Supertagging for a Statistical HPSG Parser for Spanish

Luis Chiruzzo, Dina Wonsever

Instituto de Computación
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

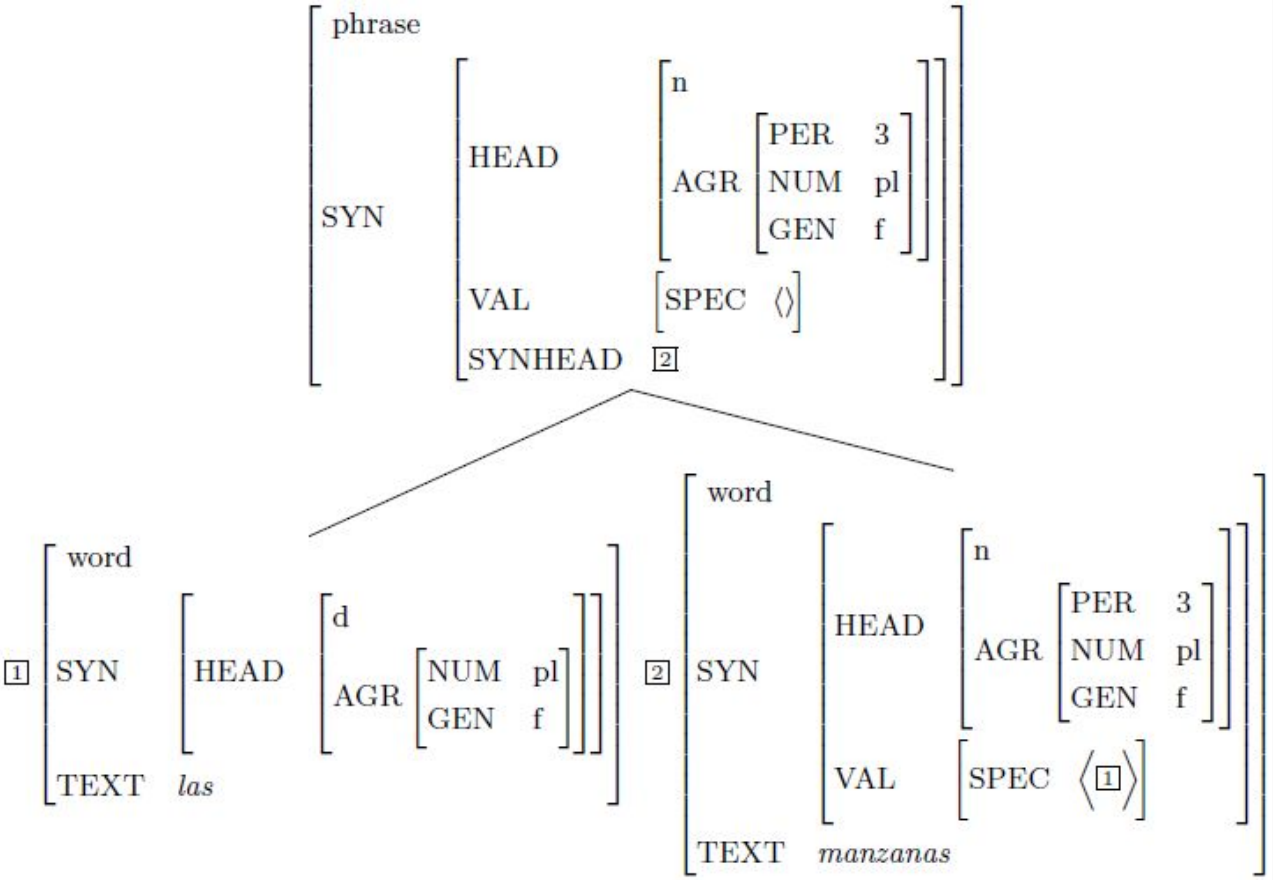
Agenda

- HPSG
- Corpus transformation
- Lexical frames
- Supertagging
- Experiments
- Conclusions

HPSG

- Rich grammar formalism
- Based on typed feature structures
- Lexicalized: words define combinatorial properties and guide the parsing process
- Few grammar rules: specifier, complement, modifier, coordinations…
- Important: to identify the head of every phrase

HPSG


phrase
SYN
HEAD
n
AGR
PER 3
NUM pl
GEN f
VAL
SPEC ⟨⟩
SYNHEAD 2
1
word
SYN
HEAD
d
AGR
NUM pl
GEN f
TEXT las
2
word
SYN
HEAD
n
AGR
PER 3
NUM pl
GEN f
VAL
SPEC ⟨1⟩
TEXT manzanas

AnCora

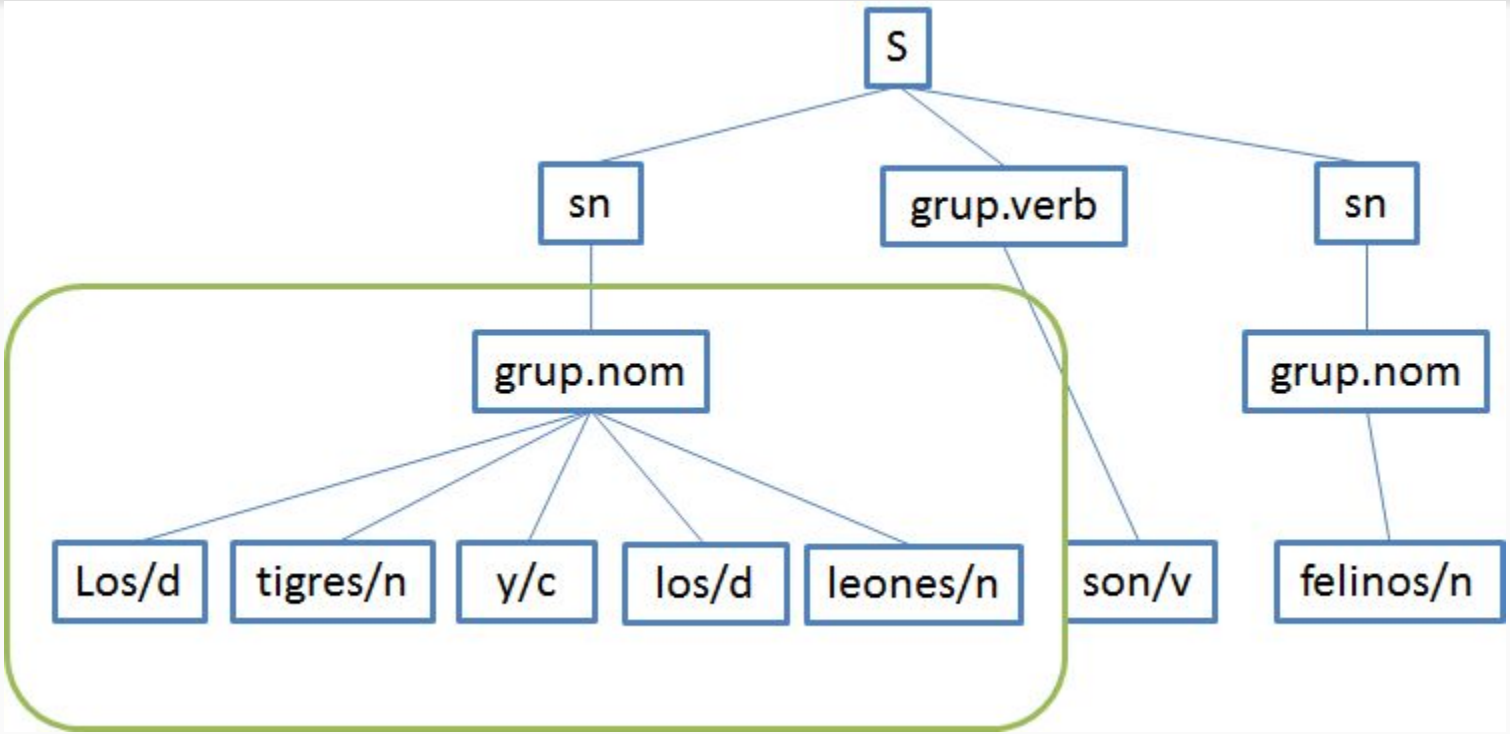
Corpus for Spanish and Catalan

500,000 words in 17,000 sentences

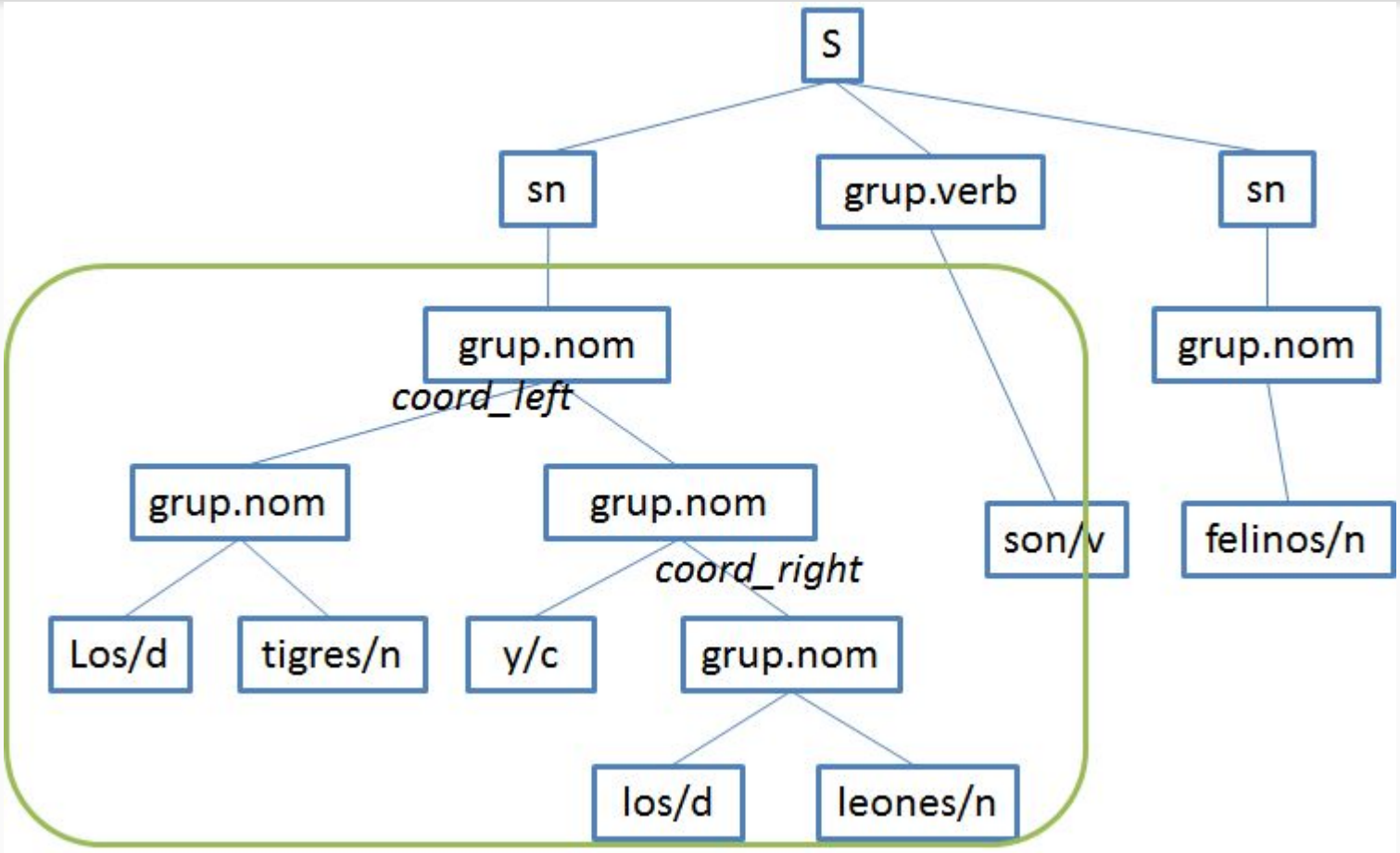
Annotated in a CFG-like formalism

Enriched with attributes: morphology, syntactic function, predicate-argument structure, wordnet senses, and many more...

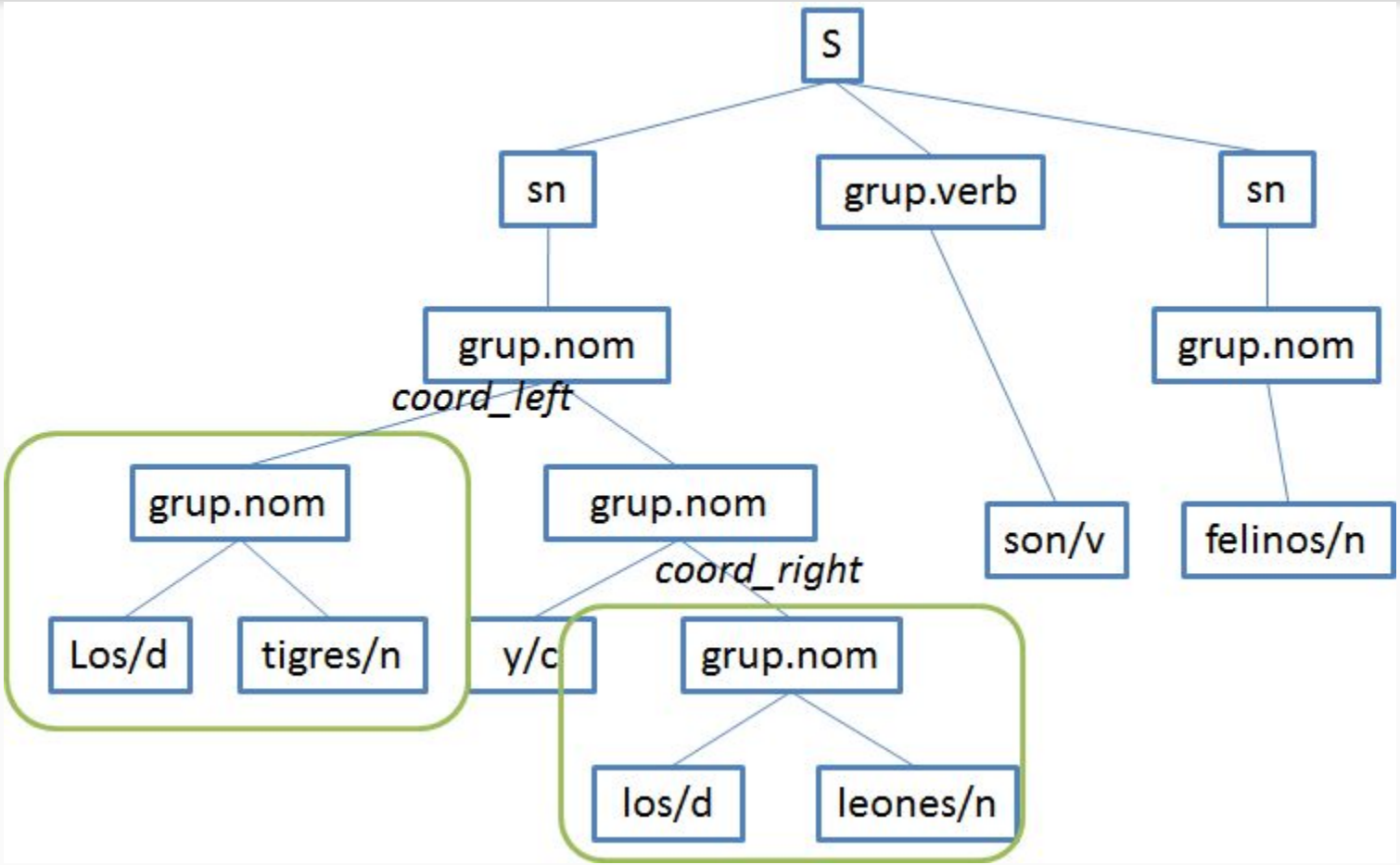
Transformation



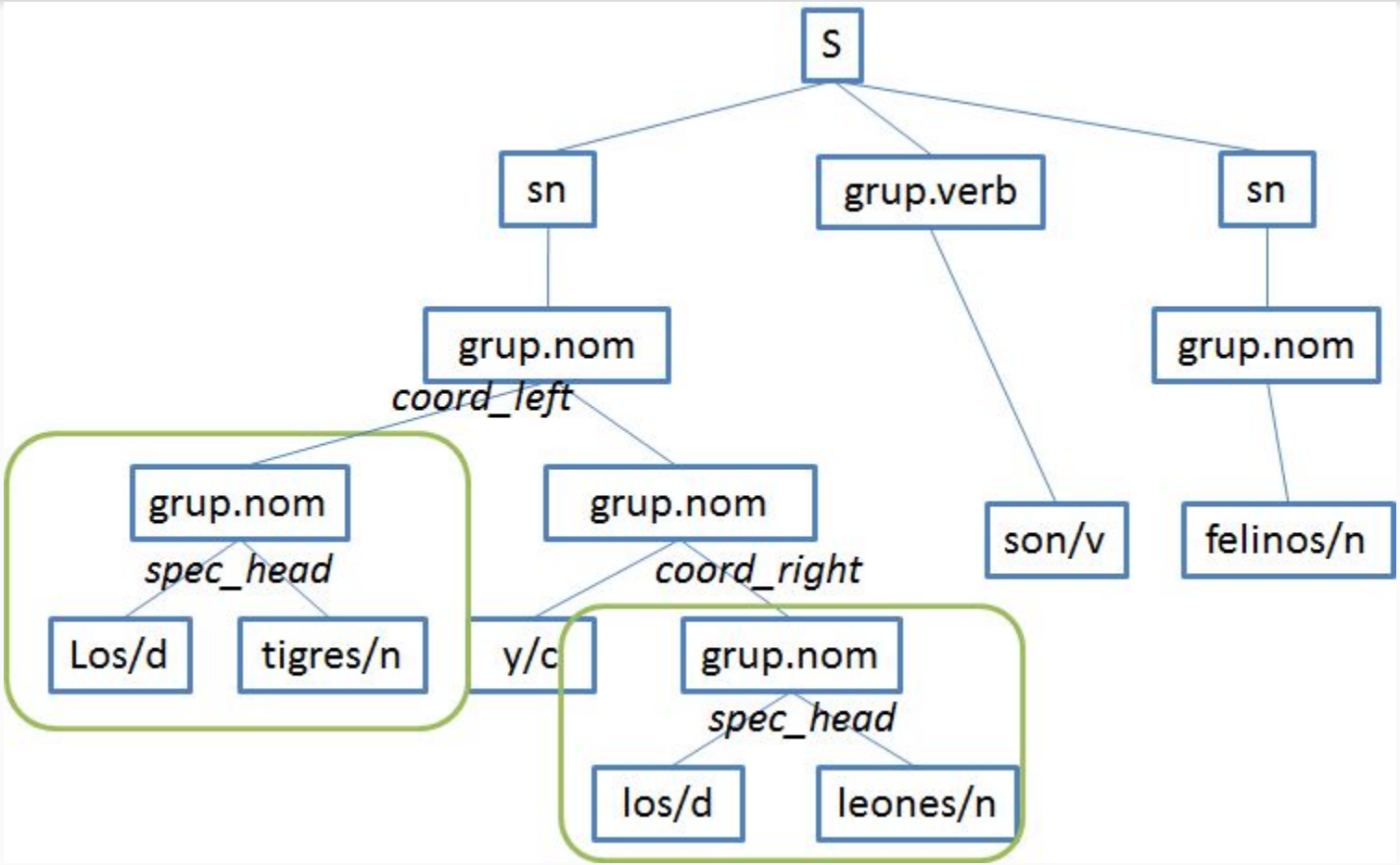
Transformation



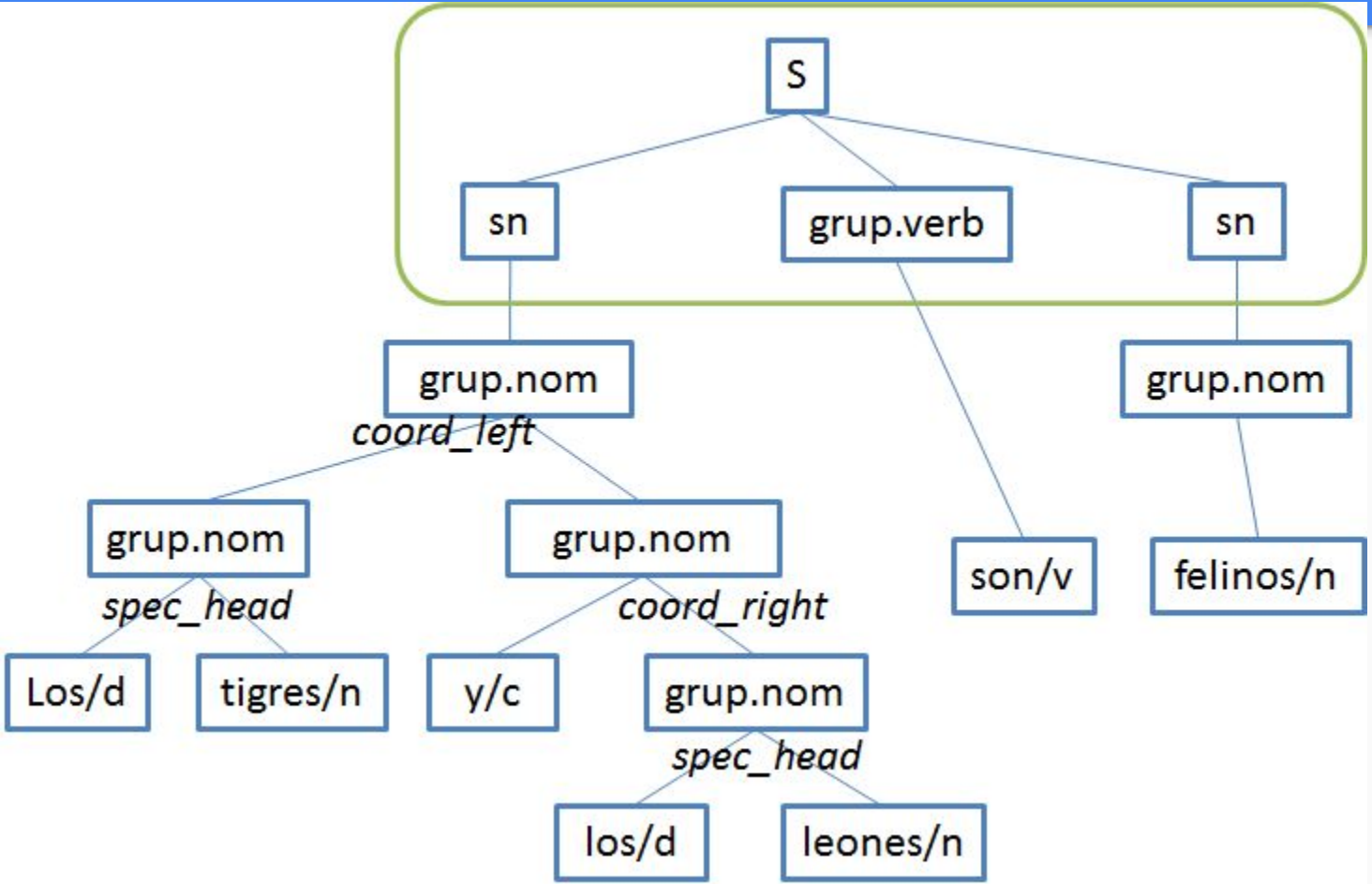
Transformation



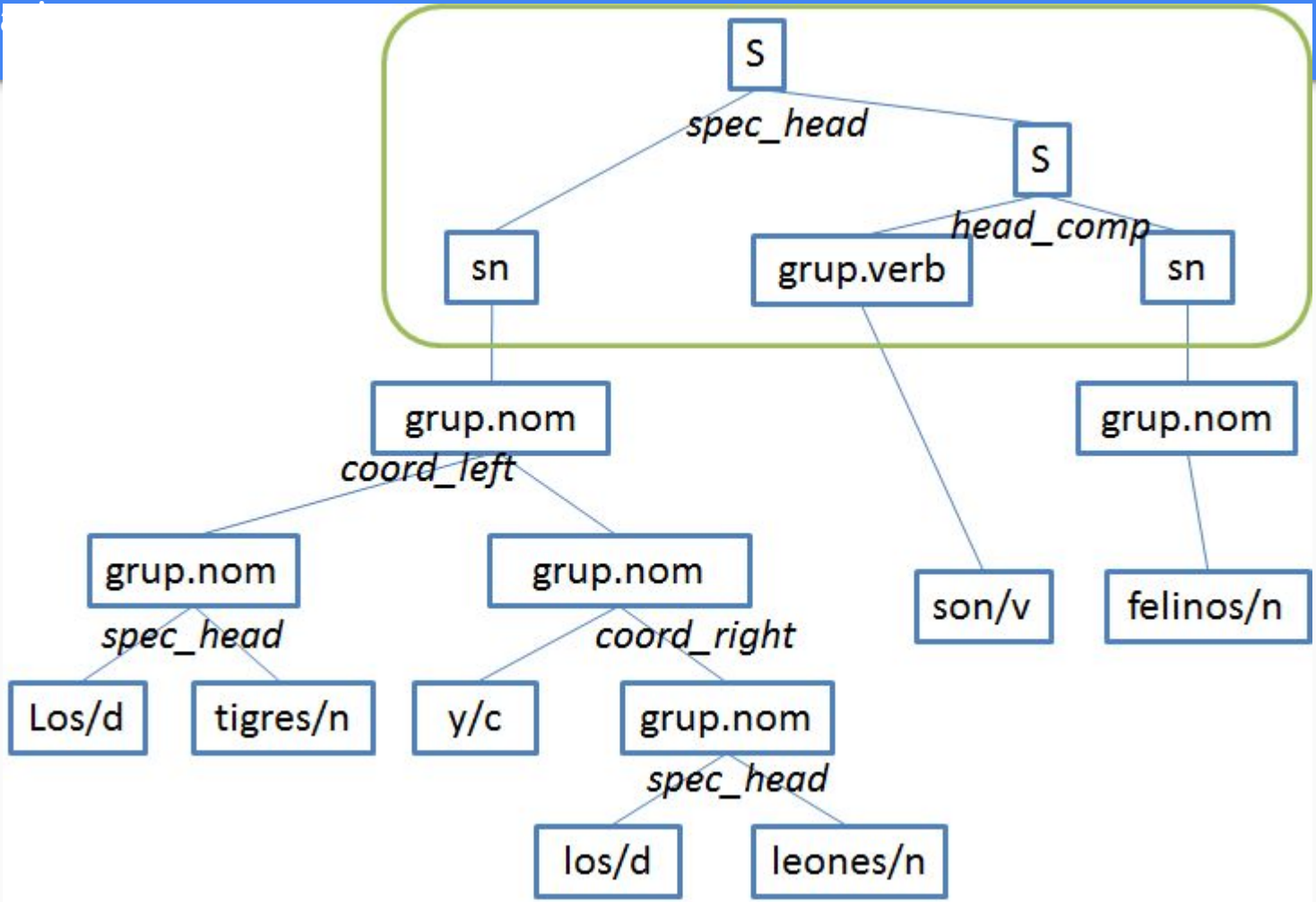
Transformation



Transformation



Transforma


S
spec_head
S
head_comp
sn
grup.verb
sn
grup.nom
grup.nom
coord_left
grup.nom
grup.nom
son/v
felinos/n
spec_head
coord_right
Los/d
tigres/n
y/c
grup.nom
spec_head
los/d
leones/n

Lexical frames

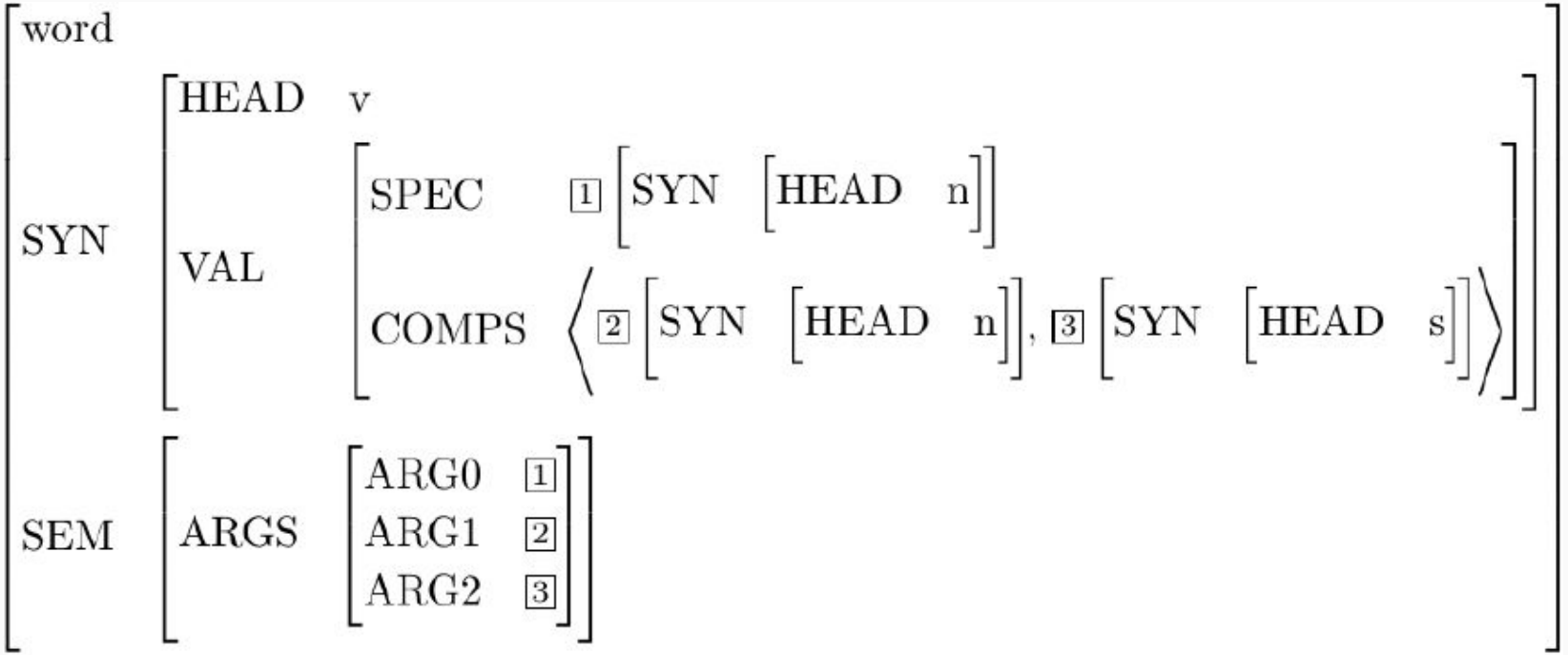
After the transformation:

- only binary or unary rules are used
- all constituents identify their heads
- all arguments (subject and complements) are tagged

Knowing the head and arguments, we create a *lexical frame* for each verb

Lexical frames

ofrecer *(to offer)* / v-arg0s_sn-arg1_sn-arg2_sp_a



Supertagging

The process of assigning the most likely lexical frames to the words of a sentence

It’s like a generalization of tagging: instead of parts-of-speech, we use more fine-grained categories

Useful to speed up the parsing process: reduces the combinatorial explosion of token combinations

Supertagging

We focused on a supertagger that could identify the appropriate lexical frame for verbs

Verbs are the most complex category in Spanish

453 different lexical frames for verbs found in the corpus

Some verbs have dozens of possible valid frames: “ir” (“to go”), “hacer” (“to do”)

Experiments

Corpus pre-processed: all lexical frames that appear fewer than 30 times replaced by a generic tag

CRF and MaxEnt models

Variants between experiments:

- context: number of tokens before and after the current word to look at (5, 7, 9 or 11)
- threshold: words that appear less than *threshold* times are replaced by an unknown token for the category (20, 30 or 50)

Experiment 1

Only lemmas as features

Trained using only CRF

- Training set: 470,000 tokens (15,600 sentences)
- Dev set: 23,000 tokens (800 sentences)
- Test set: 23,000 tokens (800 sentences)

Experiment 1

Baseline: 58.59% for verbs (top three tags)

Best CRF supertagger: context 5, threshold 30, accuracy 78.1% on dev corpus, 78.7% on test corpus (top three tags)

Experiment 2

Lemmas + POS as features

Trained using CRF and MaxEnt

As the experiments for CRF were very slow we trained on smaller versions of the corpus first

Then the configurations that looked more promising were used to train over bigger versions of the corpus

Experiment 2

corpus size	50k		106k		218k		all
threshold	30	50	30	50	30	50	30
context 5	72.70%	71.88%	75.90%	75.20%			
context 7	72.29%	72.64%	76.43%	75.85%	80.11%	79.70%	81.51%
context 9	71.88%	71.41%	77.07%	75.73%	80.11%	79.23%	81.21%
context 11	70.48%	69.72%	74.97%	74.91%			

Experiment 2

Best CRF supertagger: context 7, threshold 30, accuracy 81.51% on dev corpus, 81.35% on test corpus (top three tags)

Best MaxEnt supertagger: context 5, threshold 30, accuracy 78.90% (top three tags)

Experiment 3

The generic ‘v’ tag was assigned to verbs that should’ve had a proper subcategorization

It could be because in the training data 18% of the verbs didn’t have any argument identified

Experiment 3

We created a new version of the corpus pruning all sentences with verbs that don’t have any argument

Unfortunately, this also prunes some examples of verbs with arguments

The new corpus is roughly half the size of the original one: 260,000 tokens (10,000 sentences)

Experiment 3

Lemmas + POS, but only use sentences that have the verbal arguments annotated

Trained using CRF and MaxEnt

- Training set: 230,000 tokens (8,600 sentences)
- Dev set: 13,000 tokens (600 sentences)
- Test set: 13,000 tokens (600 sentences)

Baseline for the new corpus: 59.84% (top three tags)

Experiment 3

corpus size	60k		120k		all	
threshold	20	30	20	30	20	30
context 5	76.71%	76.39%	78.82%	78.59%		
context 7	76.71%	76.16%	80.39%	80.23%	82.35%	81.88%
context 9	75.76%	76.47%	79.06%	79.29%		
context 11	74.35%	74.04%	78.51%	78.43%		

Experiment 3

Best CRF supertagger: context 7, threshold 20, accuracy 82.35% on dev corpus, 83.58% on test corpus (top three tags)

Best MaxEnt supertagger: context 5, threshold 30, accuracy 82.35% on dev corpus, but dropped to 80.71% on test corpus (top three tags)

Conclusions

We created an annotated corpus for Spanish including the lexical frames for each verb: 453 different lexical frames

We trained a series of supertaggers using this corpus, the highest accuracy was 83.58% for verbs (95.40% for all tags) considering the top three tags

This result is training using only verbs with arguments, about half the original corpus. As the performance has not plateaued, results could be improved using more training data

Questions?

Thank you!