here in detail, the following six rows show other resulting schemes derived from other groups of examples of BDS, each of these associated with different keys for the relevant fields of BDS.

1.3.3. Frequencies

Looking at Table 1, it may also be observed that, on the one hand, many examples of BDS encode the same configurations, and, on the other, even if they encode different configurations, due to the intervention of the operations of reduction, they may result in the same schemes for the AVALON lexicon. Obviously, this ultimately means that we can consider certain schemes "more important" than others, on the grounds of their higher frequency, and, therefore, their higher possibilities of accounting for sequences of analysis. As a consequence of this, in addition to the operations of reduction, we decided to carry out a further selection of schemes according to their frequencies. So far, we have applied the following, quite simple, selection:

a) A scheme obtained from BDS enters the AVALON lexicon if its frequency is higher than 15.
b) If its frequency is lower than 15, but is higher than 10% of the frequency of a verb with frequency higher than 10.

Figure 8. Frequency selection.

But this can be, of course, easily modified. According to this selection, anyway, the only scheme of verb *abandonar* that, after the application of the operations of reduction, enters the AVALON lexicon is the second, since it has a frequency of 175, so higher than 15. The other six schemes, on the contrary, have all of them, on the one hand, frequencies lower than 15, and, on the other, lower than 19.7, which is the 10 % of the total frequency of verb *abandonar* in BDS (197).

1.3.4. Verb scheme clusters

Similarly, by looking at Table 1, we can also observe that the operations of reduction sometimes produce schemes that are partially equal, having one or more of the values for *mvtype*, *voice* or *preptype* equal. Taking this into account, in order to reduce as much as possible the size of the lexicon, we carry out what we call an operation of *clustering*, i. e. merging, of schemes. Figure 9 il-

lustrates the result of clustering the schemes that resulted after the application of the operations of reduction on the configurations collected in BDS for verb *abandonar*. The second column shows the result of the first operation of clustering, which consists of merging all the schemes that are only different with respect to their values for affix *preptype*. Next, the third and last column shows the result of the second operation of clustering, which in turn consists of merging all the schemes that, after the first operation, are only different with respect to their values for affix *mvtype*. There is still a third operation of clustering, which, quite obviously, consists of merging all the schemes that, after the second operation, are only different with respect to their values for affix *voice*. However, as there are not remaining schemes that fulfil this condition, this operation cannot be illustrated by means of verb *abandonar*.

Schemes of abandonar	Clustering of Prepositions	Verb type
S/ACT	S/ACT	S,SDO/ACT
SDO/ACT	SDO/ACT	
SDOPC/ACT/A	SDOPC/ACT/A, EN, SOBRE	SDOPC/ACT/A, EN, SOBRE
SDOPC/ACT/EN		
SDOPC/ACT/SOBRE		
S/MIDD	S/MIDD	S/MIDD
SPC/MIDD/A	SPC/MIDD/A	SPC/MIDD/A

Figure 9. Clustering of the schemes resulting from BDS for verb *abandonar*.

Obviously, if frequencies of schemes were not taken into account, what we have in the third column in Figure 9 would be the resulting schemes for verb *abandonar* in the AVALON lexicon, where each of the forms of the verb would have a set of lexicon definitions as the one that we can see in Figure 10. Final lexicon definitions for each form of abandonar in the AVALON lexicon. for the form *abandono*: each of these lexicon definitions associates *abandono*, or the form in question, with one of the verb scheme clusters obtained.

```
VerbSt ( S | SDO, ACT, FIRST, SING, PRESENT, INDICATIVE ): "abandono".
VerbSt ( SDOPC, ACT, A | EN | SOBRE, FIRST, SING, PRESENT, INDICATIVE ):
  "abandono".
VerbSt ( S, MIDD, FIRST, SING, PRESENT, INDICATIVE ): "abandono".
VerbSt ( SPC, MIDD, A, FIRST, SING, PRESENT, INDICATIVE ): "abandono".
```

Figure 10. Final lexicon definitions for each form of *abandonar* in the AVALON lexicon.

After all these operations, from 1284 verbs with frequency higher than 10, we have obtained 2236 verb scheme clusters, 1.73 average per verb, the maximum number of verb scheme clusters associated with a verb being 4.

1.4. BDS and the grammar of AVALON

In this Section we address the answer of the following question: How do we exploit the information about the schemes contained in the lexicon in the syntax of the grammar? Has BDS any additional role, beyond being the source of such schemes?

1.4.1. The AVALON grammar. The description of the syntax

In outline, the syntax of the AVALON grammar can be described as divided in two parts:

a) The first one is devoted to the structure of phrases. Among these, the verb phrase, intended as the combination of verb forms of auxiliary and/or main verbs and clitic pronouns, the structural category that underlies the clause function of PREDICATE, is formally described without any intervention whatsoever of BDS, since what we do at this level is writing all the rules necessary to account for all the theoretically possible combinations of verb forms and clitic pronouns that from one scheme can generate a configuration that can be contextualized at clause level.

b) The second part is devoted to clauses. These are described by means of a set of rules, many rules, similar to the one in Figure 11, according to which a clause consists of a subject, a predicate, a direct object and an indirect object, in the order SVDOIO.

```
Clause ( MAIN | SUBORDINATED, mood ):
  SUBJECT ( const_category1, person, gender, number ),
    PREDICATE ( SDOIO, ACT | MIDD, person, number, mood ),
    DIRECT OBJECT ( const_category2, person1, gender1, number1, person ),
    INDIRECT OBJECT ( const_category3, person2, gender2, number2, person ).
```

Figure 11. A rule for a clause in the AVALON grammar.

1.4.2. BDS and the rules of clauses in the AVALON grammar

Let's now have a look to the internal organization of the rules for clauses in the AVALON grammar. First of all, in the light of the example rule of Figure 11, we can observe that, as we already mentioned in 1.3.2., there are not optional arguments in the schemes: all the arguments represented in the scheme must indeed be made explicit in the rule, either by means of syntactic functions, as in the example rule of Figure 11, or by means of clitic pronouns. We call this property complete recall of the clause with respect to the scheme.

Secondly, if we take into account the reciprocal organization of the rules of clauses in the AVALON grammar, we can also observe that, by means of the order of the rules, priority is given to those rules that describe clauses with a

higher number of arguments. That is, if a verb is associated in the AVALON lexicon with two possible schemes, one SDO and another SDOIO, the second one will be always tried before during the process of analysis of input sequences. We call this property complete recall of the clause with respect to the set of schemes of a verb.

Finally, with respect also to the reciprocal organization of the rules, we can still observe that, for the same number of arguments specified within them, priority is given to those rules that present the arguments in the order more frequent for them. That is, priority is given, for instance, to a rule that identifies a clause as having a subject and a direct object in the order S-Verb-DO over one that identifies it as having the same arguments in the order DO-Verb-S. This point is especially relevant for our issue in this exposition, because it is at this point precisely where BDS has an important role for the description of clauses in the AVALON grammar, given the fact that BDS is the source of the data about the frequency of the order of arguments. In this respect, we also want to remark that, in a great many cases, on the basis of these data, we have even disabled rules that specified an order of arguments not documented in BDS, which had the desirable consequence of a considerable increment in the efficiency of the performance of the parser.

2. CONCLUSIONS

- BDS supplies the AVALON lexicon with the necessary data about verb schemes.
- BDS supplies the AVALON grammar with the data used in the rules of clauses about the frequency of the order of arguments.
- In general, the analysis as was performed in BDS is the analysis encoded in the AVALON grammar, given that what we ultimately want is to produce a parser to automatically further extend BDS.

3. REFERENCES

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

Koster, Cornelis H. A. (1991): Affix Grammars for Natural Languages. In Albas, H. and B. Melichar: Attribute Grammars, Applications and Systems, International Summer School SAGA, Prague, June, 1991, Lecture Notes in Computer Science 545. Heidelberg: Springer-Verlag: 469–484.

Mas, Concepción, Guillermo Rojo (2004): Design, Construction and Exploitation of the "Base de datos sintácticos del español actual (BDS)". This volume.

Muñiz, Eva M.ª, et al. (2003): Description and Exploitation of BDS: A Syntactic Database about Verb Government in Spanish. In: Angelova, Galia et al. (eds. 2003): *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP–2003), Borovets – Bulgaria, 10–12 September 2003.* Shoumen: Incoma, pp. 297–303.

Rojo, Guillermo (1995): La Base de Datos Sintácticos del español actual. In: Español Actual: 59: 15–20.

Rojo, Guillermo (2001): La explotación de la Base de Datos Sintácticos del español actual. In Kock, Josse De: Lingüística con corpus (=De Kock, Josse De: Gramática española. Enseñanza e investigación, I, 7). Salamanca: Universidad de Salamanca.

Santalla, Mª. Paula (2002): A Formal Grammar of Spanish for Phrase-level Analysis Applied to Information Retrieval. Santiago de Compostela: Servicio de Publicaciones de la Universidad de Santiago.