Exploitation de corpus analysés syntaxiquement pour les expressions polylexicales : extraction de motifs syntaxiques et détection d'expressions polylexicales

Yoann Dupont [1]

[1]LIFO, 45067 Orléans CEDEX 2, France

Plan

Cadre

- on souhaite améliorer la reconnaissance des MWEs dans les textes :
  - faciliter l'écriture de règles pour des grammaires
  - avoir des moyens de trouver de nouvelles MWEs
  - améliorer la couverture des lexiques existants
- MWEs correspondent à des sous-arbres en dépendances contigus (Ramisch, Cordeiro, et al. 2018)
- but $\rightarrow$ à partir d'exemples de MWEs :
  - extraire des patrons de reconnaissance depuis un corpus analysé
  - déduire des règles générales + règles particulières
  - écrire règles grammaire avec règles générales
  - écrire entrées lexiques avec règles particulières

Plan

Le corpus

- Europarl-fr : transcriptions de discours du parlement européen, 41.5M mots
- fr-Wikisource : textes littéraires, 64M mots
- fr-wiki : dump Wikipédia français, 180M tokens
- total : 285.5M mots
- déjà analysés en dépendances par FRMG (De La Clergerie 2010)
- besoin de détecter des candidats : utilisation d'un lexique

Le lexique

- Wiktionnaire, le dictionnaire en ligne
- extraction des "locutions" Wiktionnaire
- couverture plus large que LeFFF (Sagot 2010)
- classification selon leur catégorie syntaxique

type	compte
adjectivales	1367
adverbiales	2657
conjonctives	147
interjectives	440
nominales	33988
prépositives	287
verbales	6499

Plan

Extraction de patrons depuis un lexique

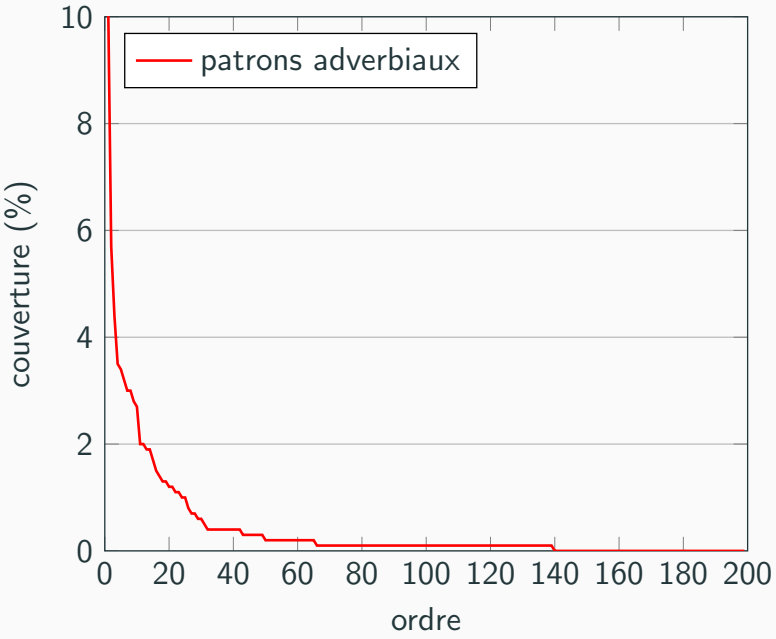
- extraction sur Europarl
- recherche occurrences en utilisant les lemmes et mots (ex: "serrer ses dents")
- patron = sous-arbre de dépendances contigus (Ramisch, Cordeiro, et al. 2018)
- sous-arbres non lexicalisés : labels de dépendances et PoS des nœuds
- contexte dans les arbres : arcs entrants/sortants reliés à un token de la MWE
- couverture d'un patron estimée à partir des nombres d'occurrences des lemmes

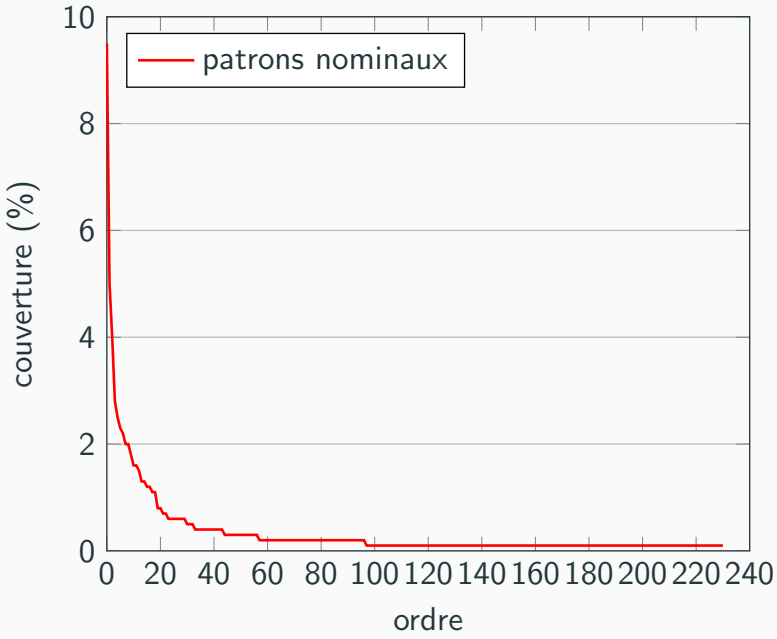
type	compte différents	compte sans contexte
adverbiales	477	22
nominales	9253	110
prépositives	5052	23
verbales	89709	492

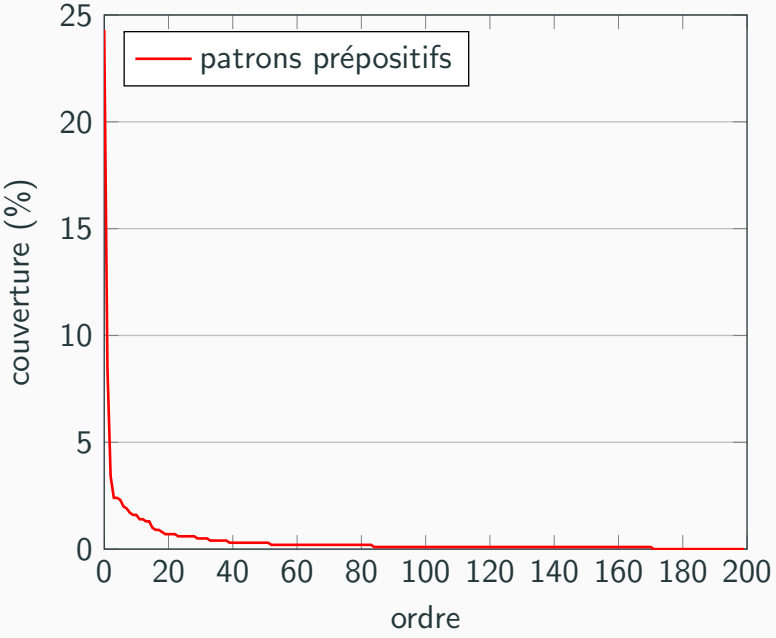
Requêter un corpus

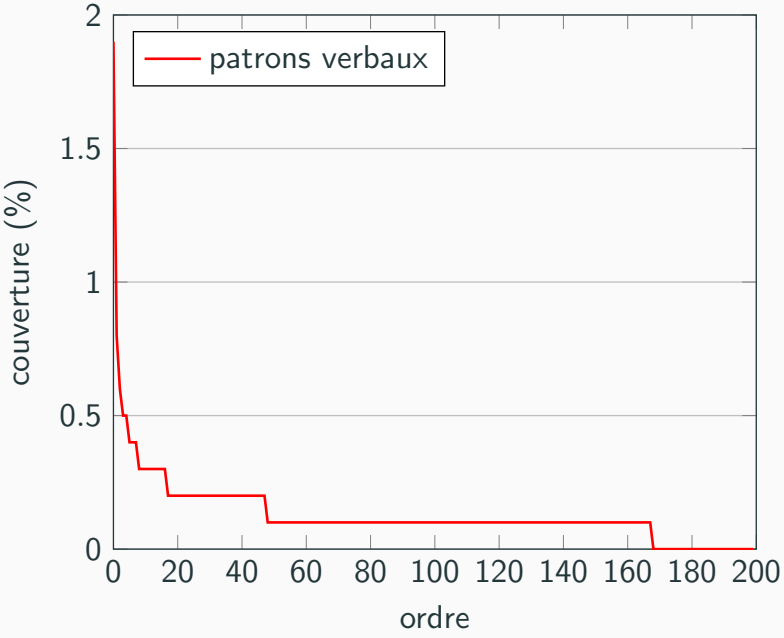
- DSL pour représenter les patrons en dépendances
- besoin d'informations précises et de gérer format non-inclus dans mwetoolkit (Ramisch, Villavicencio, and Boitet 2010)
- voir en quelle mesure intégrable dans mwetoolkit
- exemple :
  - verbe avec objet nominal saturé : "serrer les dents"
  - patron : (mwe v (object nc (det det)))
  - match : (mwe v[serrer] (object nc[dents] (det det[les])))
- large couverture : capture relativisation et passivation

Couverture des patrons

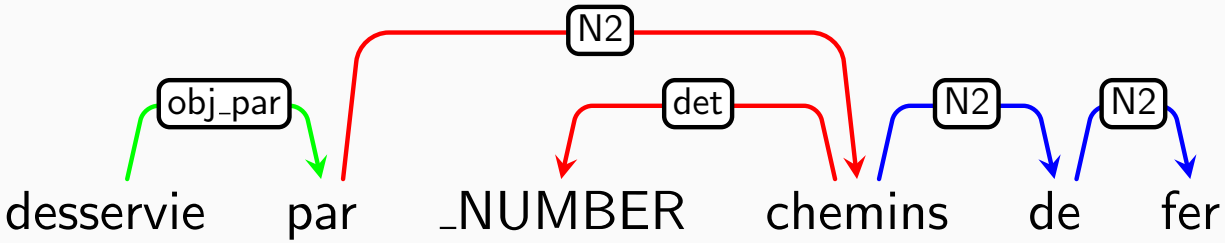


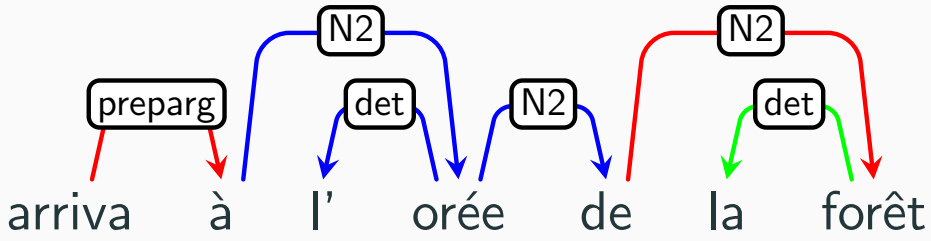


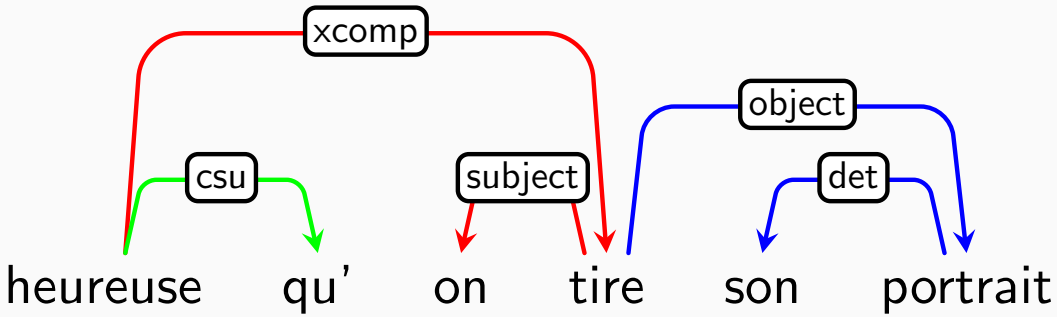




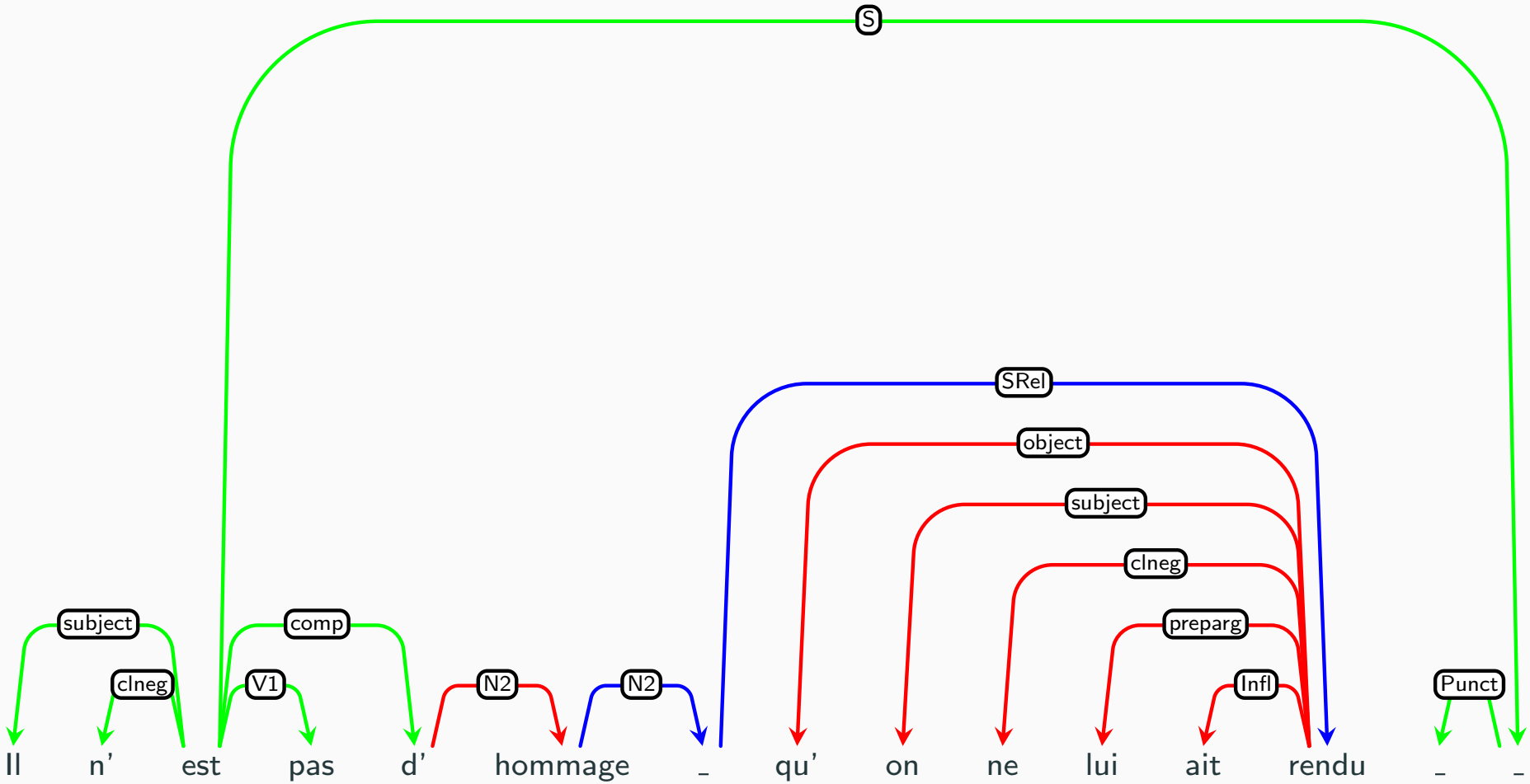
Quelques exemples



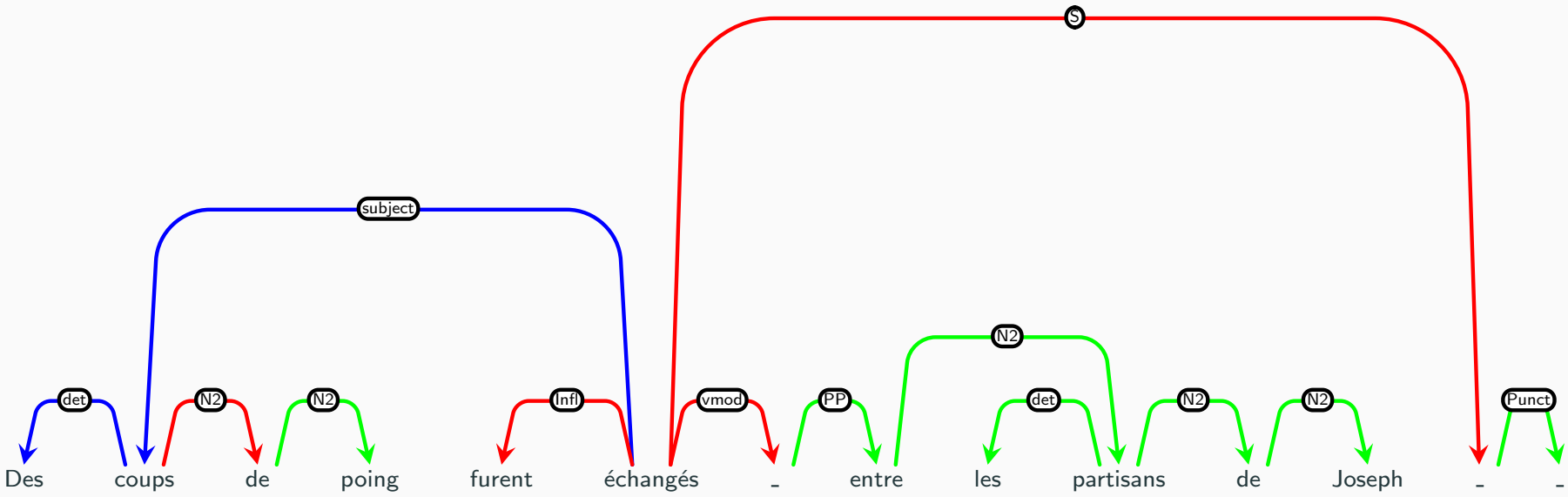




Quelques exemples – relativisation



Quelques exemples – passivation



Extraction de candidats depuis des patrons

- extraction:
  - patrons créés sur Europarl-fr
  - nouveaux candidats extraits sur WkS (limiter biais du corpus et erreurs de parsing)
- filtrage:
  - calcul de PMIs sur EP+WkS+frwiki avec mwetoolkit (Ramisch, Villavicencio, and Boitet 2010)
  - suppression des éléments si :
    - nombre d'occurrences dans le grand corpus $= 0$
    - nombre d'occurrences $<$ nombre moyen d'occurrences d'une MWE
    - PMI $<$ PMI moyenne

Résultat de l'extraction

catégorie	total	différents	après filtrage
adverbiales	695 k	33.7 k	563
nominales	146 M	4.7 M	13632
prépositionnelles	159 M	641 k	7279
verbales	22 M	2.4 M	7274

- beaucoup de MWEs sont matchées plusieurs fois (contextes différents)
- bruit demeure important
- besoin d'une meilleure gestion du contexte
- ordonnancement des patrons du plus précis au plus général

Plan

Génération d'entrées pour un lexique

- avec patrons $\rightarrow$ on génère des règles
- avec informations contextuelles $\rightarrow$ modifications applicables (adverbes, etc...)
- recoupement des informations pour règles plus précises (accord, saturation)
- informations trop précises pour la grammaire (ex: lemmes) $\rightarrow$ lexique
- nouvelles entrées dans LeFFF intentionnel

Génération d'entrées pour le LeFFF

catégorie	lemme (FRMG)	→	entrée LeFFF simplifiée
(adv)	non moins que	→	non moins que; inv-e
(nc)	parti politique	→	parti politique; nc-2m
(nc)	chemin de fer	→	chemin de fer; nc-2m+inv+inv
(prep)	à l'orée de	→	à l'orée de; prep-e
(v)	tirer son portrait	→	tirer Poss0 portrait
(v)	avoir du temps devant lui	→	avoir du temps devant Lui-0

- encore expérimental
- besoin de lier MWE à son patron

Plan

Conclusions et perspectives

- conclusions :
  - moyen simple d'extraire des patrons de reconnaissance
  - large couverture : relativisation, passivation
  - découverte de MWEs absentes des lexiques (Wiktionary et LeFFF)
  - génération d'entrées pour les lexiques par recherche sur corpus et élagage
- perspectives :
  - sélection des traits spécifiques à grammaire/lexique
  - intégration de traits de contexte négatifs
  - amélioration de la génération d'entrées dans LeFFF
  - rapprochement avec mwetoolkit
  - PMIs sur corpus autre que Europarl et Wikisource

bibliographie i

Éric Villemonte De La Clergerie. “Convertir des dérivations TAG en dépendances”. In: *TALN 2010*. 2010.

Carlos Ramisch, Silvio Cordeiro, et al. “Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions”. In: *the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. 2018, pp. 222–240.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. “Mwetoolkit: a framework for multiword expression identification.”. In: *LREC*. Vol. 10. Valletta. 2010, pp. 662–669.

Benoît Sagot. “The LeFFF, a freely available and large-coverage morphological and syntactic lexicon for French”. In: *LREC*. 2010.