Looking Behind the Scenes of Syntactic Dependency Corpus Annotation: Towards a Motivated Annotation Schema of Surface-Syntax in Spanish

Alicia Burga[1], **Simon Mille**[1], **and Leo Wanner**[1,2]
[1]Universitat Pompeu Fabra [2]ICREA, Barcelona
firstname.lastname@upf.edu

Abstract

Over the last decade, the prominence of statistical NLP applications that use syntactic rather than only word-based shallow clues increased very significantly. This prominence triggered the creation of large scale treebanks, i.e., corpora annotated with syntactic structures. However, a look at the annotation schemata used across these treebanks raises some issues. Thus, it is often unclear how the set of syntactic relation labels has been obtained and how it can be organized so as to allow for different levels of granularity in the annotation. Furthermore, it appears questionable that despite the linguistic insight that syntax is very much language-specific, multilingual treebanks often draw upon the same schemata, with little consideration of the syntactic idiosyncrasies of the languages involved. Our objective is to detail the procedure for establishing an annotation schema for surface-syntactic annotation of Spanish verbal relations and present a restricted set of easy-to-use criteria which facilitate the decision process of the annotators, but which can also accommodate for the elaboration of a more or a less fine-grained tagset. The procedure has been tested on a Spanish 3,500 sentence corpus, a fragment of the AnCora newspaper corpus.

1 Introduction

Over the last decade, the prominence of statistical Natural Language Processing (NLP) applications (among others, machine translation parsing, and text generation) that use syntactic rather than only word-based shallow clues increased very significantly. This prominence triggered, in its turn, the creation of large scale treebanks, i.e., corpora annotated with syntactic structures, needed for training of statistical algorithms; see, among others, the Penn Treebank (Marcus et al., 1993) for English, the Prague Dependency Treebank (Hajič et al., 2006) for Czech, the Swedish Talbanken05 (Nivre et al., 2006), the Tiger corpus (Thielen et al., 1999) for German, and the Spanish, Catalan, and Basque AnCora treebank (Taulé et al., 2008). Even though this is certainly a very positive tendency, a look at the annotation schemata used across the treebanks of different languages raises some issues. Thus, despite the linguistic insight that syntax is very much language-specific, many of them draw upon the same more or less fine-grained annotation schemata, i.e., sets of syntactic (dependency) relations, with little consideration of the languages themselves. Often, it is unclear how the individual relations in these sets have been determined and in which linguistic theory they are grounded, and occasionally it is not obvious that the annotation schema in question uses only syntactic (rather than also semantic) criteria.

Our objective is to detail the process of elaboration of an annotation schema for surface-syntactic verbal relation annotation of Spanish corpora,[1] which has already been used to annotate a 3,500 sentence corpus of Spanish. The corpus is a fragment of the AnCora corpus which consists of newspaper material.

Our work draws heavily on the principles of the Meaning-Text Theory (MTT) as far as the nature of dependency in general and (surface-) syntactic dependency in particular are concerned.

In the next section, we analyze the state of affairs in some of the well-known dependency treebanks and justify why we set out to write this paper. In Section 3, we present the notion of surface-syntactic structure and the general principles of dependency as defined in MTT. Section 4 outlines the annotation schema we propose and the principles used to distinguish between different relations. Section 5, finally, summarizes the paper and draws some conclusions.

[1]"Surface-syntactic" is used here in the sense of the Meaning-Text Theory (Mel'čuk, 1988).

2 A Glance Behind the Scenes

It is well-known that surface-syntactic relations (SSyntRels) as usually used in dependency treebanks are language-specific. A dependency relation annotation schema should thus, on the one hand, facilitate the annotation of all language-specific syntactic idiosyncrasies, but, on the other hand, offer a motivated generalization of the relation tags such that it could also serve for applications that prefer small generic dependency tag sets. However, as already mentioned above, in a number of dependency treebanks containing corpora in different languages, the same arc tag set is used for all languages involved—no matter whether the languages in question are related or not. For instance, AnCora (Taulé et al., 2008) contains the related Spanish and Catalan, but also Basque; the treebank described in (Megyesi et al., 2008) contains Swedish and Turkish, etc. This makes us think that little work has been done concerning the definition of the relation labels. In general, for all parallel and non-parallel treebanks that we found—the Czech PDT2.0-PDAT (Hajič et al., 2006) and (Hajič and Zemánek, 2004)) and PCET (Čmejrek et al., 2004), the English-German FuSe (Cyrus et al., 2003), the English-Swedish LinEs (Ahrenberg, 2007), the English Penn Treebank (Marcus et al., 1993), the Swedish Talbanken (Nivre et al., 2006), the Portuguese Bosque (Afonso et al., 2002), the Dutch Alpino (Van der Beek et al., 2002), etc.—the justification of the choice of dependency relation labels is far from being central and is largely avoided. This may lead to the conclusions that the selection of the relations is not of great importance or that linguistic research already provides sets of relations for a significant number of languages. Each of these two conclusions is far from being correct. In our work, we found the question of the determination of SSyntRels very crucial, and we observed the lack of an appropriate description of the language through a justified description of the SSyntRels used even for languages for which treebanks are available and widely used.

In MTT, significant work has been carried out on SSyntRels—particularly for English and French. Thus, (Mel'čuk and Percov, 1987; Mel'čuk, 2003) present a detailed inventory of SSyntRels for English, and (Iordanskaja and Mel'čuk, 2009) suggest criteria for establishing an inventory of labeled SSyntRels headed by verbs as well as a preliminary inventory of relations for French. However, we believe that both inventories are not thought for large scale corpus annotation to be used in statistical NLP in that the criteria are generally difficult to apply and do not separate enough surface-syntactic phenomena from the phenomena at other levels of the linguistic description. For instance, one important distinction in (Iordanskaja and Mel'čuk, 2009) is whether a dependent is actantial or not—in other words, if a dependent forms part of the definition of its governor or not—, which is, however, a clear semantic distinction.

We attempt to avoid recourse to deep criteria. Instead, we replace deep criteria by a list of strictly syntactically motivated, easy-to-use criteria in order to make their application efficient on a large scale, and detail the process from the very beginning. This list is as reduced as possible, but still sufficient to capture fine-grained idiosyncrasies of Spanish. Obviously, we intensely use the cited works on SSyntRels in MTT as a source of inspiration.

3 MTT Guide to SSynt Dependencies

The prerequisite for the discussion of the compilation of a set of SSyntRels for a particular language is a common understanding of (i) the notion of a surface-syntactic dependency structure (SSyntS) that forms the annotation of a sentence in the corpus; (ii) the principles underlying the determination of a dependency relation, i.e., when there is a dependency relation between two lexical units in a sentence, and what is the direction of this dependency, or in other words, who is the governor and who is the dependent. In the presentation of both, we widely follow (Mel'čuk, 1988).

3.1 Definition of SSyntS

In MTT, an SSyntS is defined as follows:

Definition 1 (Surface-Syntactic Structure, $SSyntS$) *Let L, G_{sem} and R_{ssynt} be three disjunct alphabets, where L is the set of lexical units (LUs) of a language $\mathcal{L}$, G_{sem} is the set of semantic grammemes, and R_{ssynt} is the set of names of surface-syntactic relations (or grammatical functions).*

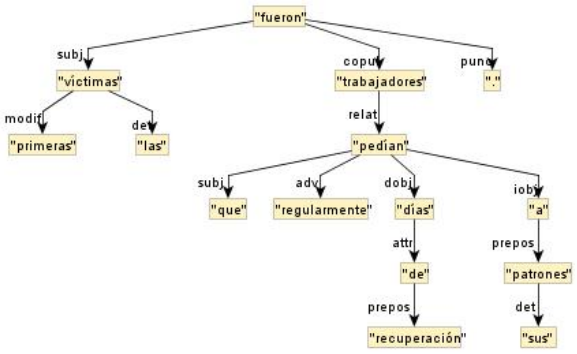

Figure 1: SSyntS of the sentence *Las primeras víctimas fueron trabajadores que pedían regularmente días de recuperación a sus patrones.* 'The first victims were employees who regularly asked days-off to their bosses'

An SSyntS of $\mathcal{L}$, S_{SSynt}, is a quintuple over $L \cup G_{sem} \cup R_{ssynt}$ of the following form:

$$S_{SSynt} = \langle N, A, \lambda_{l_s \to n}, \rho_{r_s \to a}, \gamma_{n \to g} \rangle$$

where

- *the set N of nodes and the set A of directed arcs (or* branches*) form an unordered dependency tree (with a source node n^s and a target node n^t defined for each arc),*
- *$\lambda_{l_s \to n}$ is a function that assigns to each $n \in N$ an $l_s \in L$,*
- *$\rho_{r_s \to a}$ is a function that assigns to each $a \in A$ an $r_s \in R_{ssynt}$,*
- *$\gamma_{n \to g}$ is a function that assigns to the name of each LU associated with a node $n_i \in N$, $l_i \in \lambda_{n \to g}(N)$, a set of corresponding grammemes $G_t \in G_{sem}$.*

For illustration, consider the SSyntS of a Spanish sentence in Figure 1.[2]

We are particularly interested in the assignment of surface-syntactic relation labels to the arcs (i.e., the function $\rho_{r_s \to a}$). These labels are of the nature as used by many other treebanks: 'subject', 'direct/indirect object', 'copulative', 'modificative', 'determinative', 'adverbial', etc, i.e., *grammatical functions*. We want to determine when to use each of them and how to build the tag set such that it can be enriched or reduced in a prescribed way under clearly defined conditions. For instance, in Figure 1, the indirect object of the verb *pedían* 'asked$_{PL}$' is introduced by a preposition *a* 'to'. However, in Spanish, direct objects can also be introduced by it. So, obviously, looking at the units of the sentence is not enough to establish the dependency relations.

Each relation has to be associated with a set of central properties. These properties must be clearly verifiable. For instance, a direct object is cliticizable by an accusative pronoun, an indirect object by a dative pronoun, and every relation must have one type of dependent that can be used with any governor.

3.2 Principles for Determination of SSynt-Dependencies

The central question faced during the establishment of the SSyntS as defined above for each sentence of the corpus under annotation is related to:

- the elements of A: when is there a dependency between two nodes labeled by the LUs l_i and l_j and what is the direction of this dependency,
- the elements of R_{ssynt}: what are the names of the dependencies, how they are to be assigned to $a \in A$, and how they are to be distinguished,

or, in short, to the determination of SSynt-Dependencies. In what follows, we address this question in terms of two corollaries.

Corollary 1 (Dependency between nodes) *Given any two unordered nodes n_1 and n_2, labeled by the LUs l_1 and l_2 respectively, in the sentence S of the corpus, there is a dependency between n_1 and n_2 if either*

(a) in order to position l_i in S, reference must be made to l_j, with $i, j = 1, 2$ and $i \neq j$ (linear correlation criterion)

and

(b) between l_i and l_j or between syntagms of which l_i and l_j are heads ($i, j = 1, 2$ and $i \neq j$), a prosodic link exists (prosodic correlation criterion)

or

(c) l_i triggers agreement on l_j ($i, j = 1, 2$ and $i \neq j$) (agreement criterion)

[2] The nominal node labels reflect the number (*víctimas* 'victims', *trabajadores* 'workers', *patrones* 'bosses') only to facilitate the reading; semantic plural is encoded as a grammeme in terms of an attribute/value pair on the node: *number=PL*. Note also that we consider each node label to be a disambiguated word, i.e., lexical unit (LU). For details on grammemes, see (Mel'čuk, 2006).

Thus, in *John has slept well today*, *John* has to be positioned before the auxiliary *has* (or after in a question) and a prosodic link exists between *John* and the syntagm headed by *has*. This means that *John* and *has* are likely to be linked by a dependency relation. *Well* has to be positioned compared to *slept* (not compared to *has*), hence there is a dependency between *slept* and *well*.

With respect to agreement, we see that the verb is *has* and not *have*, as it would be if we had *The boys* instead of *John*. This verbal variation in person, which depends on the preverbal element, implies that a dependency links *John* and *has*.

Once the dependency between two nodes has been established, one must define which node is the governor and which one is the dependent, i.e., the direction of the SSynt arc that links those two nodes. The following corollary handles the determination of the direction of the dependency:

Corollary 2 (Direction of a dependency relation) *Given a dependency arc a between the nodes n_1 and n_2 of the SSyntS of the sentence S in the corpus, n_1 is the governor of n_2, i.e., n_1 is the source node and n_2 is the target node of a if*

(a) the passive valency (i.e., distribution) of the group formed by the LU labels l_1 and l_2 of n_1/n_2 and the arc between n_1 and n_2 is the same as the passive valency of l_1 (passive valency criterion)

or

(b) l_1 as lexical label of n_1 can be involved in a grammatical agreement with an external element, i.e., a label of a node outside the group formed by LU labels l_1 and l_2 of n_1/n_2 and the arc between n_1 and n_2 (morphological contact point criterion)

If neither (a) nor (b) apply, the following weak criteria should be taken into account:

(c) if upon the removal of n_1, the meaning of S is reduced AND restructured, n_1 is more likely to be the governor than n_2 (removal criterion),

(d) if n_1 is not omissible in S, it is more likely to be the governor than n_2 (omissibility criterion),

(e) if l_2 as label of n_2 needs ("predicts") l_1 as label of n_1, n_2 is likely to be a dependent of n_1 (predictability criterion).

As illustration of the passive valency criterion,[3] consider the group *the cats*. It has the same distribution as *cats*: both can be used in exactly the same paradigm in a sentence. On the other side, *the cats* does not have the distribution of *the*. We conclude that *cats* is the head in the group *the cats*. It is important to note that, for instance, in the case of prepositional groups, the preposition does not have its own passive valency since it always needs an element directly after it. It does not prevent the passive valency criterion from applying since, e.g., the distribution of *from* [*the*] *house* is not the same as the distribution of *house*. It is the presence of the preposition that imposes on the group a particular distribution.

The morphological contact point criterion is used as follows: considering the pair *sólo felinos* in *sólo felinos ronronean* 'only felines$_{PL}$ purr$_{PL}$', *felinos* is the unit which is involved in the agreement with an external element, *ronronean*. As a consequence, *felinos* is more prone to be the governor of *sólo*.

We illustrate the omissibility criterion in Section 4.2, but do not elaborate on the removal criterion nor on the predictability criterion; for more details see (Mel'čuk, 1988).

3.3 Labelling the dependencies

With the two corollaries from above at hand, we should be able to state when there is a dependency arc between two nodes, and which node governs which other node. Now, labels to the dependency arcs need to be assigned. The assignment may be very intuitive and straightforward (as, e.g., the assignment of *subject* to the arc between *caen* 'fall' and *bolas* 'balls' in *bolas caen*, lit. 'balls fall', or the assignment of *object* to the arc between Sp. *tiran* 'throw' and *bolas* 'balls' in *tiran bolas*, lit. '[they] throw balls') or less clear (as, e.g., the assignment of a label to the dependency arc between *caen* 'fall' and *bolas* 'balls' in *caen bolas*, lit. '[it] falls balls': is it the same as in *bolas caen*, namely *subject* or a different one?).[4]

[3] For the definition of the notion "passive valency", see (Mel'čuk, 1988).

[4] We do not encode linear order in the SSyntRels: in practice, this allows us to limit the tagset size. However, it does not mean that some relations do not impose a particular linear order between the governor and dependent (see Section 4.2). The dependency tree as such remains unordered.

The following corollary addresses the question whether two given dependency arcs are to be assigned the same or different labels:

Corollary 3 (Different labels) *Be given an arc* a_1 *and an arc* a_2 *such that*
• a_1 *holds between the nodes* $n_{s_{a1}}$ *(labeled by* $l_{s_{a1}}$*) and* $n_{st_{a1}}$ *(labeled by* $l_{t_{a1}}$*), with the property set* $P_{a_1} := \{p_{a1_1}, p_{a1_2}, \ldots, p_{a1_i}, \ldots, p_{a1_n}\}$,
• a_2 *holds between the nodes* $n_{s_{a2}}$ *(labeled by* $l_{s_{a2}}$*) and* $n_{st_{a2}}$ *(labeled by* $l_{t_{a2}}$*), with the property set* $P_{a_2} := \{p_{a2_1}, p_{a2_2}, \ldots, p_{a2_j}, \ldots, p_{a1_m}\}$

Then, $\rho_{r_s \to a}(a_1) \neq \rho_{r_s \to a}(a_2)$*, i.e.,* a_1 *and* a_2 *are assigned different labels, if*

(a) $\exists p_k : (p_k \in P_{a_1} \wedge p_k \notin P_{a_2}) \vee (p_k \in P_{a_2} \wedge p_k \notin P_{a_1})$ *and* p_k *is a central property*

or

(b) one of the following three conditions apply; cf. (Mel'čuk, 1988):

1. semantic contrast condition: $l_{s_{a1}}$ *and* $l_{s_{a2}}$ *and* $l_{t_{a1}}$ *and* $l_{t_{a2}}$ *are pairwise the same word-forms, but either* $l_{s_{a1}}$ *and* $l_{s_{a2}}$ *or* $l_{t_{a1}}$ *and* $l_{t_{a2}}$ *have different meanings.*

2. prototypical dependent condition (quasi-Kunze property)*: given the prototypical dependents* d_{p_1} *of* a_1 *and* d_{p_2} *of* a_2*, when* $l_{t_{a1}}$ *in* $l_{s_a} - a_1 \to l_{t_{a1}}$ *is substituted by* d_{p_2} *the grammaticality of* $l_{s_{a1}} - a_1 \to l_{t_{a1}}$ *is affected or when* $l_{t_{a2}}$ *in* $l_{s_{a2}} - a_2 \to l_{t_{a2}}$ *is substituted by* d_{p_1} *the grammaticality of* $l_{s_{a2}} - a_2 \to l_{t_{a2}}$ *is affected.*

3. SSyntRel repeatability criterion: *If* $l_{t_{a1}}$ *and its dependency* a_1 *from* $l_{s_{a1}}$ *can be repeated and* $l_{t_{a2}}$ *and its dependency* a_2 *from* $l_{s_{a2}}$ *cannot (or vice versa).*

Condition (a) entails first of all that a relation should have clear properties associated to it. Associating properties to a relation is exactly what means to define a relation. This can only be done in opposition to other relations, which means that this is the result of numerous iterations after the inspection of numerous examples. As a consequence, paradoxically, the list of properties of a relation is one of the last things which is defined.[5]

[5] A restricted property set of the *direct objectival* relation in Spanish includes: the direct object (1) is cliticizable (2) by an accusative pronoun, (3) can be promoted, (4) does not receive any agreement, and (5) is typically a noun.

The semantic contrast condition (b1) states that for a given relation and a given minimal pair of LUs, there must not be any semantic contrast; the arc orientation has to be the same for both members of the minimal pair, and the deep-morphologic representation should be different (different possible orders or different case on the dependent for instance). Both pairs have the property to be able to occupy the same syntactic role in a sentence. Consider the two LUs *matar* 'kill' and *gatos* 'cats': they can form an ambiguous sentence *Matan gatos*, lit. 'Cats kill'/'[They] kill cats'. The ambiguity cannot be explained by the difference of meaning of the components of the sentence (since they are the same). Hence, the semantic contrast criterion prevents both dependencies to be the same; in one case, *gatos* is subject, and in the other case, it is object of *matar*.

The semantic contrast condition does not apply to *una casa* 'indefinite + house' / *una casa* 'one + house' because *una* does not have the same meaning (i.e., is not the same lexeme) in both cases.

The quasi-Kunze criterion (b2) states that any SSyntRel must have a prototypical dependent, that is, a dependent which can be used for ANY governor of this SSyntRel; see (Mel'čuk, 2003). Consider, for illustration, $poder{-}\text{R}{\to}caer$ 'can fall' vs. $cortar{-}\text{R}{\to}pelo$ 'cut hair': it is not possible to have an N as dependent of *poder* 'can' nor an V_{inf} as dependent of *cortar* 'cut'. More generally, no element of the same category can appear below both *poder* and *cortar*. This implies that the prototypical dependents in both cases do not coincide, so it is not the same relation.

The SSyntRel repeatability criterion (b3) indicates that a particular SSyntRel should be, for any dependent, either always repeatable or never repeatable. If one dependent can be repeated and another one cannot, then we have two different relations. In a concrete case, we can start with the hypothesis that we have ONE relation R for which we want to know if it is suitable to handle two dependents with different properties (in particular, two different parts-of-speech). If the same relation R can be used to represent the relation, for instance, between a noun and an adjective, and, on the other side, between a noun and a numeral quantifier, R should be either repeatable or not repeatable in both cases. We observe

that R is repeatable for adjectives but not for quantifiers and conclude, thus, that R should be split in two relations (namely 'modifier' and 'quantificative').

4 Towards a SSynt Annotation Schema

In Section 3, the general principles have been presented that allow us to decide when two units are involved in a dependency relation and who is the governor. Furthermore, some generic cases have been identified in which it seems clear whether a new relation should be created or not. With these principles at hand, we can set out for the definition of a motivated SSynt annotation schema. To be taken into account during this definition is that (a) (unlike the available MTT SSyntRel sets,) the schema should cover only syntactic criteria; (b) the granularity of the schema should be balanced in the sense that it should be fine-grained enough to capture language-specific syntactic idiosyncrasies, but be still manageable by the annotator team (we are thinking here of decision making and inter-agreement rate). The latter led us target a set of 50 to 100 SSyntRels.

4.1 Principles for the criteria to distinguish between different relations

The following properties are particularly important:
• **Applicability:** The criteria should be applicable to the largest number of cases possible. For instance, a head and a dependent always have to be ordered, so a criterion implying order can be applied to every relation whatever it is. One advantage here is to keep a set of criteria of reasonable size, in order to avoid to have to manage a large number of criteria which could only be applied in very specific configurations. The other advantage in favouring generic criteria is that it makes the classification of dependency relations more readable: if a relation is opposed to another using the same set of criteria, the difference between them is clearer.
• **Visibility:** When applying a criterion, an annotator would rather *see* a modification or the presence of a particular feature. Indeed, we try to use only two types of criteria: the ones that transform a part of the sentence to annotate—promotion, mobility of an element, cliticization, etc.—, and the ones that check the presence or absence of an element in the sentence to annotate (is there an agrement on the dependent? does the governor impose a particular preposition? etc.). In other words, we avoid semantically motivated criteria. The main consequence of this is the absence of opposition complement/attribute as discriminating feature between syntactic relations.
• **Simplicity:** Once the annotator has applied a criterion, he/she must be able to make a decision quickly. This is why almost all criteria involve a binary choice.

All of the resulting selected criteria presented below have been used in one sense or the other in the long history of grammar design. However, what we believe has not been tackled up to date is how to conciliate in a simple way fine-grained syntactic description and large-scale application for NLP purposes. In what follows, we present a selection of the most important criteria we use in order to assign a label to a dependency relation. Then, we show how we use them for the annotation of a Spanish corpus with different levels of detail.

4.2 Main criteria to distinguish between different relations

• **Type of linearization:** Some relations are characterized by a rigid order between the head and the dependent (in any direction), whereas some others allow more flexibility with respect to their positioning. Thus, e.g., the relations that connect an auxiliary with the verb imply a fixed linearization: the auxiliary (head) always appears to the left of the verb (dependent):
He comido mucho.'[I] have eaten a-lot' /
**Comido he mucho*.

On the other hand, even if Spanish is frequently characterized as an SVO language, the relation 'subject' does allow flexibility between the head and the dependent:
Juan come manzanas. 'Juan eats apples'/
Come Juan manzanas./Come manzanas Juan.

Given that it is possible to apply this criterion to all the relations, the linearization criterion is very relevant to our purposes.
• **Canonical order:** As just stated, some relations are more flexible than others with respect to the order between head and dependent. When the order is not restricted, there is usually a canonical order. Thus, although it is possible to have a postverbal subject, the canonical order between the subject and

the verb is that the former occurs at the left of the latter. On the other hand, the relations introducing the non-clitic objects have the opposite canonical order, i.e. the object appears at the right of the verb.

• **Adjacency to the governor:** There are some relations that require that the head and the dependent are adjacent in the sentence, and only accept a very restricted set of elements to be inserted between them, but there are some other relations that allow basically any element to appear between them. We believe that the fact to keep a dependent very close in the sentence is an important syntactic feature. All the relations involving clitics belong to the first type, and a relation such as determinative belongs to the second type:

Cada día, lo miraba. 'Every day, [I] watched it'/
**Lo cada día miraba.*
El hombre bueno. lit. 'The guy good' /
El buen hombre.

• **Cliticization:** Concerning only elements for which the order between the verbal head and its dependent is not restricted, an important criterion refers to the possibility for the dependent to be replaced or duplicated by clitic pronouns. Thus, the relation *indirect object* allows cliticization, as opposed to the *oblique object* that does not:

$Miente-$IObj$\rightarrow a$ *Carla. / Le miente. / A Carla le miente.*
lit. '[He] lies to Carla.' / '[He] to-her lies.' / 'To Carla [He] to-her lies .'

$Invierte-$OblObj$\rightarrow en$ *bolsa. / *La invierte. / *En bolsa la invierte.*
lit. '[He] inverts in stock-market.' / '[He] in-it inverts.' / 'In stock-market [He] in-it inverts .'

• **Promotion/demotion:** Promotion and demotion refer to the possibility of becoming, respectively, a closer or a further argument in a parallel sentence. Thus, the dependent of the relation *direct object* can be promoted to the dependent of the relation *subject* in a passive sentence (and, from the opposite point of view, the subject can be demoted to the dependent of the relation *agent* in a passive sentence):

Juan compuso las canciones./ Las canciones fueron compuestas por Juan.
'Juan wrote the songs' / 'The songs were written by Juan'

Cliticization and promotion/demotion can only be applied if the head is a finite verb and from this perspective, do not seem comply with the Applicability principle. However, since there are many different relations that can appear below a verb, this is not totally true. In addition, they are very efficient with respect to the other two principles, Visibility and Simplicity.

• **Agreement:** Agreement appears when head and dependent share some morphological features, such as gender, number, person, etc., which one passes to the other. The agreement actually depends on two parameters: on the one hand, the target of the agreement must have a Part of Speech which allows agreement, and on the other hand, the dependency relation itself must allow it. For example, the *copulative* relation allows agreement, but if the dependent is not an adjective, it is not mandatory: *Pedro y Carla son relajados* 'Pedro and Carla are relaxed$_{\mathrm{PLU}}$' as opposed to *Pedro y Carla son una pareja*, 'Pedro and Carla are a couple$_{\mathrm{SING}}$'. Inversely, the past participle in the perfect analytical construction is intrinsically prone to agreement (as shows the second example that follows), but the relation does not allow it: *Carla está perdida*, 'Carla is lost$_{\mathrm{FEM}}$' as opposed to *Carla ha perdido* 'Carla has lost$_{\mathrm{noFEM}}$'. This is why the notion of prototypical dependent is important (see next paragraph): if a relation licences agreement, it doesn't mean that any dependent must have agreement, but that there is always agreement for its prototypical dependent.

There are different types of agreements allowed by a syntactic relation:

- dependent agrees with head: $sillas-$modificative$\rightarrow rotas$ 'chairs broken$_{\mathrm{FEM.PL}}$',
- head agrees with dependent: $Juan\leftarrow$subject$-viene$ 'Juan comes',
- dependent agrees with another dependent: $Juan\leftarrow$subject$-parece-$copulative$\rightarrow$ *enfermo* 'Juan seems sick$_{\mathrm{MASC.SG}}$'.

When there is agreement, secondary criteria can be applied, concerning the type of inflection of the agreeing element: in some cases, the agreement can vary, in some cases it cannot (see the opposition between *subject* and *quotative subject* in the next subsection).

• **Prototypical dependent:** As mentioned in Section 3, every relation must have a prototypical dependent. This criterion is more useful for designing

the set of dependency relations than for assigning a tag to a relation, since it involves a generalization over a large number of cases, which are not accessible during the process of annotation. However, it can be used during annotation as well, especially in order to infirm/confirm a relation: if a dependent of a SSyntRel cannot be replaced by the prototypical dependent of this relation, then the relation should be changed. It can also be useful when looking for a relation in the hierarchical representation of the criteria (see Figure 2), for instance in combination with the Agreement criterion: if the pair $son-??\rightarrow pareja$ in the sentence *Pedro y Carla son una pareja* 'Pedro and Carla are a couple$_{\mathrm{SING}}$' has to be annotated, although there is no visible agreement, the native speaker annotator has the knowledge that the typical dependent in that case for that verb is an adjective and then should consider that an agreement is usually involved.

• **Part-Of-Speech of the Head:** The actual PoS of the governor is relevant in that there are very few syntactic dependents that behave the same with heads of differerent syntactic categories once a certain level of detail has been reached in the annotation. As a consequence, we decided to separate the tags of our tagset by PoS of the governor.

• **Governed Preposition/ Conjunction/ Grammeme (P/C/G):** There are some relations that require the presence of a preposition, a subordinating conjunction or a grammeme. For instance, the relation *oblique object* implies the presence of a preposition which has no meaning to introduce the dependent (*viene* **de** *comer* 'he/she has just eaten'), and the relation *subordinate conjunctive* requires the presence of a feature in the verb indicating that it is finite.

• **Dependent omissibility:** This syntactic criterion is defined within an "out-of-the-blue" context, given that otherwise it is very difficult to determine whether or not a dependent is omissible: it is always possible to create pragmatic contexts whereas the dependent can be perfectly omitted. There are two cases: on the one hand, relations such as *prepositional* always require the presence of the dependent and, on the other hand, relations as *modifier* do not require the presence of the dependent. Consider:
Juan viene $para-$prepos$\rightarrow trabajar$. / 'Juan comes to work' / 'Juan comes to'
**Juan viene para.*
Tiene $sillas-$modif$\rightarrow verdes$. / *Tiene sillas.*
lit. '[He] has chairs green' / '[He] has chairs'

4.3 Application of the Schema to Spanish

We organized all the criteria into a tree-like hierarchy so that if an annotator identifies a pair a governor/dependent, but wonders which relation holds between the two, he only has to follow a path of properties that leads to the relation. The order in which the criteria are applied is only important for a generalization over the relations, since it allows to keep close in the graphical representation the relations that have the same type (see Figure 2).

Due to space restrictions, we only present in this paper a part of the hierarchy, namely, the relations headed by a verb which do not impose a rigid order between governor and dependent; our complete hierarchy contains 70 different arc labels and covers the annotation of a 100,000 word corpus. We use here nine criteria: removability of dependent, possible cliticization, agreement type, inflection type, PoS of prototypical dependent, behaviour to promotion, presence of governed P/C/G, presence of quotes, presence of parentheses or dashes. With this level of detail, we get sixteen different relations; c.f Figure 2.

In the following, we give an example for each relation; the governor of the relation appears in bold uppercase, the dependent in bold lowercase:
-*adjunctive*: **Vale, VAMOS**! lit.'Ok, let's-go!'
-*adverbial*: **Hoy PASEO** lit.'Today I-go-for-a-stroll'
-*copulative*: El gato **ES negro** 'The cat is black'
-*direct objectival*: **CONSTRUYEN una casa** lit.'They-build a house'
-*indirect objectival*: Les **MOLESTA** el ruido **a los peces**, lit. '(to-them) bothers the noise (to) the fish', 'The fish are bothered by the noise'
-*modificative adverbial*: **Llegados** a ese extremo, el trabajo se **VUELVE** insoportable lit.'Arrived-MASC-PL to that extremity, the work becomes unbearable'
-*object completive*: Pedro **CONSIDERA tontos** a los gatos lit.'Pedro considers stupid to the cats '
-*object copredicative*: Pedro **VE felices** a los gatos lit. lit.'Pedro sees happy to the cats', 'Pedro sees the cats happy'

-*oblique objectival*: **PASA de Pedro** lit.'He-ignores from Pedro'
-*quasi subjectival*: **LLUEVE(N) ranas**, lit. 'it/they-rain(s) frogs'
-*quotative copulative*: La pregunta **ERA** '**Va** a volver?' lit.'The question was/ 'Is-he-going to come-back?"
-*quotative direct objectival*: 'Dogs' **SIGNIFICA "perros"** (' "Dogs" means "perros"
-*quotative subjectival*: **"Dogs" SIGNIFICA** "perros" ' "Dogs" means "perros" '
-*subjectival*: **Pedro CORRE** 'Pedro runs'
-*subject completive*: La frase **RESULTA buena** lit.'The sentence turns-out fine'
-*subject copredicative*: Pedro **VUELVE feliz** lit.'Pedro comes-back happy'

By selecting only a few criteria, it is possible to diminish the number of relations and thus, by doing so, to tune the level of detail of the annotation. For example, keeping only four of the nine criteria presented above, we end up with only five relations, instead of sixteen:

1. Cliticization: *objectival (type 1)*
2. No Cliticization
   2.1 Dep not Removable: *completive*
   2.2 Removable Dep.
      2.2.1 Prototypical Dep.=N
         2.2.1.1 Dep. controls Agreement *subjectival*
         2.2.1.2 No Agreement *objectival (type 2)*
      2.2.2 Prototypical Dep.=A/Adv *adverbial*

Figure 2 summarizes the use of some criteria for Spanish and shows the correspondence between the fine-grained relations and generalized relations (rightmost side of the figure). On the left side, each intermediate node corresponds to the application of one criteria, and the leaves are the SSyntRels. The path from the root of the tree to one leaf thus indicates a list of properties of this relation. Within the brackets, some properties are listed which are entailed by the criterion they appear next to. For example, the *Canonical Order* (CO Right/Left) can always be predicted by a particular property: for instance, all elements that can be cliticized are usually linearized on the right of their governor. If *Canonical Order* is not mentioned for a relation, it is because there is no canonical order, as it is the case for three adverbial relations (*modificative adverbial*, *adjunct*, and *adverbial*). Obviously, every relation usually has many more properties than those listed in this hierarchy.

Although we use only syntax-based criteria, it is possible to reach the semantic level by indicating whether the dependent of a relation is acounted for in the valency of its governor (no (-), actant I, actant II, etc.), which is indicated by the numbers in the column to the right of SSYNTRELS.[6] This helps for generalizing the relations, as illustrated on the right side of the figure. This second hierarchy, over relations, is similar to those proposed by, among others, (De Marneffe et al, 2006) or (Mille and Wanner, 2010).

5 Conclusions

Even if there are dependency corpora in different languages and some of them widely used for NLP applications, it is not yet clear how the set of syntactic relations can be obtained and how it can be organized so as to allow for different levels of granularity in the annotation. In this paper, we attempt to fill this gap by detailing the procedure for establishing a tagset for Spanish verbal relations. We present a restricted selection of easy-to-use criteria which facilitate the work of the annotators, but which also can accommodate for the elaboration of a more or less fine-grained tagset. An advantage of such hierarchical schema is its potential application to any other language, although it is possible that some criteria are not needed anymore for a specific language (e.g., linearization for order-free languages) or, on the contrary, that new syntactic criteria are needed. We already successfully began to apply this method to a radically different language, namely, Finnish, and are annotating a 2,000 sentence corpus with a restricted set of about 25 relations.

The use of the fine-grained tagset and the application of the hierarchized criteria for the annotation of a 100,000 word corpus has proven feasible.

References

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. Floresta sintá(c)tica: A treebank for Portuguese. In *Proceedings of LREC 2002*, Las Palmas de Gran Canaria, Spain.

[6] We actually have a version of our corpus with such valency information (to be released).

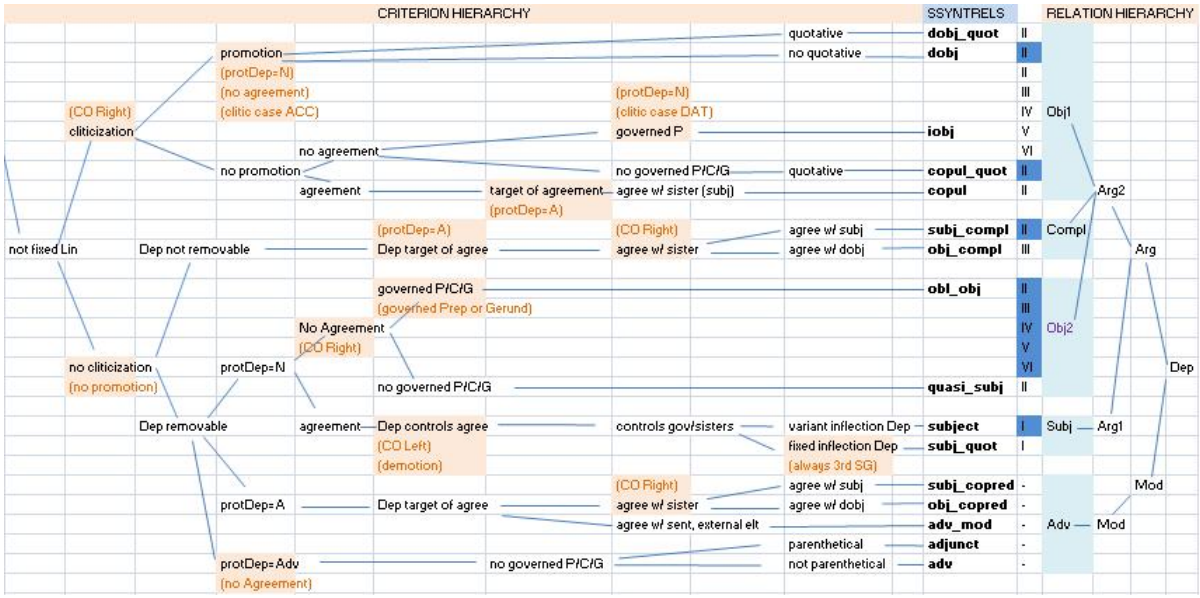

Figure 2: A partial hierarchy of syntactic criteria and a possible generalization of relations

L. Ahrenberg. 2007. LinES: An English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, Tartu, Estonia.

L. Cyrus, H. Feddes, and F. Schumacher. 2003. FuSe–a Multi-Layered Parallel Treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*, pages 213–216, Växjö, Sweden.

M-C. De Marneffe et al. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC 2006, Genova, Italy.*

J. Hajič and P. Zemánek. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117, Cairo, Egypt.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, and Z. Žabokrtský. 2006. Prague Dependency Treebank 2.0.

L Iordanskaja and I. Mel'čuk. 2009. Establishing an Inventory of Surface-Syntactic Relations: Valence-Controlled Surface-Syntactic Dependents of the Verb in French. In Alain Polguère and Igor Mel'čuk, editors, *Dependency in Linguistic Description*, Studies in Language Companion, pages 151–234. John Benjamins Publishing Company.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

B. Megyesi, B. Dahlqvist, E. Pettersson, and J. Nivre. 2008. Swedish-Turkish Parallel Treebank. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*, Marrakesh, Morocco.

I.A. Mel'čuk and N.V. Percov. 1987. *Surface Syntax of English*. John Benjamins Publishing Company.

I.A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

I.A. Mel'čuk. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, H.-W. Eroms L. Eichinnger, P. Hellwig, H.J. Herringer, and H. Lobin, editors, *Dependency and Valency. An International Handbook of Contemporary Research*, volume 1, pages 188–229. W. de Gruyter.

I.A. Mel'čuk. 2006. *Aspects of the Theory of Morphology*. Mouton De Gruyter, Berlin.

S. Mille and L. Wanner. 2010. Syntactic Dependencies for Multilingual and Multilevel Corpus Annotation. In *Proceedings of LREC 2010, Valletta, Malta.*

J. Nivre, J. Nilsson, and J. Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, Genova, Italy.

M. Taulé, M.A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the LREC-2008*, Marrakesh, Morocco.

C. Thielen, A. Schiller, S. Teufel, and C. Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit

STTS. Technical report, Institute for Natural Language Processing, University of Stuttgart.

L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. 2002. The Alpino Dependency Treebank. In *Proceedings of Computational Linguistics in the Netherlands CLIN 2001*.

M. Čmejrek, J. Hajič, and V. Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal.