[i] The Spanish and Catalan corpus were annotated with constituents and functions, whereas the Basque corpus was annotated with dependencies.
[ii] http://framenet.icsi.berkeley.edu
[iii] HUM 2004-21127-E
[iv] For this task, a steady version of the Spanish, Catalan and Basque versions of EuroWordNet was used.
[v] The field EN, the English translation of the sentence, is not included in the real lexical entry.
[vi] The role tagset appears in Table 1.
[vii] The child goes for a walk in the park vs. The mother walks the baby in the park.
[viii] Translations are literal in order to maintain the Spanish construction.

Constituent Order in Spanish: a Syntactically Annotated Corpus-Based Study

M. Pilar Valverde,[†] M. Paula Santalla[†] and Montserrat Civit*

[†]Universidad de Santiago de Compostela
Departamento de Lengua Española
{mpilarv,fempsr}@usc.es

*Universitat de Barcelona
CLiC, Centre de Llenguatge i Computació
civit@thera-clic.com

1 Introduction

The increase in the quantity of available syntactically annotated corpora or banks in the last years allows linguists to study some syntactic questions fi new perspective ([8] and [10], for example), since corpora offer both real d check hypothesis and data about the frequency of syntactic phenomena. This of studies are relevant not only in linguistic theory but also in order to de automatic parsers. In this paper we present the results of a corpus-based lir tic research work about functional constituent order in Spanish [11]. This syn knowledge is essential in the automatic identification of the syntactic function make up the clause in Spanish (along with other kind of information like agree and prepositions), and therefore for syntactic parsers, found in the majority (tural language processing tools.

2 The corpora

In Spanish there exist three syntactically annotated corpora: the Base de I Sintácticos del español actual (BDS) [9], the UAM Spanish treebank [7] an Cast3LB corpus [4]. To carry out our investigation we have at our disposal th and the third corpora: the Base de Datos Sintácticos del Español actual (BDS the Cast3LB corpus, since the UAM Spanish treebank is not publicly availabl

far we have been consulting the BDS and we have partially tested the results with Cast3LB.

BDS, developed by the research group *Sintaxis del Español* (Spanish Syntax) of the University of Santiago de Compostela is constituted by 1,500,000 words. But BDS is not an annotated corpus or a treebank, in which syntactic tags are inserted in the text. It is instead a syntactic database that contains the syntactic data that correspond to the analysis, done by hand, of the clauses that appear in the corpus. The database is made up of 160,000 records each with 63 fields of information. Each record of the database contains the syntactic analysis of one clause and it stores information about the clause as a whole, the verb and the functional constituents identified within the clause. The syntactic information is encoded by means of a large set of hierarchically organised numerical keys, so the information can be easily retrieved with different degrees of detail. This corpus is addressed to elaborate a Spanish Dictionary of Verb Structure and Government.

The Cast3LB corpus is a treebank developed by the universities of Barcelona, Politècnica de Catalunya, Politècnica de Valencia and University of Alicante. It consists of 100,000 words morphologically, syntactically and semantically annotated and constitutes a resource addressed both to develop NLP applications and to be useful for linguistic research.

3 Constituent order in Spanish

3.1 Functional constituents

The two corpora available have been syntactically annotated at two different levels: the constituent level and the functional level. Our research deals with the order of constituents that play a central function in the clause. In Spanish this functions are generally considered to be the following:[1]

1. Subject: *La gente* comienza a abandonar el templo. (People start to leave the temple)
2. Predicate: La gente *comienza a abandonar* el templo. (People start to leave the temple)
3. Direct object: La gente comienza a abandonar *el templo*. (People start to leave the temple)
4. Indirect object: He de contárselo *a Hortensia*. (I have to tell it to Hortensia)

[1]The examples that appear in this paper have been extracted from BDS.

5. Prepositional complement: Nunca pensó *en sus hijos*. (He never thought about his children)
6. Predicative complement of the subject: Ella es *especial*. (She is special)
7. Predicative complement of the direct object: Nadie te considera *un viejo*. (Nobody considers you to be an old man)
8. Agent complement: El primer ministro fue recibido también *por Felipe González*. (The Prime Minister was received by Felipe González)

We have ignored the rest of constituents than can appear in the clause: optional predicative complements, circumstances, sentence adjuncts, textual elements, modifiers and vocatives (for a definition, see [3]). Circumstances can appear almost everywhere and in different number. We have also ignored clitics pronouns, since they present a fixed order: they appear immediately before or after (and attached to) the verb.

3.2 Free constituent order?

Spanish is said to be a "free" constituent order language, although this description is not very appropiate. Constituent order is not free in the strict sense of the word because, on the one hand, not all the theoretically possible orders are found and, on the other hand, orders are not equivalent and equally frequent at all. There are some factors, of which we are going to examine some, that limit the number of possible orders.

3.3 Factors that determine constituent order

In Spanish, as in the majority of languages, constituent order can be explained from a pragmatic point of view ([5] and [12]), because what is spoken about appears in the first position of the clause, and new information appears after. Most linguistic studies about order focus on pragmatic factors, but syntactic factors can frequently explain order phenomena as well. In this respect, the fact that Spanish usually follows the order "subject + verb + complements" stands out.

Syntactic factors have been underestimated and the few studies about syntactic influence on constituent order in Spanish that exist ([6], [2] and [1]) are not based on corpus data or are not exhaustive, due to the lack of available syntactically annotated corpora. In our study we focus on syntactic factors that influence constituent order, since, on the one hand, it proves to be useful to improve automatic parsing and, on the other hand, this kind of information can easily be extracted from corpora.

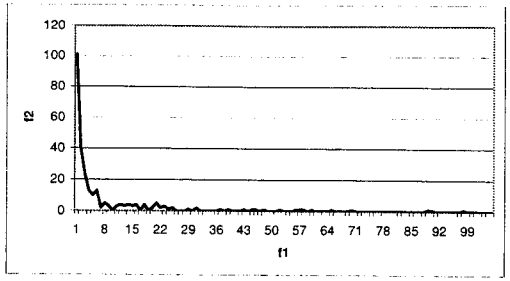

Figure 1: Quantity of orders (f2) with a given frequency (f1)

Specifically, we study the influence of two clausal characteristics on the flexibility of clauses. We define "**flexibility**" as the capacity of changing the order of constituents of a given clause, and therefore flexibility is proportional to the quantity of orders in which a clause can appear.

As we are going to see, some **types of clauses** (independent, subordinated, finite, non finite, etc.) and some **types of voices** (active, middle, passive, etc.) are more flexible than others. And more flexible ones present more orders in BDS than less flexible ones.

However, we have to take into account that there exist other syntactic factors that can determine constituent order (like the category of the constituent or the verbal scheme). For the moment we have put them aside.

4 Results

4.1 The orders and their frequency

In Spanish a given set of functions can usually appear in several different orders. But all the orders are not possible, as we can see if we compare theoretically possible orders with orders found in the corpus.

Moreover, among orders found in BDS, some are more frequent than others. In the field of the order of constituents, like in other linguistic phenomena, there are very few frequent individuals and a lot of not very frequent individuals, as is shown in Figure 1, where quantity of orders that are found with a given frequency is represented. More than a half of the orders documented in BDS appear only between one and five times, and the thirty six most frequent orders account for 97% of clauses. Therefore, the order is not so "free" as it could appear at first glance. There are, in fact, quantitative limitations.

4.2 Factors that determine the flexibility

We can establish that the quantity of orders in which a type of clause or voice can appear, and therefore the flexibility of the clause, depends on three formal characteristics:

1. **The number of elements that can take part in the order (m)**. The less the elements that can take part in the order, the less the different orders in which they can appear. As we have pointed out in section 3.1, we study the position of eight different functions. However, some function are not compatible with certain clauses and voices, which makes them less flexible than others.

2. **The number of elements that make up the clause (n)**. The less elements that make up the clause, the less different orders in which they can appear. In BDS clauses are made up mainly by one, two or three elements, although four or five elements are possible, too, so we consider that a clause can be made up by between one and five elements. However, certain clauses and voices especially tend to be made up by a high or a low number of elements, which makes them more or less flexible than others.

3. **The presence in the clause of a function that occupies always the same position (f)**. Some clauses and voices contain a function that appears almost always in the same position, which reduces the mobility of the rest of functions and therefore the number of orders in which they can appear.

4.3 The type of clause

We are going to see that some clauses are more or less flexible than others according to some of the three factors that we have presented in the previous section.

4.3.1 Classification of clauses

We follow a general classification of types of clauses in Spanish:

1. Independent
   (a) Simple (O.DIE): *La gente comienza a abandonar el templo.* (People start to leave the temple)
   (b) Interrogative *wh-* (O.Q): *¿Por qué no me buscaste?* (Why didn't you look for me?)
   (c) Imperative (O.IMP): *Abre la puerta* (Open the door)

2. Bipolars: each one of the two members of adversative, concessive, conditional, causal, consecutive and comparative clauses.

   Aceptó mi decisión, pero se quejó (He accepted my decision, but he complained about it)
   Ella tenía fuertes dolores aunque apenas comía (She had severe aches although she hardly ate)
   Si quieres, te acompaño (If you want, I come with you)
   Sería injusto, porque él no le ha abandonado (It would be unfair, because he hasn't left her)
   Olía tan mal que poca gente se le acercaba (He smelt so bad that few people approached to him)
   Yo la sufro más de lo que crees (I suffer from it more than you think)

3. Subordinated

   (a) Finite
      i. *That* clauses (S.F.C.que): Cuando intenté concentrarme advertí *que estaba temblando.* (When I tried to concentrate I noticed that I was shaking)
      ii. *Whether* clauses (S.F.C.si): Yo no sé *si la conocerás* (I don't know whether you will know her)
      iii. Interrogative *wh-* (S.F.C.Q): No sé *de qué me hablas* (I don't know what you are talking me about)
      iv. Adverbial (S.F.A): *Cuando llegué*, la función había empezado. (When I arrived, the performance had already started)
      v. Relative (S.F.R): ¿Es algo *que pueda adivinar*? (Is it something that I can guess?)

   (b) Non finite
      i. Infinitive (S.NF.C): Nos encanta *oírlo.* (We love to hear it)
      ii. Gerund (S.NF.A): Teo la hace callar *besándola.* (Teo gets her to be quiet kissing her)
      iii. Participle (S.NF.P): Estoy *cansado de la enseñanza* (I am tired of teaching)
      iv. Relative (S.NF.R): Tengo algo *que contarle* (I've got something to tell him)
      v. Interrogative *wh-* (S.NF.Q): Y no sabían muy bien *de qué hablar* (And they didn't know very well what to talk about)

4.3.2 The number of elements that can take part in the order (m)

Most clauses are compatible with every one of the eight functions considered (see section 3.1). However, some clauses are incompatible with certain functions, i.e., some functions dont't appear in certain clauses. Specifically, in BDS imperative clauses are incompatible with agent complement; subordinated non finite participle clauses are incompatible with direct object and predicative complement of the direct object; subordinated non finite relative and interrogative clauses are incompatible with subject and agent complement and subordinated non finite interrogative ones are incompatible with indirect object, too. Therefore, taking into account the number of elements that can take part in the order (m), imperative clauses and subordinated non finite participle, relative and interrogative clauses are less flexible than the other types of clauses.

4.3.3 The number of elements that make up the clause (n)

The number of constituents that make up the clause varies according to the type of clause. In BDS clauses are primarily made up by two constituents. However, there are some exceptions: subordinated finite relative clauses are mainly made up by three constituents and imperative and subordinated non finite participle clauses are made up mainly by one constituent. If we consider that clauses principally made up by two or three constituents are more flexible and that clauses made up by one or two constituents are less flexible, we can conclude that imperative, subordinated non finite (except relative ones) and subordinated finite adverbial clauses are less flexible than the others.

4.3.4 The presence in the clause of an element with a fixed position (f)

According to the data of BDS, predicate is the first function of the clause in more than 95% of cases in imperative and subordinated non finite (infinitive, gerund and participle) clauses, and subject appears inmediately after the verb in the same percentage in subordinated non finite participle clauses. Therefore, according to this factor, non finite (infinitive, gerund and participle) and imperative clauses are less flexible than the other types of clauses.

4.3.5 More or less flexible clauses

According to the three formal factors examined so far, we can conclude that some clauses are less flexible than others: subordinated non finite clauses, imperative clauses and subordinated finite adverbial clauses. In Table 1 we show how each

Type of clause	m	n	f	Result
O.DIE				
O.Q				
O.IMP	x	x	x	x
B				
S.F.C.que				
S.F.C.si				
S.F.C.Q				
S.F.A		x		x
S.F.R				
S.NF.C		x	x	x
S.NF.A		x	x	x
S.NF.P	x	x	x	x
S.NF.R	x			x
S.NF.Q	x	x		x

Table 1: More or less flexible clauses and the reason for it

factor causes each type of clause to be more or less (x) flexible, and the result: more ore less (x) flexible clauses.[2]

Moreover, we have checked that flexibility is proportional to the quantity of orders: in BDS less flexible clauses present less orders than more flexible clauses.

4.4 Voice

Certain voice structures determine the order of constituents in the clause, in the sense that some voices are more or less flexible than others. As we have seen in section 4.2, the degree of flexibility of a given clause depends on three formal factors, that we are going to examine here with respect to voice.

4.4.1 Classification of voices

We follow a general classification of types of voices in Spanish:

1. Personal voice: is compatible with the appearance of a syntactic subject.

   (a) Active personal (AP): the verb may appear together with one or two functional clitic pronouns.

[2]See section 4.3.1 to check the correspondence between acronyms and types of clauses.

   La gente comienza a abandonar el templo. (People start to leave the temple)

   (b) Middle personal (MP): the verb appears together within a verbal clitic pronoun with the same person and number of the subject (and the verb):

   El juego se acabó. (The play finished)

2. Impersonal voice: is incompatible with the appearance of a syntactic subject.

   (a) Active impersonal without *se* (AI): the verb is inherently incompatible with a subject.

   No va a llover esta noche. (Tonight it is not going to rain)

   (b) Active impersonal with *se* (AIS): the verb appears together with the impersonal clitic pronoun *se* and it is incompatible with the subject.

   No se puede circular en moto con este tiempo. (You can't drive a motorbike in this weather)

   (c) Middle impersonal (MI): the verb appears together with the verbal clitic pronoun *se* and it is incompatible with a subject.

   ¿De qué se trata? (What's it about?)

3. Passive: the subject and the agent complement correspond respectively to the direct object and the subject of the active voice.

   (a) Periphrastic (PP): the verb is formed by a past participle or by the appropiate tense and person of the auxiliary *ser* and a past participle that agrees in number and gender with the subject.

   El primer ministro fue recibido también por Felipe González. (The Prime Minister was received by Felipe González)

   (b) With *se* (PS): the verb appears together with the passive clitic pronoun *se*, which agrees with the subject and is third person singular or plural.

   El proceso se habrió en Poitiers a fines de julio de 1956. (The trial was opened in Poitiers at the end of july 1956)

4.4.2 The number of elements that can take part in the order (m)

The number of elements that can take part in the order varies according to the type of voice: impersonal voices are incompatible with subject and passive voices are incompatible with direct object and predicative complement of the direct object. Therefore, these voices are less flexible than personal voices. Middle impersonal voice is compatible only with three functions, so this is the less flexible voice with respect to this factor.

Voice	m	n	f	Result
AP				
APS				
AI	x	x		x
AIS	x	x		x
MP				
MI	x	x	x	x
PP	x		x	x
PS	x			x

Table 2: More or less flexible voices and the reason for it

4.4.3 The number of elements that make up the clause (n)

Clauses in personal or passive voice are mainly made up by two or three constituents, but clauses in impersonal voice are mainly made up by one or two constituents. Therefore, personal and passive voices are more flexible than impersonal voices, according to this factor.

4.4.4 The presence in the clause of an element with a fixed position (f)

The predicate and the prepositional complement are placed in fixed position in the middle impersonal voice, while the agent complement does it in the passive periphrastic voice.

4.4.5 More or less flexible voices

According to the three factors examined, impersonal and passive voices are less flexible than personal voices. Middle impersonal voice is the less flexible among all of them, since the three factors examined cause a low flexibility in this voice, as is shown in Table 2[3].

5 Comparison between BDS and Cast3LB

We have tested with Cast3LB some of the results obtained from the BDS. Specifically, we have checked that in the two corpora non finite clauses are less flexible than the other types of clauses, according to the three parameters (m, n, f) examined.

[3]See section 4.4.1 to check the corresponde between acronyms and types of voices.

6 Conclusions and further work

In this paper we have seen, on the one hand, that a low number of very frequent orders accounts for a very high number of clauses. On the other hand, we have proved, with the help of BDS, the influence of two syntactic factors on constituent order, type of clause and voice: less protoypical clauses and voices have a much more fixed order than more protoypical ones. This kind of information is relevant for linguistic theory, since studies about constituent order focus only on pragmatic factors or are not based on empirical data. And is also essential in the development of automatic parsers, aimed at identifying the functional constituents that make up the clause in Spanish.

We can establish the following objectives:

1. Study of other syntactic factors like the category of the constituent, which can determine the position that this constituent occupies in the clause.
2. Definition of less flexible constructions and their most frequent orders.
3. Examination of the more flexible constructions and the factors that determine the order of their constituents.
4. Verification with Cast3LB and its extension (CESS-ECE) of the results obtained from BDS.
5. Establishment of a hierarchy between the different factors that determine constituent order.
6. Improvement of a formal grammar with the help of this linguistic knowledge.

References

[1] Bosque, I. and V. Demonte (1999) *Gramática descriptiva de la lengua española*, Espasa Calpe, Madrid.

[2] Butt, J. and C. Benjamin (1988) Word order. In *A New Reference Grammar of Modern Spanish*, Edward Arnold, London, pages 394-400.

[3] Civit, M. (2004) Guía para la anotación de las funciones sintácticas de cast3lb: un corpus del español con anotación sintáctica, semántica y pragmática. Technical report, CLIC. Available: http://www.clic.fil.ub.es/personal/civit/publicacions.html

[4] Civit, M. and M.A. Martí (2004) Building Cast3LB: a Spanish Treebank. In *Research on Language and Computation*, Vol. 2, Num. 4, Springer Science+Business Media B.V, pages 549–574.

[5] Contreras, H. (1978) *El orden de palabras en español*, Cátedra, Madrid.

[6] Fernández Ramírez, S. (1986) El orden de las palabras: la posición del sujeto. In Bosque, I. (ed.), *Gramática española (4): El verbo y la oración*, Arcolibros, Madrid, pages 430–462.

[7] Moreno, A., S. López, F. Sánchez and R. Grishman (2002) Developing a Spanish Treebank. In Abeillé, A. (ed.), *Building and Using syntactically annotated corpora*, Kluwer, Dordrecht.

[8] Oostdijk, N. and P. de Haan (1997) Clause patterns in Modern British English: A corpus-based (quantitative) study. In *ICAME Journal*, Vol. 18, pages 41–79.

[9] Rojo, G. (2001). La explotación de la Base de datos sintácticos del español actual (BDS) In De Kock, J. (ed.), *Lingüística con corpus. Catorce aplicaciones sobre el español*, Universidad de Salamanca, pages 255–286.

[10] Rojo, G. (2003) La frecuencia de los esquemas sintácticos clausuales en español. In *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, Arcolibros, Madrid, pages 413–424.

[11] Valverde, M.P. (2005) Estudio sobre el orden de las funciones en español basado en corpus. Master's Thesis, Universitat de Barcelona, Departament de Lingüística General.

[12] Zubizarreta, M. L. (1999) Las funciones informativas: tema y foco. In Bosque, I. and V. Demonte (ed.), *Gramática descriptiva de la lengua española (3): Entre la oración y el discurso*, Espasa Calpe, Madrid, chapter 64, pages 4215–4244.

Parser evaluation across Text Types

Yannick Versley
Seminar für Sprachwissenschaft
Universität Tübingen
E-mail: versley@sfs.uni-tuebingen.de

1 Introduction

When a statistical parser is trained on one treebank, one usually tests it on another portion of the same treebank, partly due to the fact that a comparable annotation format is needed for testing. But the user of a parser may not be interested in parsing sentences from the same newspaper all over, or even wants syntactic annotations for a slightly different text type. Gildea (2001) for instance found that a parser trained on the WSJ portion of the Penn Treebank performs less well on the Brown corpus (the subset that is available in the PTB bracketing format) than a parser that has been trained only on the Brown corpus, although the latter one has only half as many sentences as the former. Additionally, a parser trained on both the WSJ and Brown corpora performs less well on the Brown corpus than on the WSJ one.[1]

This leads us to the following questions that we would like to address in this paper:

- Is there a difference in usefulness of techniques that are used to improve parser performance between the same-corpus and the different-corpus case?
- Are different types of parsers (rule-based and statistical) equally sensitive to corpus variation?

To achieve this, we compared the quality of the parses of a hand-crafted constraint-based parser and a statistical PCFG-based parser that was trained on a treebank of German newspaper text.

[1] Ratnaparkhi (1998) made a similar experiment on the "Magazine & Journal Articles", "General Fiction" and "Adventure Fiction" sections from the Brown Corpus and notes that the 5-7% loss in accuracy for his maximum-entropy parser cannot be accounted for by inherent parsing difficulty since training on the same amount of text yields similar (within 1%) results for the same-corpus case. Roark and Bacchiani find in their adaptation experiments that their parser on the Brown corpus and testing on WSJ texts yields a 9% loss in comparison to testing on the Brown corpus and a 10% loss in comparison to training on a similar amount of WSJ text.