The IULA Spanish LSP Treebank

Montserrat Marimon

The Spanish LSP IULA Treebank

- Developed in the framework of the European project METANET4U.
- Using the DELPH-IN processing framework (SRG).
- Corpus:
  - 1,000 sentences translated from the Penn Treebank
  - About 40,000 sentences, technical corpus (sentence length range from 4 to 30 words).

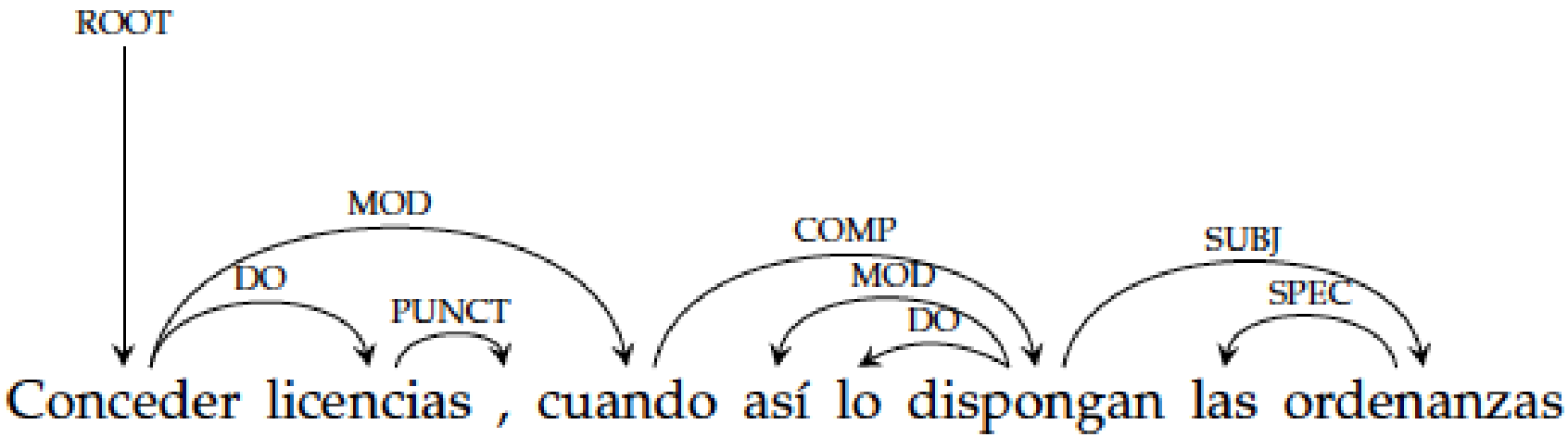
The Spanish LSP IULA Treebank

	Number of sentences
Law	6,091
Economy	3,048
Computing Science	6,770
Environment	4,414
Medicine	19,779
Total	41,102

The Spanish LSP IULA Treebank

- Automatic selection of parsed sentences for treebank construction.
- Method: parser ensemble approach using full agreement among the MaxEnt parse selection model and a dependency parser (MaltParser), trained on the same corpus.
- Comparison between them is performed on the dependency structures that we obtain by converting derivation trees.

```
(hd-ad_c
   (hd_optc_c
      (hd-cmp_c
         (vmn0000 (conceder_v-np-ppa "conceder"))
         (hd-nbar_c
            (hd-pt_c
               (ncfp000 (licencia_n "licencias"))
                (fc (comma_pt ",")))))))
   (hd-cmp_c
      (cs (cuando_p-cl "cuando"))
      (fl-hd_c
         (hd_advnp_c
            (nc00000 (asi_av "así")))
         (hd_sb_c
            (cl-hc_c
               (pp3msa00 (lo_pr "lo"))
               (vmsp3p0 (v_acc_dlr (disponer_v-np "dispongan"))))
            (sp-hd_c
               (da0fp0 (el_d "la"))
               (nccp000 (ordenanza_n "ordenanzas")))))))
```


ROOT
MOD
DO
PUNCT
COMP
MOD
DO
SUBJ
SPEC
Conceder licencias , cuando así lo dispongan las ordenanzas

Tags	Grammatical functions	Syntactic categories
ROOT	Root	S, VP, NP, AP, PP, ADVP
SPEC	Specifier	D, ADV
MOD	Modifier	AP, NP, PP, ADVP, VP, S
NEG	Negation	ADV
COMP	Complement (of N, ADJ, ADV, PREP)	PP
SUBJ	Subject	NP, VP, S
DO	Direct Object	NP, VP, S, PP[a] (for human DOs)
IO	Indirect Object	PP[a], NP
OBLC	Oblique Object	PP
BYAG	By agent complement	PP[por]
ATR	Attribute	AP, PP, ADVP
PRD	Predicative complement	AP, PP
OPRD	Object predicative complement	AP, PP
PP-LOC	Locative prepositional complement	PP, ADVP
PP-DIR	Directional prepositional complement	PP, ADVP
ADV	Adverbial complement	ADVP
IMPM	Impersonal marker	NP
PASSM	Passive marker	NP
PRNM	Pronominal marker	NP
AUX	Auxiliary	V
VOC	Vocative	NP
SUBJ-GAP	Subject in a gapping construction	NP, VP, S
COMP-GAP	Complement in a gapping construction	NP, PP, AP, ADVP, VP, S
MOD-GAP	Modifier in a gapping construction	NP, PP, ADVP
COORD	Coordination	NP, PP, AP, ADVP, VP, S
CONJ	Conjunction	C
PUNCT	Punctuation	PT

The Spanish LSP IULA Treebank

Experiments & results

- Training set: 13,901 sentences & test set: 1,428 sentences (from 4 to 20 words).
- We trained the MaxEnt model and the MaltParser and ran each of the models on the test set.
- We compared the outputs of the two models and selected only identical analyses that were produced by both of them.

The Spanish LSP IULA Treebank

Results of the parse selection model (up) and the MaltParser (down) as labeled attachment scores, unlabeled attachment scores, label accuracy score, and exact syntactic match.

	LAS	*UAS*	*Label Accur Score*	*Exact Synt Match*
Parse selection model	95.38%	96.80%	97.62%	61.02%
MaltParser	91.99%	95.02%	94.54%	43.09%

The Spanish LSP IULA Treebank

- The performance of our parser ensemble approach was measured related to the set of selected parses; the set of sentences that were predicted to be correctly parsed because they showed exact syntactic match when compared the MaltParser output and the converted parse selected by the MaxEnt model.

The Spanish LSP IULA Treebank

- Out of 1,428 sentences in the test set, 445 parses were selected; i.e. 31.16% of the test sentences were predicted to be correctly parsed.
- Out of the 445 selected parses, 403 sentences were indeed correctly parsed (exact match); thus the **precision of the detector was 90.56%.**
- Out of 864 correctly parsed sentences by the SRG, 403 were in the set of selected sentences; thus the **recall of the predictor was 46.64%**.

The Spanish LSP IULA Treebank

- We compared our ensemble approach with a baseline selection method that used a threshold on the probability computed by the MaxEnt model.
- We computed the ratio between the probabilities of the two highest-ranked analyses: a very high ratio would indicate that the parse ranked first had a large advantage over the others, while if the ratio was close to 1, both the first and the second analyses would have similar probabilities, indicating a lower confidence of the model in the decision.

The Spanish LSP IULA Treebank

Results of the baseline for different threshold values.

threshold	*%of selected sentences*	*accuracy of selected sentences*
2	46.90%	60.72%
3	34.67%	71.14%
3.6	31.12%	74.14%
5	26.41%	79.95%
10	18.26%	85.79%
20	10.53%	92.46%

The Spanish LSP IULA Treebank

- Experiment 2
  - We parsed 6,800 sentences (sentence length from 10 to 23 words) and 1,452 sentences were selected by the ensemble of parsers with and accuracy of almost 90%.