Igor Bolshakov, Alexander Gelbukh, Grigori Sidorov. *On Subject-Predicate Agreement in Spanish*. Proc. CICLing-2000, International Conference on Intelligent Text Processing and Computational Linguistics, February 13-19, 2000, Mexico City. CIC-IPN, ISBN 970-18-4206-5, pp. 59–67.

On Subject-Predicate Agreement in Spanish

Igor A. Bolshakov
Alexander F. Gelbukh
Grigori Sidorov

In the paper, subjectival relation in Spanish is studied. It connects the grammatical predicate of a clause with its subject. Possible forms of subjectival subtree are enumerated. The rules of morphological agreement between the subject and the predicate are given using a formal metalanguage. It is supposed that the su bject always is marked with the values of the grammatical categories that participate in the agreement. The rules for deducing these categories for coordinated phrases are formulated.

1 INTRODUCTION

Dependency grammars describe intra-sentence structures through dependency relations between different wordforms. As all languages, Spanish has an important syntactic relation that connects the head of a sentence called grammatical predicate (the governor in the syntactic relation) with the subject of the same sentence (the dependent). Some information on Spanish grammar can be found in [Seco 1968, Esbozo 1998, Constantino 1977].

Any syntactic relation in dependency grammars can be thought of as an arrow pointed from the governor to the dependent. The Meaning ⇔ Text theory ascribes a label to each such arrow. For the syntactic relation under consideration, this label is just subjectival, though, within the Meaning ⇔ Text theory [Mel'čuk 1998], this relation was usually called *predicative*. However, the terms *predicate* and *predicative* are rather ambiguous. Thus, we were quite ready to adopt the novel term *subjectival* relation, when it has been proposed recently in [Mel'čuk 1999], just in the same meaning.

The objective of this paper is to formulate the formal rules of agreement that permit to detect the presence of the subjectival relation during text analysis. To solve this task it is necessary to describe all possible forms that can have subject in Spanish. The forms of the predicate are not so interesting for us in this paper because the predicate is the subordinate part in agreement. Usually we ignore the form of the predicate in our further discussion.

Although there are patterns of rules for subjectival relation elaborated within the framework of Meaning ⇔ Text theory for English [Mel'čuk & Pertsov 1987] and French [Apresian 1984], we prefer to use a slightly more economic formal language.

2 FORMS OF THE SUBJECT IN SPANISH

It is important for our paper what surface forms of the subject exist in Spanish sentence. The syntactic subject in Spanish may be expressed by noun or its syntactic equivalent. Let us consider the latter case in more details.

- A personal pronoun: *Nosotros queremos bailar.*
- A pronoun of other type with substantival properties: *Nadie te llama*. *Los míos la quieren*; *¿Quién lo duda?*
- A numeral: *Esos dos entran en la sala.*
- A standalone element of a reference: *La (4) indica* ...
- A full formula: $E = mc^2$ *es la ley teórica.*
- Any string of symbols, not obligatorily belonging to this language, in a definition construction introduced with the verbs *significar, querer decir, definirse*, etc.: *Shoh significa amor.*
- A wordform of any part of speech (usually with adjective properties) heading the elective construction: *la más grande de las niñas*; *el mejor de los alumnos; una de las demandas; ninguno de los diarios; cinco de aquellos soldados.*
- A substantivized lexeme of any original part of speech except the verb with a resulting generalized meaning induced by the "neutral article" *lo*, like *lo suficiente, lo pasado, lo nuevo, lo largo, lo mío, lo cerca*: *Todo lo nuevo es invencible*.
- A grammatically substantivized verb in infinitive optionally marked with the article *el* or adjectival pronoun as a determinant: *El fumar es dañino*. *Es necesario complementar esta acción con otras.* After substantivization, the verb can preserve usual verbal dependencies: *El fumar permanentemente es muy peligroso*; *El resolver ambigüedades es tarea muy importante.* Also, the substantivized verb may acquire attribute(s) in singular masculine: *Aquel murmurar extraño se oía por todas partes*. Such use of verbs is specific for Spanish. The cases when the infinitive is fully substantivized and thus is mentioned in dicitonaries are rather rare (*deber*, *placer, poder*, *saber*, etc.).
- A grammatically substantivized interjection with the article *el* or *los*: *Los ayayay de los niños fueron muy frecuentes.*
- A grammatically substantivized participle denoting a person, with a definite or indefinite article: *Los invitados han propuesto muchas preguntas. Una acusada fue trasladada a la oficina.* Some of these participles are given in

the dictionaries, mainly in the meaning of adjectives. Here we see the further step of conversion, namely, substantivization of these adjectives.

- An anaphorically substantivized participle or adjective with any definite article: *Las habitaciones sencillas se ofrecieron a los turistas sin pareja; los acampañados se instalaron en las habitaciones comunes. Había muchos vestidos allí; los rojos eran más caros.* The implied entities (*turistas* in the first example, and *vestidos* in the second) can be restored in such cases only based on a broader context.

- *lo, le,* or *que* used in the conjunctive role to introduce a subordinate clause (the latter is given below in square brackets): *Lo* [*que conozco bien*] *parece indispensable para ...*; *El* [*cómo lo consiguió*] *nadie lo sabe. Es bien conocido que* [*el procesamiento es muy difícil*]*; Se supone que* [*fue grabado dos meses después*]. The dependency arrow is pointed here from the conjunctive word to the head of subordinate clause, i. e. to its own predicate verb (underlined within brackets).

- The head of a coordinated phrase of two or more nouns or of any syntactic equivalents of noun enumerated above: *bienes y servicios; la libertad, la igualdad y la fraternidad*; *tú y yo; IPN* y *UNAM; (4), (5) y (6); un hombre y más de diez mujeres; el procesamiento de muchas etapas y el resolver ambigüedades; ni tú ni yo; mamá o papá.* The first (underlined) component of a coordinated phrase is considered the root of the dependency subtree. It is joined to a next component of the phrase or to the conjunction through coordinative relation, rather specific in its properties within the entire system of dependency relations.

- The head of a direct speech fragment in quotation marks supplied with a determinant (i.e., an article or a possessive pronoun in the short form) and/or an adjective in masculine: *Su "déjeme dinero" enojó a Juan; Un estruendoso "dámelo" se oyó en la casa.*

- A quantitative group including a distant pair of prepositions like *a ... de, hasta ... desde* and different measuring units: *Se requiere* [*desde un minuto hasta unas horas*] *para detener el sangrado.* The underlined preposition proved to be the root of the subtree corresponding to the subject group given in brackets. As to quantitative groups with *más de ...*, *menos de ...*, *mas que ...*, *menos que ...*, or with a distant pair of prepositions *de ... a ...*, *entre ... y ...*, *desde ... hasta ...*, having coinciding measuring units ([*Más de doce*] *mesas se colocarán en el salón;*[*Mas que una*] *mujer está allá;* [*Entre 15 y 20*] *personas se reunirán aquí;* [*De tres a ocho*] *hombres seleccionarán para el trabajo...*,), they depend on underlined nouns, which are the roots of entire subjectival syntactic groups.

- A head of direct speech in quotation marks in the context of a predicative verb with pronominal clitic *se* in the special passivizing meaning and the succeeding colon: *Se dice: "Quien no trabaja no come"*.
- A head of a subordinate clause with the personal verb**.** This clause (given in each following example in the square brackets and with the underlined root) has usually the form of an embedded question: *Es bien conocido,* [*en cuales manos están los recursos de nuestro país*]; [*Quien no trabaja*] *no come.* Thus, the dependency arrow is pointed from the predicate root of the main clause tree directly to the predicative root of the subordinate clause subtree.
- A head of a subordinate clause with infinitive: *Es bien conocido* [*qué hacer ahorita*]. The dependency arrow is pointed from the predicate root of the main clause tree to the infinitive that is the root of the subordinate clause subtree.

The forms of a predicate of Spanish sentence are also rather diverse: finite verb, construction with auxiliary verbs, predicate with nominal part expressed by words of different parts of speech, prepositional phrases with noun or infinitive, or subordinate clause, but in the context of this paper the only significant thing is the possibility to have noun in the nominal part (see 3.3.)

3 SUBJECT-PREDICATE AGREEMENT

We use the following natural notation[1] for grammatical categories that participate in the agreement and their values. We introduce a linear order relation describing the hierarchy of the values of each category. We consider that each category has an unmarked value which is denoted by underscoring.

person (PRS):	1PRS > 2PRS > <u>3PRS</u>
gender (GND):	<u>M</u> > F
number (NMB):	PL > <u>SG</u>

Let us denote s as subject and p as predicate. For a category C we denote the value of the category of a word x as $C(x)$, e.g. NMB(x), GND(x), PRS(x), NMB(s)=NMB(p).

We assume that though the subject in Spanish has diverse forms of representation and the categories mentioned above sometimes are not expressed on the surface level (e.g., infinitive) or are expressed as a combination of values (e.g.,

[1] Categories: PRS - person, GND - gender, NMB - number. Values of PRS: 1PRS - first person, 2PRS - second person, 3PRS - third person. Values of GND: M - masculine, F - feminine. Values of NMB: SG - singular, PL - plural. Also we use abbreviation CP for a coordinated phrase and QG for a quantitative group.

coordinated phrase (CP)), still the subject always is marked with a value of each category.

The unmarked value of each category is used to determine the value of a category for the subject if the value does not exist at the surface level (for example, PRS for the infinitive). In this case the unmarked category should be used for the agreement rules.

In case of the subject expressed by CP it is necessary to determine a value of each category for CP (to substitute the combination of different values with one value). To solve this problem we use the hierarchy of values in accordance with the following convention: CP has the maximum value that its members have in the hierarchy. In case of the NMB the situation is more complicated (see below).

Additional properties of $C(x)$ in each case are presented in the corresponding sections below.

Using the notation and conventions introduced above we deduce very simple and natural rule for determination of the subject-predicate agreement which can be implemented in a computer system without any changes:

$C(p) = C(s)$

that means that the values of a category C of the subject s and the predicate p are equal.

3.1 Agreement in person

There are no additional conditions for PRS(p).

Examples for PRS(p):

1. *El padre cree en ella*. (3PRS)
2. *Pablo, María y Sergio se quedan aquí.* (3PRS)
3. *Fumar está prohibido*. (3PRS)
4. *Pedro y ella van al teatro.* (3PRS)
5. *Tú y yo nos quedamos aquí.* (1PRS > 2PRS)
6. *Él y yo nos quedaremos allí.* (3PRS < 2PRS)
7. *Tú o María os quedais aquí.* (2PRS > 3PRS)

3.2 Agreement in gender

Additional condition for GND(p): if the grammatical and the semantic gender of x are different or x has no grammatic gender, then the semantic gender is used (e.g., *Su Alteza* or personal pronouns (*yo, tú*) have GND(x) = M or GND(x) = F, which depends on the sex of a referent person) .

Examples for GND(p):

8. *El edificio es alto*. (M)
9. *La casa es alta*. (F)
10. *El edificio y la casa son altos*. (M > F)
11. *Fumar está prohibido*. (M)
12. *Tú estás cansado/a*. (semantic gender)
13. *Su Alteza está disgustado/a*. (semantic gender)

3.3 Agreement in number

Additional condition for NMB(p): if s is collective noun and NMB(s)=SG and p has nominal part expressed by noun y and NMB(y)=PL then NMB(p)=PL (See example 16). Generally speaking, this is a special kind of sentences where it is supposed the identity of the collective noun (like *crew*) and its members (like *sailors*).

In case of the quantitative group (QG) it is necessary to take the value of the category of the nearest to the predicate noun (see examples 17, 18, 19, 20, 21).

The rules for determination of the category of CP in case of NMB(*CP*) are more complicated than simply application of the hierarchy of the values of NMB as in the cases above. These rules are presented in the table below. The rules in the table should be applied starting from the end of the table (from the more specific to the more general). If the condition of the rule is fulfilled then the checking mechanism should stop. Nevertheless in case of the rules that have the mark *optional* (rules 6 and 7), the checking mechanism should not stop after processing of the rule even if it was successful and continue handling the rest rules.

In case of CPs which has several members with different conjunctions it is necessary to check the different hypothesis of the upper-level conjunction and apply corresponding rules (see examples 33 and 34).

In the table if there is no value in the field "Condition" it means that there are no special condition. If there is no value in the field "CP" it means that the type of CP is not important. The value *y-type* in the field CP is a phrase separated only with comas or conjunctions *y* (the analogous definitions exist for *o-type* CP and *ni-ni-type* CP).

N	CP	*s*	*p*	NMB	Example	Optional
1.				PL	24	
2.		Condition S1		SG	27	

N	CP	*s*	*p*	NMB	Exam-ple	Op-tional
3.	*y*-type CP	Condition S2		SG	30	
4.			Condition P1	SG	26, 25	
5.		Condition S3		PL	35, 36, 22, 23	
6.	*o*-type CP or *ni-ni*-type CP			SG	31, 32	✓
7.			Condition P2	SG	29	✓

where:

Condition S1:
- CP has only 2 parts, and
- these two parts are conversives, and
- the first part has a definite article or possessive pronoun (thus marking the whole phrase).

Condition S2:
- all parts of the CP are referentially identical (denote the same object), and
- all parts of the CP are in SG.

Condition S3:
- all parts of the phrase are infinitives, and
- coordinated phrase has only 2 parts, and
- they have opposite meaning,

or
- all parts of the phrase are infinitives, and
- each member of the phrase has a definite article,

or
- at least one part of the phrase is personal pronoun.

Condition P1:
- *p* has nominal part, which is an abstract noun or an article "lo"

Condition P2:
- *p* precedes the CP.

Examples for NMB(*p*):

14. *Una multitud entró al estadio.*
15. *Escasísima cantidad de obras tiene vigencia permanente.*
16. *La tripulación del barco son marineros jóvenes.*

17. *Mas que cinco mujeres están ...*
18. *Mas que una mujer está ...*
19. *Se requiere desde un minuto hasta unas horas para ...*
20. *Se requieren desde unos minutos hasta una hora para ...*
21. *Desde unos minutos hasta una hora se requiere para...*
22. *María o yo iremos al concierto.*
23. *Maria y él fueron al concierto.*
24. *La preparación y la vocación contribuyeron...*
25. *La libertad, la igualdad y la fraternidad fue el lema de la Gran Revolución Francesa.*
26. *Amor, amistad, aire sano para la respiración moral, luz para el alma, simpatía, fácil comercio de ideas y de sensaciones era lo que...*
27. *La compra y venta está prohibida.*
28. *La compra y la venta están prohibidas.*
29. *Reinaba / Reinaban constantemente tal desorden y algarabía...*
30. *Su esposa y amiga lo acompañó en todos sus viajes.*
31. *Lo que el héroe o la heroína pensaría / pensarían...*
32. *Lo que ni el héroe ni la heroína pensaría / pensarían ...*
33. *[María o Julia] y Pedro van al concierto.*
34. *[Juan y Pedro] o María va / van al concierto.*
35. *Trabajar mucho y descansar poco perjudican la salud.*
36. *El fumar y el beber son perjudiciales.*
37. *Fumar y beber es perjudicial.*
38. *Me gusta cantar y bailar.*

4 CONCLUSIONS

The rules applicable during text analysis for detection of the subjectival relation using the subject-predicate agreement criterion have been discussed. It has been postulated that the subject is marked with values of the categories that participate in the agreement (PRS, NMB, GND). The rules for the calculation of these values for the CPs have been introduced. Different possible forms of the subject in Spanish have been described. These rules can supposedly be translated into formal rules for synthesis and analysis of Spanish texts, either in the Meaning ⇔ Text framework or in other formalisms, such as, e.g., HPSG [Sag & Wasow 1999].

REFERENCES

Mel'čuk, I. A. *Dependency Syntax: Theory and Practice.* SUNY Publ., New York, 1988.

Mel'čuk, I. *Dependency in linguistic description*. Benjamins Publ., Amsterdam, 1999 (to appear).

Mel'čuk, I. and N. Pertsov. Surface syntax of English. A Formal Model within Meaning – Text Framework. Benjamins Publ., Amsterdam, 1987.

Apresian, Yu. D. Lingistic base of the system of French-Russian Automatic Translation "ETAP-1". IV. French syntactical analysis (in Russian). Institute of Russian language. Preprint No. 159, 1984.

Seco, Manuel R. *Gramática española*. Madrid, 1968.

Esbozo de una nueva gramática de la lengua española. Real Academia Española, Madrid, 1998.

Álvarez Constantino, J. *Gramática funcional del español.* México D.F, 1977.

Sag, I., and Thomas Wasow. *Syntactic Theory: A formal Introduction*. Stanford: CSLI Publications, 1999.

Igor A. Bolshakov is professor and researcher at the Laboratory of Natural Language and Text Processing of the Center for Computing Research (CIC), National Polytechnic Institute (IPN), Av. Juan de Dios Bátiz s/n, esq. Mendizabal, Zacatenco, 07738, Mexico D.F., Mexico. He has more then 60 publications in the field of computational linguistics. He can be reached at igor@cic.ipn.mx.

Alexander F. Gelbukh is the head of the same laboratory. He is author of about 90 publications in the field of computational linguistics, including computational morphology, syntax, semantics. He can be reached at gelbukh@cic.ipn.mx or gelbukh@earthling.net, see also http://www.cic.ipn.mx/~gelbukh.

Grigori Sidorov is professor and researcher at the same laboratory, author of more than 30 publications in the same field. He can be reached at sidorov@cic.ipn.mx, see also http://www.cic.ipn.mx/~sidorov.

The work was done under partial support of CONACyT, REDII, and SNI, Mexico. We thank our native speaker informants who helped us to check the Spanish examples: Álvaro de Albornoz Bueno, Ana Celia Campos Hernández, and especially Sofía Natalia Galicia Haro.