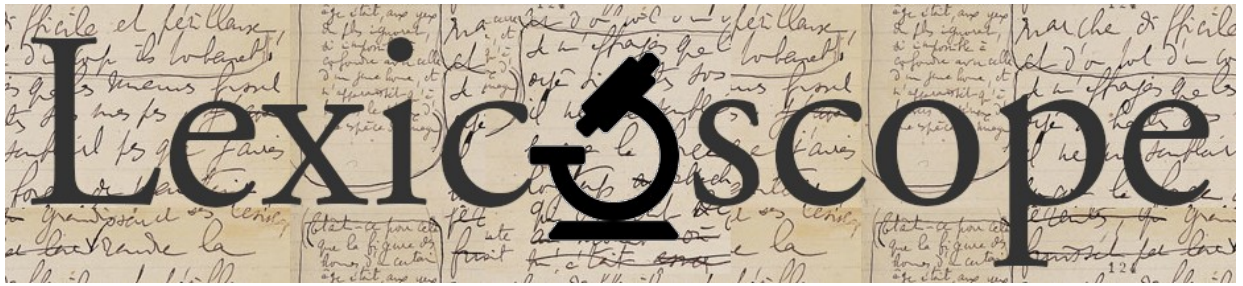
Modèles de cooccurrences collostructionnels pour l'étude des expressions préfabriquées de l'oral

LLcD - Atelier Cooccurrence et marquage discursif

Sorbonne Université, 10 septembre 2024

Olivier Kraif

Sommaire



1. Contexte

Les Phrases préfabriquées de l’oral (PPI)

- Projet ANR Préfab (piloté par A. Tutin, Lidilem)

Le projet PREFAB (Constructions des phrases PREFABriquées dans les interactions langagières) vise à inventorier et à modéliser les constructions des phrases préfabriquées des interactions (PPI), à partir de corpus variés (Tutin & Grossmann, 2023 ; Pausé, et al. 2022).

Parmi les expressions productives des interactions, nous nous intéressons aux phrases préfabriquées, qu’on définit comme des énoncés complets, « prêts à l’emploi » pour les locuteurs dans des contextes énonciatifs spécifiques et dont la lexicalisation est contrainte (Ex. : “ça marche !” “comment dirais-je ?” “tu peux le dire.” “c’est OK”)

Les Phrases préfabriquées de l’oral (PPI)

Une perpective constructionnelle :

“(...) ces phrases préfabriquées de surprise présentent souvent une forme de régularité syntaxique et sémantique. On relève par exemple des paradigmes récurrents comme tu plaisantes/tu rigoles/tu veux rire… ou je reste sans voix / bouche bée. (...)
On pourrait considérer ainsi les paradigmes précédents comme des « phrasèmes constructionnels » définis comme des associations structures-sens comportant des éléments fixes et des éléments variables (Dobrovol’skij & Pöppel 2022, Mellado Blanco 2022).

Tutin & Grossmann (2023 : 146)

2. Modèles de cooccurrence

La cooccurrence comme concept central

- Approche **contextualiste** et **multidimensionnelle** :
  - *Extended units of meaning* (Sinclair, 1991) : *collocation, colligation, semantic preference, semantic prosody*
  - Grammaire de construction (Goldberg, 2006), collostructions (Gries, Stefanovitsch, 2003, Gries 2019), constructicons (Lyngfelt, Borin, Ohara, Torrent 2018)
- Les contextes sont appréhendés à travers les structures hiérarchiques des **dépendances syntaxiques** :
  - Corpus analysés en dépendances (Universal dependency)
  - Notion de cooccurrence syntaxique plus riche que la cooccurrence de surface (Evert 2007, Bartsch & Evert, 2014)

Questions de recherche

- Pour des expressions souvent très figées comme les PPI, le modèle de cooccurrence syntaxique est-il adapté ?
    - Coût de mise en oeuvre
    - Bruit important (pour l’oral)

- Des modèles de patterns surfaciques (p.ex. “C’est pas + ADJ”) ne sont-il pas suffisants ?

- Comment combiner ces deux approches ?

3. Méthodologie

Méthodologie

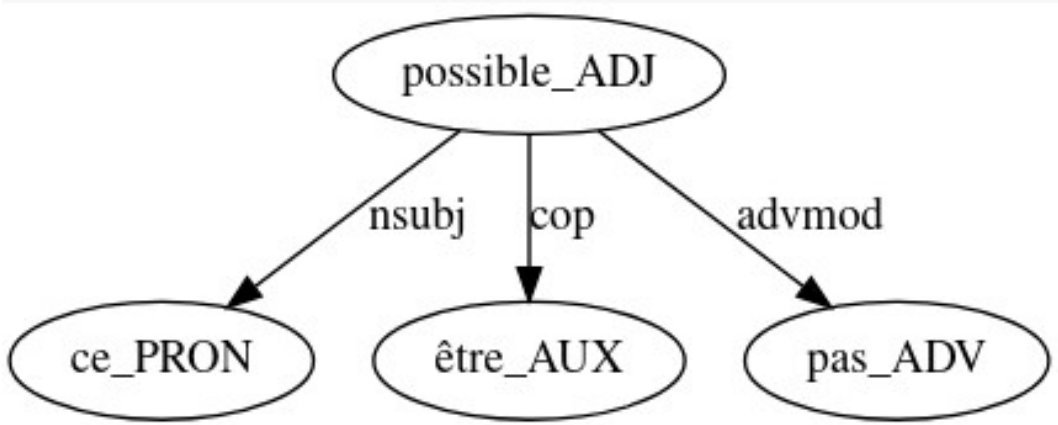
- Construction d’un corpus analysé syntaxiquement en dépendances
  - Corpus de romans (Phraséorom)
  - Corpus oraux (TCOF, MPF, CEFC Oral, ESLO2)
  - Corpus d’interactions écrites (Wiki discussion)
- Approche outillée : plateforme du Lexicoscope 2.0 (Kraif, 2019)
  - Concordances
  - Cooccurrences
  - Wordketches
  - Index hiérarchiques
  - Recherche de patterns
  - …

Choix du pattern

- On s’intéresse aux PPI correspondant au pattern : C’EST PAS + ADJ
  - De nombreuses classes de PPI suivent ce moule constructionnel et remplissent différentes fonctions interactionnelles :
    - *c’est pas vrai / possible / croyable* : expression de la surprise
    - *c’est pas étonnant / surprenant / nouveau* : expression de la non-surprise
    - *c’est pas idiot / con / bête* : expression de l’accord vis-à-vis d’une idée nouvelle
  - La négation assume différents contenus : attente contrariée, surprise, etc. Elle se présente souvent comme une litote. Dans tous les cas, elle marque une certaine expressivité (surprise, ironie, etc.).
  - Prosodie sémantique complexe

Le langage TQL

- Le Lexicoscope permet d'écrire des requêtes syntaxiques grâce au langage TQL (Tree Query Language)
- Ces requêtes peuvent être **générées par le système**
  - L'utilisateur donne un exemple : p.ex.
  - « c'est pas possible »
  - La requête (ou les requêtes) est généré en fonction de ce qui est trouvé dans le corpus
  - Cette requête exprime un mélange de contraintes lexicales et syntaxiques



<l=ce,c=PRON,#1>&&<l=être,c=AUX,#2>&&<l=pas,c=ADV,#3>&&
<l=possible,c=ADJ,#4>::(advmod,4,3) (cop,4,2) (nsubj,4,1)

Les patterns

- TQL permet d’exprimer deux types de patterns

1) Des **patterns syntaxiques**, impliquant des relations

<l=ce,c=PRON,#1>&&<l=être,c=AUX,#2>&&<l=pas,c=ADV,#3>&& <c=ADJ,#4>::
(advmod,4,3) (cop,4,2) (nsubj,4,1)

2) Des **patterns surfaciques** définissant seulement une séquences de tokens (comme le langage CQL définit dans SketchEngine)

<l=ce,c=PRON,#1> <l=être,c=AUX,#2> <l=pas,c=ADV,#3> <c=ADJ,#4>

3) Dans un tel pattern on peut éventuellement définir un « gap »

<l=ce,c=PRON,#1> <l=être,c=AUX,#2> <l=pas,c=ADV,#3> <>{0,3} <c=ADJ,#4>

Les patterns

- Ces patterns permettent d'étudier des **collostructions** :

« the basic idea of [Collostructional Analysis] is really only an extension of the notion of collocation, i.e. the co-occurrence of (most often) two lexical items, to one sense of the notion of colligation, namely the **co-occurrence of words** with **patterns** (Hunston & Francis, 1999) or **constructions** (Goldberg, 1995, 2006). » (Gries, 2019)

- On peut ainsi comparer les résultats de deux modèles collostructionnels

1) avec les **patterns syntaxiques** des collostructions filtrées par des contraintes syntaxiques
2) avec les **patterns surfaciques** des collostructions définies par la cooccurrence de surface dans un certain empan

Corpus

Pour comparer les résultats de ces deux types de requêtes, nous avons utilisé divers corpus oraux, comportant des enregistrements de situations variées : conversation, interaction avec des services, prise de parole, réunion, etc.

Noms du corpus	Période	Nombre de documents	Nombre de tokens
CEFC	1980 - 2015	903	3 507 512
ESLO2	2005 - 2013	336	1 528 769
MPF	2010 - 2019	92	977 669
TCOF2	2003 - 2015	133	426 788
Total		**1464**	**6 440 738**

4. Comparaison entre requêtes syntaxiques et surfaciques

Première comparaison : sans gap

On compare la requête syntaxique à une requête surfacique sans insertion possible

Nécessité d'ajouter une **contrainte de séquence** pour le *pas* lié à certaines relations *reparandum :*

- Sinon on trouve : *pas c'est possible, pas c'est vrai, pas c'est sûr, etc.*
- De même pour éliminer les inversions : *est-ce pas possible ?*

<l=ce,c=PRON,#1>&&<l=être,c=AUX,#2>&&<l=pas,c=ADV,#3>&&
<c=ADJ,#4>::(advmod,4,3) (cop,4,2) (nsubj,4,1) && prec(#1,#2,#3,#4)

Première comparaison : sans gap

f(tree) = 4179
f(flat) = 3270

- 21,7 %

On perd des occurrences liées le plus souvent à des modulations adverbiales à fonction euphémisante

c’est pas mal → c’est pas trop mal
c’est pas drôle → c’est pas très drôle
c’est pas facile → c’est pas toujours facile, c’est pas tout le temps facile
c’est pas grave → c’est pas bien grave

Autre type d’insertion : *c’est pas joli → c’est pas du joli joli*

Certaines expressions n’apparaissent dans le corpus que sous la forme euphémisée : *c'est pas très jojo* [crfp/PRI-NAR-1.conllu_75]

Première comparaison : sans gap

Notons que toutes les occurrences ne correspondent pas à des PPI

*donc **c'est pas** si **vieux** que ça ça fait*
*il y a de bon c'est plus au moins euh même si **c'est pas passionnant** y a toujours ce sont*
*mais je trouve que bon c'est dommage parce que **c'est pas accessible** à tous*
*et qu'est-ce que vous faites comme études ? si **c'est pas indiscret***
*et c'est là qu'on s'est aperçu que **c'était pas marrant** d'être à son compte*
*je trouve que **c'est pas créatif***

Pour éliminer les intrus on peut s'appuyer sur des paramètres :

- internes : par exemple l'utilisation de l'indicatif présent pour le verbe *être*
- externes : par des indices contextuels, p.ex. enracinement de l'adjectif (relation *root*)

Deuxième comparaison : avec gap

On compare la requête syntaxique à une requête surfacique avec insertion possible

<l=ce,c=PRON,#1><l=être,c=AUX,#2><l=pas,c=ADV,#3> <>{0,3} <c=ADJ,#4>

f(tree) = 4179
f(flat-gap) = 4898

+ 17 %

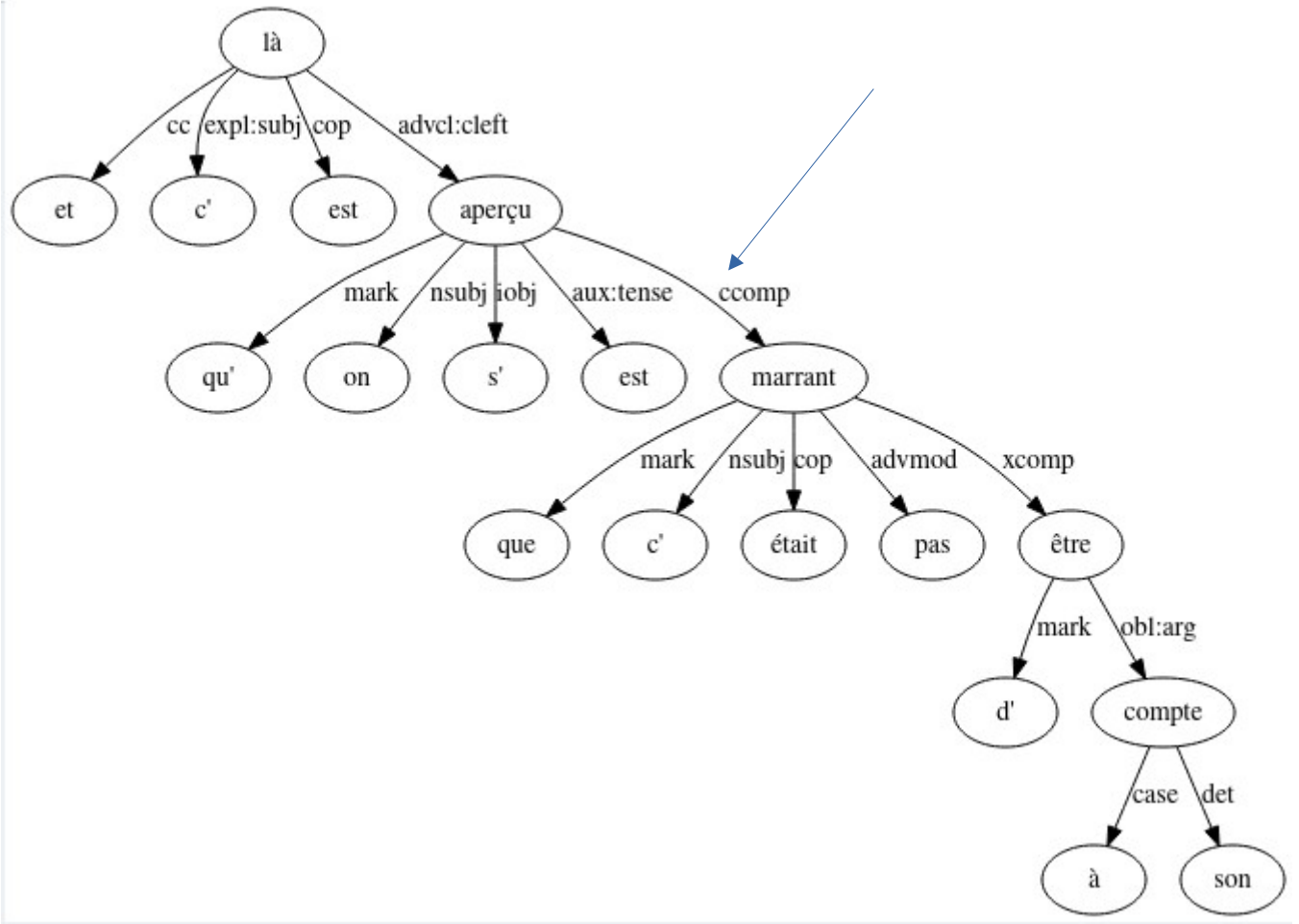
La plupart de ces nouveaux candidats correspondent à du bruit (où l'adjectif est souvent rattaché à un nom) :

mais ***c'est pas*** *une passion* ***première*** *quoi*
donc euh je trouve que ouais non ***c'est pas*** *une* ***super*** *chose quoi*
c'est pas *leur* ***premier*** *film*
c'est pas *une omelette toute* ***simple***
tant que ***c'est pas*** *tant que ça reste* ***vivable***
c'est pas *la* ***même*** *ambiance*

5. Paramétrage externe des requêtes syntaxiques

1/ Ancrage à la racine

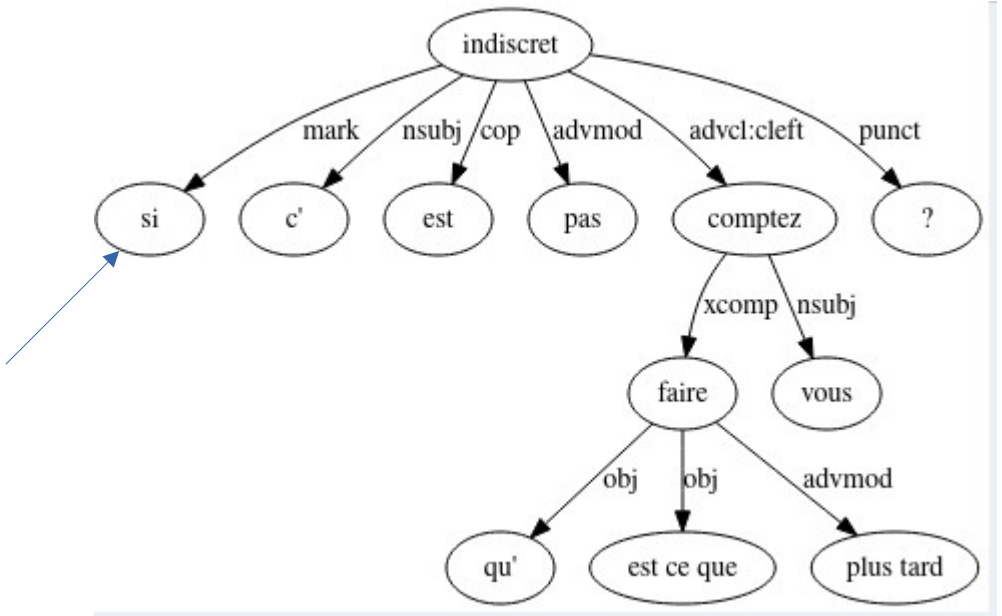
et c'est là qu'on s'est aperçu que **c'était pas marrant** d'être à son compte [CINE/ESLO2_CINE_1215.conllu_320]



1/ Ancrage à la racine

- Présence d’une conjonction de subordination liant le segment à un segment qui précède ou qui suit

si **c'est pas indiscret** qu'est ce que vous comptez faire plus tard ?
[CINE/ESLO2_CINE_1215.conllu_555]



2/ Absence de subordonnée de l’adjectif

- Quand l’adjectif est complémenté par une subordonnée : relations obl:mod, csubj, xcomp, advcl

  - euh passer une communication euh au milieu du brouhaha de vingt personnes **c'est pas** obligatoirement **agréable** pour celui qui est à l' autre bout [DIA/ESLO2_DIA_1225.conllu_1034]
  - hm je pense c' est du baratin parce que **c'est pas possible** de vivre comme ça [ENT/ESLO2_ENT_1001.conllu_956
  - donc **c' était pas** très **compliqué** de savoir où on voulait aller se poser [ENT/ESLO2_ENT_1005.conllu_289]

Mise en oeuvre des filtres syntaxiques

- Requête modifiée :
  - <l=ce,c=PRON,#1>&&<l=être,c=AUX,#2>&&<l=pas,c=ADV,#3>&&<c=ADJ,#4>::(advmod,4,3) (cop,4,2) (nsubj,4,1) **(root,4) (!mark,4,5)(!advcl|xcomp|obl:mod,4,6)** && prec(#1,#2,#3,#4)

Tableau de réalisation

Réalisation	Fréquence	Dispersion	Surface moyenne
ce_PRON être_AUX pas_ADV grave_ADJ	443	244	4.07
ce_PRON être_AUX pas_ADV possible_ADJ	267	175	4.06
ce_PRON être_AUX pas_ADV pareil_ADJ	189	122	4.17
ce_PRON être_AUX pas_ADV vrai_ADJ	187	132	4.12
ce_PRON être_AUX pas_ADV évident_ADJ	162	117	4.19
ce_PRON être_AUX pas_ADV facile_ADJ	137	96	4.42
ce_PRON être_AUX pas_ADV cher_ADJ	72	51	4.61
ce_PRON être_AUX pas_ADV bon_ADJ	49	40	4.35
ce_PRON être_AUX pas_ADV compliqué_ADJ	47	43	4.43
ce_PRON être_AUX pas_ADV même_ADJ	46	36	5.17
ce_PRON être_AUX pas_ADV normal_ADJ	44	35	4.16
ce_PRON être_AUX pas_ADV beau_ADJ	35	28	4.4
ce_PRON être_AUX pas_ADV terrible_ADJ	35	29	4.14

Cooccurrence syntaxique

Cooccurrences

Expression	Collocatif	Cooccurrences (via une relation)	Nombre de relations avec le pivot	Nombre de relations avec le collocatif	Nombre total de relations	Dispersion (file)	Relations	Mesure d'association
ce_PRON être_AUX pas_ADV grave_ADJ	mais_CCONJ	107	556	69170	11247632	86	cc	550.641
ce_PRON être_AUX pas_ADV grave_ADJ	bon_INTJ	39	556	17115	11247632	35	discourse	225.26
ce_PRON être_AUX pas_ADV possible_ADJ	non_ADV	42	315	51242	11247632	28	advmod discourse	208.014
ce_PRON être_AUX pas_ADV grave_ADJ	hein_INTJ	41	556	27746	11247632	35	discourse	202.306

Cooccurrence syntaxique

- Le tableau des cooccurrences de notre construction montre une forte attraction avec des interjections (ou marqueurs discursif assimilés) : *hein*, *bon*, *mais*, *non*, … via la relation *discourse*

- Pour mieux cibler les PPI, on peut ajouter ce paramètre contextuel :

    - <l=ce,c=PRON,#1>&&<l=être,c=AUX,#2>&&<l=pas,c=ADV,#3>&&<c=ADJ,#4>::(advmod,4,3) (cop,4,2) (nsubj,4,1) (root,4) **(discourse,4,5) (!mark,4,6)(!advcl|xcomp|obl:mod,4,7)** && prec(#1,#2,#3,#4)

Cooccurrence syntaxique

- Il reste un peu bruit dans les occurrences, qu'il faudrait quantifier

oui	c'	est pas toujours possible
et puis bon	c'	est toujours euh pas mixte
et puis bon	c'	est toujours euh pas mixte
	c'	est pas grave au revoir
ouais	c'	est pas bête
	c'	est pas très objectif quoi
c' est c' est	c'	est pas facile hein
ah non	c'	est pas grave du tout
ah non	c'	est pas grave du tout
non	c'	est pas vrai

6. Conclusion

Conclusion

- Nous avons illustré comment une **analyse collostructionnelle** peut être menée grâce au langage **TQL** (tableau de réalisations, cooccurrences, …)

- Même sur un corpus oral présentant de mauvaises performances d'annotation en dépendances, les requêtes syntaxiques semblent obtenir le **meilleur équilibre entre bruit et silence**.

- Les requêtes syntaxiques fournissent des **paramètres contextuels utiles** pour caractériser l'autonomie de la PPI :
  - enracinement
  - absence de subordination

- Les marqueurs identifiés par la relation « discourse », exprimant l'étonnement, l'exclamation, l'opposition, l'accord, etc. constituent un critère utile de filtrage

- Une évaluation quantitative sur un corpus de PPI de référence devra être conduite pour valider ces hypothèses

Merci !

Références bibliographiques

Evert, S. (2008). Corpora and collocations, in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin, 1212–1248.
Gries, S. Th. (2019) 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, 24(3):385-412

Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D. (2004), "The Sketch Engine", *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France, p. 105–116.
Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. In Max Silberstein (dir.), Langue française, N° 203, Armand Colin, p. 67-82.
Pausé, M.-S., Tutin, A., Kraif, O., Coavoux, M. (2022). Extraction de Phrases Préfabriquées des Interactions à partir d'un corpus arboré du français parlé : une étude exploratoire. *Congrès Mondial de Linguistique Française - CMLF 2022*, Jul 2022, Orléans, France.
Rainsford T., Heiden S. (2014). « Key Node in Context (KNIC) Concordances: Improving Usability of an Old French Treebank", *in Actes de la 4e édition du Congrès Mondial de Linguistique Française* (CMLF). Berlin, 19-23 juillet 2014, Vol. 8, 2707-2718.

Seretan, V. (2011) . *Syntax-Based Collocation Extraction.* Text, Speech and Language Technology, Springer.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8(2)., 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Tutin, A., Kraif O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents, in A. Tutin & F. Sitri (coord.) *Phraséologie et genres de discours : patrons, motifs, routines, Lidil* n° 53, Grenoble : Ellug.