

Grew-match tutorial

Bruno Guillaume

LORIA / Inria Nancy Grand-Est

Sémagramme meeting

February 16, 2021

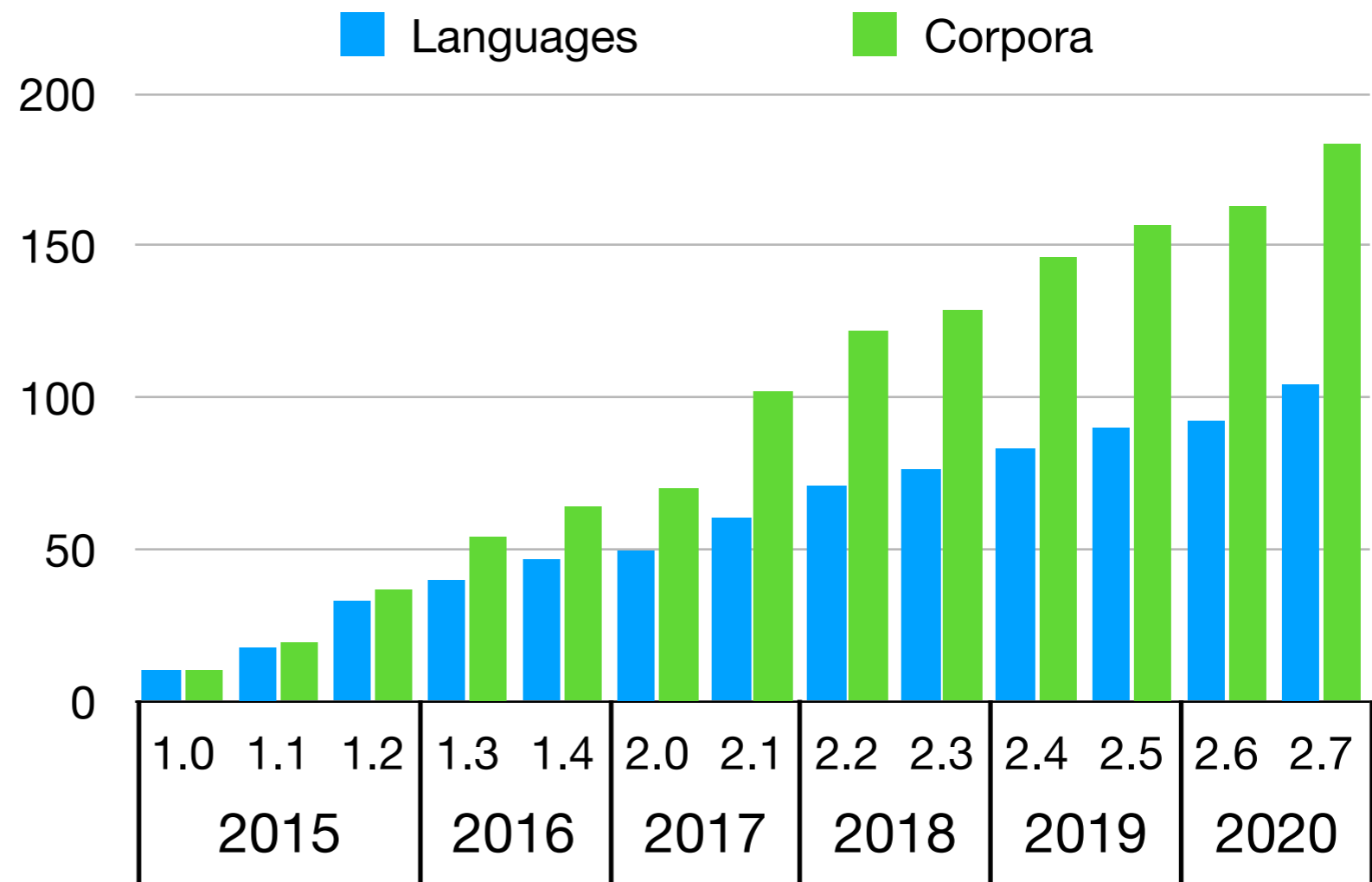
Universal Dependencies

Universal Dependencies (UD):

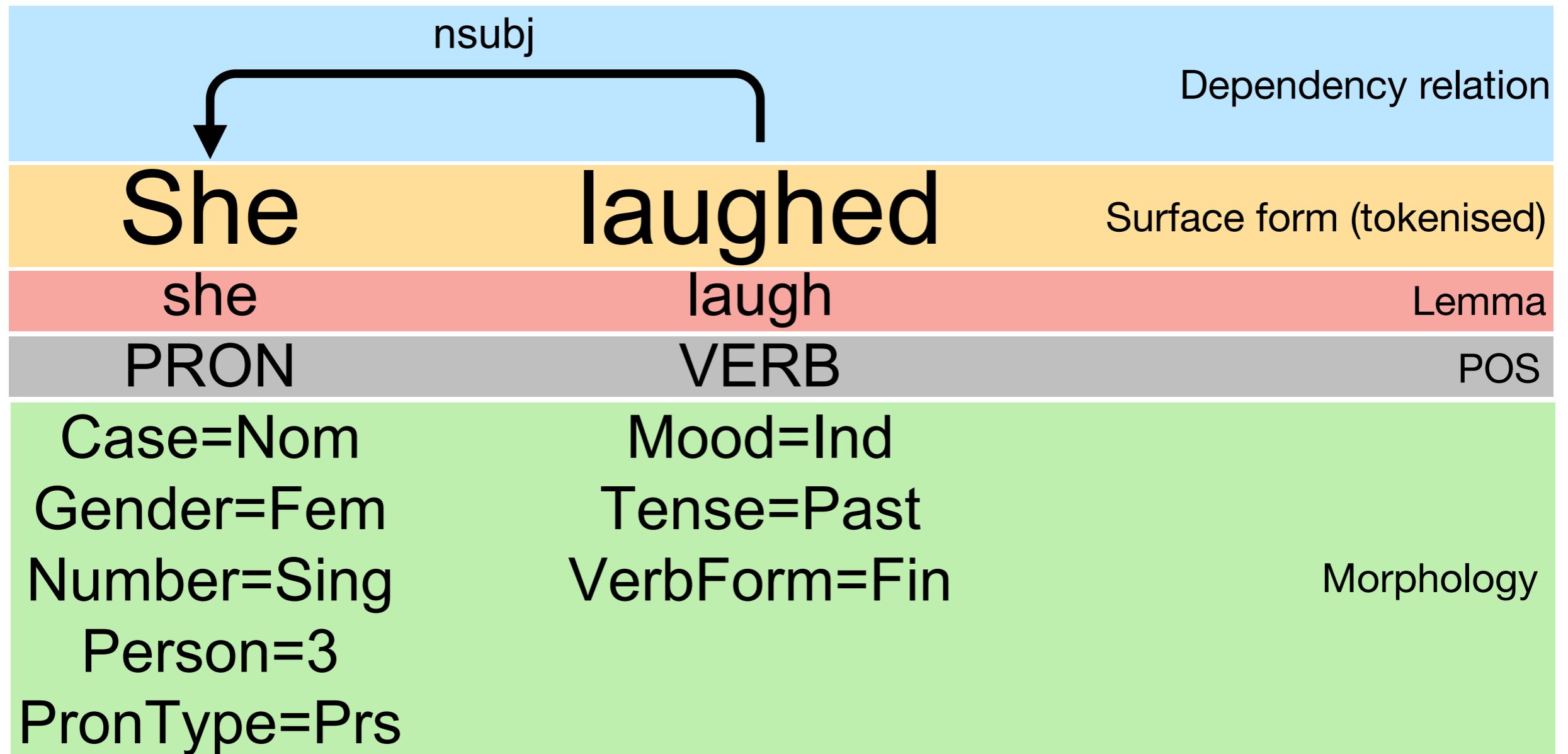
- ▶ Collaborative project of “universal” dependency annotations
- ▶ Version 2.7: 104 languages, 183 corpora



<http://universaldependencies.org>



Universal Dependencies



Grew & Grew-match

grew

<https://grew.fr>

- ▶ Graph rewriting based software
- ▶ Graph rewriting adapted to NLP specificities
- ▶ Developed in the Sémagramme team (with G. Bonfante) in the last 10 years
- ▶ More on this in an upcoming tutorial...

grew-match

<http://match.grew.fr>

- ▶ In Grew: rule = (pattern, commands)
- ▶ In Grew-match, we use only the “pattern” part
- ▶ (graph, pattern) \Leftrightarrow list of occurrences
- ▶ (corpus, pattern) \Leftrightarrow list of occurrences

Grew-match

- ▶ Available online on a large set of corpora
 - ▶ All (S)UD corpora
 - ▶ Data from Sequoia, Parseme and Orfeo projects
- ▶ Some semantics graphs
 - ▶ AMR (The Little Prince, Bio_AMR_Corpus)
 - ▶ 10 sentences from the PMB (ongoing project...)



<http://match.grew.fr>

The screenshot displays the Grew-match web interface. On the left, a sidebar lists corpora under 'AMR' (Little_Prince, Bio_AMR_Corpus, Spec_1.2.5) and 'PMB' (pmb-3.0.0-sample_en_gld). The main area shows a search for 'Bio_AMR_Corpus' with a pattern: `N -[ARG0]-> A;`, `N -[ARG1]-> *;`, and `N -[ARG2]-> *;`. It reports 424 occurrences in 0.27s. A list of results follows, with `bel_pmid_1072_9607.86` highlighted. Below the list is a semantic graph. The graph is a hierarchical tree structure. The root node is 'label = consistent-01'. It branches into 'ARG0-of ARG2' and 'ARG1'. 'ARG0-of ARG2' further branches into 'label = show-01' and 'label = data'. 'label = show-01' branches into 'ARG1' (labeled 'label = possible-01') and 'time'. 'label = data' branches into 'time' and 'poss' (labeled 'label = previous'). 'ARG1' branches into 'label = thing' and 'ARG2-of'. 'label = thing' branches into 'mod' (labeled 'label = this') and 'ARG2-of'. 'ARG2-of' branches into 'label = result-01'. 'label = possible-01' branches into 'ARG1' (labeled 'label = translocate-01'). 'label = translocate-01' branches into 'ARG3' (labeled 'label = cytosol'), 'ARG2' (labeled 'label = membrane'), 'ARG1' (labeled 'label = and'), and 'ARG0' (labeled 'label = small-molecule'). 'label = membrane' branches into 'mod' (labeled 'label = plasma'). 'label = and' branches into 'op2' (labeled 'label = enzyme') and 'op1' (labeled 'label = enzyme'). 'label = small-molecule' branches into 'name' (labeled 'label = name op1 = PAF'). 'label = enzyme' (under 'op2') branches into 'name' (labeled 'label = name op1 = PKCe'). 'label = enzyme' (under 'op1') branches into 'name' (labeled 'label = name op1 = PKCa'). A red text note above the graph states: 'These results are consistent with our previous data showing that PAF is able to translocate PKCa and PKCe from cytosol to plasma membrane'.

UD & Grew-match

UD annotation

Grew-match syntax

nsubj

```
pattern { M -[nsubj]-> N }
```

She

laughed

```
pattern { N [form="laughed"] }
```

she

laugh

```
pattern { N [lemma="laugh"] }
```

PRON

VERB

```
pattern { N [upos=VERB] }
```

Case=Nom

Mood=Ind

Gender=Fem

Tense=Past

```
pattern { N [Tense=Past] }
```

Number=Sing

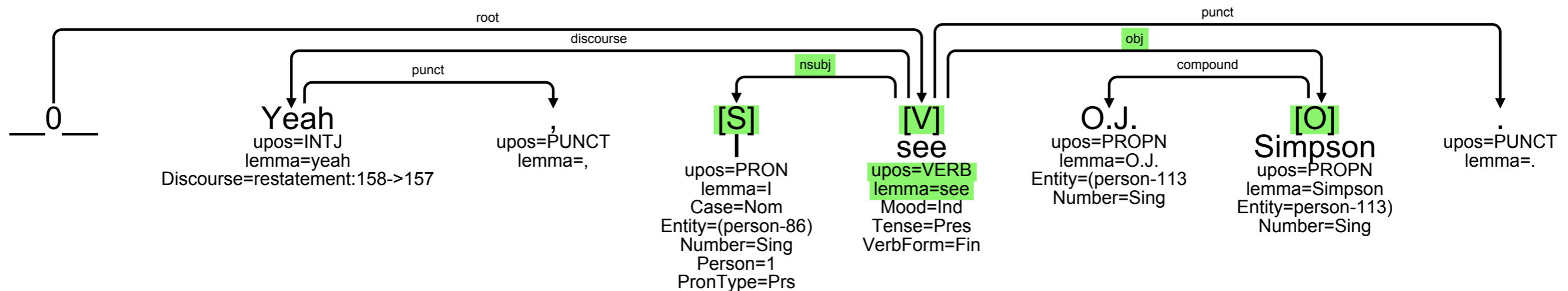
VerbForm=Fin

Person=3

PronType=Prs

Grew-match syntax

```
pattern { V [upos=VERB, lemma="see"]; V -[nsubj]-> S; V -[obj]-> O }
```



Your turn • Step 1

```
pattern { M -[nsubj]-> N }
```

```
pattern { N [form="laughed"] }
```

```
pattern { N [lemma="laugh"] }
```

```
pattern { V [upos=VERB, lemma="see"]; V -[nsubj]-> S; V -[obj]-> O }
```

```
pattern { N [lemma <> "see"] }
```

```
pattern { N [upos=VERB | AUX] }
```

```
pattern { N [upos=VERB] }
```

```
pattern { N [Tense=Past] }
```

The lemma is different from “see”

The POS is either VERB or AUX

1. In the corpus UD_English-GUM (selected by default), what are the words used with the POS “PART”?
2. In the corpus UD_French-GSD (left pane), what are the possible lemmas for POS “AUX”?
3. Chose the corpus you want, observe if it is possible to have two subjects on the same verb, two objects on the same verb.

Add `% your_name` as first line in your requests!

Grew-match • more syntax

N1 < N2 The node N1 is immediately before N2

```
pattern {  
  N1 [upos=DET]; N2 [upos=NOUN];  
  N1 < N2 }
```



N1 << N2 The node N1 is before N2

```
% left-headed nsubj  
pattern { G -[nsubj]-> D; G << D }
```



N1.f = N2.f
N1.f <> N2.f (In)equality of features

```
pattern {  
  N1 [upos=DET]; N2 [upos=NOUN];  
  N1 < N2;  
  N1.Gender <> N2.Gender }
```



without Filter out some occurrences

```
pattern { V [upos=VERB] }  
without { V -[nsubj]-> S }
```



```
pattern { N1 [upos=DET]; N2 [upos=NOUN]; N1 < N2 }  
without { N2 -[det]-> N1 }
```



Your turn • Step 2

```
pattern { V [upos=VERB, lemma="see"]; V -[nsubj]-> S; V -[obj]-> O }
```

```
pattern { N [lemma<>"see"] }      pattern { N [upos=VERB | AUX] }
```

```
pattern { N1 [upos=DET]; N2 [upos=NOUN]; N1 < N2 }
```

```
pattern { G -[nsubj]-> D; G << D }
```

```
pattern { N1 [upos=DET]; N2 [upos=NOUN]; N1 < N2; N1.Gender <> N2.Gender }
```

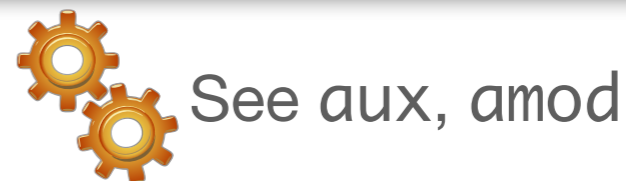
```
pattern { V [upos=VERB] } without { V -[nsubj]-> S }
```

```
pattern { N1[upos=DET]; N2[upos=NOUN]; N1 < N2 } without { N2 -[det]-> N1 }
```

1. How the trigram “in order to” is annotated in English corpora?
 - ➔ use “n-grams” from the snippets (on the right of the textarea)
2. Explore Verb/Subject number agreement in French
 - ➔ use several successive **without** clauses

Grew-match • Other stuff

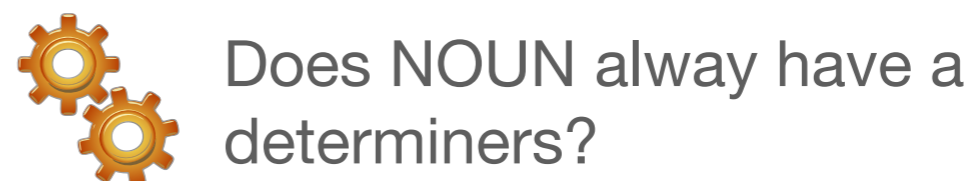
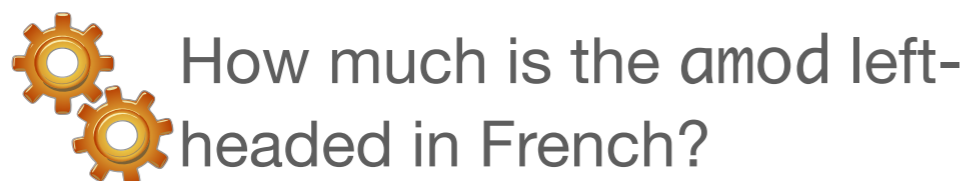
- ▶ **Relation tables:** See in one table what are the POS used for a given relation



- ▶ **Clustering:** put occurrences of a request in clusters, given a clustering key



- ▶ **Whether:** split occurrences given by the pattern in two clusters:
 - ▶ the Yes cluster where the additional constraints hold
 - ▶ the No clusters where the additional constraints do not hold (without)



How to make stats running several requests on a set of 183 corpus?

- ▶ use Grew command line interface
- ▶ dedicated web service: to be checked...

Grew-match • What's next?

- ▶ More corpora:
 - ▶ Focus on semantically annotated data (PMB...)
 - ▶ Difficulty of building readable graphical display
- ▶ Other kinds of graph used in NLP
 - ▶ Lexical database: experiment in RLF
- ▶ Multi-corpora request in the web interface
- ▶ Double clustering