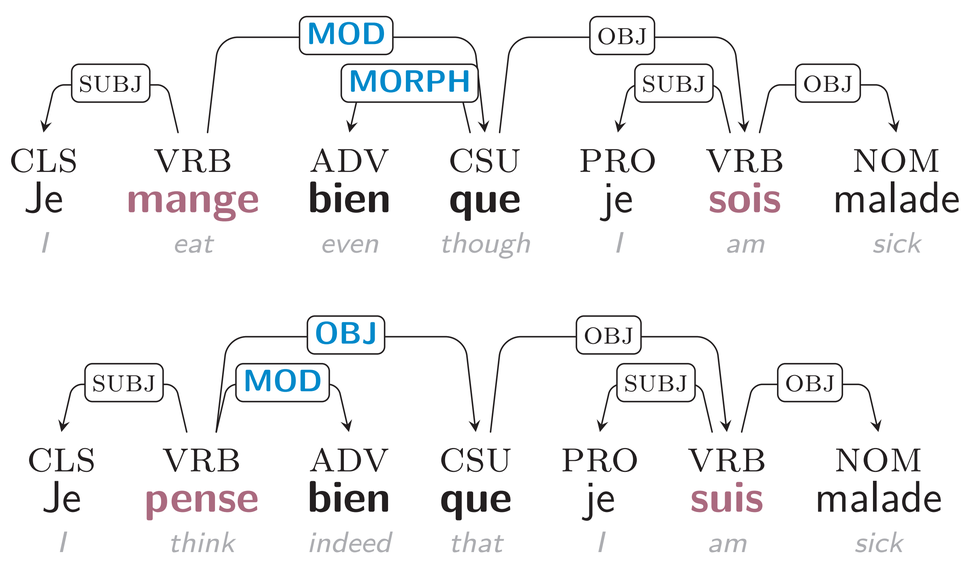
Complex Function Words in French: Representation and Disambiguation

Alexis Nasr Carlos Ramisch José Deulofeu André Valli
Aix Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille (France)
firstname.lastname@lif.univ-mrs.fr

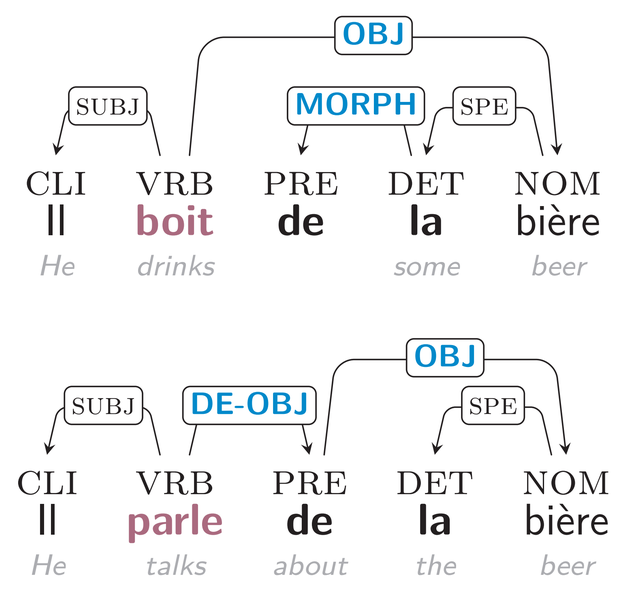
Summary

- **Complex function words** are often treated as **words-with-spaces** in parsers, although this is unrealistic when they are **ambiguous**
- We quantify the ambiguity of two types of complex function words in French: ADV+*que* **conjunctions** and *de*+DET **determiners**
- We propose a new model to represent them using special MORPH **dependency links** in **treebanks** and **parsers**
- We train and test a **graph-based probabilistic dependency parser** on a modified treebank
- We show that **joint segmentation and MWE labeling** performs better than pretokenization

ADV+*que* Constructions



de+DET Constructions



MORPH Dataset

Annotation methodology

1. Select target ADV+*que* and *de*+DET
   - Frequent in corpora
   - Syntactically ambiguous
   - Not included in larger chunks
2. Extract ∼100 sentences from FRWaC
   - Containing 1 occurrence of target
   - Well formed, 10 to 20 words
3. Annotate occurrences as MORPH/other

7 most frequent ADV+que *combinations: Proportion of complex conjunctions* (MORPH)

ADV+*que*	#Sent	Conj.	Other
ainsi que	103	76%	24%
alors que	110	88%	12%
autant que	107	86%	14%
bien que	99	37%	63%
encore que	93	21%	79%
maintenant que	120	57%	43%
tant que	98	20%	80%
Total	**730**	**56%**	**44%**

4 possible de+DET *combinations: Proportion of complex determiners* (MORPH)

de+DET	#Sent	Det.	Other
de la	138	20%	80%
de les (des)	129	77%	23%
de le (du)	136	33%	67%
de l'	136	15%	85%
Total	**539**	**36%**	**64%**

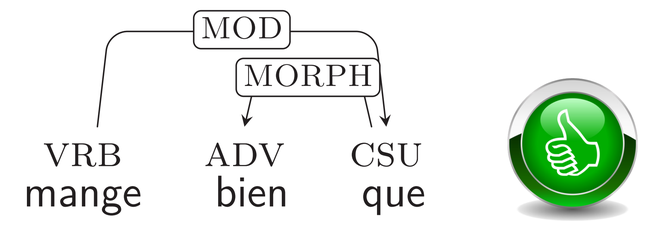
Dependency Parser Training

Parsing model

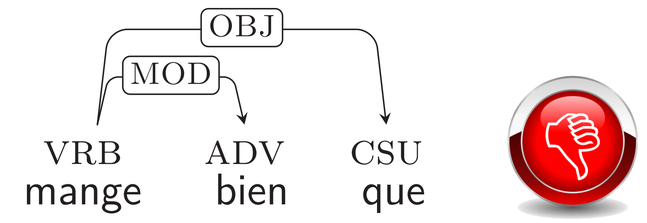
- Probabilistic 2nd-order graph-based dependency parser
- Trained on modified version of the French Treebank

$$bien\_que_{\mathrm{CSU}} \Longrightarrow bien_{\mathrm{ADV}} \overset{\mathrm{MORPH}}{\longleftarrow} que_{\mathrm{CSU}}$$

- Three feature templates with varying degrees of lexicalization:
  1. *first-order factors* : $w_2 \leftarrow w_1$
  2. *sibling factors* : $w_3 \leftarrow w_2 \rightarrow w_1$
  3. *grandchildren factors* : $w_3 \leftarrow w_2 \leftarrow w_1$



Frequent sibling-factor feature value: positive score



Implausible grandchildren-factor feature value: negative score

Results on French Treebank

	ADV+*que*	*de*+DET
LAS	88.98	89.02
UAS	90.63	90.23
# MORPH	27	145
Prec.	87.10	85.85
Rec.	100	81.12

Results on MORPH ADV+*que*

ADV+*que*	Prec.	Recall	F1
ainsi que	96.0	91.1	93.5
alors que	92.8	92.8	92.8
autant que	86.9	65.2	74.5
bien que	86.8	89.2	88.0
encore que	72.7	80.0	76.2
maintenant que	85.2	77.6	81.3
tant que	78.9	75.0	76.9
Total	**88.7**	**82.0**	**85.2**

Results on MORPH *de*+DET

de+DET	Prec.	Recall	F1
de le	72.50	64.44	68.23
de la	58.13	86.20	69.44
de les	97.36	74.00	84.09
de l'	57.14	69.56	62.74
Total	**77.00**	**73.09**	**75.00**