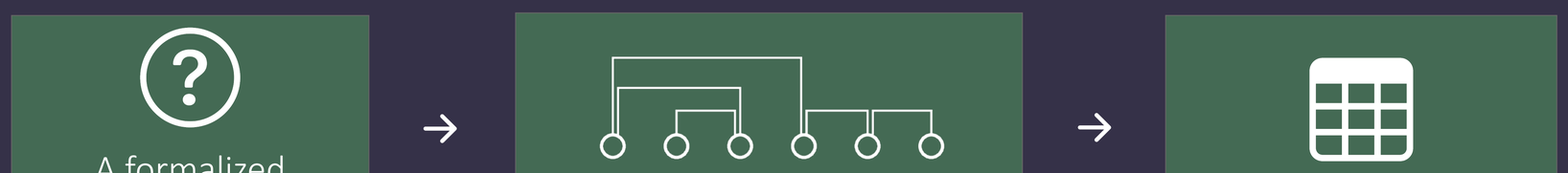
Sparse Logistic Regression with High-order Features for Automatic Grammar Rule Extraction from Treebanks

Santiago Herrera, Caio Corro and Sylvain Kahane

Graphical abstract

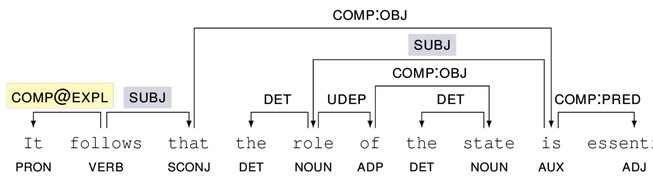


Expressive and quantitative fine-grained grammar rules for word-order and agreement

- Formalization of a syntactic rule
- ML method to extract and rank the rules
- An easy-to-interpret hierarchy of results

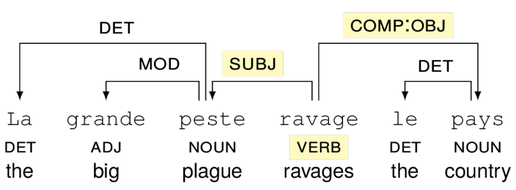
1 Grammatical descriptions

When **the subject follows the verb** in English?



	PATTERN P	ORDER	COVERAGE	PRECISION
2	siblings.rels_deep=comp@expl, siblings.upos=PRON	after	30.6	96.6
18	gov.lemma_AUX=do	after	17.4	27.4

When there is **number agreement** in French?



	PATTERN P	AGREEMENT	COVERAGE	PRECISION
1	gov.position=after_dep	yes	64.9	98.8
4	dep.upos=ADJ	yes	16.6	98.9
6	gov.in_upos=AUX|VERB, dep.rel_synt=comp:obj	no	27.4	34.5
8	dep.rel_synt=det	yes	46.1	99.7
36	gov.VerbForm=Fin, dep.rel_synt=subj	yes	13.2	98.1

2 Definition of a syntactic rule

$$\mathrm{S} \implies (\mathrm{P} \overset{\alpha\,\%}{\Longrightarrow} \mathrm{Q})\,.$$

e.g. Given all subjects (S), these are inverted (Q) when there is an expletive complement (P) in 96% of cases (α).

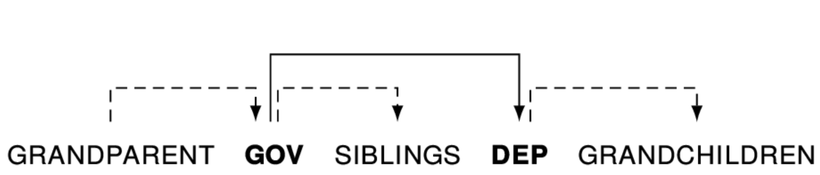
- **Probabilistic**
- **Potentially overlapping**
- **+/- Fine-grained**

3 Data and features

Work in 3 languages: **French, Spanish and Wolof**

Dependency treebanks : Universal Dependencies (UD) and Surface-Syntactic UD

The P patterns are filled in by the model using linguistic information within the following search space:



4 Rule extraction

Sparse Logistic Regression

$$P(\text{“number agreement”}|\boldsymbol{x}) = \sigma(\boldsymbol{a}^\top \boldsymbol{x} + b) \quad \text{where} \quad \sigma(w) = \frac{\exp\ w}{1 + \exp\ w}$$

Training Problem

$$\min_{\boldsymbol{a} \in \mathbb{R}^F, b \in \mathbb{R}} \frac{1}{|D|} \sum_{(\boldsymbol{x}, y) \in D} \ell(\boldsymbol{a}^\top \boldsymbol{x} + b; y) + \lambda r(\boldsymbol{a})$$

- Negative likelihood loss
- L1 Norm
- Ranks through the regularization parameter
- Sensible to less pronounced shift distributions

skglm

+ precision
+ coverage/recall
+ statistical tests

R Results

More grammar rules with more expressiveness

An (almost) hyper-parameter free method to extract and rank rules

A better adapted tool for rule mining

Acknowledgment: We thank Mathurin Massias and Badr Moufad for the help with the Skglm library. We also thank Bruno Guillaume, Guillaume Bonfante, and Guy Perrier for their insightful conversations and feedback.

LREC-COLING 2024