Autour de l'extraction d'arbres lexico-syntaxiques récurrents spécifiques : corpus, outils et méthodes

Olivier Kraif
Laboratoire Lidilem – Université Grenoble Alpes

Journée Phraseotext – jeudi 8 décembre 2016

Introduction

- Objectifs :
  - Articuler l'étude phraséologique (au sens large, cf. Tutin, Legallois, 2013) et la caractérisation des sous-genres littéraires
  - Approche *corpus driven* : découverte sans *a priori* de motifs spécifiques
  - Reprise des outils et méthodes développées dans le cadre Projet Emolex
  - Mise au point de nouvelles méthodologies
  - Élargissement des corpus : quantité et diversité des langues (ajout du latin)

Outils mis en oeuvre

Outils

- Analyse syntaxique : Xip (fr,en), Malt (la)
- Lexicoscope (anciennement EmoConc)

  (http://phraseotext.u-grenoble3.fr/lexicoscope/)

  - Profils combinatoires synthétisés par les *lexicogrammes (Tournier, Heiden 1998).* Matrice des collocatifs associés à une valeur permettant de mesurer la corrélation statistique (fréquence, loglike, t-score, etc.)
  - Cooccurrence syntaxique (relation de dépendance) et non cooccurrence de surface (Evert, 2008, Kilgariff et Tugwell, 2001)

Outils
Lexicogramme

Lexicogramme | Graphiques

Show 25 entries Search:

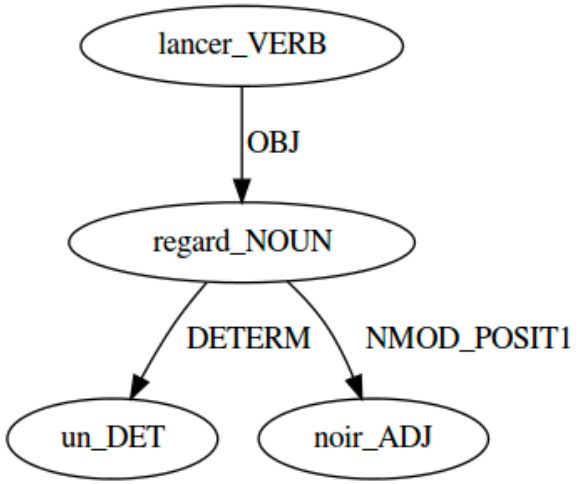
l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log
regard_NOUN	$SON_DET	DETERM_POSS NMOD_POSIT1	2570	43353	375269	32474662	7	4379,5023	1
regard_NOUN	$UN_DET	DETERM NMOD_POSIT1 ~NMOD_POSIT1 NMOD	2193	43353	444742	32474662	7	2598,6806	2
regard_NOUN	jeter_VERB	~OBJ ~SUBJ ~VMOD ~VMOD_POSIT1	516	43353	19036	32474662	7	2144,3856	3
regard_NOUN	croiser_VERB	~SUBJ_COORD ~OBJ ~SUBJ ~VMOD_POSIT1 ~OBJ_COORD	165	43353	5646	32474662	7	708,4942	4
regard_NOUN	sous_PREP	PREPOBJ	230	43353	15063	32474662	7	705,1637	5
regard_NOUN	lancer_VERB	~OBJ ~SUBJ ~VMOD_POSIT1 NMOD_POSIT1 ~OBJ_COORD	210	43353	13717	32474662	7	644,7825	6
regard_NOUN	échanger_VERB	~OBJ NMOD_POSIT1 ~SUBJ ~VMOD	94	43353	2488	32474662	7	450,6484	7
regard_NOUN	détourner_VERB	~OBJ ~SUBJ NMOD_POSIT1 ~DEEPOBJ ~SUBJ_PASSIVE ~VMOD_POSIT1	90	43353	2396	32474662	7	430,4358	8

Outils
Pivots complexes

- Notion de **pivot complexe** (Kraif &Diwersy, 2012)

  Le pivot est un sous-arbre pouvant correspondre à une expression, qu'il suffit d'entrer en mode « libre » (requêtes basées sur l'exemple, cf. Augustinus et al., 2012) :

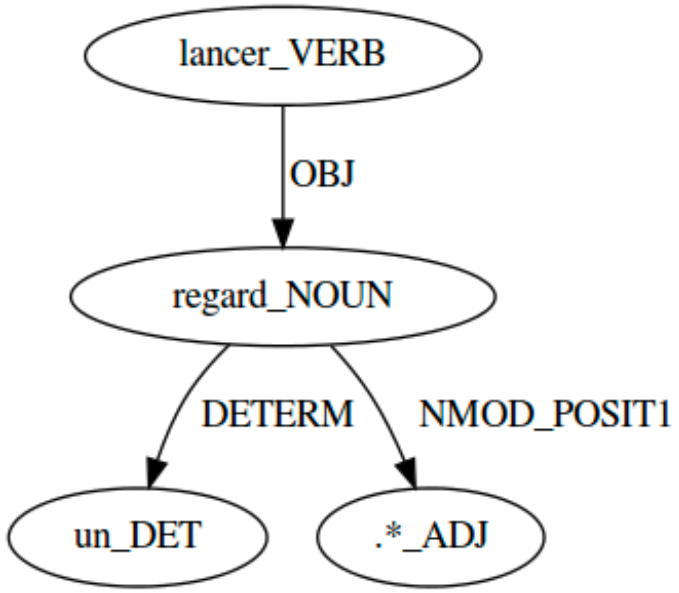
  ex. *lancer un regard noir*



Outils
Pivots complexes

- Le langage de requête permet également d'exprimer des *patterns* généraux :

```
<l=lancer,c=VERB,#1>&&<l=un,c=DET,#2>&&<l=regard,c=NOUN,#3>&&<c=ADJ,#co>::(DETERM,3,2) (NMOD_POSIT1,3,co) (OBJ,1,3)
```



Outils
Pivots complexes

- Pour un tel pivot, il est possible d'extraire les **concordances**, ou les cooccurrents syntaxiques, en précisant (ou non) la place des cooccurrents :

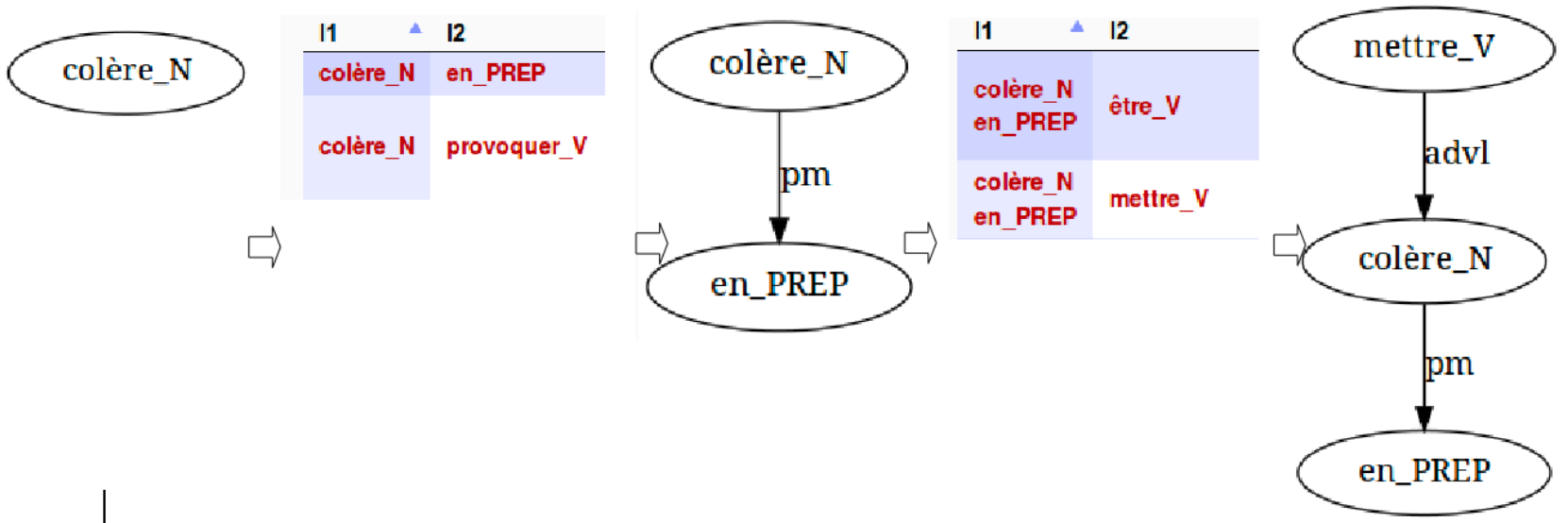
Lexicogramme | *Graphiques*

Show 25 entries Search:

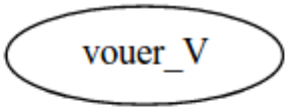
l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
lancer_VERB un_DET regard_NOUN ADJ	circulaire_ADJ	NMOD_POSIT1#3	9	97	448	32474662	2	141,6930	1
lancer_VERB un_DET regard_NOUN ADJ	bref_ADJ	NMOD_POSIT1#3	7	97	1288	32474662	1	91,6515	2
lancer_VERB un_DET regard_NOUN ADJ	noir_ADJ	NMOD_POSIT1#3	8	97	10735	32474662	3	73,0545	3

Outils
Extraction automatique d'ALR

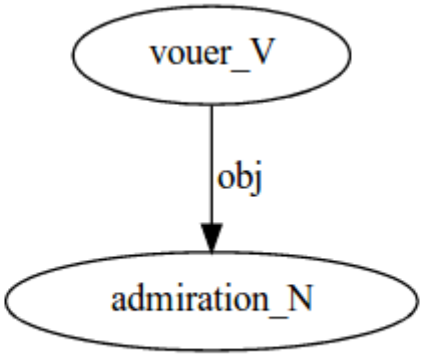
- La fonctionnalité « pivot complexe » permet d'extraire itérativement des ALR (arbres lexico-syntaxiques récurrents)



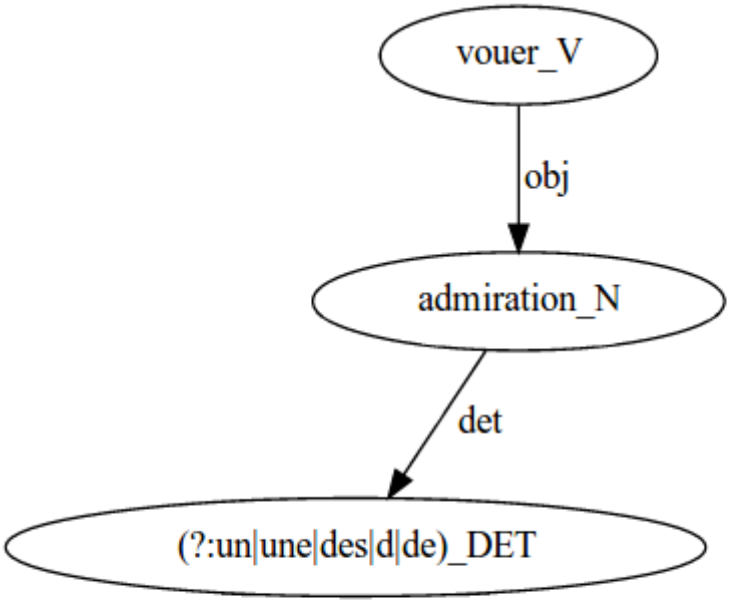
Outils
Extraction automatique d'ALR



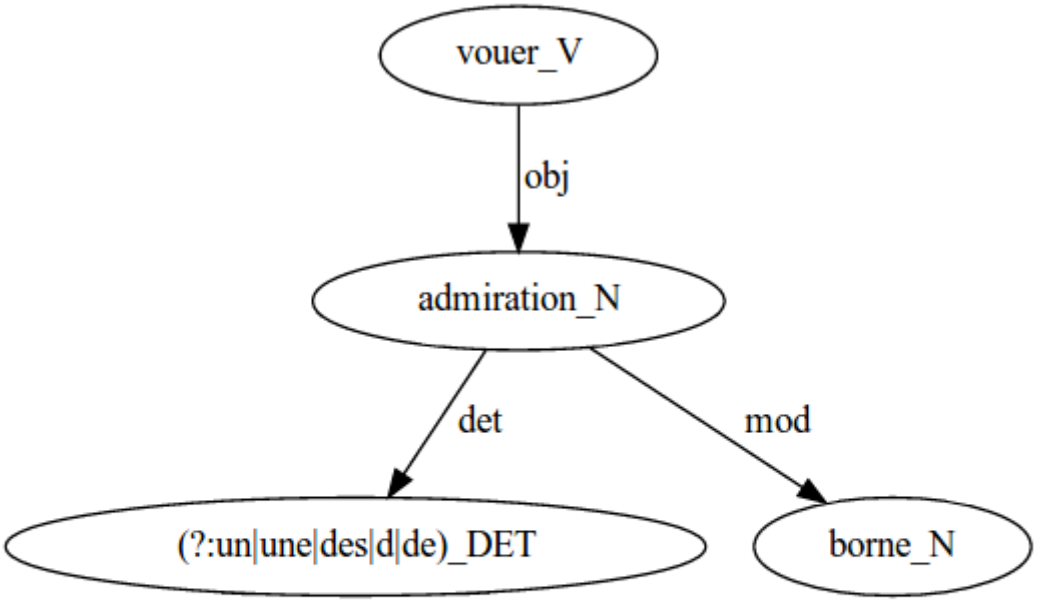
Outils
Extraction automatique d'ALR



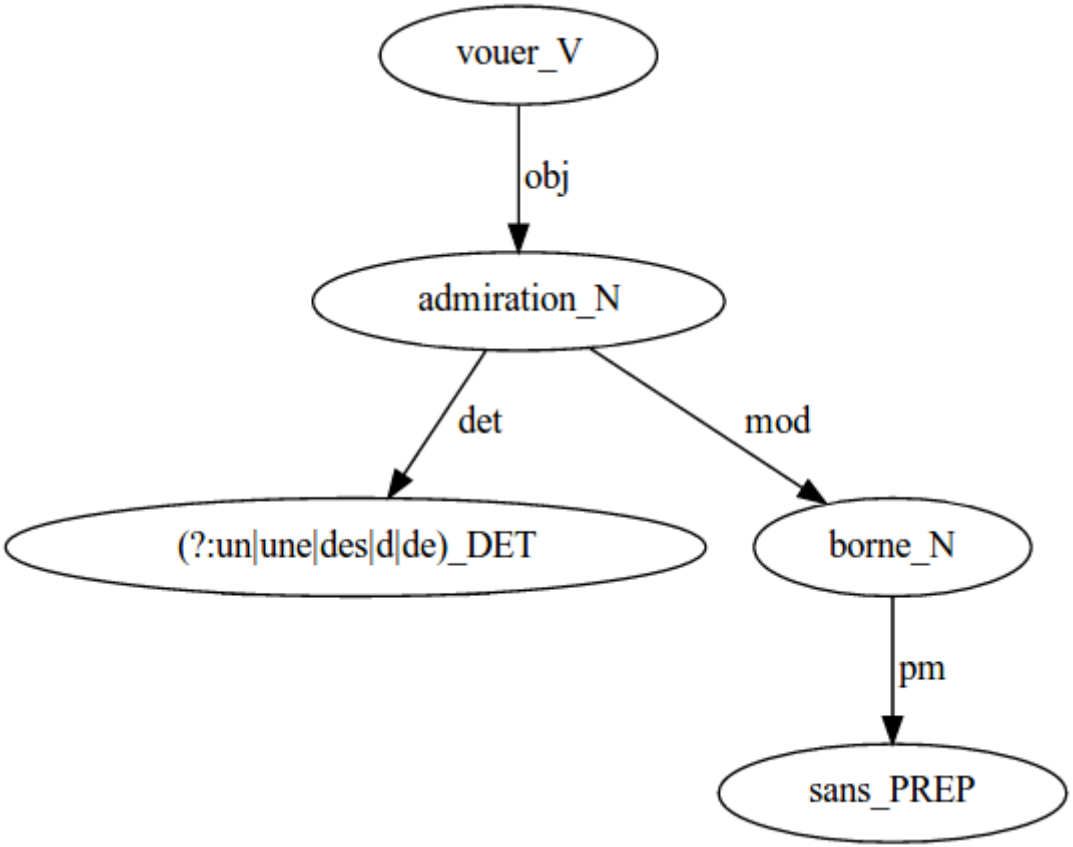
Outils
Extraction automatique d'ALR



Outils
Extraction automatique d'ALR



Outils
Extraction automatique d'ALR



Outils
Extraction automatique d'ALR

- Intérêt de la méthode :
  - Pas besoin de pré-calculer à l'avance tous les arbres récurrents du corpus (très long)
  - On peut « zoomer » sur la combinatoire de n'importe quelle expression, ou classe d'expressions, non définies *a priori,* dans une perspective *corpus-driven*
  - L'option « extraction des expressions polylexicales » permet d'obtenir les ALR significatifs (seuils de fréquence, de dispersion, de loglike) autour d'un pivot simple ou complexe

Méthodologie de comparaison de corpus

Méthodologie

- Des expressions sont surreprésentées dans certains sous-corpus :

  p.ex. *se passer la langue sur les lèvres*

  Hist : 2 occ., POL : 12 occ., SF : 1 occ., AUT : 1 occ.

- Comment mesurer rigoureusement ce déséquilibre ?

Méthodologie

- Calcul de spécificité : comparaison des fréquences relatives dans un sous corpus par rapport aux fréquences dans l'ensemble du corpus.
- On extrait le tableau de contingence :
  - f1 : la fréquence dans le sous-corpus
  - f2 : la fréquence dans le corpus complet
  - T1 : le nombre total de mots du sous-corpus
  - T2 : le nombre total de mots du corpus complet
- Test de Fisher ou calcul du rapport de vraisemblance (*log likelihood ratio*)

Méthodologie

- Observations préliminaires. On extrait tous les ALR :
  - du sous-corpus POL
  - du corpus total
- On ne retient que les ALR émergents (cf. motifs séquentiels émergents, Quiniou et al. 2012)
  - Loglike > 3.84
  - Dispersion ≥ 3

Observations (préliminaires)

Observations

- On obtient des listes d'ALR hors contexte :

Key	f1	f2	disp	loglike
secouer_VERB<tête_NOUN<le_DET>>	531	885	7	174,87
lancer_VERB<regard_NOUN>	152	193	7	133,75
lancer_VERB<regard_NOUN<un_DET>>	140	178	7	122,68
poser_VERB<main_NOUN<le_DET>,épaule_NOUN<sur_PREP>>	53	51	7	106,65
hocher_VERB<tête_NOUN<le_DET>>	529	983	7	100,82
composer_VERB<il_PRON,numéro_NOUN>	81	96	6	87,89
regarder_VERB<montre_NOUN<son_DET>>	113	156	7	76,1
coup_NOUN<téléphone_NOUN<de_PREP>>	129	188	7	72,14
côté_NOUN<autre_ADJ,rue_NOUN<de_PREP,le_DET>>	45	48	7	67,52
tour_NOUN<le_DET,pièce_NOUN<de_PREP>>	27	21	6	65,83
sortir_VERB<pièce_NOUN<de_PREP,le_DET>>	56	67	6	59,06
photo_NOUN<le_DET,sur_PREP>	84	116	6	56,51
mordre_VERB<lèvre_NOUN<le_DET>,se_PRON>	91	134	6	49,04
passer_VERB<main_NOUN,visage_NOUN<sur_PREP>>	39	45	6	45,87
froncer_VERB<sourcil_NOUN<le_DET>>	172	304	7	42,92
fourrer_VERB<poche_NOUN<dans_PREP>>	53	72	6	37,64
pousser_VERB<porte_NOUN<le_DET>>	119	203	7	35,26

Observations

- Ces ALR émergents jouent un rôle heuristique :
  - pistes à suivre pour des analyses plus approfondies des *motifs* (p.ex. *Sur la scène de crime...*, cf. Gonon et al. 2016)
  - retour au texte nécessaire
- Ils esquissent des catégories générales prototypiques du sous-genre étudié :
  - Des éléments thématiques prévisibles : *crime, tueur, police, victime, médecin, sang…*
  - Ou moins prévisibles : *voiture, photo, montre, téléphone, ...*

Observations

- Beaucoup d'expressions centrées sur les personnages
  - Attitudes physiques : *un signe de tête, secouer la tête, lancer un regard, poser la main sur l'épaule, hocher la tête, se mordre la lèvre, passer la main sur le visage, froncer les sourcils, ...*
  - Rapport aux accessoires, objets de communication, artefacts : *composer le numéro, regarder sa montre, coup de téléphone, sur la photo, fourrer dans la poche, etc.*
- Expressions spatiales et de mouvement également fréquentes : *de l'autre côté de la rue, le tour de la pièce, tourner les talons, pousser la porte, kilomètres à la ronde, à pleine vitesse*

Perspectives

Perspectives

- Développements de nouvelles fonctionnalités dans le cadre du projet PhraséoRom
  - Souplesse dans le choix des sous-corpus
  - Méthodologie de comparaison automatisée (lexicogrammes de spécificités)
  - Extraction (automatique ?) de motifs multidimensionnels :
    - Lexèmes
    - Traits morpho-syntaxiques
    - Classes sémantiques

Merci de votre attention