On the interplay between complex function words and parsing

Carlos Ramisch

Presented at ACL 2015

Alexis Nasr, José Deulofeu, André Valli

UFRGS, October 20, 2015

Outline

Complex function words

Function words

Words that act as **grammatical functions**, like prepositions, determiners, conjunctions, pronouns

Multiword expressions

Idiosyncratic combinations of 2 or more words that must be **interpreted as a unit** in linguistic analysis

Complex function words

Multiword expressions that act as function words, for example :

- **Complex prepositions**: *en face de*, *de par*, *à part*
- **Complex determiners**: *de la*, *beaucoup de*
- **Complex conjunctions**: *même si*, *ainsi que*, *si bien que*

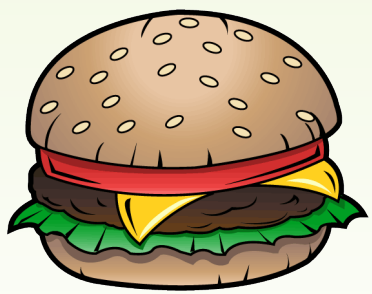
ADV+*que* constructions

Je mange **bien que** *je n'aie pas faim*

I eat although I am not hungry

Je pense **bien que** *je n'ai pas faim*

I think indeed that I am not hungry



de+DET constructions

Il boit **de la** *bière*

He drinks some beer

Il parle **de la** *bière*

He talks about the beer

Motivation

- **Ambiguity**
  - both interpretations are frequent
  - requires syntax information (e.g. verb valency)
  - generally dealt with during tokenisation
- **Frequency**
  - frWaC corpus - 1.6G words :
    - ADV+*que* present in 2.1% of sentences
    - *de*+DET present in 48.6% of sentences
  - *des* - 7^{th} most frequent word

Outline

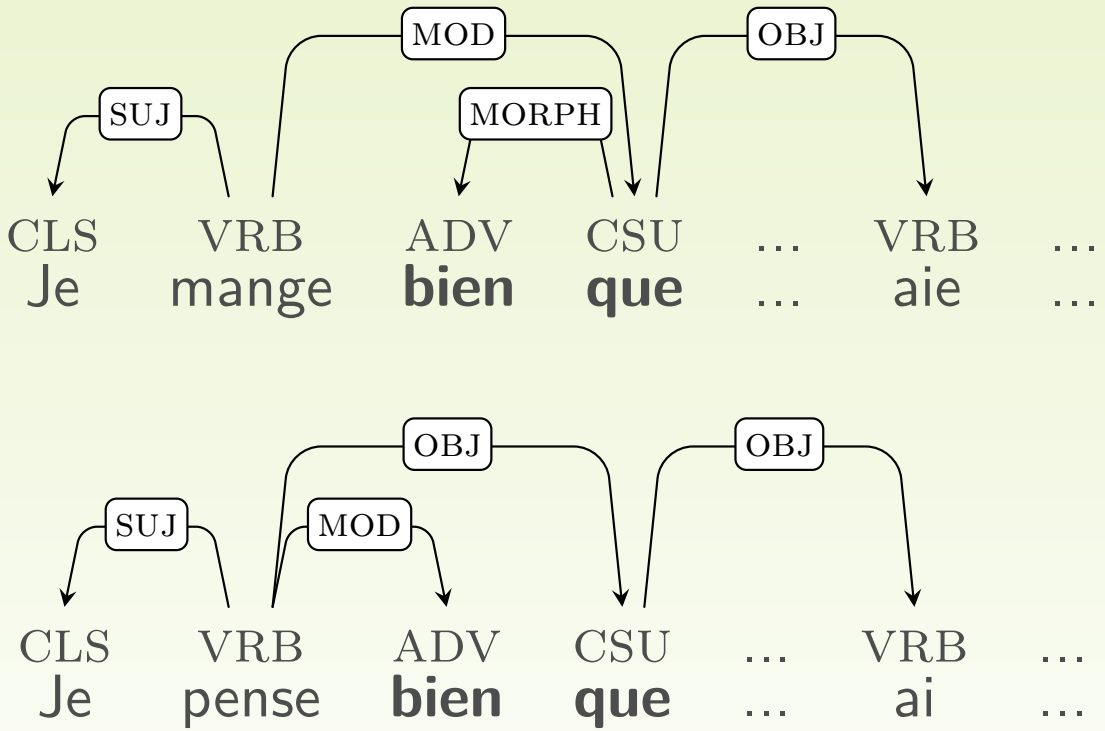
Main idea

- Detection of ADV+*que* & *de*+DET from tokeniser to parser
- Words systematically tokenised as separated units
- Introduction of MORPH link for complex function words
- Transform a **segmentation** into a **linking** problem
  1. Modify training data

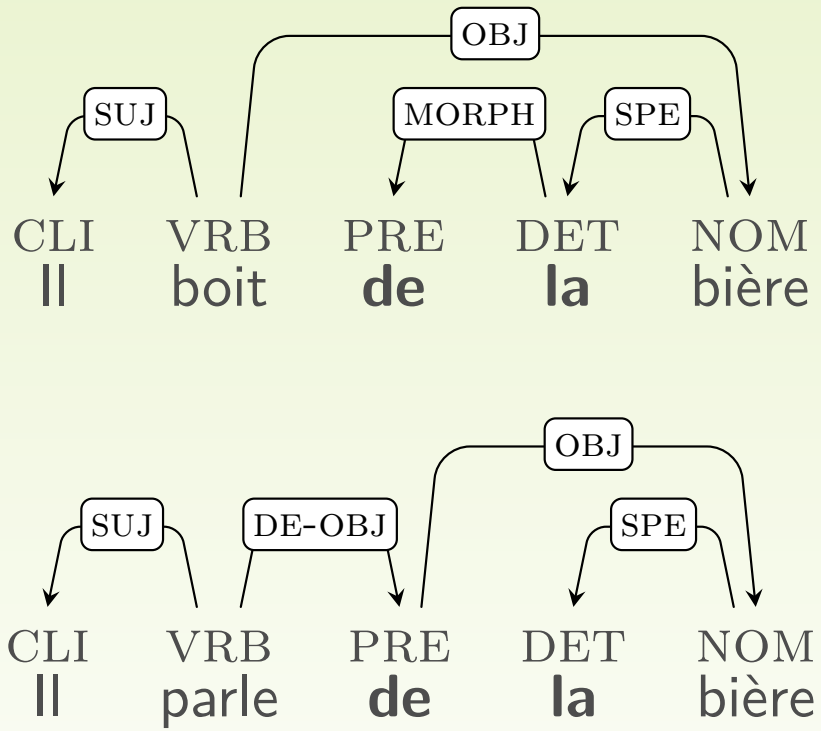
$$\textit{bien\_que}_{\text{CSU}} \Longrightarrow \textit{bien}_{\text{ADV}} \overset{\text{MORPH}}{\longleftarrow} \textit{que}_{\text{CSU}}$$

  2. Include syntactic resources into parser

ADV+*que* constructions



de+DET constructions



Parser

- Probabilistic linking model, trained on **treebank**
- Given a sentence $W = w_1 \dots w_l$, the parser looks for the dependency tree $\hat{T}$ of W that maximizes the score s:

$$\hat{T} = \underset{T \in \mathcal{T}(W)}{\text{argmax}} \sum_{F \in \mathcal{F}(T)} s(F)$$

- $\mathcal{T}(W)$: set of all possible dependency trees for sentence W
- $\mathcal{F}(T)$: set of all relevant subparts, called *factors*, of tree T,
- $s(F)$: score of factor F, estimated from data

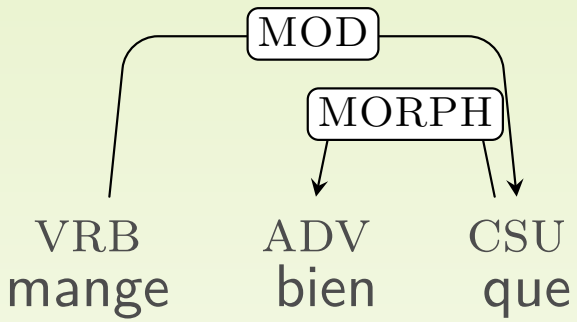
Factors $\mathcal{F}(T)$

1. *first-order factors* : 

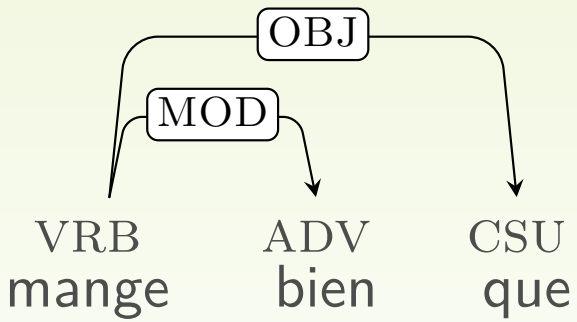
2. *sibling factors* : 

3. *grandchildren factors* : 

Example



Frequent sibling-factor feature value: positive score



Implausible grandchildren-factor feature value: negative score

External resource: syntactic lexicon

- Types of arguments accepted by each verb

manger	-QUE	-DE
penser	+QUE	-DE
boire	-QUE	-DE
parler	-QUE	+DE

Factor templates

1	G.SF	G.POS	D.FCT	D.POS	
2	G.SF	G.POS	D.FCT	D.POS	GD.POS
3	G.SF	G.POS	D.FCT	D.LEM	GD.POS

Outline

French Treebank

- Syntactically annotated news articles from *Le Monde* [Abeillé et al., 2003]
- **Dependency trees** version [Candito et al., 2009]
- 9,881 sentences (278K words) **training**
- 1,235 sentences (36K words) **test**
- Rules to transform flat representation of ADV+*que* and *de*+DET constructions into MORPH-**linked individual tokens**

MORPH dataset

- ∼100 sentences per target annotated by 2 native speakers

ADV+*que*	#sent	conj.	other	#occur
ainsi	103	76.7	23.3	498,377
alors	110	88.2	11.8	291,235
autant	107	86.0	14.0	39,401
bien	99	37.4	62.6	156,798
encore	93	21.5	78.5	18,394
maintenant	120	55.8	44.2	16,567
tant	98	20.4	79.6	168,485
Total	730	56.4	43.6	1,189,257

de+DET	#sent	det.	other	#occur
le (du)	136	33.1	66.9	16,609,049
la	138	21.0	79.0	10,849,384
les (des)	129	77.5	22.5	23,395,857
l'	136	16.9	83.1	8,204,687
Total	539	36.5	63.5	59,058,977

Dicovalence Lexicon

- subcategorization frames of $\sim 3,700$ French verbs[1]
- Number of verbs in DicoValence per value of subcat feature.

-QUE	+QUE	-DE	+DE
3,814	356	3,450	720

[1] http://bach.arts.kuleuven.be/dicovalence/

Evaluation

- Use of MACAON tool suite
- Standard LAS and UAS on FTB-test
- Precision, recall and F1 of MORPH prediction on MORPH dataset
- Comparison with majority baseline and Stanford parser [Green et al., 2013][2]

[2] http://nlp.stanford.edu/software/lex-parser.shtml

Outline

Results for ADV+*que* I

ADV+*que*	base-line	Green et al. (2013)	Without SF		
			Prec.	Recall	F1
ainsi que	76.7	81.44	96.00	91.14	93.50
alors que	88.2	95.10	92.78	92.78	92.78
autant que	86.0	92.00	86.95	65.21	74.53
bien que	37.4	55.22	86.84	89.18	88.00
encore que	21.5	64.52	72.72	80.00	76.19
maintenant que	55.8	87.01	85.24	77.61	81.25
tant que	20.4	90.91	78.94	75.00	76.92
Total	56.4	83.06	88.71	**82.03**	85.24

Results for ADV+*que* II

	With SF		
ADV+*que*	**Prec.**	**Recall**	**F1**
ainsi que	95.94	89.87	92.81
alors que	93.81	93.81	93.81
autant que	86.66	70.65	77.84
bien que	91.66	89.18	90.41
encore que	92.85	65.00	76.47
maintenant que	90.91	74.62	81.96
tant que	82.35	70.00	75.67
Total	**91.57**	81.79	**86.41**

Results for *de*+DET I

de+DET	base-line	Green et al. (2013)	Without SF		
			Prec.	Recall	F1
de le	66.9	56.96	72.50	64.44	68.23
de la	79.0	22.83	58.13	86.20	69.44
de les	22.5	87.72	97.36	74.00	84.09
de l'	83.1	18.55	57.14	69.56	62.74
Total	63.5	44.37	77.00	73.09	75.00

Results for *de*+DET II

	With SF		
de+DET	**Prec.**	**Recall**	**F1**
de le	85.41	91.11	88.17
de la	81.25	89.65	85.24
de les	98.70	76.00	85.87
de l'	64.51	86.95	74.07
Total	**86.70**	**82.74**	**84.67**

Orchestration

1. Deal with "easy" MWEs before parsing
2. Deal with ambiguous ones during parsing
3. Deal with semantically transparent MWEs after parsing

Orchestration

1. Deal with "easy" MWEs before parsing
2. Deal with ambiguous ones during parsing
3. Deal with semantically transparent MWEs after parsing

How to determine which MWEs are easy?

→ lexical resources

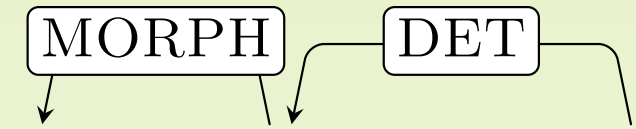
CPRE and CCONJ lexicon

1. Theoretical ambiguity
   $\Longrightarrow$ Can you have accidental cooccurrence?
2. Practical ambiguity $\Longrightarrow$ How frequent is accidental cooccurrence?

Outline

Conclusions

- Proposed solution is more precise than usual pipeline
- Linguistic representation is more accurate
- Future: generalisation of MORPH link
- Future: inclusion of more sophisticated information



de les questions ?

References I

Abeillé, A., Clément, L., and Toussenel, F. (2003).
Building a treebank for french.
In Abeillé, A., editor, *Treebanks: building and using parsed corpora*, pages 165–168. Kluwer academic publishers, Dordrecht, The Netherlands.

Candito, M., Crabbé, B., Denis, P., and Guérin, F. (2009).
Analyse syntaxique du français : des constituants aux dépendances.
In *of Traitement Automatique des Langues Naturelles*, Senlis, France.

Green, S., de Marneffe, M.-C., and Manning, C. D. (2013).
Parsing models for identifying multiword expressions.
Comp. Ling., 39(1):195–227.