# AnCora-UPF: A Multi-Level Annotation of Spanish

**Simon Mille**[1] **Alicia Burga**[1] **Leo Wanner**[1,2]
[1] Natural Language Processing Group, Pompeu Fabra University, Barcelona, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA)
firstname.lastname@upf.edu

## Abstract

There is an increasing need for the annotation of multiple types of linguistic information that are rather different in their nature, e.g., word order, morphological features, syntactic and semantic relations, etc. Quite frequently, their annotation is combined in a single structure, which not only results in inadequate annotations of treebanks and consequent low-quality applications trained on them, but also is deficient from a theoretical (linguistic) perspective. We present a new corpus of Spanish annotated on four independent levels, morphology, surface-syntax, deep-syntax and semantics, as well as the methodology that allows for obtaining it with fewer cost while maintaining a high inter-annotator agreement.

## 1 Introduction

There is an increasing need in stochastic dependency-oriented NLP applications (among them, semantic role labeling or semantic analysis, sentence generation, abstractive summarization, etc.) to deal not only with syntactic, but also with semantic information. This need implies that dependency treebanks must be annotated with both syntactic and semantic information, as, e.g., the Prague Dependency Treebank (PDT) 2.0 for Czech (Hajič, 2004; J.Hajič et al., 2006) and the Italian Syntactic-Semantic Treebank (S.Montemagni et al., 2003). However, most of the widely-known treebanks contain only one layer of annotation, namely the syntactic one; see, e.g., the dependency version of the Penn TreeBank (Johansson and Nugues, 2007) for English, Talbanken05 for Swedish (Nilsson et al., 2005), and SynTagRus for Russian (Apresjan et al., 2006). To also offer semantic annotation, some corpora have been enriched a posteriori by semantic information; cf., e.g., Penn Treebank/PropBank (Palmer et al., 2005)/NomBank (Meyers et al., 2004) or Ancora (Taulé et al., 2008). The disadvantage of such amendments is that they risk to intermingle syntactic and semantic information in the same annotation scheme, which then negatively affects the applications trained on them. This is true in particular for Natural Language Generation: see for instance (Bohnet et al., 2010) and the first Surface-Realization Shared Task (Belz et al., 2011), who both needed to separate semantic and syntactic annotations for their experiments.

In this paper, we propose a genuinely multilevel corpus annotation scheme for Spanish and discuss a sample annotation of the corpus (Ancora-UPF), the current version of which contains 3,513 sentences (100,892 tokens).[1]

## 2 The layers in our annotation

Our annotation intends to ensure that (i) a level of representation does not percolate into another one, and (ii) the annotation is complete in order to allow for easy automatic processing at each layer. Following the levels of the linguistic model in the Meaning-Text Theory (Mel'čuk, 1988), we annotate four different layers on top of the sentence level: morphological, surface-syntactic, deep-syntactic, and semantic.

### 2.1 Morphological layer

The morphological layer is a simple chain of surface lexical units bearing morpho-syntactic information. Surface lexical units are all the items of the vocabulary, that is, words as they appear in any monolingual dictionary, and their inflected variants. In Table 1, all possible values of all morpho-syntactic features used in our annotation are detailed. In addition to features such as gender and number, we use three different tagsets for Part-of-Speech: a coarse-grained one, *dpos*, and two fine-grained ones: *pos* and *spos*. The difference between *pos*, which is a subset of the PoS tagset from the Penn TreeBank set (Marcus et al., 1993), and *spos* is minor, although, for instance, in parsing

[1] It includes the 3,510 sentences that AnCora comprised at the time we launched this project back in early 2008, and three additional sentences we used for early tests. For downloads, see http://www.taln.upf.edu/content/resources/495.



*Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 217–226,
Prague, August 27–30, 2013. 

| Features | Possible values | # |
|---|---|---|
| **dpos** | A, Adv, N, V | 88,873 |
| **spos** | adjective, adverb, auxiliary, conjunction, copula, determiner, foreign_word, formula, interjection, interrogative_pronoun, noun, number, percentage, preposition, pronoun, proper_noun, punctuation, relative_pronoun, roman_numeral, verb | 100,892 |
| **pos** | CC, CD, DT, IN, JJ, N, NN, NP, PP, RB, SYM, UH, VB, VH, VV, WP, formula | 100,892 |
| **id** | 1 to $\infty$ | 100,892 |
| **surface form** | any | 100,892 |
| **lemma** | any | 100,892 |
| **gender** | C, FEM, MASC | 41,735 |
| **number** | PL, SG | 53,608 |
| **mood** | IMP, IND, SUBJ | 8,116 |
| **person** | 1, 2, 3 | 8,132 |
| **tense** | FUT, PAST, PRES | 8,070 |
| **finiteness** | FIN, GER, INF, PART | 11,176 |

Table 1: Morpho-syntactic features

experiments reported upon in (Ballesteros et al., 2013), *spos* performed better than *pos* for labeled relation attachment. Table 2 shows the repartition of the morpho-syntactic features that not all nodes carry, while Table 3 allows for visualizing the difference between the two fine-grained part-of-speech tags.[2]

| FEAT | V | N | Adj | Det | Pro | Other |
|---|---|---|---|---|---|---|
| finiteness | 99.91 | 0.01 | 0.06 | 0 | 0 | 0.02 |
| gender | 2.02 | 46.72 | 14.31 | 32.33 | 4.37 | 0.25 |
| mood | 99.95 | 0.01 | 0 | 0 | 0 | 0.04 |
| number | 16.74 | 36.57 | 15.15 | 27.1 | 4.25 | 0.19 |
| person | 99.98 | 0.01 | 0 | 0 | 0 | 0.01 |
| tense | 99.98 | 0 | 0 | 0 | 0 | 0.02 |

Table 2: Distribution of features (%)

| pos | spos |
|---|---|
| CC | **conjunction** |
| CD | cardinal number |
| DT | determiner |
| IN | **conjunction**<br>preposition |
| JJ | adjective |
| NN | common noun |
| NP | proper noun |
| PP | personal pronoun |
| RB | adverb |
| SYM | punctuation<br>percentage |
| UH | interjection |
| VB | **auxiliary**<br>copula |
| VH | **auxiliary** |
| VV | verb |
| WP | interrogative pronoun<br>relative pronoun |
| Formula | formula |
| - | foreign word |

Table 3: Correspondences between *pos* and *spos*

[2]There are only 88,873 *dpos* features because punctuations do not receive any.

### 2.2 Surface-syntactic (SSynt) layer

This layer is annotated with unordered dependency trees in which labelled dependencies link pairs of surface lexical units. Thus, the nodes have a one-to-one correspondence with the nodes of the morphological level. The 47 language-specific surface-syntactic relations used for the annotation of this layer are given and briefly explained in Table 4.[3] In the corpus, 14 of these relations occur more than a thousand times; these are, from the most frequent to the less frequent: *prepos*, *det*, *punc*, *adv*, *modif*, *subj*, *obl_obj*, *dobj*, *conj*, *coord*, *aux_phras*, *attr*, *copul*, and *relat*. Depending on the application, one can need more or less tags in the annotation. In order to allow for tuning the granularity of the tagset, we organized the relations in a hierarchy (see (Mille et al., 2012) for illustration).

### 2.3 Deep-syntactic (DSynt) layer

The structures at this layer are dependency trees in which labelled dependencies link pairs of *deep* lexical units. To the lexical units, deep-syntactic grammemes are assigned. The deep-syntactic dependency relations (cf. Table 5) are language-independent and thus also more abstract than the surface-syntactic ones. In our corpus, the deep-syntactic layer contains only 66,980 nodes since all punctuation signs and functional nodes have been removed. In the following, the four particular cases of node-removal are listed.[4]

**(a) Governed elements**

The presence of a governed preposition is imposed by the subcategorization ("valency") characteristics of its head, as, e.g., the appearance of TO in *give it TO your friend*), in the sense that the preposition TO is required by 'give'. TO in itself is here void of own meaning and should thus not appear in the deep-syntactic structure. This is different in, for instance, *to go INTO/IN FRONT OF/NEXT TO/. . . your house*, where the preposition is meaningful (even though it is governed) and thus should appear in the deep-syntactic structure. The depen-

[3]So far, we do not have special relations for ellipses; we add a syntactic empty node in order to deal with "impossible" dependencies only in case of what is commonly known as "gapping" and "right-node-raising".

[4]Some nodes are also added in the deep-syntactic structure. Thus, when there is an empty subject, we introduce a node with the person and number information as first argument of the verb (since the verb takes that information for being inflected), and when necessary link that new node to another one with a coreference relation.

| DepRel | Distinctive properties |
|---|---|
| abbrev | abbreviated apposition |
| abs_pred | non-removable dependent of an N making the latter act as an adverb |
| adv | mobile adverbial |
| agent | promotable dependent of a participle |
| analyt_fut | Prep *a* governed by future Aux |
| analyt_pass | non-finite V governed by passive Aux |
| analyt_perf | non-finite V governed by perfect Aux |
| analyt_progr | non-finite V governed by progressive Aux |
| appos | apposed element |
| attr | right-side modifier of an N |
| aux_phras | multi-word marker |
| aux_refl | reflexive Pro depending on a V |
| bin_junct | for binary constructions |
| compar | complement of a comparative Adj/Adv |
| compl1 | non-removable adjectival object agreeing with subject |
| compl2 | non-removable adjectival object agreeing with direct object |
| compl_adnom | prepositional dependent of a stranded Det |
| conj | complement of a non-coordinating Conj |
| coord | between a conjunct and the element acting as coordination conjunction |
| coord_conj | complement of a coordinating Conj |
| copul | cliticizable dependent of a copula |
| copul_clitic | cliticized dependent of a copula |
| det | non-repeatable left-side modifier of an N |
| dobj | verbal dependent that can be promoted or cliticized with an accusative Pro |
| dobj_clitic | accusative clitic Pro depending on a V |
| elect | non-argumental right-side dependent of a comparative Adj/Adv or a number |
| iobj | dependent replaceable by a dative Pro |
| iobj_clitic | dative clitic Pro depending on a V |
| juxtapos | for linking two unrelated groups |
| modal | non-removable, non-cliticizable infinitive verbal dependent |
| modif | for Adj agreeing with their governing N |
| num_junct | numerical dependent of another number |
| obj_copred | adverbial dependent of a V, which agrees with the direct object |
| obl_compl | right-side dependent of a non-V element introduced by a governed Prep |
| obl_obj | prepositional object that cannot be demoted, promoted or cliticized |
| prepos | complement of a preposition |
| prolep | for clause-initial accumulation of elements with no connectors |
| punc | for non-sentence-initial punctuations |
| punc_init | for sentence-initial punctuation |
| quant | numerical dependent which controls the number of its governing N |
| quasi_coord | for coordinated elements withthe no connector |
| quasi_subj | a subject next to a grammatical subject |
| relat | finite V that modifies an N |
| relat_expl | adverbial finite clause |
| sequent | right-side coordinated adjacent element |
| subj | dependent that controls agreement on its governing V |
| subj_copred | adverbial dependent of a V agreeing with the subject |

Table 4: 47 dependency relations used at the surface-syntactic layer

dents involved in the following SSynt-relations are concerned: *agent*, *compar*, *dobj*, *iobj*, *obl_compl*, and *obl_obj*. We also remove all subordinating conjunctions *que* 'that' when they introduce an argument of a predicate.

**(b) Auxiliaries**

An auxiliary is a functional element and therefore should not appear as such in a "deep" structure. However, it expresses semantic grammatical significations, namely tense (past: *haber* 'have' + past participle; future: *ir* 'go' + preposition *a* 'to' + infinitive), aspect (progressive: *estar* 'be' + present participle) or voice (passive: *ser* 'be' + past participle). These significations must be reflected in the deep-syntactic structure. For this purpose, corresponding attributes have been introduced to capture tense, aspect and voice: 'tense' for tense (with as possible values *present*, *future* and *past*); 'tem_constituency' for aspect (with as possible values *simple*, *progressive*, *perfect*, *perfect progressive*); and the attribute 'voice', with the values *active* or *passive*. However, since there are two ways to realize passive voice in Spanish (one with an auxiliary and one with a reflexive pronoun), the mapping between a deep-syntactic verb with "voice=passive" and its superficial counterpart is not straightforward.

**(c) Determiners**

Definite *el* 'the' and indefinite *un* 'a' determiners (at least) should be removed from the deep-syntactic annotation: they indicate degrees of givenness, and in this respect account for a part of the information and coreference structures. The determiners can be replaced by attribute/value pairs assigned to the governing noun (*given* VS *new*). However, we are conscious that there is no reliable way to identify automatically the givenness of nouns, since there is no systematic correlation between the presence or the absence of a determiner for a noun and its givenness. A manual annotation of givenness would be needed for some tasks; for instance, for a generator to learn correctly how to deal with the introduction of determiners in a superficial structure. For now, we only annotate definiteness on nouns so as to encode the presence of a definite or indefinite determiner at the surface. All other determiners (demonstrative, possessives, etc.) are kept in the deep annotation because they can encode more than mere givenness: possessives can receive any edge in deep-syntax since they can stand for a modifier (*su silla*

'his/her chair') or an argument (first argument: *su traducción* 'his/her translation (of something)'; second argument: *su elección* 'his/her election (by someone)', etc.) of the governing noun. The determiners that are maintained in DSynt receive the dependency relation *ATTR*.

**(d) Relative Pronouns**

Relative pronouns with antecedent should be substituted by their antecedent in the deep-syntactic structure, and a coreference link added between them. Given how we annotate relative clauses (see Figure 1), we can always find the antecedent of the pronoun as the governor of the *relat* relation.

| DepRel | Short description |
|---|---|
| **I** | first argument |
| **II** | second argument |
| **III** | third argument |
| **IV** | fourth argument |
| **V** | fifth argument |
| **VI** | sixth argument |
| **APPEND** | backgrounded modifier |
| **ATTR** | regular modifier |
| **COORD** | coordinate |
| **coref** | special coreference relation |

Table 5: 9 dependency relations used at the deep-syntactic layer

The deep-syntactic grammemes comprise the features of the more superficial layers (see Table 1), and some additional features specific to this level (see Table 6). We see that the feature(s) *id_ssynt* store the correspondence between the DSynt node and one or more SSynt nodes.

| DSynt Feature | Possible values |
|---|---|
| **coref_id** | 1 to $\infty$ |
| **definiteness** | DEFINITE \| INDEFINITE \| N/A |
| **id_ssynt1** | 1 to $\infty$ |
| **id_ssynt2** | 1 to $\infty$ |
| **id_ssyntn** | 1 to $\infty$ |
| **tem_constituency** | SIMPLE \| PROGRESSIVE \| PERFECT \| PERFECT PROGRESSIVE |
| **voice** | ACTIVE \| PASSIVE |

Table 6: Additional (compared to SSynt) grammemes used in the DSynt annotation

### 2.4 Semantic (Sem) layer

A semantic structure is an acyclic predicate-argument graph. The nodes at the semantic level in our corpus are the same as the nodes at the deep-syntactic level. In other words, in the first version of the corpus, we do not generalize the word labels: different words which have identical meanings keep a different label in semantics. However, we add six different types of meta-nodes in order to encode information stored as feature/values in the previous layers or to connect non-predicative units to the rest of the structure:[5]

***ROOT***: it has only one argument, and simply indicates which node of the semantic structure is the most important; it directly relates with the main node of the sentence, that is, usually, the main verb of the main clause.

***TENSE***: the first argument is by convention the event, and the second argument indicates whether it was in the past, is in the present, or will be in the future.

***NUMBER***: following the same model as *TENSE*, the first argument is the semantic number, and the second argument is the value *SINGULAR* or *PLURAL*. Note that this should concern semantic number only, and not lexical number. For instance, the number of the word *paro* 'unemployment' in Figure 1d is *lexical*; it cannot vary. As a result, it should not be an argument of a node *NUMBER*. However, in this version of the corpus, all nouns receive a number.

***TEM_CONSTITUENCY***: again, the first argument is by convention the event, and the second argument indicates whether it is progressive, perfect, both or none.

***ELABORATION***: this meta-node is used to connect to the semantic graph those non-predicative deep-syntactic nodes that receive the relations ATTR or APPEND. The node ELABORATION takes the dependent as its second argument, and the governor as its first one. It is mainly used in the case of apposition. In Figure 1c, there are two predicative attributes, *este* 'this' and the head of the relative clause, *engrosar* 'make swell'; in both cases, their syntactic governor is their first argument and, therefore, no ELABORATION node is needed to connect them to the semantic structure. However, in some appositive constructions, for instance, the apposed element cannot take its DSynt governor as argument: in *Pipo, mi perro* 'Pipo, my dog', we have Pipo-*ATTR*→perro, and *perro* is not a predicate. An extra node is therefore needed to connect it to the structure. The attributive relation in this case stands for the fact that the governor is the name given to the dependent; subsequently,

[5] Meta-nodes are shown in upper case in Figure 1d, while regular nodes are in lower case.

we should have at the semantic level 'Pipo'←*2*-NAME-*1*→'perro'. However, since we did not undertake a manual revision of the semantic layer as yet, we use for now the generic label *ELABORATION* in all cases, considering that the second argument somehow elaborates on the first one.

***POSSESS***: as already mentioned in Section 2.3, when the possessive determiner is not an argument, it usually stands for a possession relation between the governor, which will be the second semantic argument, and the dependent, which will be the first one.[6]

These predicates are called "meta" because they encode information that is necessary at the semantic level of representation, but that should not be considered the same as other nodes, since they should not be realized as words in the final sentence. If we would not differentiate one type of node from the other, Figure 1d could result in a sentence like "The document, the number of which is singular, suggests in a present time that ...".[7] Finally, the semantic features are (i) a unique individual ID, (ii) an ID that indicates the correspondence with DSynt nodes, and (iii) an attribute that encodes the definiteness of some nouns.

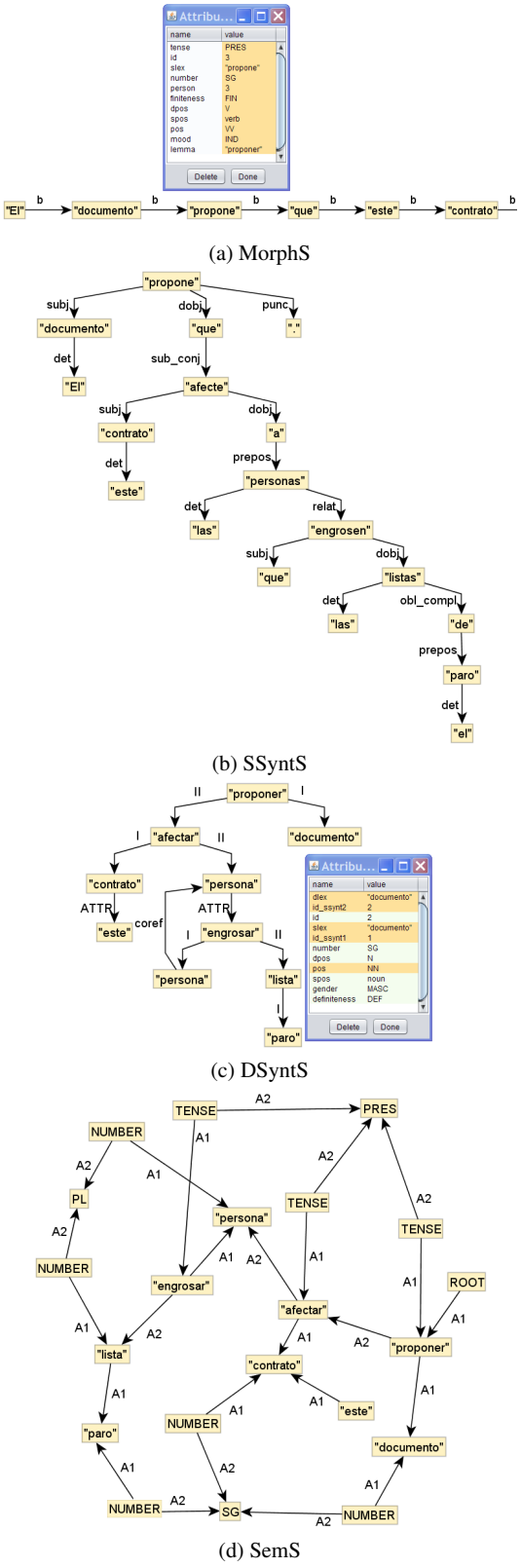
The nomenclature of predicate-argument relations is given in Table 7;[8] an example of each annotation level is shown in Figure 1.

| DepRel | Short description |
|---|---|
| **1** | first argument |
| **2** | second argument |
| **3** | third argument |
| **n** | nth argument |

Table 7: Predicate-argument relations used at the semantic layer

## 2.5 Format

In order to facilitate the processing of the superficial layers of the annotation, the sentence, morphological and surface-syntactic layers are pre-




(a) MorphS

(b) SSyntS

(c) DSyntS

(d) SemS

Figure 1: The four levels of annotation for the sentence *El documento propone que este contrato afecte a las personas que engrosen las listas del paro* 'The document suggests that this contract affect the persons who make the unemployment lists swell'

[6]These three last meta-nodes are not shown in Figure 1d in order to make the figure more readable.

[7]Technically, the information encoded until now in the semantic structure is still not sufficient to regenerate the sentence as it was on the surface: the information structure also constrains the realization of the semantic graph. However, as we consider the superimposing of an information structure on a semantic network as a different task, this is out of the scope of this paper.

[8]Note that unlike the semantic annotation of PTB/PB, the semantic structure in MTT has transparent semantic frames, in the sense that no difference is made between external or internal arguments.

sented in a single standard 14-column CoNLL file. The deep-syntactic layer is also provided in a separate CoNLL file, while the semantic layer is presented in the HFG format used in the Surface-Realization Shared Task in 2011 (Belz et al., 2011). The different layers are connected thanks to the IDs of the nodes.

## 3 Multilayered annotation in practice

Annotating such a corpus manually can seem too costly at the first sight. In this section, we show that a solid theoretical framework and the use of adequate tools can allow for significant reduction of the manual workload.

### 3.1 The advantages of our theoretical framework

As already mentioned, our annotation model is strongly influenced by the Meaning-Text Theory (Mel'čuk, 1988). Its rich stratification facilitates a clear separation of different types of linguistic phenomena and thus a straightforward handling for various NLP-applications. Equivalent annotations for other theoretical frameworks can be easily derived from our representations—which is why we believe that MTT in general has considerable advantages. But on top of that, the MTT model is a transductive model (Kahane, 2003). This means that it also provides the instruments for the mapping of a representation at a given level to the corresponding representations at the adjacent levels. This has an interesting consequence as far as corpus annotation is concerned: starting from a given stratum and a manually created mapping grammar (the coverage does not need to be broad at first), the annotations at the adjacent strata can be easily obtained, and they can on their turn be used to derive the annotations at the next strata, and so on. In other words, with a corpus of SSyntSs, it is straightforward to derive parallel corpora of DSyntSs and SemSs using an adequate tool, such as the graph transducer MATE (Bohnet et al., 2000). The process of annotation can be reduced to a minimal manual revision of automatically created structures.

For the surface-syntactic annotation, we use our detailed annotation schema that allows for relatively easy dependency relation identification, based on easy-to-use criteria. The annotation schema has been defined taking into account that (a) the schema should cover only criteria that are related to the *syntactic* behaviour of the nodes; (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.[9] The latter led us target a set of around 50 SSyntRels. For details on when we establish a dependency between two nodes as well as its direction, and to see which criteria we used for labelling dependencies, see (Burga et al., 2011).

### 3.2 Annotation of the morphological and surface-syntactic layers

The dependency treebank from which we started is AnCora-DEP-ES in its 2008 version (Taulé et al., 2008). The surface-syntactic annotation procedure comprised two stages: (1) an automatic projection of the annotations of the sentences from AnCora onto rudimentary surface-syntactic structures (see (Mille et al., 2009) for more details); (2) multiple manual revisions of the structures obtained in Stage 1. For the revision work carried out by a small team of trained annotators, the graph editor of the graph transducer MATE was used (Bohnet et al., 2000).

To facilitate the annotation of the deeper levels, we split 14 of the relations shown in Table 4 into more fine-grained relations which also encode predicate-argument information. Those labels are used to derive automatically rather complete deep-syntactic structures (see Section 3.3), but are not retained in the surface-syntactic annotation, which only includes the 47 original labels. That is, in order to label the dependencies, the annotator has to follow the syntactic guidelines, and when annotating some of the relations in the DepRel column of Table 4, add or not a suffix to the label, based on three criteria:

**(1)** What is the configuration of the underlying predicate-argument structure? (5 DepRel→ 25)

For the DepRel *iobj*, *iobj_clitic*, *obl_compl*, *obl_obj*, the goal is to associate to the dependent a slot in the valency frame of its governor: by convention, we number the argument slots from 0 to 5, although they correspond to the first to the sixth arguments. For this, we asked the annotators to (i) consider the definition of the predicate, which can only be complete if all its arguments are mentioned, and (ii) evaluate the importance of each ar-

[9] We refer here, first of all, to decision making and inter-agreement rate.

gument with respect to this predicate, which allows for assigning them a slot in its valency. At the first glance, the task may appear subjective and thus difficult. However, the very large majority of predicates have between one and three arguments. This makes the task easier, especially for verbs, for which the subject (in active voice) is always considered the first argument,[10] and the direct object the second. In case of oblique or indirect objects or oblique complements (see Table 4 for more details), the decision can be harder to make. But the high inter-annotator agreement rate obtained for the task (see Section 3.5) indicates that the intuition of the annotators coincides to a large extent. Consider, for example, the predicate *proponer* 'suggest': its definition would be something like "an entity E1 giving an idea I to another entity E2 for E2 to consider I". In other words, *proponer* has three arguments, E1, I, E2; E1 and I are almost never omitted, which makes them higher in the argument hierarchy than E2, and the entity "who does" is considered more important than what is done. As a result, we have E1=Arg1 (subject), I=Arg2 (direct object), and E2=Arg3 (oblique object 2).

In addition to object and complement DepRel, the reflexive auxiliary *aux_refl* tag is subdivided into four groups: direct (the pronoun is the second argument of the verb and has a coreference link with its subject), indirect (same as direct but the pronoun is third argument), passive (the pronoun is not an argument but triggers an inversion of first and second arguments in the DSyntS), and lexical (the pronoun is just a part of the verb's lemma).

**(2)** Is the dependent parenthetical?(6 DepRel→ 12)

This criterion is used in order to distinguish between two levels of modification for basic modifiers, one being closer to the governor than the other. For instance, the *adv* DepRel below a verb indicates the presence of a circumstancial element related to the verb itself, while the *adjunct* DepRel indicates that the circumstancial operates at the sentence level: (normalmente←*adjunct*-corre-*adv*→[cada dia] 'usually he runs every day'). For nominal governors (*appos*, *attr*, *modif*, *quant*, *relat*), the descriptive extension is usually granted to groups separated by a comma from their head.

**(3)** Is the dependent quoted? (3 DepRel→ 6)

In simple terms, it is the group formed by the dependent and all its dependents surrounded by quotation marks, which indicate an actual quotation. Consider, for illustration, the difference between *dijo "me voy"* 'he said "I'm going"' (quote), and *¡Mira, el "presidente" llega!* 'Look, the "president" is arriving!', in which the quotation marks are a stylistic way of making fun of someone. Three DepRel are concerned: *subj*, *dobj* and *prepos*.

As a result, instead of 14 DepRel, the annotator has to consider 43, that is, 29 more. So far, this gives us 76 different tags (47 + 29). In addition, we further split for testing reasons (which we do not have space to detail in this paper) the label *conj* into *sub_conj* and *compar_conj*, and added a third label *restr* when splitting the DepRel *adv*. Thus, the total tagset which represents the base of our annotation process comprises **79 different tags**. We refer to this tagset as the "Annotation SSynt DepRel" tagset (SSynt $\text{DepRel}_A$).

As for the annotation at the morphological layer, it was mostly derived automatically from the AnCora annotation.

### 3.3 Annotation of the deep-syntactic layer

As mentioned in Section 2, the deep-syntactic layer has the form of an unordered dependency tree. The edges encode explicit valency relations, and also coordination and modifications, while only meaning-bearing units are accepted as nodes. Multi-word expressions are fused into single nodes. Sentence-internal coreferential links are superimposed on the annotation. All surface-syntactic relations (except *det*, see Section 2.3) have a direct correlation with deep-syntactic configurations.

Taking this into account, together with the syntactic properties of each DepRel (e.g., *obl_obj* points to a governed preposition, i.e., to a functional node which does not carry any meaning on its own), the mapping between SSynt and DSynt can be largely automatic (for instance, the DSyntS shown in Figure 1c has required no manual modification, although this is not always the case). The workload of the annotator is reduced to (i) addition of coreferences between nodes of the same sentence, (ii) definition of the argument slot of possessive pronouns when necessary, and (iii) repair of possible erroneous rule applications. There are currently 129 rules in the SSynt-DSynt mapping grammar, and its coverage is not yet complete,

[10] This is why there is no extension 0 for verbal relations (iobj, iobj_clitic, obl_obj), and also why by default we start numbering the arguments from the second.

as some very specific configurations are still not taken into account. However, we intend to expand the coverage as much as possible in the future. An average-length sentence (around 30 nodes) takes an annotator around one and a half minutes to process (while without the automatic annotation derivation it takes her/him about 10 minutes).

| SSynt | DSynt | SSynt | DSynt |
|---|---|---|---|
| abbrev | ATTR | iobj_clitic1 | II |
| abs_pred | ATTR | iobj_clitic2 | III |
| adjunct | APPEND | iobj_clitic3 | IV |
| adv | ATTR | iobj_clitic4 | V |
| adv_mod | ATTR | iobj_clitic5 | VI |
| agent | I | juxtapos | APPEND |
| analyt_fut | - | modal | II |
| analyt_pass | - | modif | ATTR |
| analyt_perf | - | modif_descr | APPEND |
| analyt_progr | - | num_junct | COORD |
| appos | ATTR | obj_copred | ATTR |
| appos_descr | APPEND | obl_compl0 | I |
| attr | ATTR | obl_compl1 | II |
| attr_descr | APPEND | obl_compl2 | III |
| aux_phras | - | obl_compl3 | IV |
| aux_refl_dir | II | obl_compl4 | V |
| aux_refl_indir | III | obl_compl5 | VI |
| aux_refl_lex | - | obl_obj1 | II |
| aux_refl_pass | - | obl_obj2 | III |
| bin_junct | ATTR | obl_obj3 | IV |
| compar | II | obl_obj4 | V |
| compar_conj | II | obl_obj5 | VI |
| compl1 | II | prepos | II |
| compl2 | III | prepos_quot | II |
| compl_adnom | ATTR | prolep | APPEND |
| coord | COORD | punc | - |
| coord_conj | II | punc_init | - |
| copul | II | quant | ATTR |
| copul_clitic | II | quant_descr | APPEND |
| copul_quot | II | quasi_coord | COORD |
| det | any | quasi_subj | I |
| dobj | II | relat | ATTR |
| dobj_clitic | II | relat_descr | APPEND |
| dobj_quot | II | relat_expl | APPEND |
| elect | ATTR | restr | ATTR |
| iobj1 | II | sequent | ATTR |
| iobj2 | III | sub_conj | II |
| iobj3 | IV | subj | I |
| iobj4 | V | subj_copred | ATTR |
| iobj5 | VI | | |

Table 8: Mapping of the 79 SSynt DepRel$_A$ onto DSynt DepRel

Table 8 indicates that some SSynt DepRel$_A$ are not mapped to any DSynt DepRel. This is due to the fact that some nodes (namely the functional ones) are removed from the deep-syntactic structure. The idea is that from the perspective of Natural Language Generation (NLG) from abstract structures, the system will only have access to non-linguistic data; see, e.g., (Bouayad-Agha et al., 2012). This implies that a system that generates statistically from those abstract representations MUST be able to learn when to introduce functional words (i.e., words that carry a grammatical content, but no own lexical meaning). Therefore, a corpus claimed to be suitable for training statistical NLG modules should always take this into account. In addition, the removal of functional nodes allows the generators to deal with different surface realizations when several realizations are possible (e.g., *give something to Mary* VS *give Mary something*). Having in parallel two layers, one with all the words, and one without the functional words, is one way to provide the basis for statistical models.

| SSynt DepRel$_A$ | Changes in DSynt |
|---|---|
| analyt_fut | remove Gov and Dep<br>add tense=FUT |
| analyt_pass | remove Gov<br>invert I and II<br>add voice=PASS |
| analyt_perf | remove Gov<br>add tense=PAST |
| analyt_progr | remove Gov<br>add tem_constituency=PROGR |
| aux_refl_dir | replace node label with antecedent's<br>add coreference between I and II |
| aux_refl_indir | replace node label with antecedent's<br>add coreference between I and III |
| aux_refl_lex | remove Dep<br>add *se* at the end of Gov's lemma |
| aux_refl_pass | remove Dep<br>invert I and II<br>add voice=PASS |
| det | **IF Dep=el\|un**<br>remove Dep<br>add definiteness=DEF/INDEF |
| | **IF Dep=possessive**<br>replace node label with antecedent's<br>edit DSynt DepRel<br>add coreference link with antecedent |
| | **IF Dep=other**<br>map *det* to *ATTR* |
| dobj/iobj1-5/obl_compl0-5<br>obl_obj1-5 | remove Dep if governed preposition |
| relat/relat_descr | replace node label with antecedent's<br>add coreference link with antecedent |
| . . . _conj | remove Dep if governed preposition |

Table 9: More complex SSynt to DSynt mappings

Since, as we have seen in Section 3.3, not all surface-syntactic nodes are mapped to the deep-syntactic level, some configurations imply non-typical equivalences. Table 9 completes Table 8 by summarizing all mappings of SSynt DepRel$_A$ to something else than a single DSynt DepRel.

### 3.4 Annotation of the semantic layer

Since in the deep-syntactic layer all grammatical units are removed from the structure, the mapping to a connected acyclic graph entirely composed of predicate-argument relations that connect any meaning-bearing unit used in the sentence (which includes DSynt nodes and some additional meta-nodes) is much easier. A different mapping grammar from the one detailed in Section 3.3 can transform the deep-syntactic structure in Figure 1c into a semantic structure shown in Figure 1d.

During this second mapping, all nodes from the deep-syntactic structure are transferred, except

nodes which have a coreference relation with another node. Only one node that stands for all coreferring nodes appears in the semantic structure; all edges that point to a node which is removed are transferred to that one node.[11]

Most relations can be derived in a straightforward way: Roman numerals map to Arabic numerals, and *ATTR*, *APPEND* and *COORD* edges are inverted and relabelled with *1* when the DSynt dependent is a predicate. Otherwise, we introduce meta-predicates like, for instance, *ELABORATION* or *POSSESS* in order to connect the equivalent of the DSynt dependent to the graph (see Section 3.4).

In the procedure of obtaining the annotation at the semantic layer, the mapping grammar does all the work, and there is no need for manual revision at all.

### 3.5 Inter-annotator agreement

Due to the still preliminary nature of our deep-syntactic and semantic annotations, we evaluated the inter-annotator agreement so far only for the surface-syntactic annotation. However, we used the 79-relation tagset, which facilitated the automatic derivation of the deeper annotations; see Section 3. This tagset thus allows us to indirectly obtain the deep layer inter-annotator agreement (while the 47-relation tagset gives us the SSynt-layer inter-annotator agreement)—with the exception of possessive determiners, which are mapped to a variety of different deep-syntactic relations (ATTR, I, II, etc.). Therefore and given that possessive determiners represent only 1% of the total number of dependencies in the corpus, we decided not to take them into account in the deep evaluation.

To obtain the material for the inter-annotator agreement evaluation, we parsed with Bohnet's parser (Bohnet, 2009), trained on the surface-syntactic annotation the *lingüística* 'linguistics' wikipedia page,[12] 72 sentences in total (2,443 tokens). Two annotators then post-edited in separate sessions every sentence using the 79-tag tagset as described in Section 3.2. Drawing upon the surface-syntactic tag hierarchy described in (Mille et al., 2012), the resulting two annotations were further generalized to 47, 31 and 15 tags, such that we obtained parallel annotations for four different annotations.

Taking one annotation of each pair as gold standard and the other as "predicted", we ran the CoNLL'08 evaluation and calculated the LAS. The results are displayed in Table 10.

| | 79 | 47 | 31 | 15 |
|---|---|---|---|---|
| UAS (%) | 96.15 | 96.15 | 96.15 | 96.15 |
| LAS (%) | 89.40 | 92.26 | 92.51 | 92.80 |

Table 10: Inter-annotator agreement.

Since the successive mappings from 79 to 15 DepRel only concern the edge labels, it is normal that the Unlabeled Attachment Score remains the same for all tagsets. As expected, the agreement rate correlates with the number of tags in the tagset. Thus, we reached 89.4%, including predicate-argument identification 92.26%, with the 47 DepRel given in Table 4 in Section 2.2, and up to 92.8% with the reduced tagset of 15 DepRels. All inter-annotator agreement figures oscillate around the 90% threshold recommended in the OntoNotes project (Hovy et al., 2006).

## 4 Conclusions and future work

In this paper, we report on the results of the annotation of a Spanish corpus, in which the different levels of annotation are clearly separated. We show that thanks to a sound theoretical framework and appropriate tools, it is possible to reduce the manual workload and, at the same time, achieve a high inter-annotator agreement rate on all evaluated levels (more than 92% for syntax and more than 89% for syntax and semantics). These figures are largely due to the fact that the criteria that define each dependency relation have been carefully selected and are exclusively linguistically motivated. However, the 3-point difference between semantic and syntactic tagsets confirms that predicate-argument structures are less easily identifiable than syntactic dependencies since the criteria that define them are not as straightforward as syntactic criteria. In the future, we aim to augment the size of our tree bank, work on improving the predicate-argument identification, and add the dimension of the information structure. Both the treebank and all resources developed during the annotation (guidelines, software, etc.) will be made available to the community.

[11] Our mapping grammar actually has a parameter that allows for keeping the coreferring nodes separated in the SemS. This can be useful for experiments on information structure.

[12] Prior to parse it, the page has been cleaned.

## Acknowledgements

We would like to thank warmly Bernd Bohnet, Roberto Carlini, Gabriela Ferraro, Kim Gerdes, Antónia Martí, and Igor Mel'čuk. It is because of them that this work became possible.

## References

J. Apresjan, I. Boguslavsky, B. Iomdin, L. Iomdin, A. Sannikov, and V. Sizov. 2006. A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *Proceedings of LREC*, pages 1378–1381.

M. Ballesteros, S. Mille, and A. Burga. 2013. Exploring morphosyntactic annotation over a spanish corpus for dependency parsing. In *Proceedings of DepLing*.

A. Belz, M. White, D. Espinosa, E. Kow, D. Hogan, and A. Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at ENLG*, pages 217–226.

B. Bohnet, A. Langjahr, and L. Wanner. 2000. A development environment for an MTT-based sentence generator. In *Proceedings of INLG*, pages 260–263.

B. Bohnet, L. Wanner, S. Mille, and A. Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of COLING*, pages "98–106".

B. Bohnet. 2009. Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of CoNLL-2009*.

N. Bouayad-Agha, G. Casamayor, S. Mille, M. Rospocher, H. Saggion, L., and L. Wanner. 2012. From Ontology to NL: Generation of Multilingual User-Oriented Environmental Reports. In *Proceedings of NLDB*, Groningen, The Netherlands.

A. Burga, S. Mille, and L. Wanner. 2011. Looking Behind the Scenes of Syntactic Dependency Corpus Annotation: Towards a Motivated Annotation Schema of Surface-Syntax in Spanish. In *Proceedings of DepLing*, pages 104–114.

J. Hajič. 2004. Complex corpus annotation: The prague dependency treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT/NAACL*, pages 879–884, USA, June.

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. Linguistic Data Consortium, Philadelphia.

R. Johansson and P. Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA*, pages 105–112, Tartu, Estonia, May 25-26.

S. Kahane. 2003. The Meaning-Text Theory. In *Dependency and Valency. Handbooks of Linguistics and Communication Sciences*, volume 1-2. De Gruyter.

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.

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

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*.

S. Mille, L. Wanner, V. Vidal, and A. Burga. 2009. Towards a rich dependency annotation of Spanish corpora. In *Proceedings of SEPLN*, San Sebastian, Spain.

S. Mille, A. Burga, G. Ferraro, and L. Wanner. 2012. How does the granularity of an annotation scheme influence dependency parsing performance? In *Proceedings of COLING*, Mumbai, India.

J. Nilsson, J. Hall, and J. Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proceedings of NODALIDA*, pages 119–132.

M. Palmer, P. Kingsbury, and D. Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31.

S.Montemagni, F. Barsotti, and M. Battista et al. 2003. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, *Building and Using Syntactically Annotated Corpora*, pages 189–210.

M. Taulé, M. A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC*, Marrakech, Morocco.