

# Syntactic Dependencies for Multilingual and Multilevel Corpus Annotation

Simon Mille<sup>1</sup>, Leo Wanner<sup>1,2</sup>

<sup>1</sup>DTIC, Universitat Pompeu Fabra, <sup>2</sup>ICREA  
C/ Roc Boronat, 138, 08018 Barcelona, Spain  
simon.mille@upf.edu, leo.wanner@upf.edu

## Abstract

The relevance of syntactic dependency annotated corpora is nowadays unquestioned. However, a broad debate on the optimal set of dependency relation tags did not take place yet. As a result, largely varying tag sets of a largely varying size are used in different annotation initiatives. We propose a hierarchical dependency structure annotation schema that is more detailed and more flexible than the known annotation schemata. The schema allows us to choose the level of the desired detail of annotation, which facilitates the use of the schema for corpus annotation for different languages and for different NLP applications. Thanks to the inclusion of semantico-syntactic tags into the schema, we can annotate a corpus not only with syntactic dependency structures, but also with valency patterns as they are usually found in separate treebanks such as PropBank and NomBank. Semantico-syntactic tags and the level of detail of the schema furthermore facilitate the derivation of deep-syntactic and semantic annotations, leading to truly multilevel annotated dependency corpora. Such multilevel annotations can be readily used for the task of ML-based acquisition of grammar resources that map between the different levels of linguistic representation – something which forms part of, for instance, any natural language text generator.

## 1. Introduction

The relevance of syntactic dependency annotated corpora for Language Engineering is nowadays unquestioned. Several well-known dependency treebanks are already available; cf., for instance, the Prague Dependency Treebank (PDT, Hajič *et al.*, 2006), the dependency versions of the Penn Treebank (e.g. Mitchell *et al.*, 1993 and Li *et al.*, 2003), the AnCora treebank (Martí *et al.*, 2007), the Russian MTT-treebank (Apresjan *et al.*, 2006) and some others. Still, a broad debate on the optimal set of dependency relation tags and its application - and language-specificity, respectively - independence did not take place yet. As a result, largely varying tag sets of a largely varying size are used in different annotation initiatives. This is, without doubt, mainly due to the fact that annotation of dependency structures is quite a recent trend, and the annotation of corpora in different languages as part of the same endeavor even more so. However, to a certain extent, this is also due to the fact that so far dependency annotation schemata have often been created with a specific application in mind – in particular, analysis (cf., for instance, the CoNLL competition) – instead of attempting to accommodate for a large range of applications and a number of different languages. Our work is intended as a contribution to the solution of this problem. In what follows, we report on our experience of the annotation of corpora with surface-syntax dependency structures (Mille *et al.*, 2009) as known from the Meaning-Text Theory, MTT (Mel'čuk, 1988) and propose a hierarchical annotation schema that accommodates for both fine-grained language-specific dependency structures and a generic picture of abstract dependency relations. The former are needed if the corpus is intended, for instance, for use in corpus-based text generation, while the latter may serve better when the corpus is to be used for training in parsing applications.

## 2. On the nature of dependency relations

Theoretical linguistic studies show that the nature and diversity of dependency relations that hold between lexical units in a sentence are not language-independent. Rather, quite often, a language or a group of languages reveal some peculiarities that require the introduction of specific tags. For instance, in Catalan, Galician and Italian, the article combines with the possessive pronoun: Cat. *la meva mare*, lit. ‘the my mother’ vs. Gal. *a miña nai* vs. It. *la mia madre*, while in Spanish, French, etc. it does not: Sp. *\*la mi madre*, Fr. *\*la ma mère*. In principle, if they combine, both the article and possessive pronoun could be considered determiners (as, in fact, does PDT). However, this would not capture their idiosyncrasy with respect to repetition (only one article per NP is admissible, while several possessive pronouns can occur) and order (they cannot be permuted).

In a series of multilingual dependency treebanks, the same dependency relation tag set is used for each language. It is the case, for instance, in the AnCora dependency treebank released in three languages, namely Spanish, Basque and Catalan, and in the Swedish-Turkish parallel treebank (Megyesi *et al.*, 2008). In general, for all parallel treebanks that we could inspect – PDT2.0-PDAT (Hajič *et al.*, 2006, 2004), PCET (Čmejrek *et al.*, 2004), FuSe (Cyrus *et al.*, 2003), LinEs (Ahrenberg, 2007), etc. –, the justification of the choice of dependency labels is far from being central or is even largely avoided. In our work, we found this question very crucial. Thus, we observed that the choice of tags varies across languages (in the sense that distinct tags are required for distinct languages) and across applications (in the sense that depending on the application, a tag set needs to be more or less fine-grained). Thus, in the framework of corpus-based text generation, it is essential to capture such idiosyncratic dependencies as discussed above for Catalan, Galician and Italian, while in the framework of corpus-based parsing technologies, often more generic (and thus smaller) dependency tag sets are preferred.

Ideally, a dependency relation annotation schema would, on the one hand, facilitate the annotation of all language-specific syntactic idiosyncrasies, but, on the other hand, also offer a motivated generalization of the tags such that it could also serve for applications that prefer small generic dependency tag sets. In the next section, we present the proposal for such a schema. The proposal is based on our work on Spanish, with an occasional contrastive look at Catalan, English, Finnish, Galician, and Swedish.

### 3. Towards a generic annotation schema

As mentioned in Section 1, our annotation schema draws upon the surface-syntactic dependency relation repertoire from the MTT. Therefore, before we present the schema, we introduce the notion of surface-syntactic structure.

#### 3.1 The surface-syntactic structure

The surface-syntactic structures (SSyntSs) are one of the two types of syntactic dependency structures in MTT (cf. also Section 4 below). That is, they follow the properties of syntactic dependency as established in MTT (Mel'čuk 1988): (1) they hold between individual lexemes of the sentence, rather than constituents, (2) they are binary, such that each of them relates two and only two word forms, and (3) they are antisymmetric, antireflexive and antitransitive, which means that for each pair of syntactically connected lexemes, one and only one can be governor and one and only one can be dependent, and that a lexeme governing another lexeme cannot govern the dependent(s) of the latter. Two other important properties are: (4) the connectedness of the syntactic tree and (5) the uniqueness of the governor, meaning that each lexeme but the root has exactly one governor.<sup>1</sup>

SSyntSs captures fine-grained grammatical functions of the lexemes in a sentence. The repertoire of SSyntS functions is considerably more detailed than the repertoire in PDT and AnCorra, which introduce only the main grammatical functions (*subject, object, adverbial, apposition, etc.*) and a number of punctuation and sentence markup tags, and even considerably more detailed than Talbanken05 (Nivre *et al.*, 2006), whose level of detail is mainly due to the distinction of morpho-syntactic categories involved in dependencies. Consider, for illustration, a sample SSyntS in Figure 1:

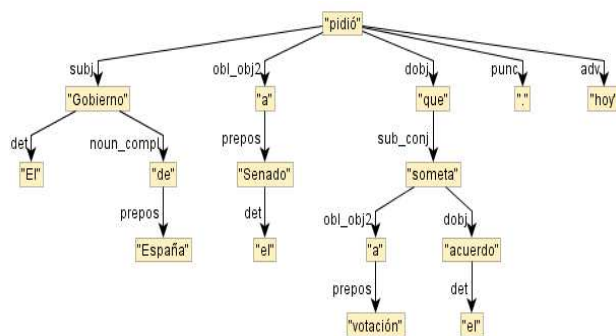


Figure 1: A sample SSyntS

<sup>1</sup> The root has, by definition, no governor.

The SSyntS represents the sentence *El Gobierno de España pidió hoy al Senado que someta a votación el acuerdo*, lit. ‘The Government of Spain asked today to-the Senate to submit to vote the agreement’.

(Mel'čuk, 2003) contains a preliminary set of SSyntS relations for English, which we used as inspiration for our own set of grammatical functions in Spanish and other languages we worked with.

#### 3.2 A proposal of an annotation schema

Figure 2 displays our hierarchical annotation schema that is based on a generalization of surface-syntactic dependency relations, mainly of Spanish.

The annotation schema should be seen as being twofold: On the one side, it contains purely syntactic dependencies, organized in three main groups, *complement, non-complement* and *auxiliary*. *Complement* and *non-complement* are subdivided into further subgroups that roughly correspond to what we referred to above as “main grammatical functions”: *subject, direct object, adverbial, modifier, etc.* Those functions represent the first level of detail in our annotation; their number is around 12 (they are presented in capital letters in Figure 2). The second level consists of all children of the first-level functions, and this is where the small differences between languages become visible. For instance, following the example from above, only the “determiner” relation is needed in Spanish, while for Galician, Italian or Catalan, a further relation like “possessive determiner” would be added at this level. For Spanish, we have so far 57 second-level syntactic arcs, which are those that are found in the ready-to-use annotation of the surface-syntactic level.

On the other side, our schema contains dependency tags that reflect fine-grained semantico-syntactic distinctions (see the rightmost framed part in Figure 2) – adding up to a total of 69 dependency tags<sup>2</sup>. For instance, although the reflexive auxiliary *se* displays only one syntactic behavior (in that it acts as a clitic of the verb that governs it), it can reflect a variety of semantic realities. Thus, it can indicate the presence of the passive voice of the verb it is the dependent of, be a marker of reflexiveness, beneficiary, or even emphasis. In other words, a single purely syntactic *reflexive auxiliary* relation corresponds to four semantic subtypes: *passive, direct, indirect, and lexical*, which are needed to reconstruct the semantic valency of the verbal predicate. Another example of this kind is the subset of relations *oblique\_object*:<sup>3</sup> in Spanish, an indirect object of an active verb can be its second, third, or fourth argument (the syntactic subject generally being the first one). The semantic valency slot that is occupied by the object is indicated by the number that follows the relation name *oblique objectival*; the first, second and third object respectively occupy the second, third, and fourth semantic slot in the valency pattern of the verbal predicate.

<sup>2</sup> In the case of semantic annotation, the semantic tags are used instead of the second-level tags to which they are associated.

<sup>3</sup> An oblique object is an object that is pronominalized by an indirect pronoun and introduced by a preposition.

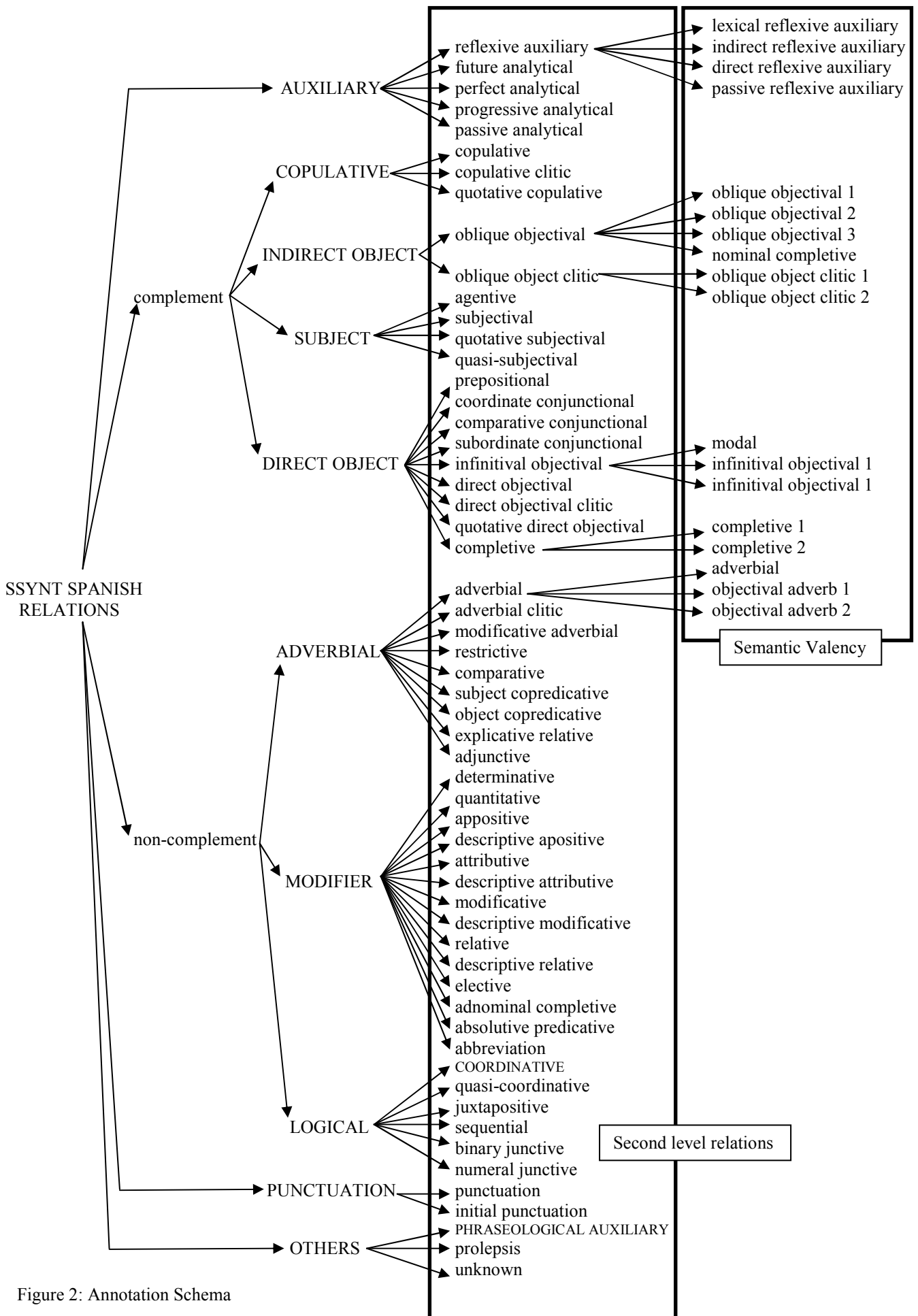


Figure 2: Annotation Schema

These semantico-syntactic distinctions enable us to extract valency dictionaries and eventually deduce deeper, semantically-oriented, annotation schemas, contributing thus to the creation of a multilevel (surface-syntactic, deep-syntactic and semantic) annotation of corpora (see also Section 4).

The schema presented in Figure 2 is not the first attempt to define this kind of hierarchy. For instance, DeMarneffe *et al.* (2006) suggest a hierarchy which can be used for annotating dependency treebanks converted from constituency treebanks – such as, e.g., the Penn treebanks. They use 48 relations, but many of them reflect categorial rather than purely syntactic distinctions. As a consequence, the accuracy of the annotation obtained from such a hierarchy can only be limited. Bolshakov (2002) presents a classification of dependency labels for Spanish which, as our schema, follows Mel’čuk’s (2003) model. However, Bolshakov’s classification is based almost exclusively on semantic valency criteria. As a result, it does not clearly separate syntactic and semantic relations.

### 3.3 Applying the annotation schema

Currently, we are in the process of annotating a number of corpora in accordance with the annotation schema presented in the previous subsection.

Our corpus of Spanish is the AnCora corpus. The first version of the SSynt treebank has been obtained by an automatic mapping of about 3500 sentences of the original AnCora annotation (Martí *et al.* 2007) to the SSynt-level annotation. The obtained annotation has been revised manually in a first iteration. Right now, we are in the process of the second (and final) revision, which is performed by two expert annotators. Since there is only a very small share of really problematic cases, two experts suffice to reduce the inconsistencies in the corpus to the minimum.

The tree bank of 3,500 sentences will serve us as a gold standard reference, which will be extended either by the entire AnCora corpus (about 14,000 sentences) or by another newspaper corpus.

We follow the same strategy as described above to obtain an annotated Swedish corpus. In this case, we started from the Talbanken05 corpus (Nivre *et al.*, 2006). The automatic mapping of the original annotation to our annotation has already been done. The manual revision iterations are about to start.

At the University of La Coruña, the annotation of a mid-size Galician corpus has been recently launched; the findings gained there continuously contribute to the revision and improvement of our annotation schema. Furthermore, we are currently about to annotate manually a Finnish corpus from the start.<sup>4</sup>

Figures 3 and 4 show an example for two of the languages mentioned above, Swedish and Finnish (a SSyntS for Spanish can be found in Section 3.1).

So far, our experience with the proposed annotation schema has been very positive. Even for languages as different from Spanish as Finnish, the adaptation of the dependency relation tag set did not pose particular problems. This offers certain evidence that the annotation schema is applicable to languages typologically different from Spanish, and, more generally, from Romance languages.

When starting with the annotation of a corpus in a new language, we begin with a reduced set of around 12 “first level” functional tags (in capital letters in Figure 2; see also next subsection) and extend this set with as many “secondary” relations as we think is necessary while looking into written data and academic grammars, using the same criteria as the ones we used for Spanish relations.

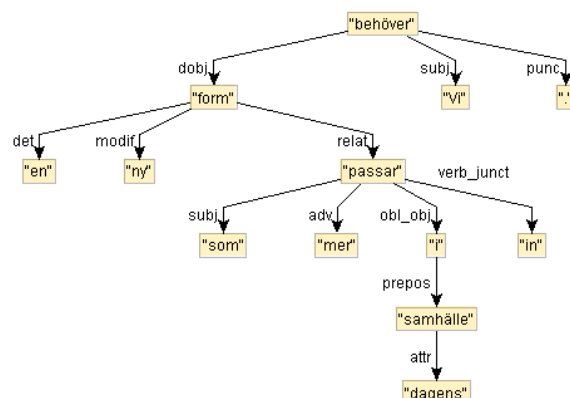


Figure 3: A sample annotation of a Swedish sentence  
Vi behöver en ny form som mer passar in i dagens samhälle.  
*We need a new form that more fits in to today's society.*

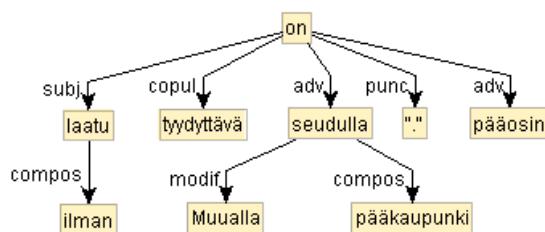


Figure 4: A sample annotation of a Finnish sentence  
“Muulla pääkaupunkiseudulla ilmanlaatu on pääosin tyydyttävä”.  
*In other parts (of)metropolitan\_area air\_quality is in general satisfying.*

## 4. From one-level to multilevel annotation

An increasing number of corpora are annotated not only with syntactic, but also with semantic information (cf., e.g., AnCora and PDT). Our goal is to annotate corpora with at least three types of structures from the multistratal MTT model (cf. Figure 5): surface-syntactic, deep-syntactic (DSyntS) and semantic (SemS). A DSyntS is a dependency tree where the nodes are deep lexical units (LUs)<sup>5</sup> and the arcs are universal

<sup>4</sup> The annotation of the Finnish corpus is done in the framework of the European project PESCaDO (FP7-ICT-248594).

<sup>5</sup> The set of deep LUs of a language L contains all LUs of L— with some specific additions and exclusions. Added are two types of “artificial” LUs: (i) symbols of lexical functions (LFs), which are used to encode lexico-semantic derivation

dependency relations that mark the actants of a predicative LU (I, II, III, ...), attributes (ATTR), appenditives (APPEND) and coordinations (COORD); cf. a sample DSyntS in Figure 6. A SemS is a predicate-argument graph with nodes labelled by semantemes and arcs labelled by the ordinal numbers of the argument relations (ordered in ascending degree of obliqueness); cf. an example of a SemS in Figure 7.

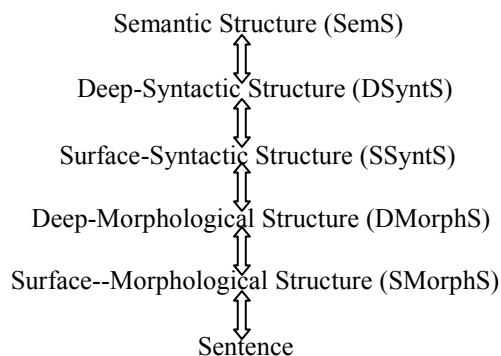


Figure 5: The MTT multi-stratal model

Thanks to the high degree of detail of the SSyntS, we are able to speed up the annotation with DSyntS and SemS. In particular, as already mentioned, our SSynt annotation subclassifies syntactic dependencies with respect to different actants. Consider, for illustration, the predicative lexemes *pedir* ‘ask’, and *someter* ‘put’<sup>6</sup> in Figure 1, which is annotated with the extended set of arcs:

- *pedir* has an actant 1 (‘subjectival’), an actant 2 (‘direct objectival’), and an actant 3 (‘oblique objectival 2’);
- *someter* has an actant 2 (‘direct objectival’), and an actant 3 (‘oblique objectival 2’); Spanish being a pro-drop language, the first actant does not have to be realized.

As mentioned in Section 3.2, an oblique object can be the second, third, fourth, etc. actant of the verb. Although all oblique objects behave the same way from the syntactic point of view and one would thus assume that there is no reason to have different edge labels at the SSynt-level, their differentiation as *obl\_obj1*, *obl\_obj2*, *obl\_obj3*, etc. (cf. Section 3.2) facilitates the association of each of them to a specific semantic valency slot, and, subsequently, to a specific deep-syntactic (*II*, *III*, *IV*, ...) or semantic (2, 3, 4, ...) arc label.<sup>7</sup> Hence, for instance, in the case of the SSyntS

that we have been using as an example in Section 3.1, we can readily derive a DSyntS shown in Figure 6 using a simple structure mapping grammar: all governed prepositions have been removed and the determiners that do not convey any other meaning than mere definiteness have been eliminated. The morpho-syntactic information (such as, e.g., verbal tense, definiteness of nouns, etc.) is encoded in terms of attribute/value structures assigned to the corresponding nodes of the DSyntS.

The DSyntS in Figure 6 is correct, although not necessarily complete after the automatic projection from SSyntS since this projection does not identify LFs, which form part of the DSyntS node label alphabet (cf. Footnote 5), such that they must be introduced into the resulting DSyntS manually;<sup>8</sup> however, the total amount of work necessary for the compilation of a DSyntSs corpus remains rather low once the SSyntSs corpus has been built.

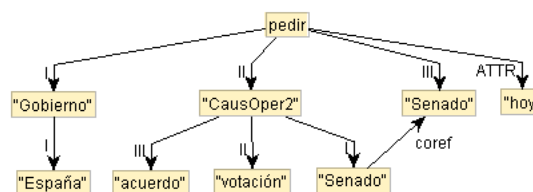


Figure 6: DSyntS for SSyntS in Figure 1

A stage further towards abstraction is the annotation of the corpus with *semantic* structures (SemSs) as shown in Figure 7. Again, once the DSyntS has been reviewed, the derivation of the associated SemS is straightforward and an automatic mapping gives good results.

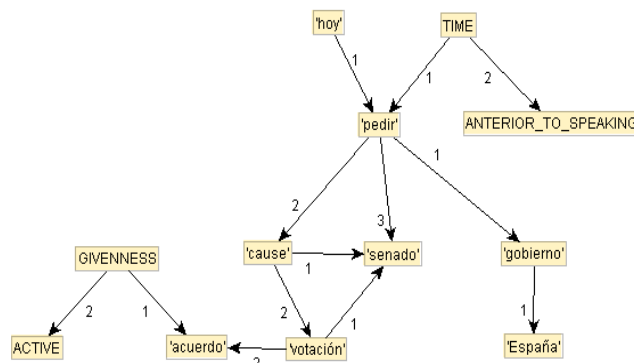


Figure 7: Automatically derived SemS

As Figure 7 shows, in contrast to the “shallow” semantic annotations as seen for instance in Propbank (Palmer *et al.*, 2005), SemSs are genuine connected predicate-argument structures. The nodes in a SemS are thus of semantic rather than of syntactic nature (they are *semantemes* in the MTT terminology). That is, all nodes

and lexical co-occurrence (Mel’cuk, 1996); (ii) fictitious lexemes which represent idiosyncratic syntactic constructions of L. Excluded are: (i) structural words, (ii) substitute pronouns and values of LFs.

<sup>6</sup> *Someter* is not always translated as ‘put’; here, it is, actually, the value of a lexical function (CausOper2 in Figure 6).

<sup>7</sup> It is important to repeat (see Section 3.2) that in the final version of the surface-syntactic corpus, all semantically motivated relation tags will not appear. Rather, they will be substituted by their respective mother tags (cf. Figure 2),

which are strictly syntactic (called “second level relations” in Section 3.1).

<sup>8</sup> The work on the automatic recognition of LFs in corpora as discussed, e.g., in (Wanner *et al.*, 2006) is still too preliminary to be used for automatic high quality annotation.

of the DSyntS – including the feature-value structures attached to the individual DSynt nodes (such as, e.g., tense) – correspond to fragments of a predicate-argument configuration.

To be noted is also a peculiarity of our current semantic annotation, which will be changed in the progress of our annotation initiative: Figure 7 shows that we also annotate as part of the SemS aspects of the information structure. Thus, the definite determiner *el* ‘the’ (*acuerdo*), which appears in the SSyntS as a node label and in the DSyntS as an attribute/value pair on the node of the noun, signals, according to Gundel’s (1988) hierarchy of Givenness, that *acuerdo* is “activated in the memory of both the Speaker and the Addressee”. In Figure 7, this is expressed by a GIVENNESS predicate whose second argument is ACTIVE<sup>9</sup> (to distinguish between genuine semantemes and semantemes that express “meta” information such as GIVENNESS, the former are written in single quotes and the latter in capital letters). In the final version of our annotation, the information structure will be annotated as a meta-structure of SemSs. In any case, the presence of information structure categories (such as GIVENNESS) at the semantic level of annotation illustrates the fact that the meaning-oriented nature of SemSs enables semantic inferences that syntactic structures do not directly allow.

## 5. The costs of the annotation

The cost of the annotation of corpora according to the schema outlined in the previous sections is acceptable. According to our estimations and based on the work that has been done so far, an adequately trained full time annotator is able to annotate with good quality fifty sentences or revise at least a hundred structures per day, using the second-level arcs shown in Figure 2. Theoretically, one annotator should then be able to annotate around 1,100 sentences per month of work (22 days/month), excluding revision cycles. Taking into account the repartition of the tasks and the discussions between the annotators, it seems reasonable to foresee, for a group of 3 annotators, an average of 2,000 completely annotated and revised structures per month. SSynt annotation is more costly, but thanks to the extended set of SSyntRels, the annotation of the other levels (DSynt and Sem) is much faster (cf. the argumentation in Section 4).

In fact, the general cost of the annotation depends on the choice of the set of arc labels: apparently, with more general relation labels, the cost is lower than with more specific relation labels.

To decide which level of annotation granularity is adequate, we need to assess, once again, what the corpus is annotated for. For instance, for training of a syntactic parser, no semantic annotation is needed, and

<sup>9</sup> Strictly speaking, the information on Givenness should be captured in a separately annotated information structure. However, given that we are not yet in the process of annotating our corpus with information structure, we allow ourselves to incorporate this information into SemSs.

even with a rather reduced set of SSynt relation labels, the results show to be satisfying. Also, the size of the annotated corpus may be smaller than, for instance, for corpus-based generation. In order to obtain a clearer picture with respect to the required size, we performed some small experiments with Bohnet’s (2009) dependency parser. The following table summarizes the results.

# of sentences in training set	470 (test set: 60)	3,500	20,000
Overall precision on labels and dependencies	76% (06/2009)	81% (prevision)	88% (prevision)

In contrast, if the application in question requires more than a merely syntactic annotation, it is more appropriate to invest more effort at the beginning in order to save time on other tasks (cf. the derivation of DSyntSs and SemSs elaborated on in the previous section and of generation resources discussed in the next section). The hierarchical annotation schema we propose offers the needed flexibility and helps to tune the cost of the annotation.

Of course, the costs of the SSynt annotation will also largely vary between different languages. For languages with a higher idiosyncrasy of the syntax, the costs will be higher. The adaptation of the annotation schema to other languages also largely depends on how closely related these languages are to the languages for which the schema has already been adjusted. An empirical study of the language’s syntax is the best way to adapt the set of relation tags.

## 6. Using the annotation to derive resources

As mentioned in the Introduction, one of the goals of our annotation schema is to support the derivation of resources for natural language generation. This includes lexical resources, and generation grammars. A generation grammar maps, generally speaking, a given input structure (most often, an abstract conceptual or semantic representation) to a well-formed sentence (or to a coherent and cohesive sequence of sentences, i.e., a text). In the multistratal MTT-framework as displayed in Figure 5, a single generation grammar maps a structure at a given level  $L_i$  ( $i = \text{semantic, deep-syntactic, ...}$ ) to an equivalent structure at the adjacent level  $L_{i+1}$ .

The main lexical information needed in such a generation model consists of: (i) the projection of the semantic valency structure of a given LU to its syntactic valency pattern, (ii) the subcategorization information of an LU.

A simple grammar defined in the development environment MATE (Bohnet et al., 2000; Bohnet and Wanner, 2010) extracts for the verb *pedir* ‘ask’ this



lexical information from the SSyntS in Figure 1 in terms of the following lists of attributes:<sup>10</sup>

```
pedir { dpos=V
  I_dpos=N I_spos=proper_noun I_rel=subj
  II_dpos=V II_spos=verb II_rel=dobj
  II_prep="que" II_mood=SUBJ
  III_dpos=N III_spos=proper_noun
  III_rel=obl_obj2 III_prep="a" }
```

The *Pedir*-attributes consist of four blocks of attribute/value pairs: the first block concerns *pedir* itself; the other three concern its actants. The *pedir*-block contains its deep part-of-speech (*dpos*). The block of the first DSynt actant contains its deep part-of-speech (noun, N) and its surface part-of-speech (*spos*): *proper\_noun*. Furthermore, it is linked by the relation “subj” to its governor. The block concerning the second DSynt actant occupies the third and fourth lines: it is a verb linked to *pedir* by a direct objectival relation ‘dobj’, such that this verb is introduced by *que* ‘that’ and is in the subjunctive mood (‘SUBJ’). Similarly, the last two lines present the information block concerning the third DSynt actant of *pedir*.

Any government pattern of any lexical unit can be stored in the dictionary, with all properties of the governed element that are required by the governor (Part-Of-Speech, mood, finiteness, etc.), and so on.

Apart from being needed in generation, such a dictionary helps in the derivation of DSyntSs from SSyntS since one of the main challenges of the SSynt-DSynt transition is to distinguish semantic prepositions from syntactic (*governed*) prepositions. Indeed, only the latter are stored in the entry for their governor (as it is the case of *a* on the last line of the figure above), whereas the former appear in the DSyntS.

For the derivation of the generation grammars we experiment with machine learning techniques. The goal is to learn from aligned structures at two adjacent levels of annotation minimal mapping rules. This is why choosing an annotation strategy that will make easier the annotation of other levels of representation is crucial, and why it is very interesting for us to introduce some semantico-syntactic arc labels on our syntactic annotation.

## 6. Conclusions

We propose a hierarchical dependency structure annotation schema that is more detailed and more flexible than the known state-of-the-art annotation schemata. The presented schema allows us to choose the level of the desired detail of the annotation and to adapt it easily to new syntactic phenomena. Thanks to the inclusion of semantico-syntactic tags, we can annotate a corpus not only with syntactic information, but also with valency information for all valency-bearing lexemes (verbs and nouns, and adjectives) as it is usually found in separate treebanks such as PropBank

and NomBank. Furthermore, this annotation schema facilitates the derivation of deeper annotations, leading to truly multilevel annotated dependency corpora.

## Acknowledgements

Many thanks to our colleagues and friends Igor Mel’čuk, Alicia Burga, Gaby Ferraro, and Anton Granvik for their invaluable contributions to the work presented here. We would also like to thank the three anonymous LREC reviewers for their insightful comments that helped to considerably improve the final version of the paper.

The work presented in this paper has been partially funded by the Spanish Ministry of Science and Innovation and FEDER (EC) under the contract number FFI2008-06479-C02-01 and by the European Commission under the contract number FP7-ICT-248594.

## References

- Ahrenberg, Lars (2007). “LinES: An English-Swedish Parallel Treebank”. In Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA, 2007).
- Apresjan, Ju., *et al.* (2006). “A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects”. In *Proceedings of LREC*. Genova, Italy, 1378-1381.
- Bohnet, B., (2009). “Efficient Parsing of Syntactic and Semantic Dependency Structures”. In *Proceedings of the Conference on Natural Language Learning (CONLL)*, Boulder, 67-72.
- Bohnet, B., A. Langjahr and L. Wanner. (2000). “A Development Environment for an MTT-Based Sentence Generator”. *Proceedings of the First International Conference on Natural Language Generation*, Mitzpe Ramon, Israel, 260-263
- Bohnet, B. and L. Wanner. (2010). “Open Source Graph Transducer Interpreter and Grammar Development Environment”. In *Proceedings of LREC, this volume*. Malta.
- Bolshakov, Igor A. (2002). “Surface Syntactic Relations in Spanish”. In *Proceedings of CICLing 2002*, Mexico City, 210-219.
- Čmejrek, M., *et al.* (2004). “Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation”, In *Proceedings of LREC*, Lisbon, Portugal.
- Cyrus, Lea, *et al.* (2003). “Fuse- a multi-layered parallel Treebank”. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.
- De Marneffe, Marie-Catherine, *et al.* (2006). “Generating Typed Dependency Parses from Phrase Structure Parses.” In *Proceedings of LREC*, Genova, Italy.
- Gundel, Jeanette. K. (1988): “Universals of topic-comment structure”. In M. Hammond, E. Moravczik and J. Wirth (eds.) *Studies in syntactic typology*. Amsterdam: John Benjamins, 209-239.

<sup>10</sup> This list of attributes corresponds to the “syntactic combinatorial zone” of a lexical entry as described in (Mel’čuk, 2006):

- Hajič, J., *et al.* (2004). "Prague Arabic Dependency Treebank: Development in Data and Tools". In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, September 2004, 110-117.
- Hajič, J. *et al.* (2006). Prague Dependency Treebank 2.0, Linguistic Data Consortium, Philadelphia.
- Li, M. *et al.* (2003). "Building A Large Chinese Corpus Annotated With Semantic Dependency". In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 2003*, 84-91.
- Martí, M.A., *et al.* (2007): "Ancora: A Multilingual and Multilevel Annotated Corpus", <http://clic.ub.edu/ancora/publications/>
- Megyesi, B., *et al.* (2008). "Swedish-Turkish Parallel Treebank". In *Proceedings of LREC*, Marrakech, Morocco, May 2008.
- Mel'čuk, I.A. (1988). *Dependency Syntax: Theory and Practice*, Albany, N.Y.: The SUNY Press.
- Mel'čuk, I.A. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: Benjamins.
- Mel'čuk, I.A. (2003). Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, L. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1, Berlin - New York, W. de Gruyter, 188-229.
- Mel'čuk, I.A. (2006). Explanatory Combinatorial Dictionary. In G. Sica (ed.). *Open Problems in Linguistics and Lexicography*. Monza, Italy: Polimetrica, 225-355.
- Mille, S., Burga, A., Vidal, V. and Wanner, L. (2009). "Towards a Rich Dependency Annotation of Spanish Corpora". In *Proceedings of SEPLN'09*, San Sebastian.
- Mitchell P. M., *et al.* (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", In *Computational Linguistics*, 19(2):313– 330.
- Nivre, J., *et al.* (2006). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of LREC*, Genova, Italy.
- Palmer, Martha, Dan Gildea, Paul Kingsbury (2005). "The Proposition Bank: A Corpus Annotated with Semantic Roles", in *Computational Linguistics Journal*, 31:1.
- Wanner L., Bohnet B., Giereth M. (2006): "What is beyond collocations? Insights from Machine Learning Experiments". In *Proceedings of the EURALEX Conference*. Turin.