Text Mining for Discourse Markers

Karolin Boczoń [1,3] Jacques Jayez [1,2]

[1]LORIA [2]ENS de Lyon [3]IDMC

13 June 2023

Ambiguity

The biggest challenge in recognizing discourse markers is their *ambiguity*.

bon, *bien*: adjective? adverb? noun? DM?
après: preposition? temporal adverb? concessive adverb?

Model

- CamemBERT [2] – large language model for French
  finetuned on
- PERCEO [1] – high-quality corpus with morpho-syntactic labels

	Bon	,	tout	est	*bon*	.
camembert-perceo	**INT**	KON	PRO:ind	VER:pres	**ADJ**	FNO
spacy	**ADJ**	PUNCT	ADV	VERB	**ADJ**	PUNCT

Table 1: Example POS-tagging comparision

Unstable results

Après	il y	a	un	problème	de	securite	.
ADV	PRO	VER	DET	NOM	PRP	NOM	FNO
Après	la	maison	*vous*	*trouvez*	un	champ	.
PRP	DET	NOM	PRO	VER	DET	NOM	FNO
Après	la	maison	*on*	*trouve*	un	champ	.
ADV	DET	NOM	PRO	VER	DET	NOM	FNO

Table 2: Adverbial and prepositional uses of *après*

Unwanted decomposition

afin de	ADV + PRP	☹
afin que	ADV + KON	☹
après tout	ADV	☺
au contraire	ADV	☺
d'abord	ADV	☺
du coup	ADV	☺

Table 3: Examples of composed discourse markers

But for *Paul a souffert du coup qu'il a reçu*, *du coup* remains ADV.

Unitex

It is possible to manually model the environment for the discourse marker use and other uses based on a native speaker's intuition.

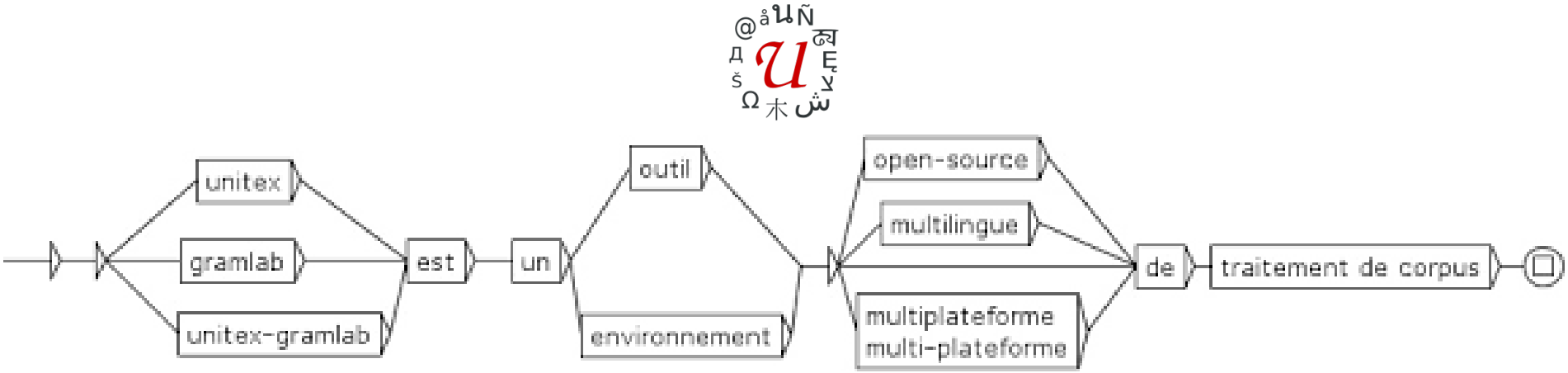


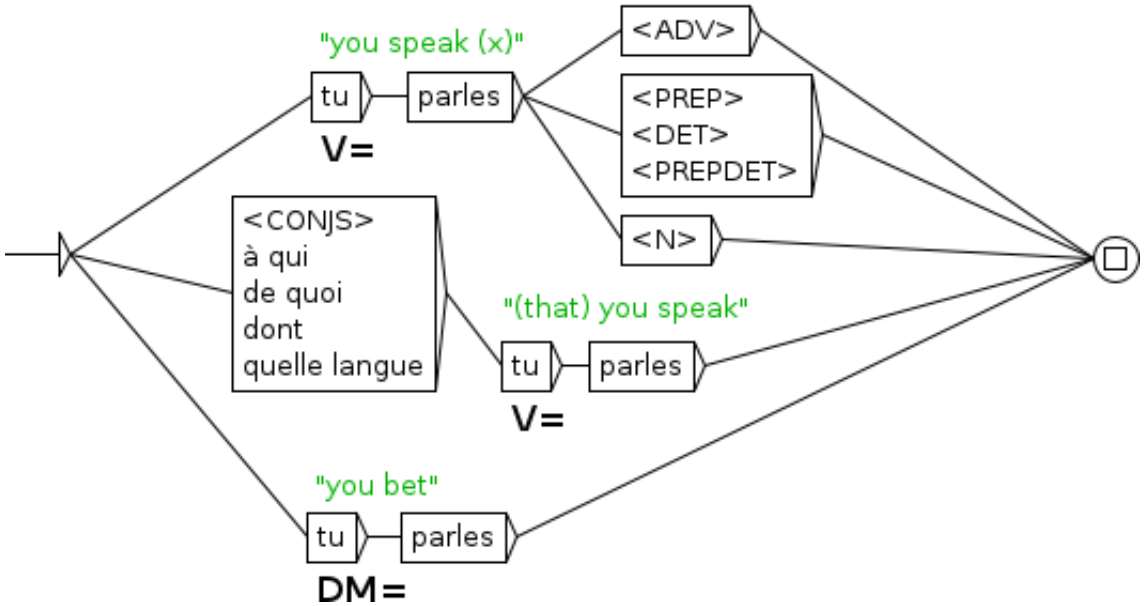
Figure 1: An example of a local grammar graph in Unitex

Example

Tu parles is a pronoun + verb sequence [=you speak] if it is

- followed by an adverb (*parler bien*, *vite*, *trop* ...)
- followed by a noun phrase (*parler [langue]*, *à qqn*, *de qqch*, *avec qqn* ...)
- preceded by a conjunction or a relative pronoun (*que*, *quand*, *à qui*, *dont* ...)

In all the other cases it should be a discourse marker [=you bet].



Example

left context	matched sequence	right context
tu te pointes devant un paysan, même si	V=tu parles bien$_{ADV}$	français, faut surtout pas
Ils sont restés sans voix : Enfin,	V=tu parles pas$_{ADV}$	sérieusement ?
Mais comment, me répond-on,	V=tu parles chinois$_{N}$	?
Lorsque	V=tu parles des$_{DET}$	règles de la radio, de ton souci
Regarde	à qui$_{\text{à qui}}$ V=tu parles	, ont répliqué ses camarades
On a vu l'accident, mais pas cette tête	dont$_{dont}$ V=tu parles	.
Quand tu te pointes devant un paysan,	même si$_{CONJS}$ V=tu parles	bien français, faut surtout
Tais-toi	quand$_{CONJS}$ V=tu parles	
Tu nous a coûté 2 millions, il faudra	que$_{CONJS}$ V=tu parles	, que tu dises où sont tes amis.
Mais, quand même, on les voit ! _ Oh !	DM=tu parles	, qu'est-ce que tu en as de plus ?
Ah, la fierté, le courage,	DM=tu parles	.
	DM=Tu parles	!
va pas en faire toute une histoire. _	V=Tu parles d'$_{DET}$	une histoire !
Ils disent qu'on est l'avenir,	V=tu parles d'$_{DET}$	un avenir !

Table 4: Results of parsing with the automaton

It is possible to manually model the environment for the discourse marker use and other uses based on a native speaker's intuition.

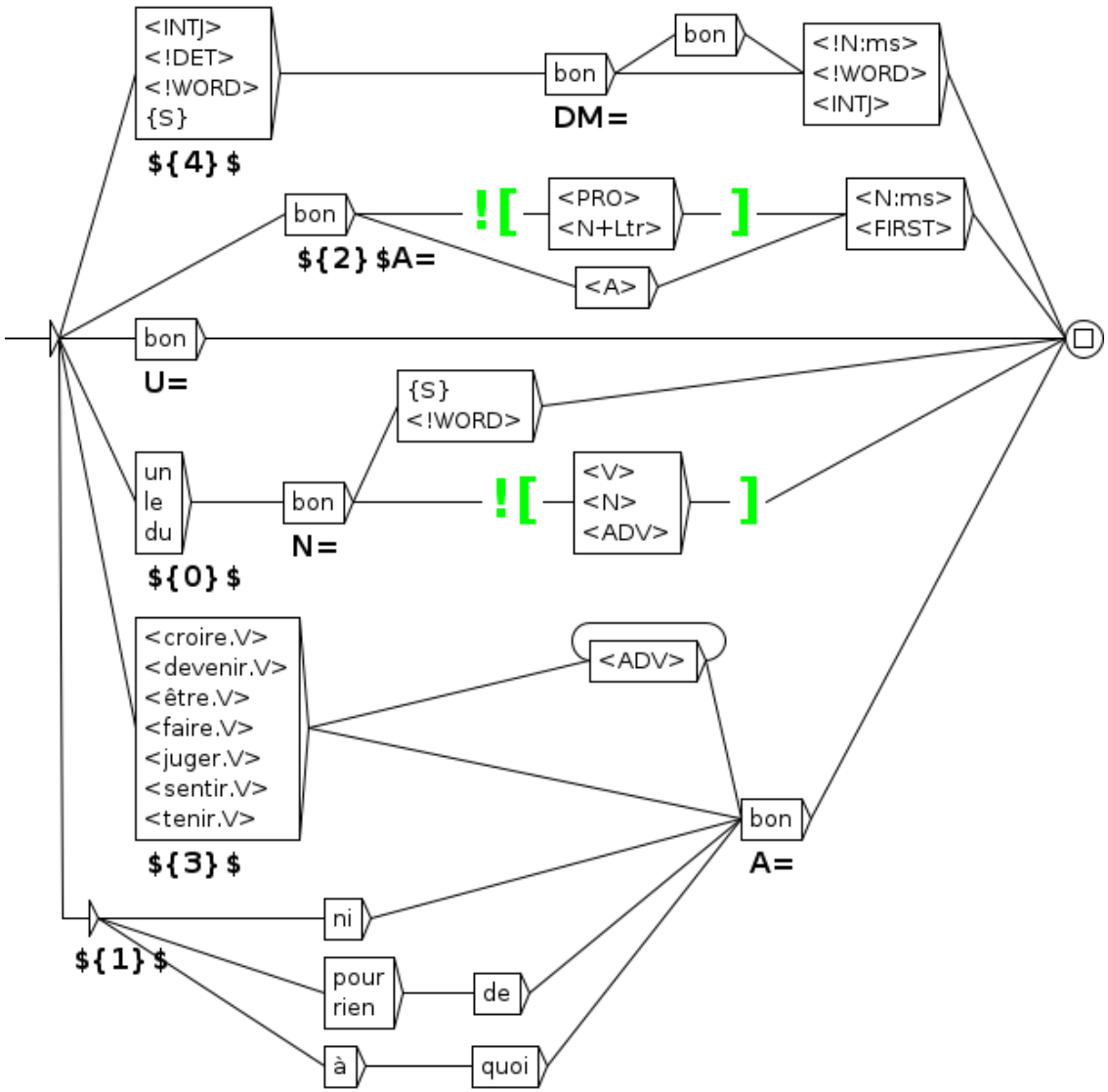


Figure 2: Local grammar for bon

But it is also tedious, time-consuming and not very robust.

A hybrid strategy

- Not always possible (results of taggers uninteresting).
- Script: tagging $\longrightarrow$ 'geometry' of mistakes $\longrightarrow$ finite automaton (FA).
- Typical mistakes
  1. Ignorance of idioms.
  2. Function/category confusion.
  3. Some black holes (DL black boxes).

A hybrid strategy, cont'd

- Example: *bon* as a DM in a Le Monde + CRFP mixed corpus.
- $\approx$ 61000 occurrences, sampled to 2000 and reviewed.
- $\approx$ 9% errors.
- Low precision (0.6), high recall (0.97, virtually no false negative).
- Uneven error distribution.

A hybrid strategy, cont'd

- In terms of types.

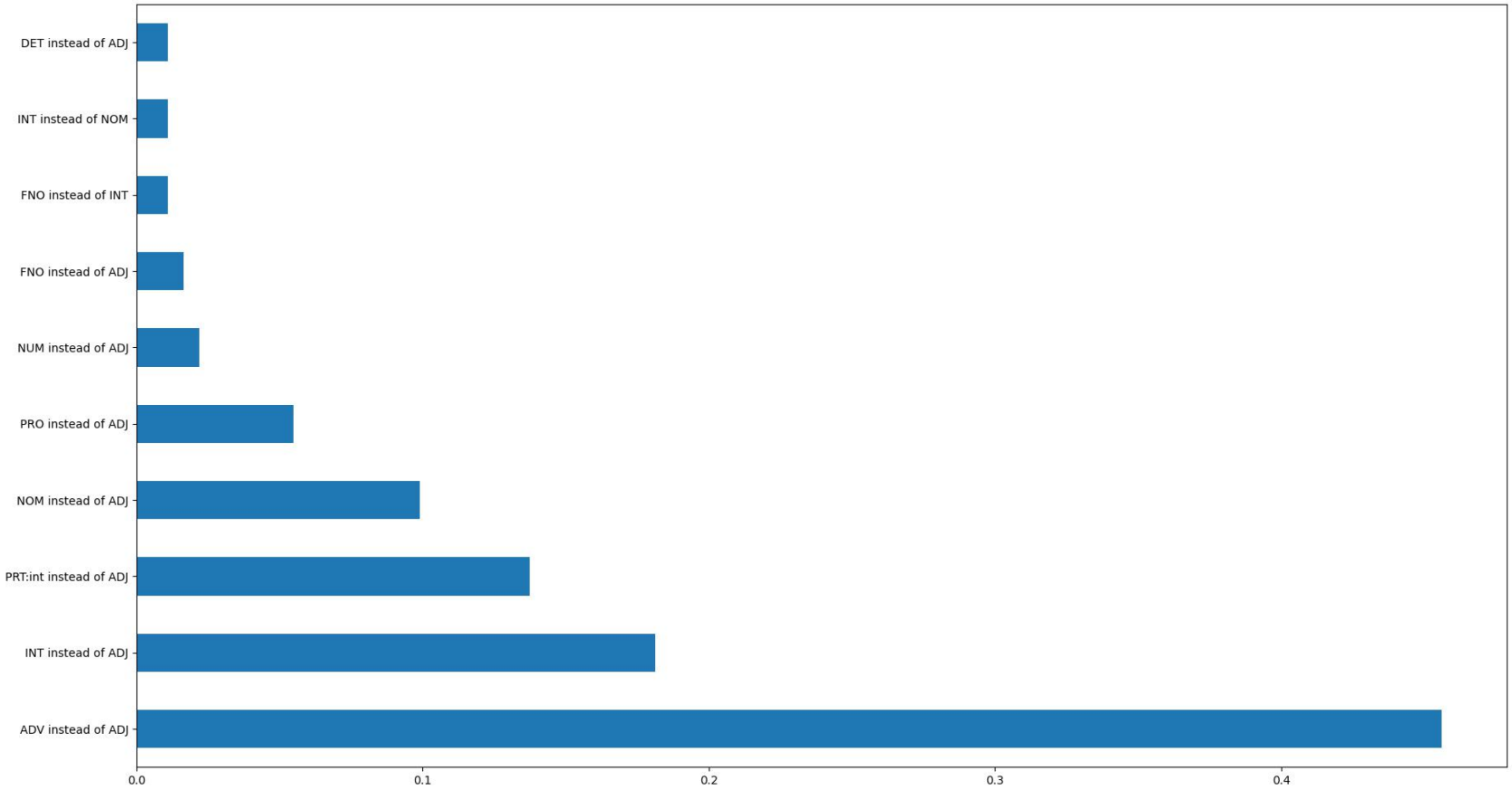


Figure 3: Error Type Distribution

A hybrid strategy, cont'd

- In terms of expressions.

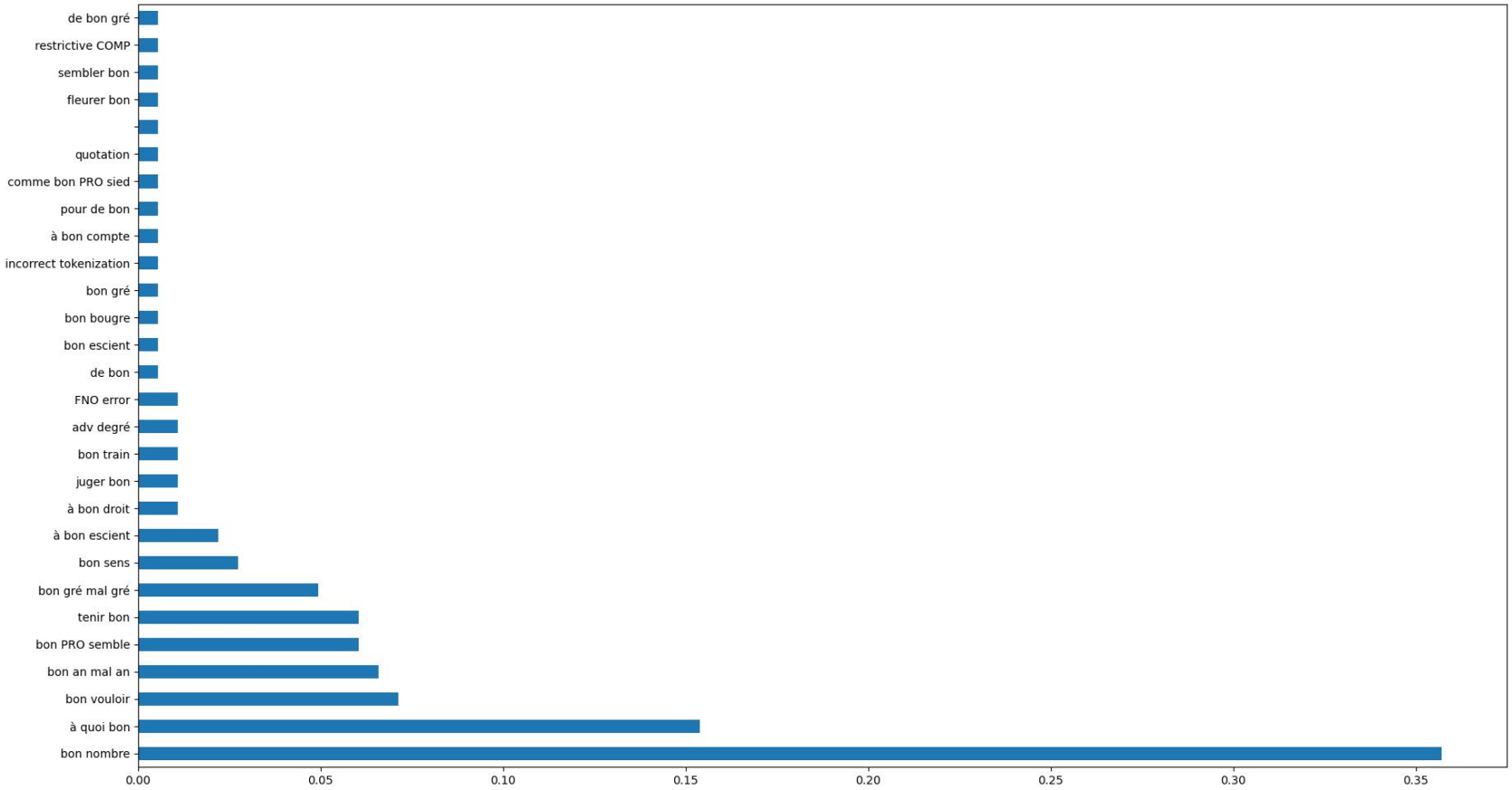


Figure 4: Error Distribution by Expression

A hybrid strategy, cont'd

- Tokenization may be a problem.

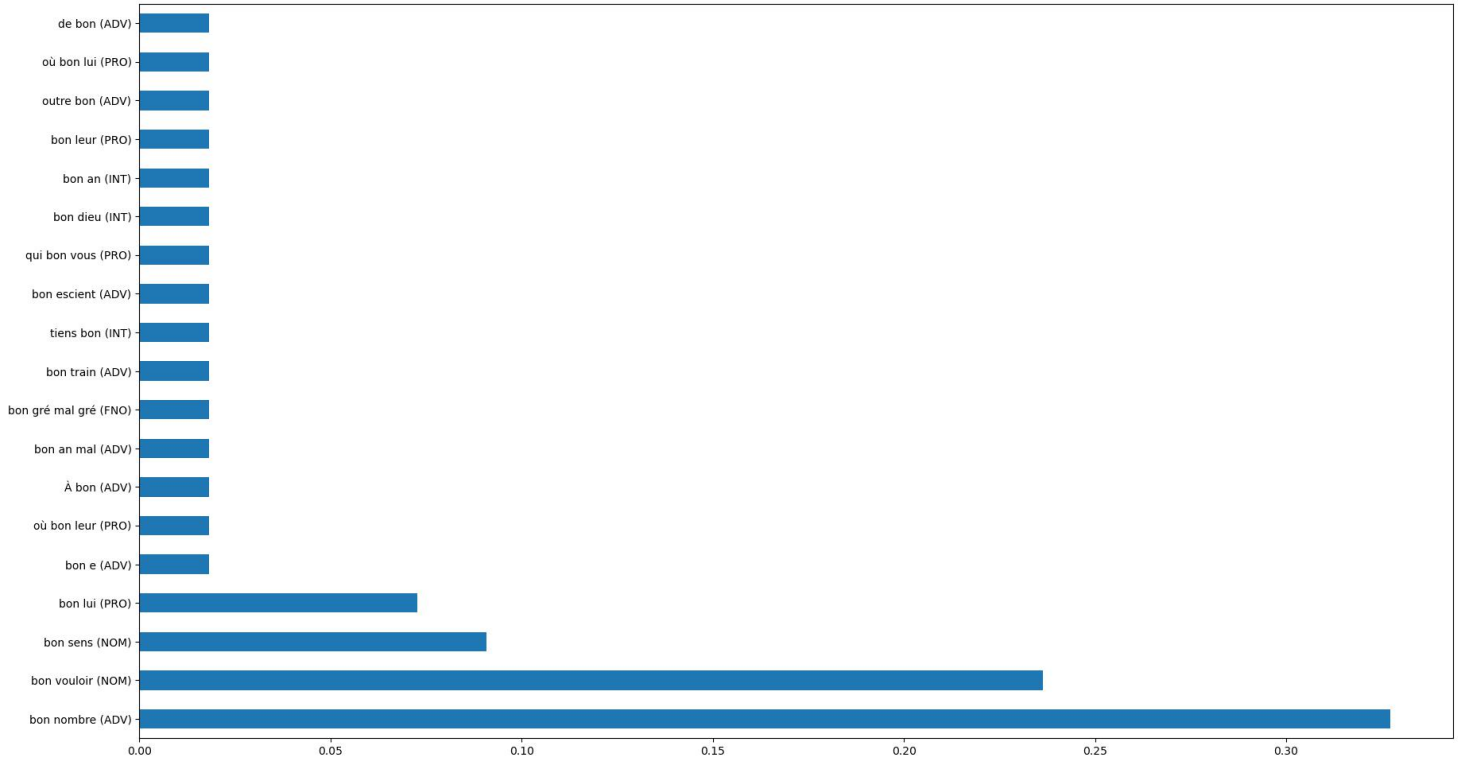


Figure 5: Particular tokenizations

A hybrid strategy, cont'd

- Using 'repulsion'.
- Environments where *bon* **cannot** be a DM.
- Unitex subgraphs, for instance degree expressions.

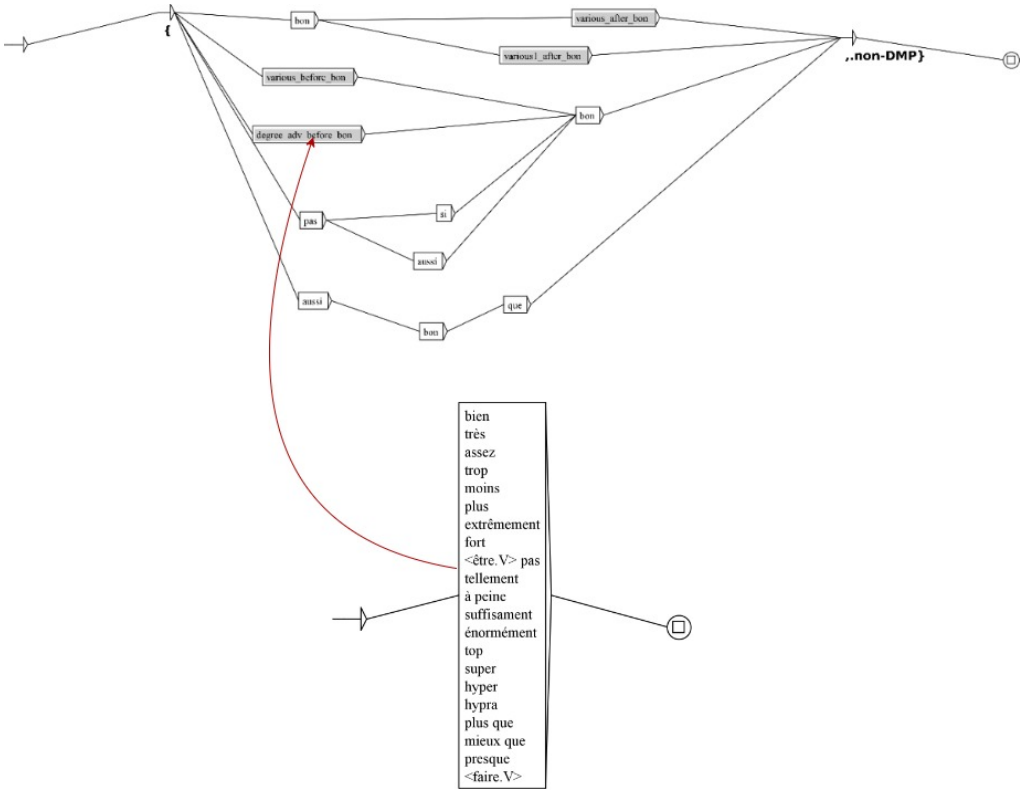


Figure 6: A subgraph in a cascade stage

A hybrid strategy, cont'd

- Final stage: keep the rest.

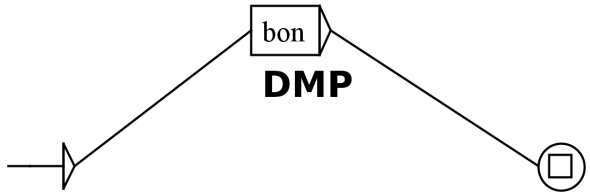


Figure 7: Last Stage (winning the bet)

A Hybrid strategy, cont'd

- Output for a subcorpus of ESLO (`eslo.huma-num.fr/`).
- 1800 results, out of 1,1511403 words, vocabulary = 17216.
- Eliminates all the previous problems.
- But not the problems with *après* or *tu parles*...

Final diagnostic

- Major problems when the category of an item depends on syntactic structure.
- *Attends le train* (V + NP) vs. *Attends le train est en retard* (V + S).
- Taggers practically useless.
- FA more flexible. Can start a cascade

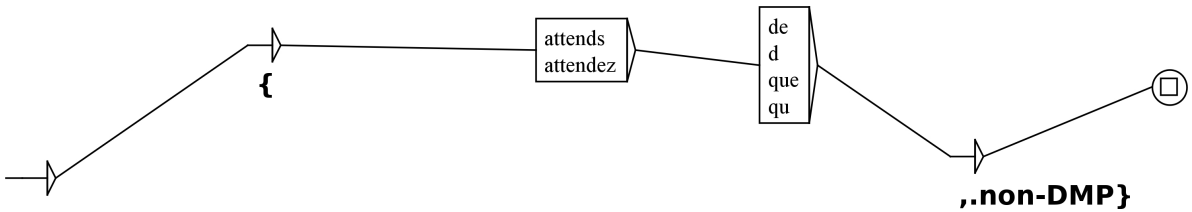


Figure 8: Initial stage

Final diagnostic

- Problem: constraints like.

If there is a finite verb on the right of *attends* and no intervening relative pronoun or complementizer, then we probably have a DM.

- Difficult (possible?) to express in Unitex.
- Easier with an external script.

Final diagnostic

- Conclusion.
  1. Stick to Unitex for non-ambiguous items.
  2. Apply the hybrid method when cascades are efficient.
  3. Use external scripts (possibly after a cascade).

References

ATILF, INIST & LIPN – "Perceo : un projet d'etiqueteur robuste pour l'ecrit et pour l'oral", 2012, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah & B. Sagot – "CamemBERT: a tasty French language model", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online), Association for Computational Linguistics, 2020, p. 7203–7219.