Motifs et romans dans le projet PhraseoRom

Jalons méthodologiques et perspectives de résultats

Sascha Diwersy, Olivier Kraif

Phraséologie et stylistique de la langue littéraire
Erlangen, 13-15 mars 2019

Plan :

Introduction
Méthodologie
Elaboration d’une mesure de similarité
Résultats
Perspectives d’application
Conclusion

Introduction

Quels motifs ?

- Ganascia, 2001 : structures lexicosyntaxiques répétées
    - Extraction de motifs caractéristiques de Madame de Lafayette

Mise en œuvre

- Longrée, Luong & Mellet, 2008 : association récurrente d'éléments de nature variée (mots, lemmes, traits catégoriels morphosyntaxiques ou autres) dans une structure linéaire.

- Legallois, Charnois et Poibeau, 2016 : séquences multidimensionnelles mêlant mots outils et catégories morphosyntaxiques :
    - Etude de clichés dans les romans sentimentaux :

PRES le NC, il V-PS -> « Rouvrant les yeux, elle retint un cri »

Mise en œuvre

Quel motifs ?

- A la fois *structurants* et *caractérisants* (Longrée et Mellet, 2013)
- pour nous : **séquences préfabriquées multidimensionnelles** dotées d'une **fonction discursive**

- Pour l'étude des romans, niveau **narratif** à prendre compte

Intéressant pour étudier l'opposition *spécificité* (style de l'auteur) vs *généricité* (normes et conventions du genre)

Méthodologie du projet PhraseoRom

- *corpus driven* (Tognini-Bonelli, 2001)
- identification automatique des structures récurrentes

Mise en œuvre

- prise en compte des annotations syntaxiques en dépendances : extraction d'arbres récurrents (ALR) plutôt que de segments répétés (ngrams)

Problème des extractions automatiques

- foisonnantes et redondantes

  ngrams : “elle s’assit sur le bord du lit” (9 occ., corpus SENT)

  elle s’assit sur le bord du lit
  elle s’assit sur le bord du
  s’assit sur le bord du lit
  elle s’assit sur le bord
  s’assit sur le bord du
  assit sur le bord du lit
  elle s’assit sur le
  s’assit sur le bord
  assit sur le bord du
  sur le bord du lit
  elle s’assit sur
  s’assit sur le
  assit sur le bord

- Comment différencier séquences autonomes vs fragmentaires ?

Mise en œuvre

Problème des extractions automatiques

- Même problème pour les ALR :
  - Variantes lexicales + extensions :

Mise en œuvre

<gravir le escalier>
<grimper le escalier>
<il gravir marche>
<il monter marche>
<il monter le escalier>
<il monter le marche>
<il gravir le marche>

Questions

- comment regrouper toutes ces occurrences sous des motifs plus abstraits ?
- comment délimiter des motifs voisins ?
- différents niveaux pour grouper ces motifs ?

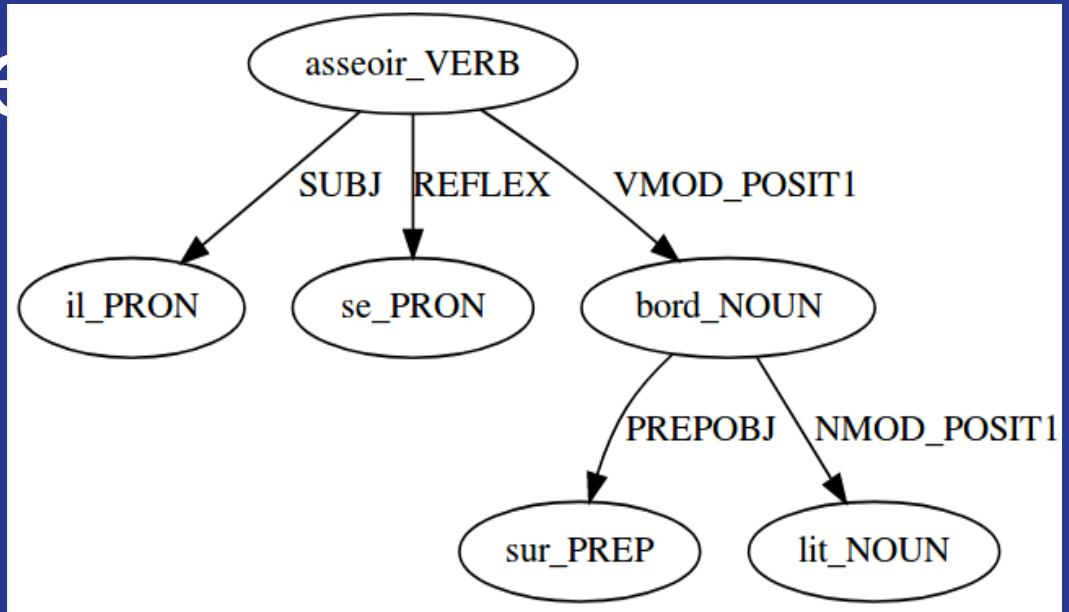
Mise en œuvre

Méthodologie du projet PhraseoRom

Extraction des Arbres Lexicosyntaxiques Récurrents (ALR, Kraif, Diwersy 2012).

Ex. : <il se asseoir sur le bord du lit>

Mise e



Mise en œuvre

Extraction des Arbres Lexicosyntaxiques Récurrents :

- extraction itérative (on augmente l'arbre à chaque itération)
- critère de **fréquence** -> pour ne retenir que les expressions récurrentes
- critère de **dispersion** (nb. d'auteurs différents) -> pour ne retenir que les expressions partagées par plusieurs auteurs
- calcul de la **spécificité** (loglike) dans chaque sous-corpus par rapport à l'ensemble du corpus

p. ex. <se laisser tomber sur le bord du lit>
-> spécifique du sous-corpus Harlequin

Mise en œuvre

Identification manuelle des motifs

1. identification d'un **noyau** commun à plusieurs ALR :

<pousser le porte>
<ouvrir le porte>
<le porte se ouvrir>
<le porte se refermer>

-> Motif spécifique à POL structurant la narration autour de l'enquête

2. identification des **extensions** et de leur fonction :

<ouvrir le porte de le chambre>
<refermer doucement le porte>

Mise en œuvre

3. Identification de motifs abstraits (combinant des traits lexicaux, syntaxiques, sémantiques). P.ex. Gonon et al., 2018 :

Vdicendi ou V de mouvement ou V d'action + motif au gérondif

Vous avez été longs ! **s'écria** Emma **en ouvrant la porte**. J'imaginais déjà le pire ! (F. Thilliez, La Forêt des ombres, 2006)

Fonction narrative -> nouvelle séquence (changement du nombre de personnages)

Problème de l’annotation manuelle : comment annoter les dizaines de milliers d’ALR à l’issue des extractions automatiques ?

Mise en œuvre

-> Elaboration d’une mesure de similarité entre ALR afin d’effectuer des regroupements et limiter la redondance

Mesure de similarité entre ALRs

Mise en œuvre

1. On assimile un ALR à une suite de triplets :

(REL ,LEMME1, LEMME2)

Exemple : *<il reprendre le parole>*

Triplets : [('DETERM_DEF', ('parole', '', 'NOUN'), ('le', '', 'DET'))
('OBJ', ('reprendre', '', 'VERB'), ('parole', '', 'NOUN'))
('SUBJ', ('reprendre', '', 'VERB'), ('il', '', 'PRON'))]

Similarité DICE : nombre de triplets communs / moyenne du nombre de triplets des ALR (0=différence complète, 1=identité) :

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

Amélioration : on ne calcule pas exactement le nombre de triplets communs, mais la somme des similarités des triplets considérés comme similaires

Mise en œuvre

2. Réduction : suppression de certains triplets considérés comme secondaires : pronom sujet, déterminant, etc.

Exemple : <*il reprendre le parole*>

Triplets : ('DETERM_DEF', ('parole', '', 'NOUN'), ('le', '', 'DET'))
('OBJ', ('reprendre', '', 'VERB'), ('parole', '', 'NOUN'))
('SUBJ', ('reprendre', '', 'VERB'), ('il', '', 'PRON'))]

Mise en œuvre

3. Comparaison des triplets

Deux conditions pour l'appariement des triplets :

4. Même structure <REL, CAT, CAT>
5. Contient une paire de noms similaires ou de verbes similaires :
   a. (N1,N2) tel que simLex(N1,N2) > seuil (0.7)
   b. (V1,V2) tel que simLex(V1,V2) > seuil (0.7)

La similarité lexicale est calculée avec Word2Vec (plongement lexical, Mikolov et al. 2013), entrainé sur le corpus LITT (282 romans, 30 000 536 mots).

Exemple de similarité :

mostSimilar("bondir")= [["sauter",0.7108582244], ["foncer",0.68430234],["propulser",0.63368963],.....]

Dans le calcul de Dice, on fait intervenir la similarité lexicale des mots appariés.

Résultats

Résultats

- **Réordonnancement des ALR**
  Méthode de tri permettant de maximiser la similarité entre 2 ALR contigus.

  -> Facilitation de l’annotation manuelle - les ALR de même noyau se suivent.

Résultats

- **Exemple 1 : Motif 12**

<marcher dans le rue>
<je sortir dans le rue>
<passer dans le rue>
<je marcher dans le rue>
<croiser dans le rue>
<il passer dans le rue>
<sortir dans le rue>

Ici on voit que c'est le prédicat "dans la rue" qui constitue le noyau le plus stable. On pourrait schématiser le motif comme :

V /déplacement/ + *dans la rue*

Résultats

- **Exemple 2 : motif 28**

<il gravir le escalier>
<je descendre le escalier>
<il descendre escalier>
<descendre le escalier>
<monter le escalier>
<gravir le escalier>
<grimper le escalier>
<il gravir marche>
<il descendre le escalier>
<il monter escalier>
<il monter le escalier>
<dévaler le escalier>

Ici, le noyau lexical est bien le nom *escalier*, mais on constate que la mesure de similarité arrive bien à récupérer une forme synonyme (*marche*). Par ailleurs, le motif incorpore différents verbes de mouvement, qu'ils indiquent la montée ou la descente. Le motif sous jacent peut s'écrire :

V /mouvement vertical/ + N /escalier/

Résultats

- **Exemple 3 : motif 102**

<il fourrer dans le poche>
<glisser dans son poche>
<il ranger dans poche>
<enfoncer dans poche>
<mettre dans son poche>
<ranger dans son poche>
<fouiller dans son poche>
<le fourrer dans son poche>
<il fouiller dans poche>
<il fourrer dans sac>
…

Là encore, c'est un prédicat qui forme le noyau lexical le plus stable : "dans sa poche", associé à différent verbe d'action, qui se répartissent sur deux acceptions /mettre/ ou /chercher/.
L'identification d'un seul ou de deux motifs dépendra des fonctions discursives identifiées dans les textes. L'inclusion de la variante "sac" dans le motif en dépend également.

V /mettre ou chercher/ + *dans*

Perspectives d’application

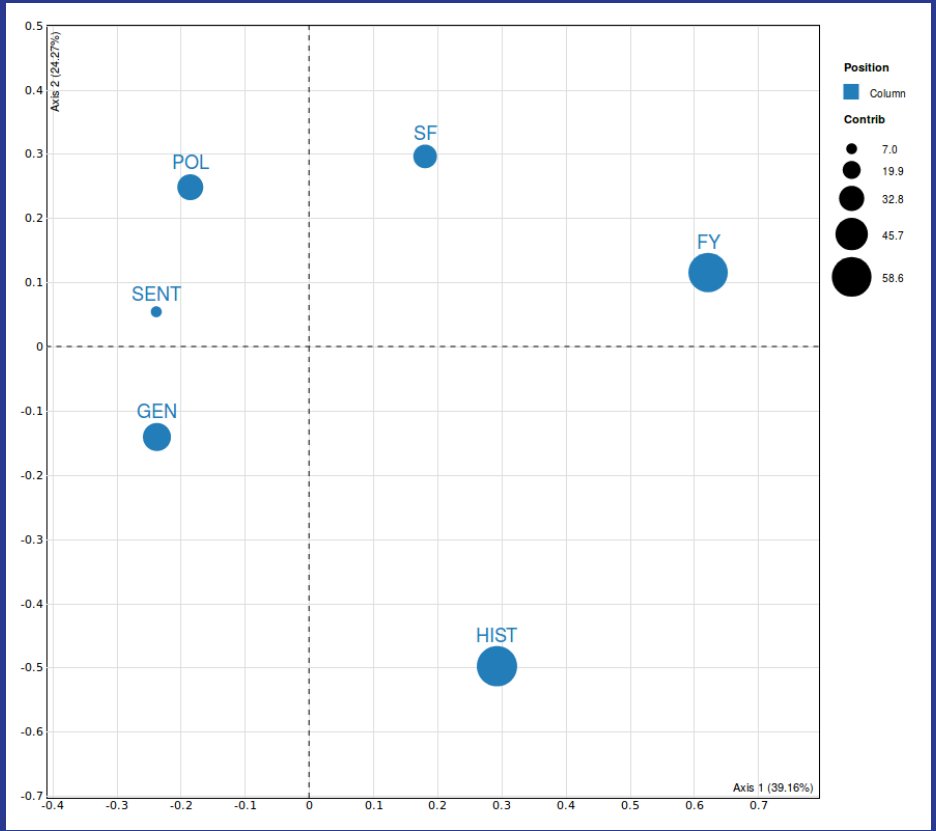
Perspectives d’application

- **Traitements textométriques**
  - Analyse factorielle des correspondances (AFC; Benzécri 1973; Lebart & Salem 1994)
  - Spécificités (Lafon 1980; Lebart & Salem 1994)

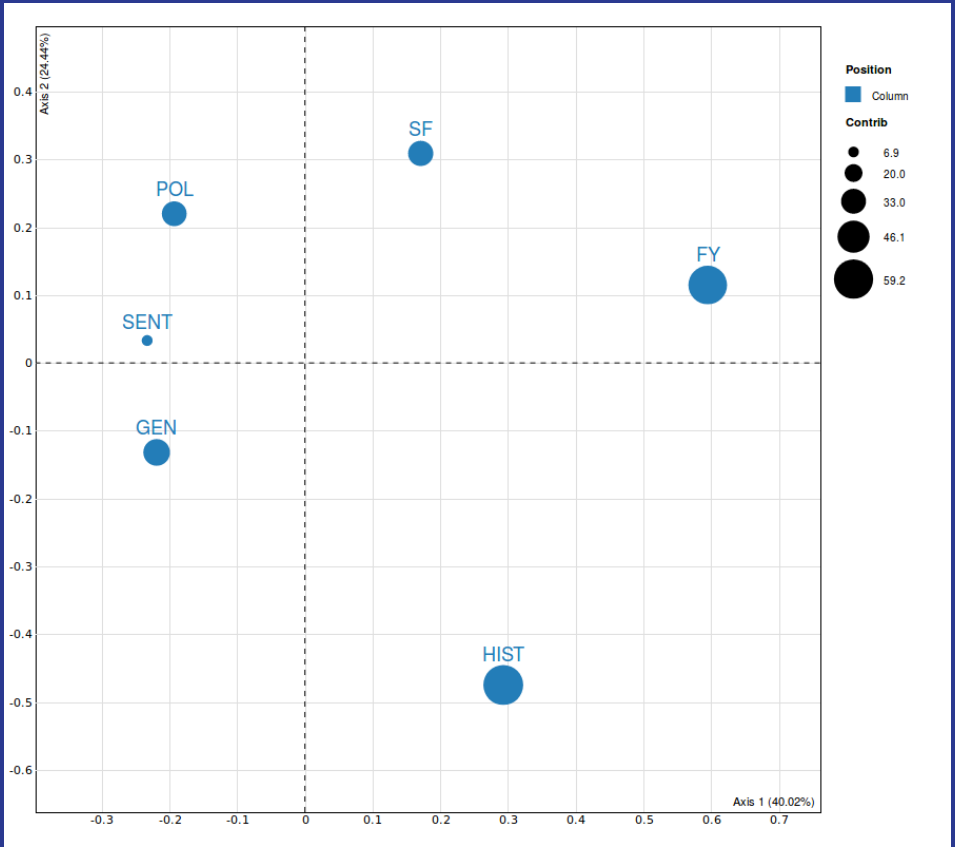
- **Réseaux d’ALR regroupés**
  - Mesures de centralité
  - Identification de “communautés” par mesures de modularité

AFC - ALR x Genres

Mise



AFC - ALR regroupés x Genres



Spécificités - ALR regroupés x Genres

ID	ALR_reg	FY	GEN	HIST	POL	SENT	SF
PG10324	<la jeune femme>	1000	-1000	-44,2686	72,3182	0,8752	0,5832
PG3562	<le jeune homme>	283,007	-77,6296	-3,7133	-7,7636	-2,3247	-5,126
PG1673	<de la forêt>	255,5144	-45,6972	-20,5139	-7,7883	-22,754	1,044
PG8013	<la jeune fille>	218,0978	-61,4423	-10,0133	-40,6123	1,2424	10,622
PG1795	<sur DET champ de bataille>	189,0628	-68,3817	34,9003	-30,4651	-24,2382	-23,9287
PG8907	<de la guilde>	168,9264	-36,1847	-10,7601	-17,9591	-9,4334	-6,688
PG5966	<la scène de crime>	-27,5939	-43,3409	-34,6596	258,05	-9,8174	-12,0065
PG2749	<composer DET numéro>	-83,6098	-14,135	-108,6601	211,3682	27,3005	-21,3847
PG8889	<de l'enquête>	-17,9354	-35,7473	-18,8489	183,687	-13,2228	-1,5132
PG7656	<c'est ça>	-53,6605	10,2013	-273,4842	180,1776	38,6008	-18,8063
PG9303	<il y a>	-236,5403	1000	-280,1429	56,0555	-23,8746	2,1809
PG92	<dire que P-DI | + P-DD>	-1000	1000	-125,1114	11,3155	26,2604	-207,3711
PG1488	<avoir l'air + ADJ | + de VINF>	-196,0484	146,8433	-262,0611	19,3904	140,3525	-24,1603

Perspectives d’application

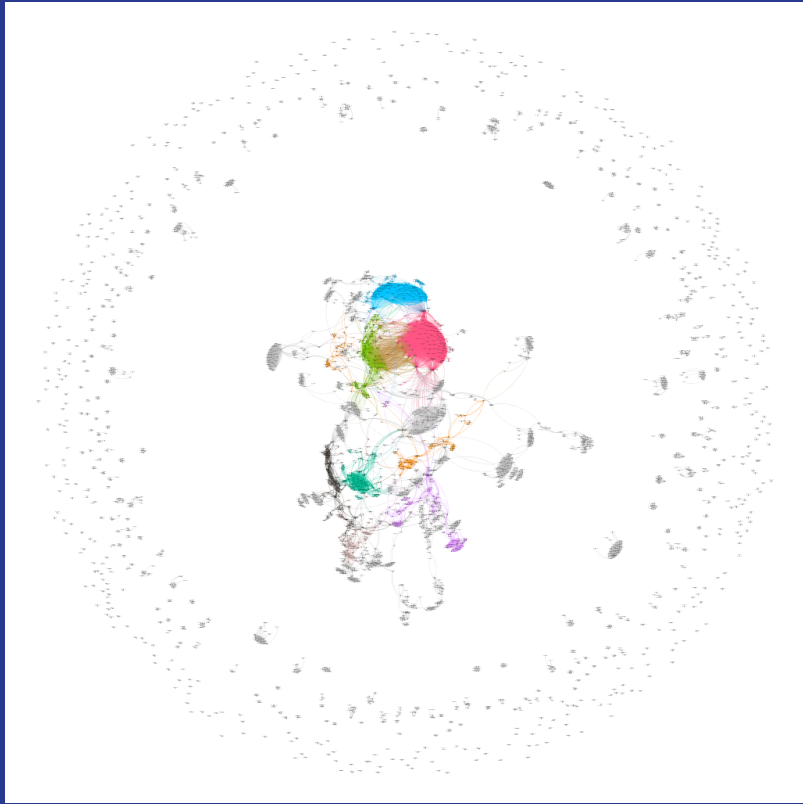
- **Problème de la délimitation des motifs**
  - N_{SUJ} *apparaître* *sur l’écran* vs N_{SUJ} *voir* N_{COD} *sur l’écran*

- **Motifs en tant qu’assemblages collocationnels ?**
- **Ressemblances de famille ?**
- **Relations réticulaires entre (sous-)motifs ?**

Perspectives d’application

- Réseaux de d’ALR regroupés, construits sur la base de bigrammes partagés
  - apparaître sur l’écran → voir sur l’écran
- Méthodes d’exploration de réseaux (logiciel *Gephi*, v. 0.9.2)
  - Identification d’ALR regroupés en tant que noeuds centraux du réseau
  - Délimitation de “communautés” d’ALR regroupés au sein du réseau

Réseau d’ALR regroupés



Perspectives d’application

- ALR regroupés centraux
  - G420 : V /déplacement/ *dans* N /loc/
    {*entrer* | *rentrer* | *pénétrer*} *dans* {*chambre* | *cuisine* | *pièce* | *salon*}

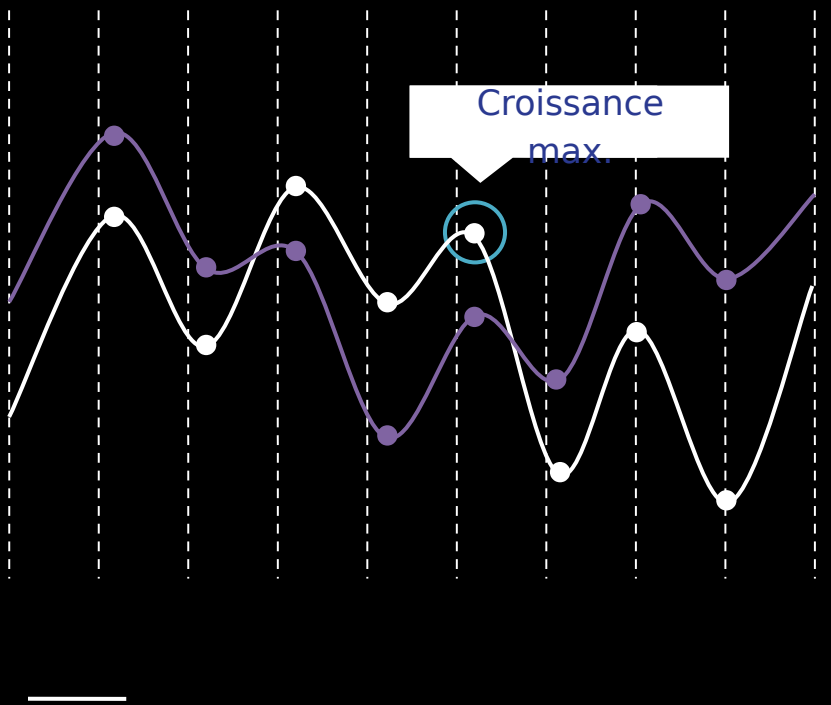
  - G1746 : *sortir de* N /loc/
    sortir de {*chambre* | *cuisine* | *pièce* | *salle*}

  - G6616 : *que* + $\hat{e}tre_{COP}$ | $avoir_{Vsup}$

Perspectives d’application

- Communautés d’ALR regroupés
  - PG10180 + PG1069 + PG1159 + PG1199 + ... + G6616 :
    - V /dicendi/ {*arguer* | *dire* | ...} + *que* P-DI
    - V /sentiendi/ {*réaliser* | *craindre* | ...} + *que* P-DI
    - *que* + *être*$_{\text{COP}}$ | *avoir*$_{\text{Vsup}}$

Conclusion



Conclusion

- La méthode de regroupement d’ALR permet de faciliter la description de motifs.
- L’utilisation d’ALR et d’ALR regroupés donne les mêmes résultats dans les tâches de classification textométrique (AFC).
- Questions ouvertes :
  - Identification (de “familles”) de motifs par exploration de réseaux d’ALR regroupés ?
  - Comment rendre plus efficace l’utilisation de ressources sémantiques dans l’extraction des ALR afin de les hybrider ?

Merci !

Ganascia J.-G. (2001). « Extraction automatique de motifs syntaxiques », *Actes de TALN 2001*, Tours, 2-5 juillet 2001.
Gonon L., Goossens V., Kraif O., Novakova I., Sorba J. (2018) Motifs textuels spécifiques au genre policier et à la littérature “blanche”. *6e Congrès Mondial de Linguistique Française*, Jun 2018, Mons, Belgique.
Kraif O., Diwersy S. (2012). « Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques », *Actes de la conférence TALN 2012*, Grenoble, 399-406.
Legallois D., Charnois T., & Poibeau T. (2016). « Repérer les clichés dans les romans sentimentaux grâce à la méthode des “motifs” », *Lidil. Revue de linguistique et de didactique des langues*, (53) : 95-117.
Longrée D., Mellet S. (2013). « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours », *Langages*, *189*, (1) : 65-79.
Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013) Distributed representations of words and phrases and their compositionality, *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Volume 2, Pages 3111-3119.
Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Mise en œuvre