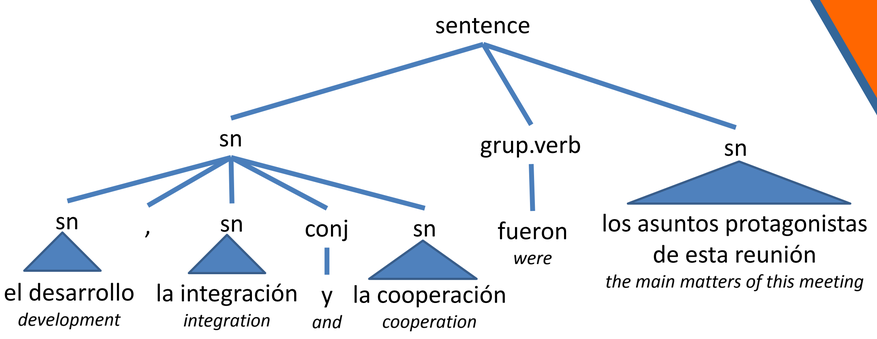
Transforming the AnCora corpus to HPSG

Luis Chiruzzo and Dina Wonsever

Grupo de PLN, InCo - Facultad de Ingeniería, UdelaR - Montevideo, Uruguay

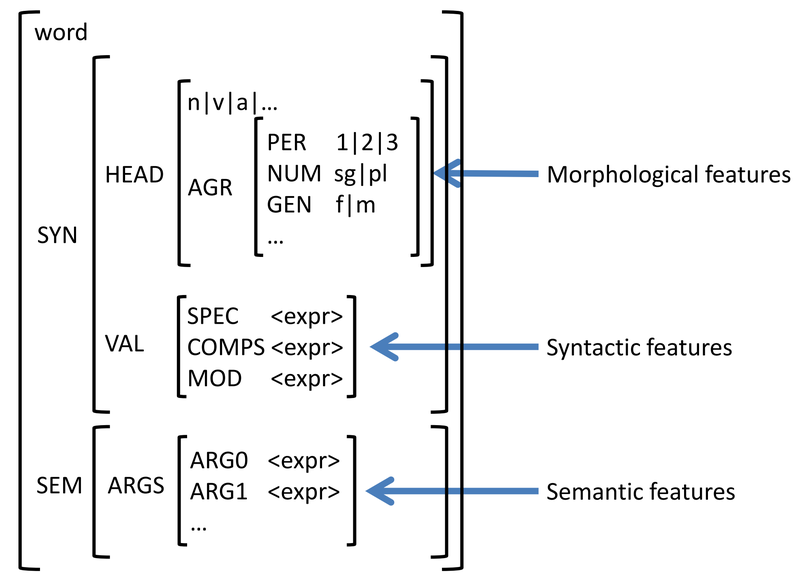


Development

The AnCora corpus has 500,000 words in 17,000 sentences, annotated in a CFG style with syntactic annotations and argument structures. We transformed them into a HPSG annotation style with the aim of creating a statistical HPSG parser for Spanish.

Our approach is based on the construction of the Enju parser, they transformed the Penn Treebank into a HPSG style corpus and learned a statistical HPSG model from it.

The aim is to build a HPSG corpus that contains morphological, syntactic and simplified semantic features. Our approach to semantics is based on the PropBank style argument structure annotations found in the corpus.



Results

The resulting corpus has only binary or unary constituents, and every inner node is annotated with its head and HPSG rule.

We manually annotated 40 sentences (779 constituents) in order to evaluate the performance of the transformation process.

The head detection rules have an overall precision of 95.3%, which climbs to 98.7% if we don't consider nodes with coordinations.

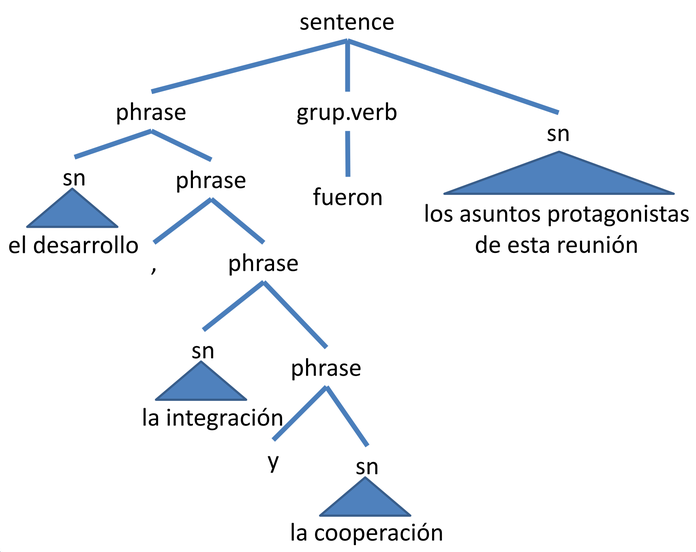
Category	Total	Correct	Precision
grup.a	9	6	66.7%
grup.adv	3	3	100.0%
grup.nom	162	154	95.1%
grup.verb	23	23	100.0%
infinitiu	3	3	100.0%
relatiu	1	1	100.0%
S	91	85	93.4%
s.a	4	3	75.0%
sa	1	1	100.0%
sadv	7	7	100.0%
sentence	40	35	87.5%
sn	220	216	98.2%
sp	207	204	98.6%
spec	8	1	12.5%
Total	779	742	95.3%

The argument classification rules have an overall precision of 92.5%. The category that is the most difficult to classify it the complements, which have a 84.95% precision.

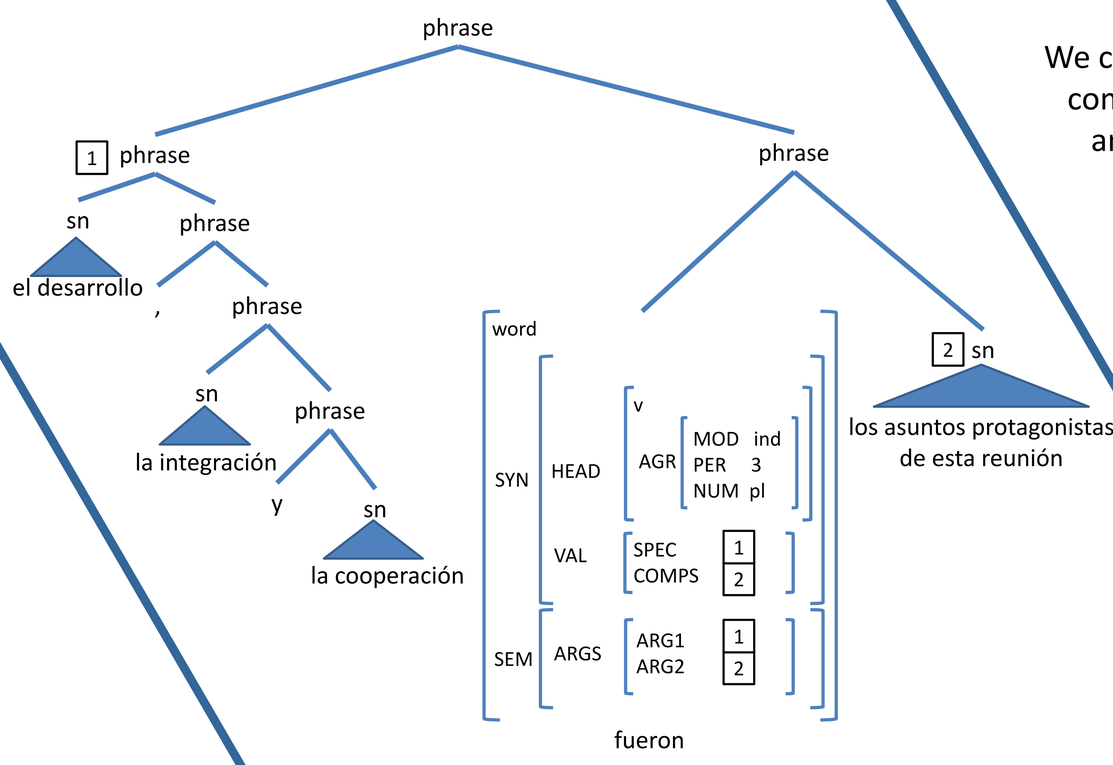
Supertagging

From this corpus we extracted a set of lexical frames that contain syntactic and semantic features. A supertagger that tags each word in a sentece with its frame was trained. It has an accuracy of 83.58% for verbs, 85.78% for nouns and 81.40% for adjectives, considering the top three tags.

The top-down stage breaks the complex structures like coordinations and relative clauses into simpler trees



The bottom-up stage identifies the head of every constituent and classifies its complements, modifiers and specifier



Transformation

Taken as a CFG, AnCora has a complex set of rules, e.g.: there are 5800 ways of building a subordinate sentence and 900 ways of building a noun phrase.

We created a process that breaks complex structures in the corpus and leaves simpler structures which we call *elementary trees* (a head surrounded by arguments). Then we wrote a set of 70 head detection and 184 argument classification rules in order to transform the simpler trees into HPSG.

The leaves are extracted as lexical frames, keeping only combinatorial features