Tibidabo: a syntactically and semantically annotated corpus of Spanish

Montserrat Marimon

Abstract

In this paper, I discuss the theoretical and practical issues raised in the development of Tibidabo, a syntactically and semantically annotated corpus of Spanish, which has been developed using the DELPH-IN corpus annotation environment and the HPSG-based grammar, SRG. I describe in detail the linguistic analysis that Tibidabo provides, focusing on various syntactic phenomena that are frequently discussed in the literature and consider alternative analyses proposed in the HPSG literature and in AnCora, another Spanish Treebank.

Keywords: Spanish, Syntactic and Semantic annotation, Treebank.

1. Introduction

Linguistically interpreted natural language corpora – treebanks – constitute a crucial resource both for theoretical linguistic investigations about language use and for practical Natural Language Processing (NLP) purposes. Thus, in past decades, there has been an increasing interest in the construction of treebanks that provide various types of linguistic annotations, like constituent structure, dependency structure or semantic structure (or those that combine two or even all three of these structures), and both theory-neutral and theory-grounded treebanks have been developed, on small and large scales, for a great variety of languages. Descriptions of annotated corpora that are available can be found in Abeillé (2003) and in the proceedings of the annual International Workshop on Treebanks and Linguistic Theories (TLT).

In the case of Spanish, the reference treebank is AnCora (Taulé et al., 2008) – a corpus of about 500,000 words (about 17,000 sentences). The annotation process was carried out in layers: first, part-of-speech tags were (fully) automatically annotated; then, named entities were marked on top of that; next, syntactic constituents were semi-automatically built and

1 Universitat Pompeu Fabra, Roc Boronat 138, 08018-Barcelona, Spain.
2 This work was funded by the Ramón y Cajal program of the Spanish Ministerio de Ciencia e Innovación.
syntactic dependencies were obtained from them by a conversion process; and finally, semantic roles and co-references were manually included in the annotations. Previously, Moreno et al. (2000) presented a smaller treebank of 1,500 sentences which were annotated, mainly by hand, with part-of-speech or phrasal categories, syntactic and semantic features (e.g., subj, obj1 and time), and constituent nodes.

This paper considers the theoretical and practical issues raised in the development of Tibidabo, a syntactically and semantically annotated corpus of Spanish. The basis of the corpus is a set of 4,000 sentences taken from newspaper texts borrowed from AnCora. Although AnCora already provides syntactic and semantic annotations, these are mainly hand-created. I believe that providing a corpus consisting of the same text annotated under a different paradigm – where the annotation criteria are enforced by a deep analysis lexical grammar instead of human annotators – may be a valuable resource for research. Such a corpus can be useful in studying the variability of human annotation, the ability of machine algorithms to capture the structures annotated in each approach, and the study of how different linguistic criteria can be mapped to one another, among many other possibilities.

The corpus has been annotated with a grammar of Spanish which covers a wide range of structures, based on Head-driven Phrase Structure Grammar (or HPSG; see Pollard and Sag, 1987, 1994) and implemented in the Linguistic Knowledge Builder (LKB) system (Copestake, 2002): Spanish Resource Grammar (SRG; see Marimon, 2013); and the corpus annotation environment of Deep Linguistic Processing with HPSG Initiative (DELPH-IN), also used in several treebank projects within this international initiative (Branco et al., 2010; Flickinger et al., 2012; Hashimoto et al., 2007; Kordoni and Zhang, 2009; Marimon et al., 2012; Marimon et al., 2014; and Oepen et al., 2002).

Previously published works on treebank projects using the DELPH-IN framework focussed on a detailed presentation of the annotation environment and a quantitative assessment of the evolution of the Redwoods treebank in three development phases, together with a brief overview of the linguistic representation formats provided by Redwoods (Oepen et al., 2002). On the other hand, Hashimoto et al. (2007) presented a large-scale and detailed database of lexical types in Japanese constructed from the Hinoki Treebank; Kordoni and Zhang (2009) briefly presented how they used the DELPH-IN framework and English Resource Grammar (ERG) (see Flickinger, 2002) to annotate the Wall Street Journal section of the PennTreebank; Branco et al. (2010) briefly reported on several automatic and manual procedures to obtain a POS-tagged corpus, a constituency TreeBank, a DependencyBank, a PropBank, and a LogicalFormBank out of CINTIL DeepGramBank developed with Portuguese Resource Grammar (Branco and Costa, 2008); Flickinger et al.

---

3 See: http://www.delph-in.net/
(2012) presented the initial version of ParDeepBank, a parallel treebank for English, Portuguese and Bulgarian selected from the PennTreebank data and translated to Portuguese and Bulgarian; and, finally, Marimon et al. (2014) presented an ensemble parse approach to detecting and selecting high-quality analyses output by SRG that aimed to increase the degree of automation in the annotation process when constructing a large treebank that contains a technical corpus (the IULA Spanish LSP Treebank, see Marimon et al., 2012).

This paper focuses on the linguistic annotations that Tibidabo provides for several syntactic phenomena that are dealt with innovative solutions w.r.t. HPSG theoretical proposals and/or other Spanish treebanks. I discuss alternative analyses proposed in the HPSG literature, and, since the basis of the annotated corpus is a set of sentences borrowed from AnCora, I also compare the annotations that these two corpora provide for the same phenomena.4 In Section 2, I go on to provide a detailed description of the annotations of syntactic phenomena and I provide a conclusion in Section 3.

2. Representation of syntactic phenomena

The linguistic annotation of Tibidabo includes the syntactico–semantic analysis that the DELPH-IN framework currently displays in the form of a binary branching phrase structure tree, which encodes constituency, and a Minimal Recursion Semantics (MRS) semantic representation (Copestake et al., 2005) which encodes structural semantics (i.e., predicate–argument relations) – both obtained from a complete syntactico–semantic analysis that is represented in a parse tree with HPSG-typed feature structures at each node. In addition, the treebank provides the annotation of dependency structures, in the CoNLL format (Buchholz and Marsi, 2006), that is generated from the derivations trees produced by the LKB system.5

This section describes in detail the syntactic and semantic annotations that Tibidabo provides for various syntactic phenomena that are frequently discussed in the literature, like argument/adjunct distinction, null elements and co-ordination, providing examples from the corpus. I discuss alternative analyses proposed in the HPSG literature, and compare the dependency structures that Tibidabo and AnCora provide for the same phenomena.

2.1 Arguments and adjuncts

---

4 Marimon (2012) briefly presented the use of SRG and the DELPH-IN framework to annotate the Spanish corpus.

5 See Marimon et al. (2014) for a brief description of the procedure.
Complements and adjuncts in SRG are modelled in terms of the traditional HPSG complement/adjunct distinction, as opposed to the so-called ‘adjunct-as-complement’ approach in HPSG (Bouma et al., 2001; and Przepiórkowski, 1999, among others) where, based on a variety of empirical phenomena (e.g., adjunct extraction and case assignment), adjuncts are added to the verb’s argument structure at the lexical level. In SRG, heads select their complements using the COMPL(EMENT)S valence feature and they are combined by the head-complement phrase structure rule, whereas adjuncts are treated as functors that select their modificees by a head feature, MOD(IFIES), and they are combined by the head–adjunct phrase structure rule.

The distinction between arguments and adjuncts in Tibidabo is made in the annotations of both the dependency structure and the semantic structure. No distinction is made between them in the constituent structure depicted in the phrase structure trees, where atomic syntactic labels for sentence constituents, such as ‘S’, ‘VP’ and ‘V’, are used and where modifiers may be adjoined to both lexical and phrasal nodes, or they may even appear between two complements, as in Example 1, since Spanish is a free constituent order language.

(1) Podríamos aplicar la frase, en un sentido más noble, a Tommy Romminger.

Can.COND.1.PL apply.INF the phrase, in a sense more noble, to Tommy Romminger.

['We could apply the phrase, in a more noble sense, to Tommy Romminger.‘]

As in AnCora, the dependency labels in the dependency structure of Tibidabo distinguish between syntactic arguments and adjuncts of the verb or verb phrase, and they also categorise the different types of verbal arguments by means of standard syntactic function labels, (e.g., Direct Object, Indirect Object, and By-Agent Complement). In the dependency structure, Tibidabo also makes the argument/adjunct distinction inside NPs, APs, PPs and ADVPs, where arguments are labelled as COMP and adjuncts as MOD. This distinction within non-verbal phrases is not made in AnCora, where only syntactic categories are annotated. Figure 2 shows the dependency labels that are assigned to the constituents of Example 1, shown in Figure 1.

==Insert Figure 1 about here==
==Insert Figure 2 about here==

Tibidabo also provides a clear distinction between arguments and adjuncts in the semantic structure, where adjuncts make their modificees their arguments, whereas these statuses are reversed for arguments. This is
illustrated in Figure 3, which shows the semantic structure of Example 1, where the arguments of the verb *aplicar* ('to apply') – x8, x7 and x6 – are labelled as ARG1, ARG2 and ARG3 in the relation of the verb, whereas the verb’s event variable – e5 – identifies the ARG1 of the preposition’s relation heading the PP that modifies the verb. In AnCorA, arguments and adjuncts are also distinguished in the semantic dependency labels. Here, semantic dependency labels for arguments are given a numerical argument (i.e., ARG0, ARG1, ARG2, ARG3 and ARG4) plus a thematic role label (e.g., agent, patient and beneficiary); adjuncts take the tag ArgM plus a label (e.g., time and manner; see Hajic *et al.*, 2009, for details).

2.2 Null subjects

Being a pro-drop language, Spanish frequently omits explicit subjects in finite clauses, as in Example 2, where the information about the person and number of the subject is encoded in the affix of the verb.

(2)  *Tiene razón.*

Have.PRES.3.SG right.

[‘(She/He) is right.’]

Null subjects in SRG are accounted for by a unary branching rule that discharges the optional SUBJ(ECT) element of a VP-node. SRG, therefore, does not assume the HPSG version of Ginzburg and Sag (2000), where null subjects are viewed as a variation of the so-called Argument Realization Principle in which the least oblique argument in the ARG-ST feature is not mapped to a VALENCE feature. Figure 4 illustrates the constituent, dependency and semantic structures that the treebank provides for null subjects with Example 2.

As can be observed, in the constituent structure, null subjects in the treebank are represented by a unary branching VP-node below the S-node. Next, in the dependency structure, where only dependencies between actual words in the sentence are marked, no elliptical element with the syntactic function subject is inserted. Note that here the treebank differs from the analysis followed in AnCorA, where null subjects are represented by null elements in the column-based format (i.e., lines where the fields for word form, lemma and fine-grained part-of-speech tag are empty). Finally, in the

---

6 Note that in the predicate-argument relations, MRS assigns the label ARG0 for the index of the predication, then, arguments are identified with the numbered arguments of their head (ARG1, ARG2, ARG3).

semantic structure, null subjects are represented by an uninstantiated external argument that encodes the agreement features of the verb, person and number, so, in the example the ARG1 of the verb’s relation is valued as third-person singular (3SG) for the PN (PERSON and NUMBER) feature.

2.3 Elliptical NPs

Different alternatives have been discussed in the HPSG literature applied to the analysis of null-headed NPs, as in Example 3, assuming: (i) null elements selected by the determiners and adjectives with which they combine (Nerbonne and Mullen, 2000); (ii) category-changing rules applied on adjectives (i.e., nominalisation rules), (Winhart, 1997); (iii) under-specified sub-types of HEAD which can head a determiner phrase (DP),8 (Nerbonne and Mullen, 2000); or (iv) a unary syntactic version of the ‘head-functor schema’ in which the SYNSEM of the mother node is partly shared with the SYNSEM of the functor’s SELECT value and the INDEX of the mother node is identified with the ARG0 of a noun–ellipsis relation (Branco and Costa, 2006).9

(3) El más fuerte conquistó a la más bella.

The most strong captivate.PAST.3.SG to the most beautiful.

[‘The strongest one captivated the most beautiful one.’]

None of these approaches has been adopted in SRG, since both null elements and unary rules applied on a nominal adjunct would have negative effects on parsing efficiency. In SRG, elided nominal heads are treated by a binary rule that allows a headless NP by combining a determiner and a nominal modifier. In this rule, the SYNSEM value of the elided head is constrained in the SYNSEM of the adjunct’s MOD value. This rule also introduces an elliptical noun relation (i.e., the ‘elliptical_n’ relation) whose ARG0 identifies the INDEX of the mother node. Figure 5 illustrates the constituent, the dependency and the semantic structures that the treebank provides for elliptical NP with Example 3.

First, in the constituent structure, headless NPS are represented by a NP node with no descendant N-node. Then, when generating the dependency structure, no elliptical element is inserted for marking the elided element, and the treebank follows the standard strategy that is

---

8 The same constituent as NP, but considering the determiner to sub-categorise for an N and resulting in a maximal DP.

9 This analysis is based on a uniform treatment of specifiers and adjuncts under the head-functor schema, which eliminates the specifier adjunct distinction by considering both functors which select their head via the feature SELECT.
usually employed to deal with empty heads in dependency corpora, also
followed in AnCora: the modifier of the elided head is chosen to become
the head and it is labelled with the syntactic function of the elided head; so,
in the example, the first modifier is labelled as SUBJ and the second one as
COMP of the preposition which takes the syntactic function DO. Finally, the
semantic structure shows the ‘elliptical_n_rel’ relation introduced by the
rule dealing with elliptical NPs. Note that the ARG0 of this relation is co-
indexed with the ARG1 of the adjunct, and, when the headless NP is an
argument of the verb, it is co-indexed with an argument’s value of the
verb’s relation; so, in the example, the ARG0 of the first elliptical relation is
correlated with the ARG1 and the ARG0 of the second one is co-indexed
with the ARG2.

2.4 Elliptical finite verbs

This section describes the annotations that Tibidabo offers for two types of
co-ordinated constructions where the verb is missing from the second
conjunct: sentence gapping and conjunction reduction (or argument cluster
do-ordination), illustrated in Examples 4 and 5, respectively (these
examples are taken from Brucart (1999) where the reader can find a
detailed description of elliptical constructions in Spanish).10

(4) a. *Luis fue a París y María a Londres.*
   Luis go.PAST.3.SG to Paris and María to London
   ['Luis went to Paris and María to London.‘]

   b. *Luis fue a París para participar en un congreso y María para
      asistir a una reunión de psicólogos.*
   Luis go.PAST.3.SG to Paris to participate in a congress and María
to attend to a meeting of psychologists.
   ['Luis went to Paris to participate in a congress and María to attend
a psychologists meeting.‘]

(5) a. *Luis enviaba rosas a Elena y gardenias a Elisa.*
   Luis send.PAST.3.SG roses to Elena and gardenias to Elisa.
   ['Luis sent roses to Elena and gardenias to Elisa.’]

   10 The current version of SRG only covers the simplest-case examples of sentence gapping
and conjunction reduction constructions (i.e., those constructions in which the conjoined
constituents are the complements and/or adjunct of the missing verb). Other elliptical
constructions such as pseudo-gapping, sluicing and right-node raising are not covered either
by SRG.
No agreement has been reached on the structure of elliptical co-ordination phenomena in HPSG (see, for example, Yatabe, 2012, for a comparison of alternative analyses). Much recent theoretical work on co-ordination described in the HPSG literature proposes to use linearisation domains (Reape, 1994) to capture a variety of elliptical co-ordination phenomena, such as conjunction reduction, gapping, right-node raising constructions, and unlike-category co-ordination (Beavers and Sag, 2004; Chaves, 2005; Crysmann, 2003; and Yatabe, 2003). Beavers and Sag (2004) propose a unique co-ordination construction that uses the DOM(AINS) list device, originally devised to allow elements in sentences to change their position, to allow shared elements to receive no prosodic expression (i.e., to enable some elements in the daughter’s DOM lists to be absent in the mother’s DOM lists). Mouret (2006), however, focusing on syntactic issues in French, proposes an alternative approach to conjunction reduction that avoids ellipsis by enabling non-standard constituents, represented as non-headed constituents, to be conjoined in the scope of some shared predicate. And, more recently, the linearisation-based ellipsis approach has also been dismissed by Levine (2011), who discussed several examples that, going beyond the simplest-case examples, did not admit any ellipsis-based solutions.

In SRG, sentence gapping constructions are dealt with a set of quaternary rules where the first daughter is the full VP, the second daughter is a conjunction lexical item, and the remaining two daughters are the subject and complement or modifier of the missing verb. This set of rules builds up an MRS representation in which the elided verb’s relation, which is always the same one as that which appears in the first conjunct, is restored and its arguments and adjunct are identified. To this end, the grammar assumes a verbal head feature, called LSYNSEM, structure shared with the verb’s SYNSEM, and, in the elliptical constructions, with the head LSYNSEM feature of the mother node. So, in these rules, both the syntactic dependents and the elementary predication of the missing verbs are identified by the mother’s LSYNSEM feature.

Figure 6 illustrates the multi-daughter constituent tree that the treebank provides for sentence gapping, with Example 6 from the corpus, together with its corresponding dependency and semantic structures.

(6) Zarrabeitia puso la rebeldía y Delgado la gallardía.

11 Similarly, for conjunction reduction constructions, SRG uses a set of the multi-daughter rules where first daughter is the full S, the second daughter is a conjunction lexical item, and the remaining two daughters are the complements and/or modifiers of the missing verb.
Zarrabeitia put.PAST.3.SG the rebellion and Delgado the gallantry.

[‘Zarrabeitia put the rebellion and Delgado the gallantry.’]

==Insert Figure 6 about here==

In the dependency structure, the co-ordinating conjunction represents the missing verb and inherits all its properties, such that subjects, complements and adjuncts are linked to it and labelled as SUBJ-GAP and COMP-GAP, and MOD-GAP. This analysis clearly differs from the one proposed in AnCora, where, in the same example, the first element in the second conjunct (i.e., the subject of the missing verb) carries the label S and it is analysed as a dependent of the verb in the first conjunct, and the second element (i.e., the DO) carries the label SN and it is analysed as a dependent of the first element of the second conjunct (i.e., the subject).\(^{12}\)

Finally, the semantic structure shows how the semantic relation of the elided verb is restored and its arguments are identified in MRS.

2.5 Co-ordination

The implementation of co-ordination in SRG is based on the implementation of co-ordination in LinGO Grammar Matrix, which adopted a binary branching approach (as opposed to a flat structure) and proposed a binary semantic relation that co-ordinates a right and a left argument (Drellishak and Bender, 2005).

Tibidabo, thus, provides a binary right-branching analysis to co-ordination in the constituent structure, as illustrated in Figure 12, with the phrase-structure tree of Example 7.

(7)  
\textit{Aumentar la tolerancia y la radicalidad.}

Increase.INF the respect and the radicalism.

[‘To increase respect and radicalism.’]

==Insert Figure 7 about here==

At this level of the structure, specifiers, arguments and/or adjuncts that are shared by both phrasal and lexical conjuncts of a co-ordination are structurally linked to the co-ordination. This is illustrated in Figure 8, which shows the constituent structures that Tibidabo provides for conjuncts that share a specifier (as in Example 8a), conjuncts sharing an adjunct (as in

\(^{12}\) Alternative analyses found in dependency treebanks are: (i) to choose one of the dependents of a headless phrase to become the head, and (ii) to introduce a zero word form (i.e., a phonetically empty head) into the dependency tree to provide an attachment point for the dependents of the ellipsis.
Example 8b) and conjuncts sharing an argument (as in Example 8c).

(8) a. *Los compañeros y compañeras.*
The.MASC.PL colleagues.MASC.PL and colleagues.FEM.PL
[‘The colleagues and colleagues.’]

b. *Un bombo y un parche rotos.*
A drum and a drumhead broken.MASC.PL
[‘A broken drum and a broken drumhead.’]

c. *Desalojo y derribo de los edificios.*
Evacuation and demolition of the buildings
[‘Evacuation and demolition of the buildings.’]

Figure 9 illustrates the semantic representation that the corpus provides for co-ordinated phrases with Example 7. Here, L-INDEX and R-INDEX identify the left and right arguments, respectively, and C-ARG identifies the INDEX of the co-ordinated phrases as well as the ARG2 in the verb’s relation.\(^\text{13}\)

In the dependency structure (Figure 10), the first conjunct is the head of the other elements, which are organised in a chain. This approach also differs from the one followed in AnCora, where the first conjunct is also treated as the head of the co-ordinated structure, but both the conjunction and the second conjunct are dependents of the first conjunct.\(^\text{14}\)

\(\text{3. Conclusions}\)

This paper presents Tibidabo: a syntactically and semantically annotated corpus of Spanish, containing 4,000 sentences, developed using the DELPH-IN framework and the Spanish HPSG-grammar SRG. Tibidabo is annotated with constituent structure, semantic structure encoded in MRS, and dependency structure depicted in the representation proposed by

\(^\text{13}\) In non-NP phrase co-ordination, the EP of co-ordinating conjunctions, in addition, identifies the handles of the co-ordinated elements by means of the features L-HDNL and R-HNDL.

\(^\text{14}\) A third approach to co-ordination in dependency treebanks is to assume that all conjuncts are headed by the co-ordinating conjunction. See Zabokrtsky et al. (2013) for a taxonomy of various formal means developed for encoding co-ordination in dependency treebanks.
CoNLL shared tasks. I have described in detail the relevant syntactic phenomena, which are frequently discussed in the literature, that illustrate the level of annotation and the linguistic knowledge contained in the corpus, and discuss alternative analyses proposed in the HPSG literature and in AnCora, another Spanish Treebank.

I believe this corpus will constitute a valuable resource both for researchers in linguistics and in NLP. In the future, I plan to use the treebank in research in comparisons of corpus annotative approaches and cross-parser evaluation.

References


Stanford, California: CSLI Publications.


Podríamos aplicar la frase, en un sentido más noble, a Tony Rominger. ['We could apply the phrase, in a more noble sense, to Tony Rominger.']
Figure 2: Dependency structure of ‘Podríamos aplicar la frase, en un sentido más noble, a Tony Rominger.’ ['We could apply the phrase, in a more noble sense, to Tony Rominger.']
Figure 3: Semantic structure of ‘Podríamos aplicar la frase, en un sentido más noble, a Tony Rominger.’ ['We could apply the phrase, in a more noble sense, to Tony Rominger.']
Figure 4: Constituent, dependency, and semantic structures of ‘Tiene razón.’ [‘S/he is right.’]
Figure 5: Constituent, dependency, and semantic structures of ‘El más fuerte conquistó a la más bella’. [‘The strongest one captivated the most beautiful one.’]
Figure 6: Constituent structure with gapping, and its corresponding dependency and semantic structures

Zarrabéitia puso la rebeldía, y Delgado la gallardía.
Figure 7: Constituent structure of ‘Aumentar la tolerancia y la radicalidad.’ [‘To increase respect and radicalism.’]
Figure 8: Constituent structures of conjuncts sharing a specifier, conjuncts sharing an adjunct, and conjuncts sharing an argument.
Figure 9: Semantic structure of ‘Aumentar la tolerancia y la radicalidad.’ [‘To increase respect and radicalism.’]
Figure 10: Dependency structure of ‘Aumentar la tolerancia y la radicalidad.’ ['To increase respect and radicalism.']