

Clustering pour grammaires catégorielles du second ordre

Noémie-Fleur Sandillon-Rezer

Directeurs de thèse Richard Moot Christian Retoré

Mixeur - TALN 2013



Objectif

- Présentation de notre nouvel algorithme d'inférence grammaticale pour grammaires AB,
- Proposition de traitement spécial pour les adverbes et modificateurs.

Rappel : Les grammaires AB

$$\frac{A/B \quad B}{A} [/E]$$

$$\frac{B \quad B \backslash A}{A} [\backslash E]$$

$$\frac{\forall X.A}{A[B/X]} [\forall E]$$

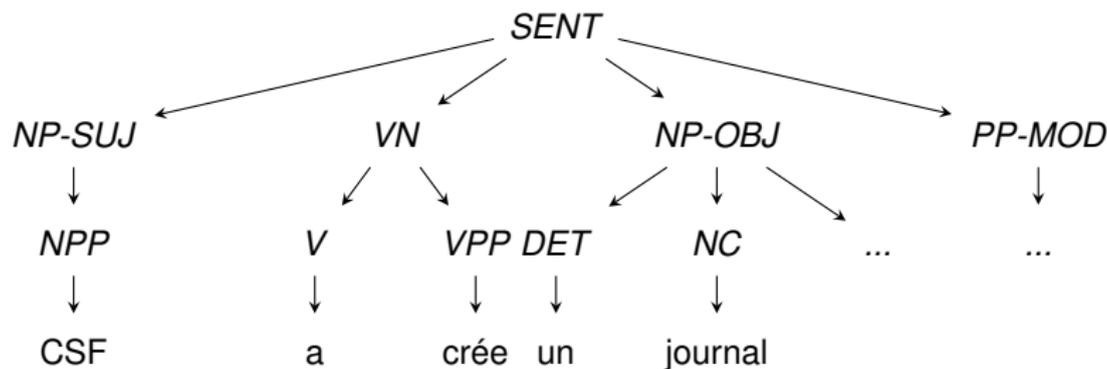


Lambek, J. : The mathematics of sentence structure (1958)

Résumé

- Méthode dans l'idée d'un apprentissage à la Buszkowski et Penn/ Kanazawa,
- Unification des variables non pas pour réduire le nombre de types par mot mais le nombre de types en tout,
- Utilisation des informations des arbres syntaxiques pour créer les arbres pris en entrée par l'algorithme,
- Utilisation de clustering hiérarchique pour guider l'unification.

Corpus de Paris VII et Sequoia



Abeillé, A., Clément, L., Toussnel, F. : Building a treebank for french (2003)

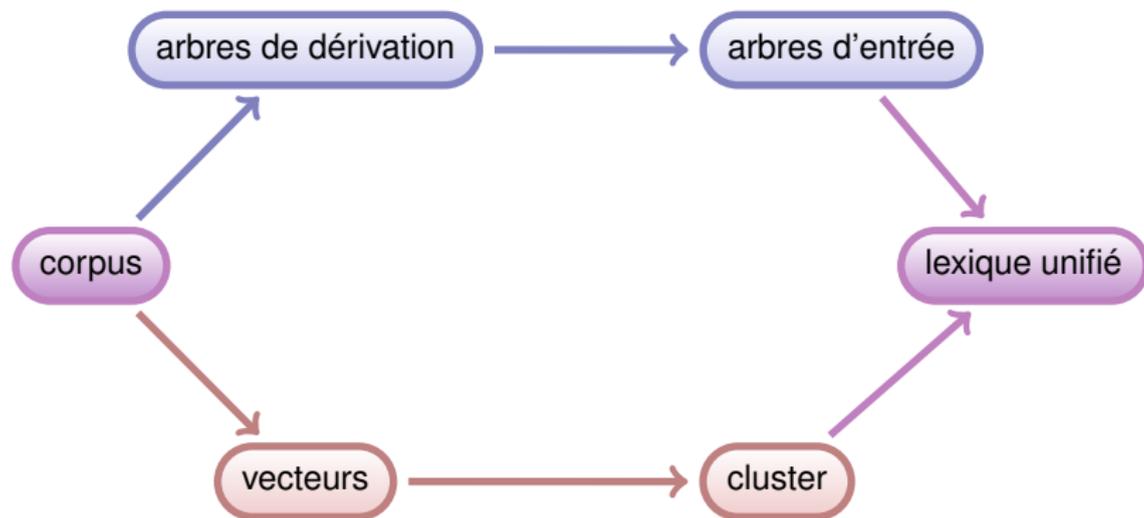


Candito, M., Seddah, D. : Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (2012)

Plan

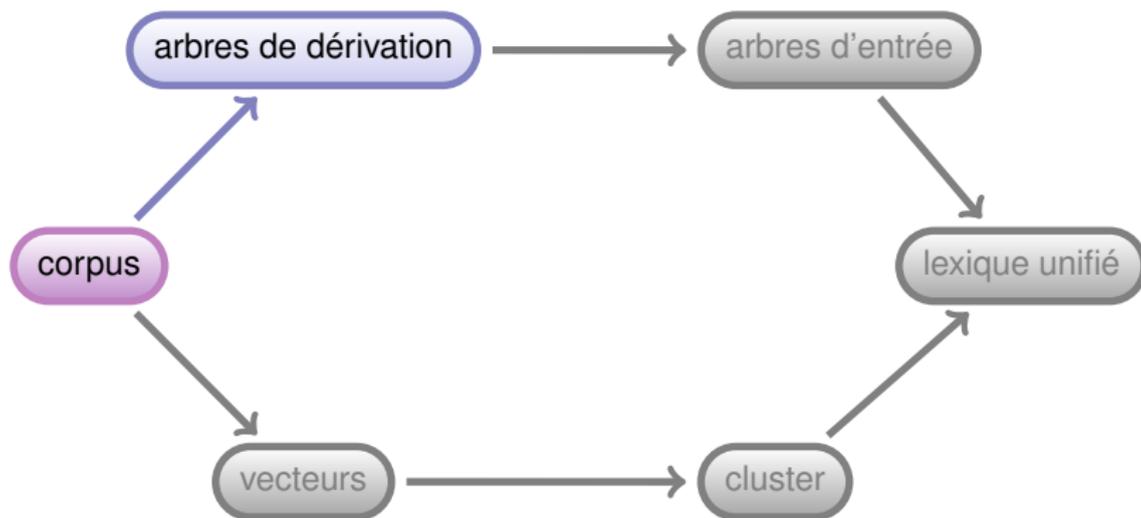
- 1 La méthode
- 2 Adverbes et autres modificateurs
- 3 Conclusion

Vue d'ensemble

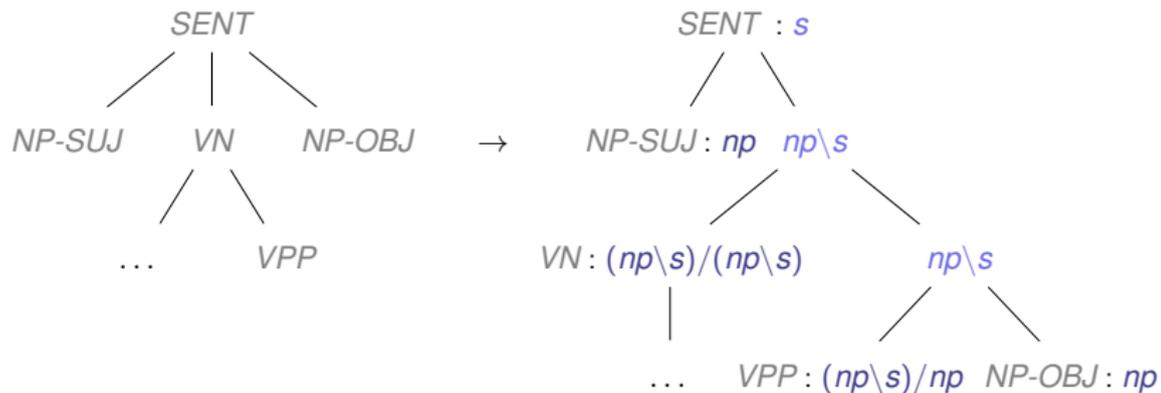


Transducteur et extraction d'arbres
Extraction de vecteurs et clustering
Unification

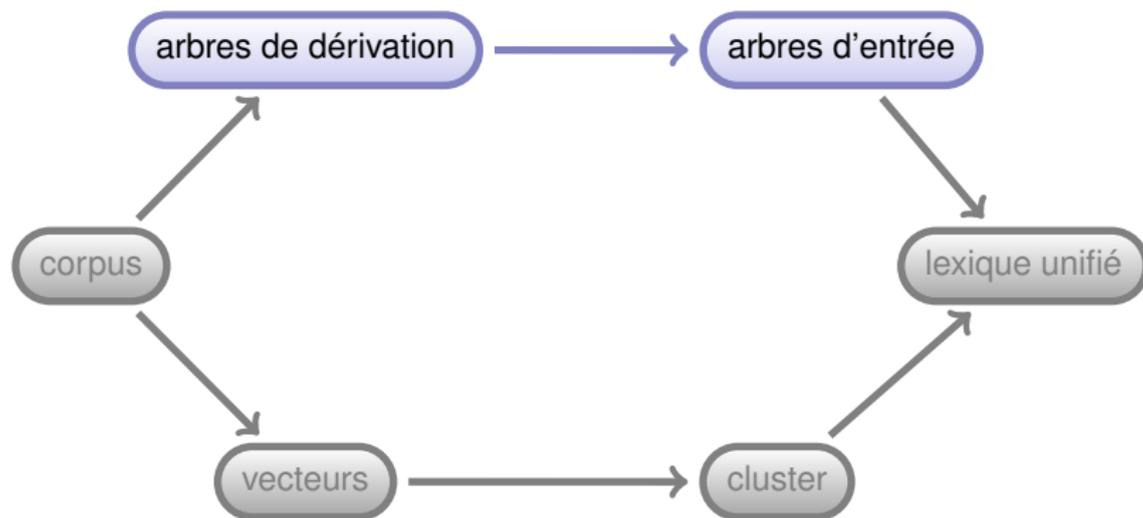
Transducteur



Transducteur



Extraction d'arbres

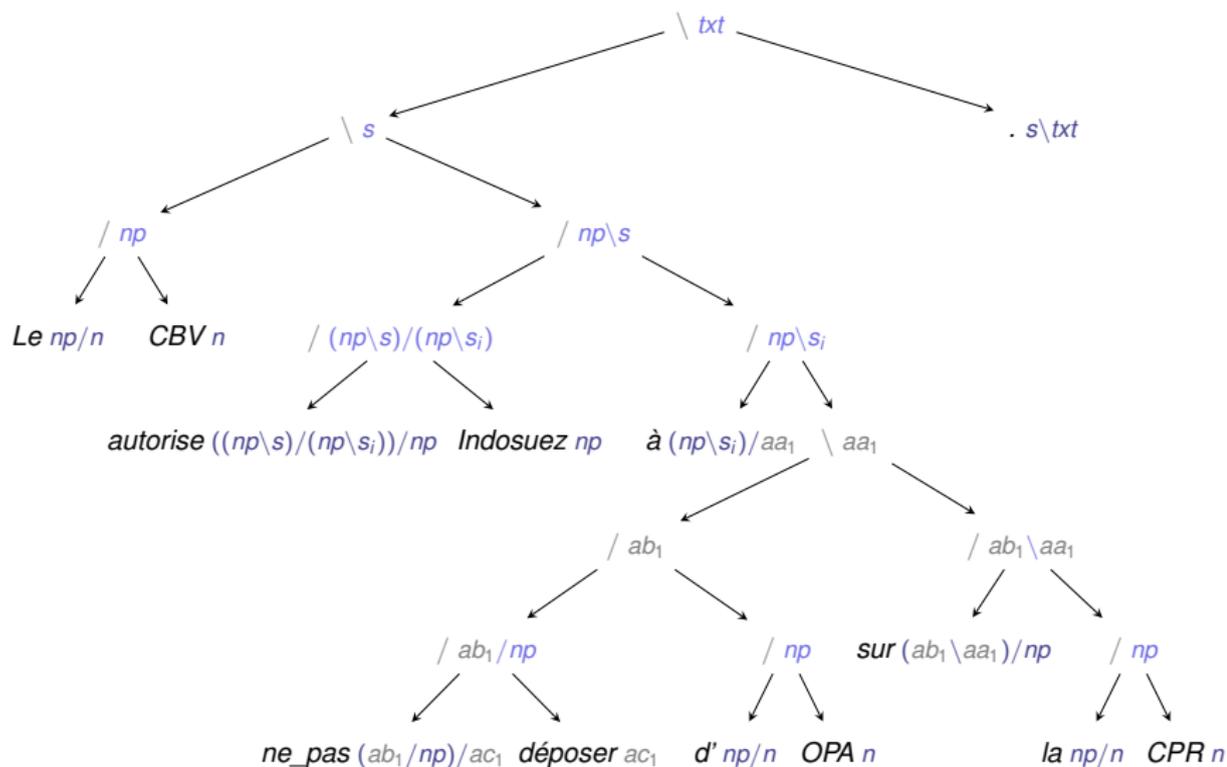


Extraction d'arbres

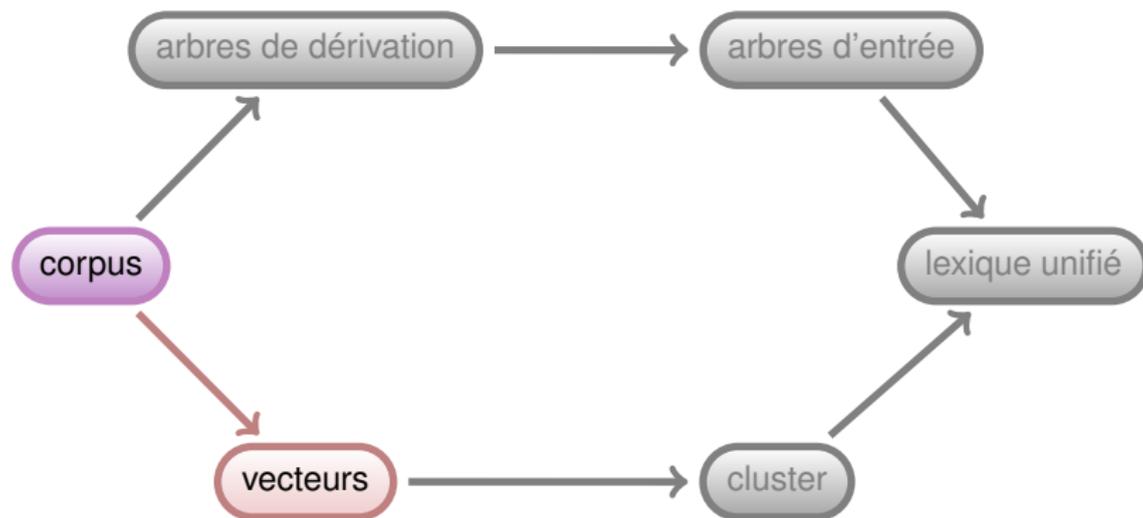
Les types sont remplacés par des variables **sauf** si le noeud est argument **et** qu'il fait partie du tableau ci-dessous.

label	type	label	type
TEXT	txt	SENT	s
NP	np	NP-ARG	np
PP	pp	AP-ARG	n\n
CLS	np	CLS-SUJ	np
NC	n	NPP	np
VPpart	np\ s_p	VPinf	np \ s_i
VPP	np\ s_p	VINF	np \ s_i

Exemple



Extraction de vecteurs



Extraction de vecteurs

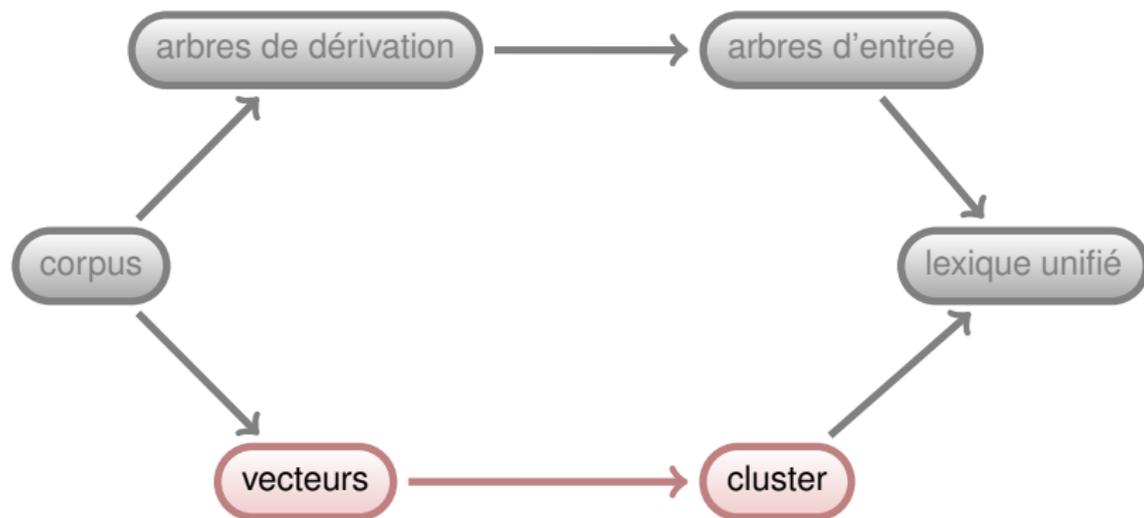
- A partir du corpus,
- 6 "dimensions" correspondant à des informations extraites du corpus,
- Transformation en vecteurs numérique ensuite.
 - 1-2 étiquettes du père et du grand père
 - 3-4 POS-tag des feuilles de droite et de gauche
 - 5-6 distance jusqu'au plus proche ancêtre commun avec les feuilles de gauche et de droite.

Passer les vecteurs dans \mathbb{Z}^n

Chaque étiquette est transformée en vecteur où seulement une ou deux dimensions possèdent la valeur 1 et le reste des coordonnées a pour valeur 0.

POS-tag	NC	DET	P	...
NC	1	0	0	0...0
DET	0	1	0	0...0
P+D	0	1	1	0...0
Other	NP	...	-ARG	-MOD
NP	1	0...0		
NP-SUJ	1	0...0	1	0
NP-MOD	1	0...0	0	1

Clustering



Clustering

- Cluster hiérarchique.
- Un cluster peut aussi bien regrouper un ensemble de cluster que des mots.
- A hauteur zéro, les clusters regroupent les mots aux vecteurs identiques.

Calcul du cluster

- Calculé avec R,
- Distance métrique Manhattan,
- Méthode de variance minimum de Ward.

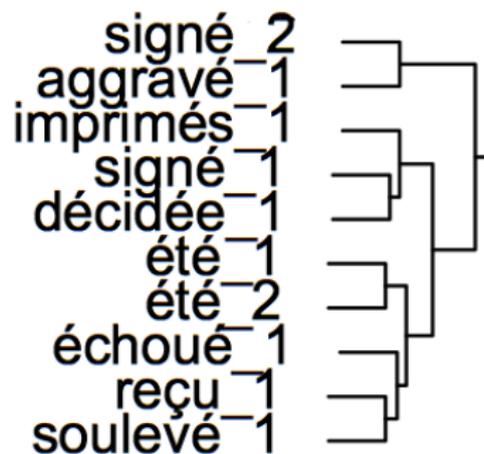


Ihaka, R., Gentleman, R. : R project (1993)

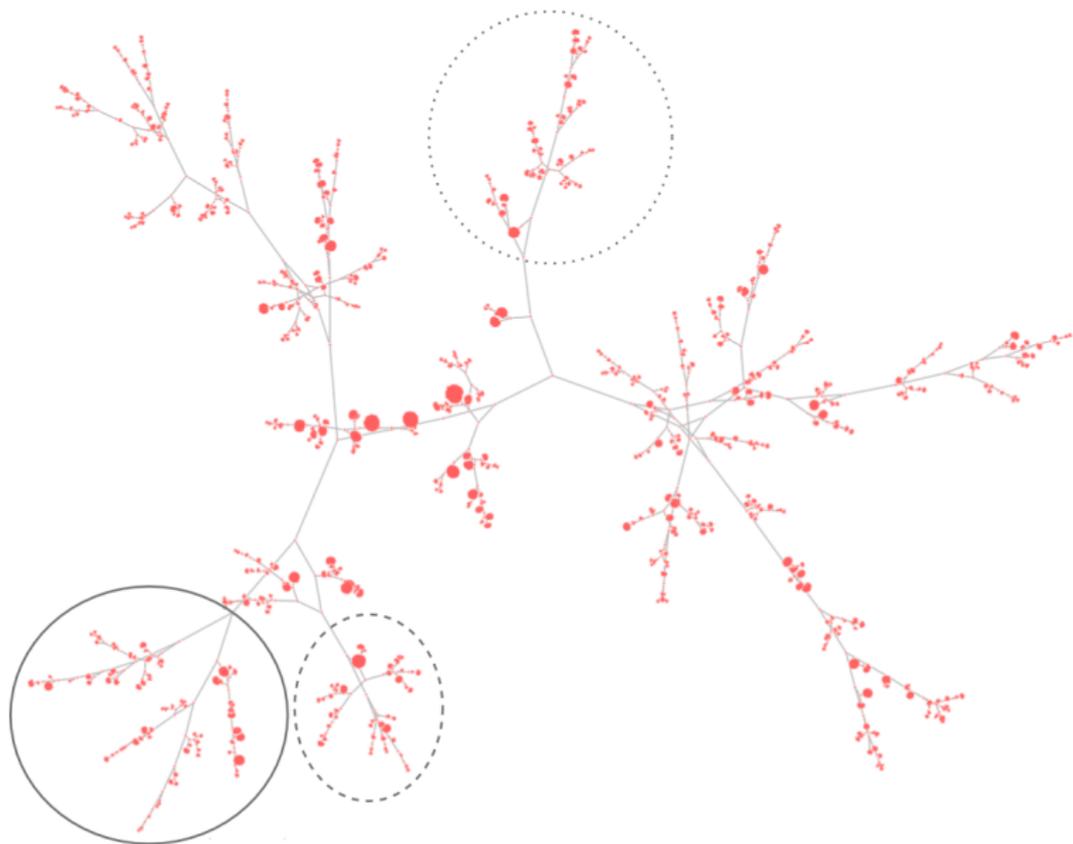


Ward, J. : Hierarchical grouping to optimize an objective function. (1963)

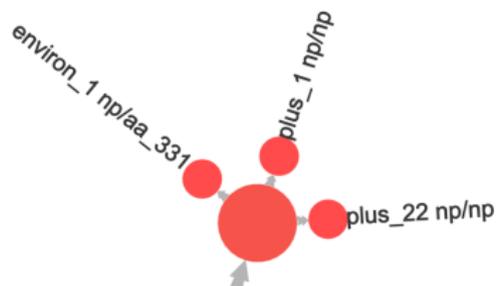
Extrait du cluster



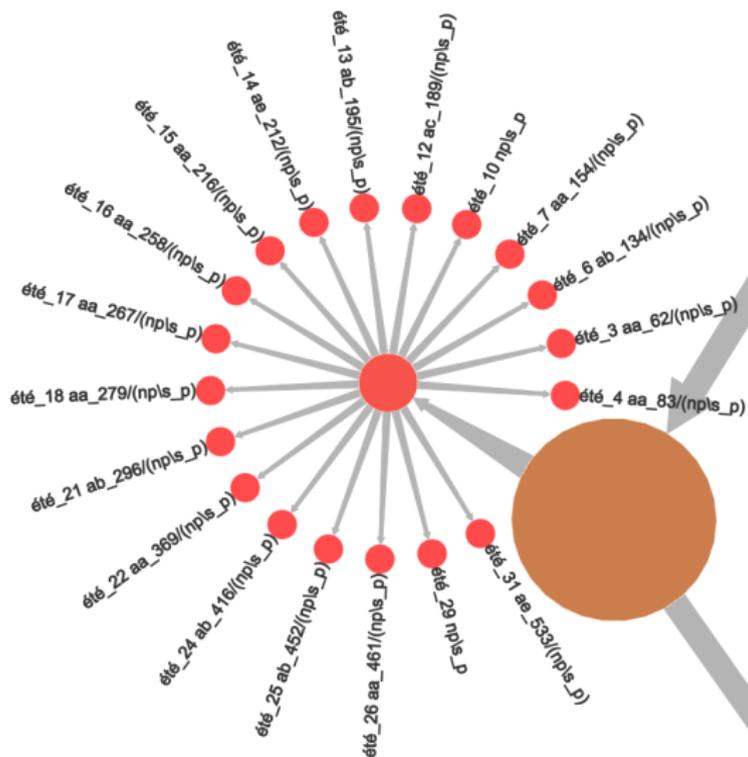
Clustering en image



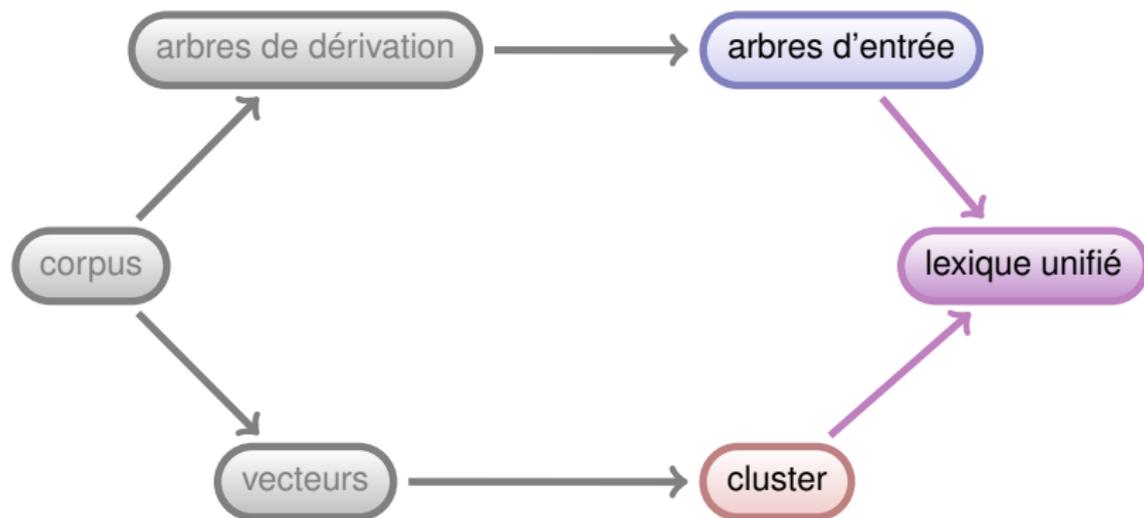
Points particulier du cluster



Points particulier du cluster



Unification



Unification

- Les variables à unifier sont extraites des arbres,
- L'unification se fait par hauteur croissante, donc par ordre de similarité des vecteurs,
- On unifie en même temps tous les clusters d'une hauteur donnée,
- Lorsque plusieurs unifications pour une variable sont possibles, utilisation d'heuristiques.

Test sur 553 phrases de Paris VII

pires erronées	569	8,7%
pires équivalentes	336	6,2%
pires identiques	4 832	85,1%
pires valides	5 168	91,3%

Plan

- 1 La méthode
- 2 Adverbes et autres modificateurs**
- 3 Conclusion

Les Modificateurs

- 1 Adverbes (encore, toutefois, davantage),
- 2 Syntagmes étiquetés "**-MOD*" ("en Espagne", "Fin Janvier").

Solution en pré-traitement

S'effectue lors de l'étape d'extraction d'arbres.



Amélioration des résultats

nombre total	5737	100 %	5737	100 %
paires erronées	569	8,7%	465	8,1%
paires équivalentes	336	6,2%	270	4,7%
paires identiques	4 832	85,1%	5002	87,2%
paires valides	5 168	91,3%	5 272	91,9%

Solution en post-traitement

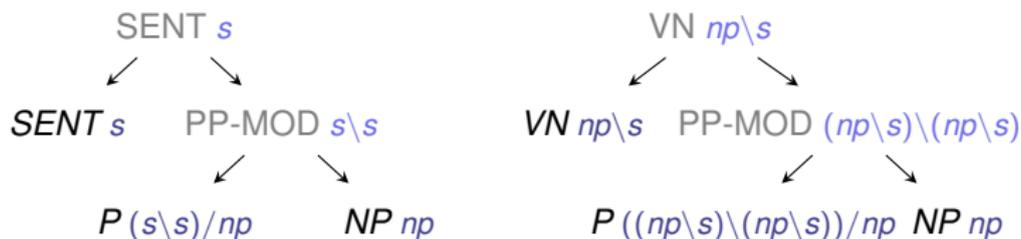
Utilisation d'un type générique X remplaçable par n'importe quel autre type du lexique.

Test manuel : sur 80 adverbes, retrait de 56,3% des incohérences.



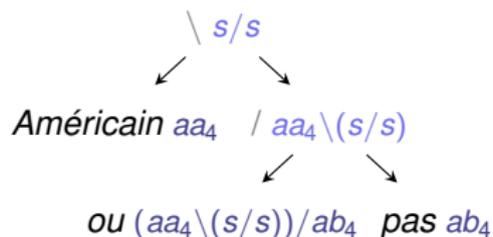
Capelletti M. and Tamburini F. : Parsing with Polymorphic Categorical Grammars (2009)

Exemple



L'unification de la préposition donnerait alors le type
 $\forall X.(X \backslash X) / np.$

Application aux coordinations



- Solution en pré-traitement moins simple à appliquer que pour les modificateurs.
- Solution en post-traitement tout à fait applicable. Les conjonctions de coordinations auraient alors le type $\forall X.(X \backslash X) / X$.

Plan

- 1 La méthode
- 2 Adverbes et autres modificateurs
- 3 Conclusion**

Conclusion

- Proposition de deux solutions pour améliorer notre algorithme.
- Solution en pré-traitement implémentée et testée.
- Solution en post-traitement à tester.
- Amélioration de deux pourcent du nombre de paires identiques
- Résultat : un lexique valide à 91,9%.

Merci de votre attention !

