# The IULA Spanish LSP Treebank: building and browsing

**Blanca Arias, Núria Bel, Beatriz Fisas, Mercè Lorente**
**Montserrat Marimon, Carlos Morell, Silvia Vázquez, Jorge Vivaldi**

Universitat Pompeu Fabra
Roc Boronat 138, 08018-Barcelona, Spain
{blanca.arias,nuria.bel,beatriz.fisas,merce.lorente,
montserrat.marimon,carlos.morell,silvia.vazquez,jorge.vivaldi}@upf.edu

**Abstract**

This paper presents the IULA Spanish LSP Treebank, a dependency treebank of over 41,000 sentences of different domains (Law, Economy, Computing Science, Environment, and Medicine), developed in the framework of the European project METANET4U. Dependency annotations in the treebank were automatically derived from manually selected parses produced by an HPSG-grammar by a deterministic conversion algorithm that used the identifiers of grammar rules to identify the heads, the dependents, and some dependency types that were directly transferred onto the dependency structure (e.g., subject, specifier, and modifier), and the identifiers of the lexical entries to identify the argument-related dependency functions (e.g. direct object, indirect object, and oblique complement). The treebank is accessible with a browser that provides concordance-based search functions and delivers the results in two formats: (i) a column-based format, in the style of CoNLL-2006 shared task, and (ii) a dependency graph, where dependency relations are noted by an oriented arrow which goes from the dependent node to the head node. The IULA Spanish LSP Treebank is the first technical corpus of Spanish annotated at surface syntactic level following the dependency grammar theory. The treebank has been made publicly and freely available from the META-SHARE platform with a Creative Commons CC-by licence.

**Keywords:** Spanish, Treebank, Dependency

## 1. Introduction

Syntactically annotated corpora *–treebanks–* constitute a crucial resource for research in quantitative and qualitative studies of a wide range of phenomena in lexis, grammar, semantics, discourse, language variation, language change, etc., as well as for natural language processing (NLP) research activities, such as training and evaluation data of data-driven parsing systems. Thus, in the past decades, there has been an increasing interest towards the construction of treebanks that provide constituent structure and/or dependency structure annotations. Descriptions of available annotated corpora can be found in (Abeillé, 2003) and in the proceedings from the annual editions of the International Workshop on Treebanks and Linguistic Theories (TLT).[1]

This paper presents the IULA[2] Spanish LSP[3] Treebank, a dependency treebank of over 41,000 sentences, of different domains (Law, Economy, Computing Science, Environment, and Medicine) and sentence length (ranging from 4 to 30 words), developed in the framework of the European project METANET4U (Enhancing the European Linguistic Infrastructure, GA270893).[4]

The aim of the IULA Spanish LSP Treebank is to contribute to the availability of parsed data in Spanish. Currently the only broadly available treebanks for Spanish are AnCora (Taulé et al., 2008) which contains 500,000 words (about 17,000 sentences) from news papers, and the UAM Spanish Treebank (Moreno et al., 2000), which contains 1,500 sentences.

The treebank has been made publicly and freely available from the META-SHARE platform with a Creative Commons CC-by licence.[5]

In what follows, we describe the methodology that we used to create the resource (cf. section 2.1), the syntactic annotations that the treebank provides (cf. section 2.2), the statistics of the treebank (cf. Section 2.3), and the treebank browser that we have developed to query the annotated corpus (cf. section 3).

## 2. Building the treebank

Figure 1 shows a summary of the methodology that we followed to build the treebank.

The dependency structures were annotated in two steps. First, we used the Deep Linguistic Processing with HPSG Initiative (DELPH-IN)[6] open-source processing framework and the publicly and freely available HPSG-based grammar (Pollard and Sag, 1994) *Spanish Resource Grammar* (SRG) (Marimon, 2013) to parse the sentences. In this first step, we used a MaxEnt based stochastic ranker (Toutanova et al., 2005) to sort the parses produced by the grammar and to reduce the forest to the 500-best trees, from which to select manually the correct parse.[7] [8] Then, we converted selected parses, represented as derivation trees, into dependency structures.

[1] http://tlt13.sfs.uni-tuebingen.de/.

[2] Institut Universitari de Lingüística Aplicada.

[3] Language for Special Purposes.

[4] http://www.metanet.eu/projects/METANET4U/.

[5] http://metashare.upf.edu and http://hdl.handle.net/10230/20408.

[6] http://www.delph-in.net.

[7] As can be observed in Figure 1, statistics are gathered from disambiguated parses.

[8] The DELPH-IN framework has also used in several treebank projects (Oepen et al., 2002; Hashimoto et al., 2007; Kordoni and Zhang, 2009; Branco et al., 2010; Marimon, 2010; Flickinger et al., 2012).

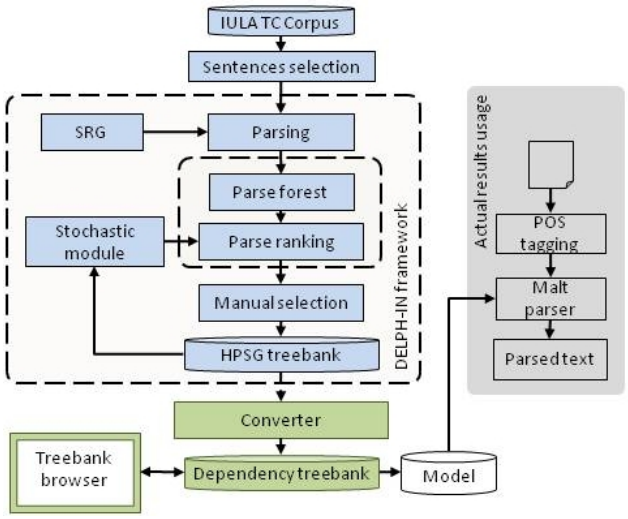



Figure 1: Methodology.

In this paper we will focus on the conversion procedure and the dependency annotations that the treebank provides. Note that Marimon et al. (2012) describe how the sentences to be annotated were selected from the IULA Technical Corpus (Cabré et al., 2006; Vivaldi, 2009), the use of the DELPH-IN framework in this treebank project, and the interannotator agreement analysis carried out to evaluate the consistency of the annotations performed by three different persons. Evaluation results using the treebank on several data-driven dependency-based parsing systems are discussed in (Padró et al., 2013).

### 2.1. The conversion procedure

### 2.2. The conversion procedure

The linguistic analysis produced by the DELPH-IN processing framework for each parsed sentence provides, together with a binary branching phrase structure tree representing constituency structure and a *Minimal Recursion Semantic* (MRS) semantic representation (Copestake et al., 2006) representing estructural semantics (i.e. predicate-argument relations), a derivation tree. For the purpose of this paper, we restrict ourselves to the derivation tree, which is the only format we considered to generate the dependency structures.

Derivation trees are encoded in a nested, parenthesized structure whose elements correspond to the identifiers of the phrase structure rules and the lexical items involved in the parsing. Phrase structure rules identify the daughter sequence and a basic dependency relation between sentence constituents, such as subject-head (sb-hd), head-complement (hd-cmp), and head-adjunct (hd-ad). Lexical items are annotated with part-of-speech information according to the EAGLES tagset for Spanish (e.g. VMIP3S0)[9] and their lexical entry identifier (e.g. *aparecer_v-pp_loc*). Figure 2 shows an example with the sentence *Un coágulo anormal que aparece en un vaso sanguíneo recibe el nombre de trombo.* ('An abnormal clot that appeears in a blood vessel is called thrombus.').

[9]Verb main indicative present third singular.

From the LKB derivation tree, we could obtain the information to generate the dependency structures that the IULA Spanish LSP Treebank provides in two formats: (i) a column-based format, in the style of CoNLL-2006 shared task (Buchholz and Marsi, 2006), and (ii) a dependency graph, where dependency relations are noted by an oriented arrow which goes from the dependent node to the head node, both illustrated in Figure 3 with the same sentence as Figure 2.

The conversion from derivation trees to dependency structures was a fully automatic and unambiguous process. A deterministic conversion algorithm made use of the identifiers of the phrase structure rules to identify the heads, the dependents, and some dependency types that were directly transferred onto the dependency structure, e.g., subject, specifier, and modifier. The identifiers of the lexical entries, which included the syntactic category of the subcategorised elements, enabled the identification of the argument-related dependency functions, e.g. direct object, indirect object, and oblique complement.

### 2.3. Syntactic annotations

Centered upon the notion of dependency, dependency-based frameworks share the basic assumption that the syntactic structure of a sentence largely resides in asymmetrical relations between a head and a dependent. They also share the analysis they provide for a core of syntactic constructions. However, there are also important differences with respect to the criteria for identifying the head and the dependent in the relations, as well as with respect to the analysis of certain types of syntactic constructions.

In this section we present the linguistic annotations that the IULA Spanish LSP Treebank provides following the dependency grammar model. We start with the presentation of the dependency relations that we have compiled, then we discuss the criteria for identifying the head and the dependent in the relations and the analysis that the treebank provides for coordination constructions and headless constructions.

#### 2.3.1. Dependency labels

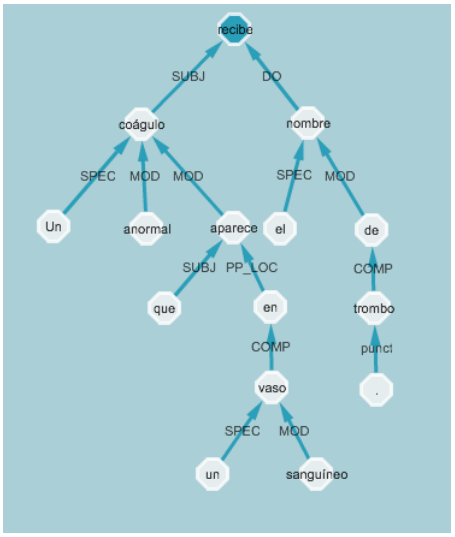
The dependency labels used in the treebank roughly correspond to the standard labels supplemented with particular tags for Spanish phenomena.[10]

- SPEC (specifier), for determiners depending on nouns and degree adverbs depending on adjectives and adverbs.
- MOD (modifier), for all types of non-subcategorized dependents with the modifying function.
- COMP (complement), for PPs governed by different non-verbal heads.
- AUX (auxiliary), for the auxiliary verb *haber* ('to have').
- The argument-related dependency relations that are governed by a verbal head are:

[10]Labels used in coordinated constructions and gapping constructions will be presented in section 2.2.2.

```
(sb-hd_c
   (sp-hd_c
      (di0ms0 (un_d "un"))
      (hd-ad_c
         (hd-ad_c
            (ncms000 (coágulo_n "coágulo"))
            (aq0cs0 (anormal_aj "anormal")))
         (fl-hd_c
            (pr0cn000 (que_pr "que"))
            (hd-cmp_c
               (vmip3s0 (aparecer_v-pp_loc "aparece"))
               (hd-cmp_c
                  (sps000 (en_p "en"))
                  (sp-hd_c
                     (di0ms0 (un_d "un"))
                     (hd-ad_c
                        (ncms000 (vaso_n-pp "vaso"))
                        (aq0ms0 (sanguíneo_aj "sanguíneo")))))))))
   (hd-cmp_c
      (vmip3s0 (recibir-np "recibir"))
      (sp-hd_c
         (da0ms0 (el_d "el"))
         (hd-ad_c
            (ncms000 (nombre_n-pp "nombre"))
            (hd-cmp_c
               (sps000 (de_p "de"))
               (hd-pt_c
                  (ncms000 (trombo_n-pp "trombo"))
                  (fp (fstop_pt "."))))))))
```

Figure 2: Derivation tree of *Un coágulo anormal que aparece en un vaso sanguíneo recibe el nombre de trombo.* ('An abnormal clot that appears in a blood vessel is called thrombus.')




| ID | FORMA | LEMA | CATEGORIA | PoS TAG | DEPENDÈNCIA | NUCLI |
|---|---|---|---|---|---|---|
| 1 | Un | un | z | Z | SPEC | 2 |
| 2 | coágulo | coágulo | n | NCMS000 | SUBJ | 10 |
| 3 | anormal | anormal | a | AQ0CS0 | MOD | 2 |
| 4 | que | que | p | PR0CN000 | SUBJ | 5 |
| 5 | aparece | aparecer | v | VMIP3S0 | MOD | 2 |
| 6 | en | en | s | SPS00 | PP_LOC | 5 |
| 7 | un | un | z | Z | SPEC | 8 |
| 8 | vaso | vaso | n | NCMS000 | COMP | 6 |
| 9 | sanguíneo | sanguíneo | a | AQ0MS0 | MOD | 8 |
| 10 | recibe | recibir | v | VMIP3S0 | ROOT | 0 |
| 11 | el | el | d | DA0MS0 | SPEC | 12 |
| 12 | nombre | nombre | n | NCMS000 | DO | 10 |
| 13 | de | de | s | SPS00 | MOD | 12 |
| 14 | trombo | trombo | n | NCMS000 | COMP | 13 |
| 15 | . | fstop | f | Fp | punct | 14 |

Figure 3: Dependency graph and column-based format of *Un coágulo anormal que aparece en un vaso sanguíneo recibe el nombre de trombo.* ('An abnormal clot that appears in a blood vessel is called thrombus.')

- SUBJ (subject).
- DO (direct object).
- IO (indirect object).
- OBLC (oblique complement).
- BYAG (by-agent complement).
- ATR (attribute).
- PRD (predicative complement).
- OPRD (object predicative complement).
- PP_LOC (locative complement).
- PP_DIR (directional complement).
- ADV (adverbial).

Verbs may also govern the following dependency labels:

- PRNM (pronominal marker), for clitic pronouns found with so-called *inherent reflexive* verbs (or pronominal verbs); i.e. verbs that require a clitic pronoun co-indexed with the subject and which lack the corresponding non-reflexive form e.g. *La industria nuclear se encuentra en crisis.* ('The nuclear industry is in crisis.').
- IMPM (impersonal marker), for the grammatical marker *se* that appears in impersonal *se*-constructions, e.g. *Se obtiene lecturas más altas o más bajas?* ('Obtained readings are higher or lower?').
- PASSM (passive marker), for the grammatical marker *se* that appears in passive *se*-constructions, e.g. *Ambas aproximaciones se comentan a continuación.* ('Both approaches are discussed below.').

#### 2.3.2. Heads and dependents

In identifying the head and the dependent in the relations, the IULA Spanish LSP Treebank annotation mostly follows syntactic criteria, and the head element is the lexical item which determines the syntactic category of the construction.

- **Noun phrases**. Nouns are the heads of NPs, and determiners are their dependents, labeled as SPEC (specifier) (See, for instance, the analysis of *un coágulo anormal* ('an abnormal clot') in Figure 3).
- **Prepositional phrases**. Prepositions are the heads of PPs and they govern their NPs, which, in turn, are represented inside the PP (i.e. nouns govern their dependents) (See the analysis of *en un vaso sanguíneo* ('in a blood vessel') in Figure 3).
- **Verb groups**. Only the auxiliary verb *haber* ('to have') takes the label AUX (auxiliary) and modal verbs are considered as heads of the verb group. In all verb groups, all dependents (subjects, complements, negative particles) are attached to the content element (the non-finite forms of the verb groups).
- **Subcategorized subordinate clauses**. In the analysis of subcategorized subordinate clauses introduced by the complementizer *que* ('that'), the complementizer lies between the two clauses: it is the head of the subordinate clause and it depends on the verb of the matrix clause.
- **Modifier subordinate clauses**. In relative clauses depending on nouns, the predicate constitutes the head of the clause and the relative pronoun is governed by the head verb and labeled according to the annotation schema, as we show in Figure 3, where the relative pronoun is the subject of the relative clause.

  Although modifier subordinate clauses can be of different types (time, cause, etc.), we only use one dependency tag –MOD– given between two verbs (main and subordinate clause) or between the main verb and a conjunction introducing the subordinate clause.
- **Coordinated structures**. The treebank follows Mel'cuk (1988)'s approach for coordination, that is, the first conjunct is the head of the other elements, which are organized in a chain; i.e. the conjunction is a dependent of the first conjunct and the second conjunct of the conjunction (in multi-conjunct coordination, the conjunction depends on the penultimate conjunct and the last conjunct on the conjunction). Coordinating conjunctions are labeled as COORD (coordinating conjunction) and conjuncts as CONJ (conjunct). In multi-conjunct coordinated constructions, we use the label ENUM (enumeration), instead of CONJ, in all but the last coordinated element
- **Headless constructions**. In elliptical noun phrases, we follow the standard strategy in dependency corpora: the modifier of the elided head is chosen to become the head of the construction and it is labeled with the syntactic function of the elided head.

  In elliptical finite verbs in e.g. gapping constructions, the coordinating conjunction represents the missing verb and inherits all its properties, such that subjects, complements, and adjuncts are linked to it, marked by the labels SUBJ-GAP, COMP-GAP, and MOD-GAP.

### 2.4. Statistics of the treebank

As we have already mentioned, the IULA Spanish LSP Treebank contains over 41,000 sentences distributed among different domains. The details about the statistics are shown in Table 1 and Table 2. It is worth mentioning that the 11.59% of the words in the treebank are tagged as verb, 25% as common noun, and 9.81% as adjectives, as we show in Table 3, which displays the relative frequency of the syntactic categories in the corpus. Finally, Table 4 gives some figures for the occurrences of the main dependency tags identified in the treebank.

| | Number of sentences |
|---|---|
| Law | 6,091 |
| Economy | 3.48 |
| Computing Science | 6,770 |
| Environment | 4,414 |
| Medicine | 19,779 |
| Total | 41,102 |

Table 1: Number of sentences of the IULA Spanish LSP Treebank distribted among different domains.

| | |
|---|---|
| Number of sentences | 41,102 |
| Number of words | 582,897 |
| Number of distinct words | 43,302 |
| Number of distinct lemmata | 16,962 |

Table 2: Statistics of the IULA Spanish LSP Treebank.

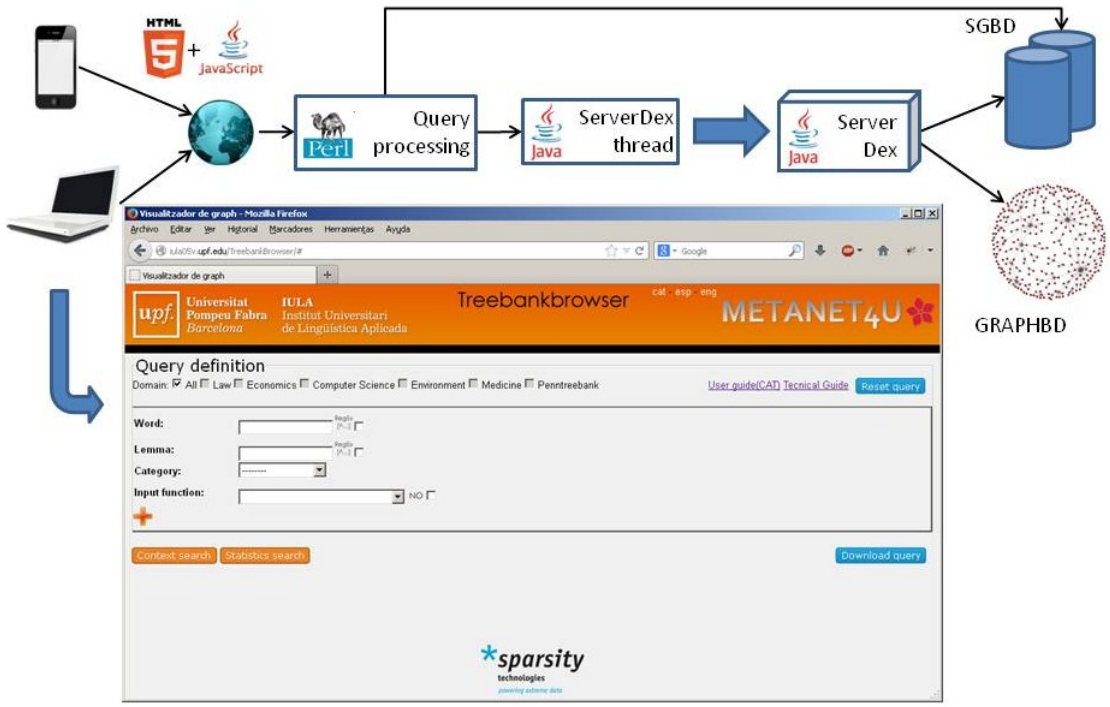



Figure 4: Browsing architecture/interface.

| | | | |
|---|---|---|---|
| Verb | 11.59% | Common nouns | 25.00% |
| Proper names | 2.02% | Adjective | 9.81% |
| Preposition | 16.25% | Adverb | 2.62% |
| Definite article | 13.26 | Determiner | 4.44% |
| Conjunction | 3.47% | Pronoun | 3.24% |

Table 3: Relative frequency of syntactic categories in the IULA Spanish LSP Treebank.

| Dep. labels | Number | Dep. labels | Number |
|---|---|---|---|
| ROOT | 41,102 | SPEC | 36,572 |
| MOD | 38,692 | COMP | 37,869 |
| SUBJ | 27,288 | ATR | 2,731 |
| DO | 20,185 | IO | 693 |
| OBLC | 5,927 | BYAG | 1,364 |
| PRD | 1,203 | PP-LOC | 727 |
| IMPM | 400 | PASM | 6,490 |
| PRNM | 1,213 | AUX | 1,787 |
| COORD | 12,614 | CONJ | 12,232 |

Table 4: Occurrences of dependency labels in the IULA Spanish LSP Treebank.

## 3. Browsing the treebank

We have developed a web application that allows to any user to query the treebank.[11] This is a java application that has been built around a graph database connected with a relational database.[12] Figure 4 shows the global architecture and the user interface of such tool.

[11] http://iula05v.upf.edu/TreebankBrowser/.

[12] See http://www.sparsity-technologies.com/. for details

The main functionalities of this tool are described in the following subsections:

### 3.1. Treebank query

The query consists of the definition of a dependency subgraph and search process as the process of searching for all the graphs in the database that include such subgraph. Each query block defines the restriccions to be applied to a given node of the query subgraph. Such restriccions include the definition of the word form/lemma, the POS tag, and the syntactic funtion of such node.

Table 5 shows an example of query building; in this case the purpose is to obtain all the sentences where the ROOT node is the verb *fabricar* ('to manufacture') and both the subject and the direct object are expressed (i.e. where somebody manufactures something).

By default, queries are done to the whole treebank. It is possible, however, to limit the query to one or more domains. Also by default, the results take the form of a list of sentences but it is also possible to obtain some statistics regarding the results found instead of the sentences. Such statistics are refered to the query blocks (Number of words/lemmas and their POS tag).

### 3.2. Result visualization

The system shows a the list of the sentences that satifies the query highlighting the nodes that satisfies the query. Figure 5 shows the list of sentences that satisfy the query illustrated in Table 5.

Optionally, for each sentence, it may also obtain each sentence in the CoNLL format or a a directed graph (as shown in Figure 3). In both cases such information may be downloaded to a local file. Full query results are also downloadable as a flat ASCII text or using the CoNLL format.

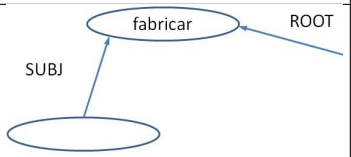
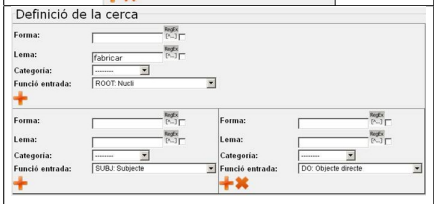
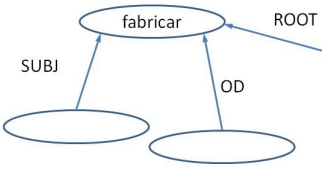
| Interface | Condition | Action |
|---|---|---|
|  |  | 1) main verb definition |
|  |  | 2) subject definition |
|  |  | 3) object definition |

Table 5: Query generation process.

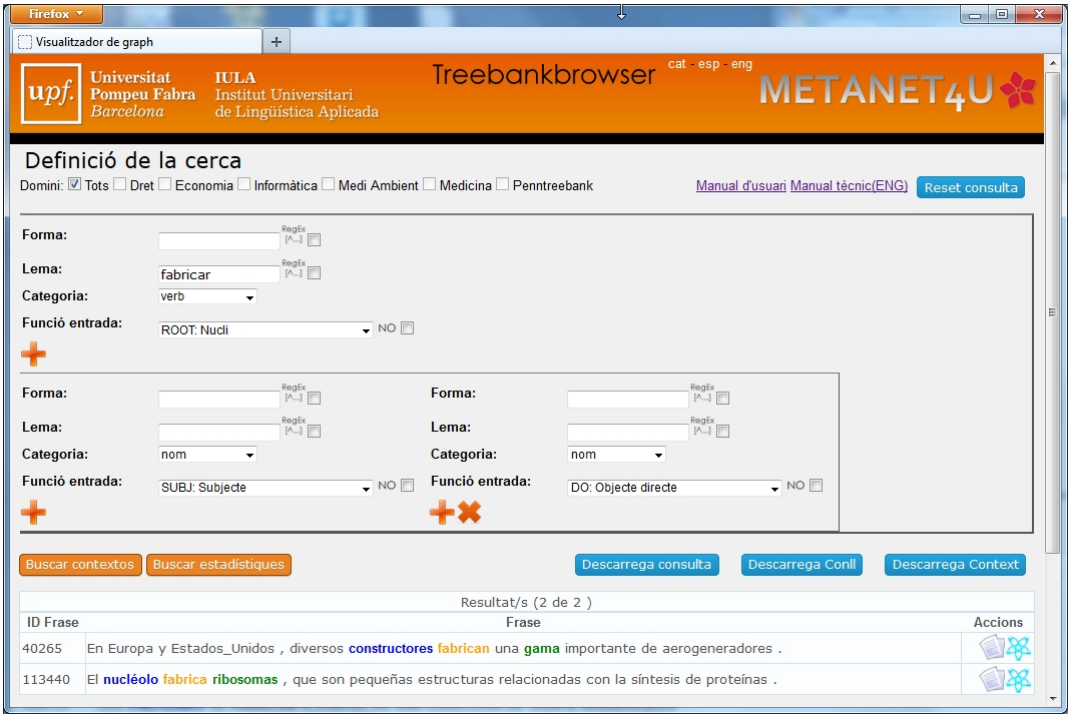



Figure 5: Results for the query shown in Table 5.

## 4. Conclusions

This paper describes the IULA Spanish LSP Treebank, a dependency treebank of over 41,000 sentences, developed in the framework of the European project METANET4U. We have described the methodology that we used to create the resource, the syntactic annotations that the treebank provides, and the treebank browser that we have developed to query the annotated corpus. In the future, we plan to add annotations of semantic role labels by extracting them from the MRS semantic representation.

## 5. Acknowledgements

This work was co-funded by the Ramón y Cajal program of the Spanish Ministerio de Ciencia e Innovación, the EU UNER - Competitiveness and Innovation Framework Program, METANET (CIP-PSP-270893), and the UPF-IULA PhD grant program.

## 6. References


Anne Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, Dordrecht, Boston and London.

António Branco, Francisco Costa, João Silva, Sara Silveira, Ségio Castro, Mariana Avelãs, Clara Pinto, and João Graca. 2010. Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In *Proceedings of LREC-2010*, La Valletta, Malta.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL-X*, New York City, USA.

M. Teresa Cabré, Carme Bach, and Jorge Vivaldi. 2006. 10 anys del corpus de l'IULA. Technical report, Institut Universitari de Lingüística Aplicada.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2006. Minimal Recursion Semantics: an Introduction. *Research on Language and Computation*, 3(4):281–332.

Dan Flickinger, Valia Kordoni, Yi Zhang, António Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Francisco Costa, and Sérgio Castro. 2012. ParDeepBank: Multiple Parallel Deep Treebanking. In *Proceedings of TLT-2012*, Lisbon, Portugal.

Chikara Hashimoto, Francis Bond, and Melanie Siegel. 2007. Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. *Language Resources and Evaluation. (Special issue on Asian language technology)*, 42(2):117–126.

Valia Kordoni and Yi Zhang. 2009. Annotating Wall Street Journal Texts Using a Hand-Crafted Deep Linguistic Grammar. In *Proceedings of LAW III*, Suntec, Singapore.

Montserrat Marimon, Beatriz Fisas, Núria Bel, Blanca Arias, Silvia Vázquez, Jorge Vivaldi, Sergi Torner, Marta Villegas, and Mercè Lorente. 2012. The IULA treebank. In *Proceedings of LREC-2012*.

Montserrat Marimon. 2010. The Tibidabo Treebank. *Procesamiento del Lenguaje Natural*, 45:113–119.

Montserrat Marimon. 2013. The Spanish DELPH-IN Grammar. *Language Resources and Evaluation*, 47(2):371–397.

Igor Mel'cuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

A. Moreno, R. Grishman, S. López, F. Sánchez, and S. Sekine. 2000. Treebank of Spanish and its Application to Parsing. In *Proceedings of LREC-2000*, Athens, Greece.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2002. LinGo Redwoods. A Rich and Dynamic Treebank for HPSG. In *Proceedings of TLT-2002*, Sozopol, Bulgaria.

Muntsa Padró, Miguel Ballesteros, Hector Martínez, and Bernd Bohnet. 2013. Finding dependency parsing limits over a large spanish corpus. In *Proceeding of the IJCNLP-2013*.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago.

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

Kristina Toutanova, Christoper D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language and Computation*, 3(1):83–105.

Jorge Vivaldi. 2009. Corpus and exploitation tool: IULACT and bwanaNet. In Asociación Española de Lingüística del Corpus, editor, *A survey on corpus-based research (CICL-09)*.