Observations diachroniques dans un corpus de presse avec le Lexicoscope

Olivier Kraif et Sascha Diwersy
Université Grenoble Alpes - Université Paul Valéry

Colloque ConCorDiaL - 13-14 oct. 2022
Lidilem & Litt&Arts, UGA

Plan

1. Contexte et objectif
2. Quelques outils grand public
3. Méthodes d'analyse diachronique
4. Présentation du Lexicoscope 2.0
5. Corpus *lmdiachro*
6. Fonctionnalité chronogramme et Étude de cas
7. Perspectives

Contexte et objectifs

Contexte et objectifs

- Intérêt des corpus arborés pour l'étude de la combinatoire
  - Cooccurrents syntaxiques (Evert, 2007)
  - Wordsketches (Sketch Engine, Lexicoscope)
  - Arbres lexicosyntaxiques récurrents (ALR)
- Nouveaux modèles performants sur différents états de langue
  - Hops (Grobol & Crabbé, 2021), BERTrade (Crabbé et al., 2022)
- Projet PhraseoRoche : corpus diachronique du 13e au 17e siècle (romans de chevalerie) (Coavoux et al. 2022).
- Enrichissement du Lexicoscope pour l'étude de ces nouvelles données textuelles en diachronie.

Outils pour l'exploration diachronique de corpus

Quelques outils grand public

- Google Trend : popularité des requêtes (depuis 2004), mais pas de fréquence d'occurrence sur le Web.
- Sketch Engine : fonction Trends

	Word	Trend	Frequency	Sample
1	γ	↗	1,033	
2	website	↗	1,343	
3	annotators	↗	13,723	
4	outperforms	↗	7,918	
5	β	↗	1,548	
6	baseline	↗	35,179	
7	challenges	↗	3,837	
8	annotator	↗	6,036	
9	download	↗	610	
10	outperformed	↗	2,196	
11	outperform	↗	4,004	
12	publicly	↗	2,428	
13	cosine	↗	4,541	

	Word	Trend	Frequency	Sample
18	outliers	↗	495	
19	leveraging	↗	896	
20	rerank	↗	290	
21	tokenised	↗	184	
22	mining	↗	4,460	
23	mined	↗	725	
24	females	↗	250	
25	toolkit	↗	3,529	
26	custom	↗	556	
27	median	↗	1,218	
28	overfitting	↗	1,486	
29	leverage	↗	1,527	
30	metrics	↗	14,021	

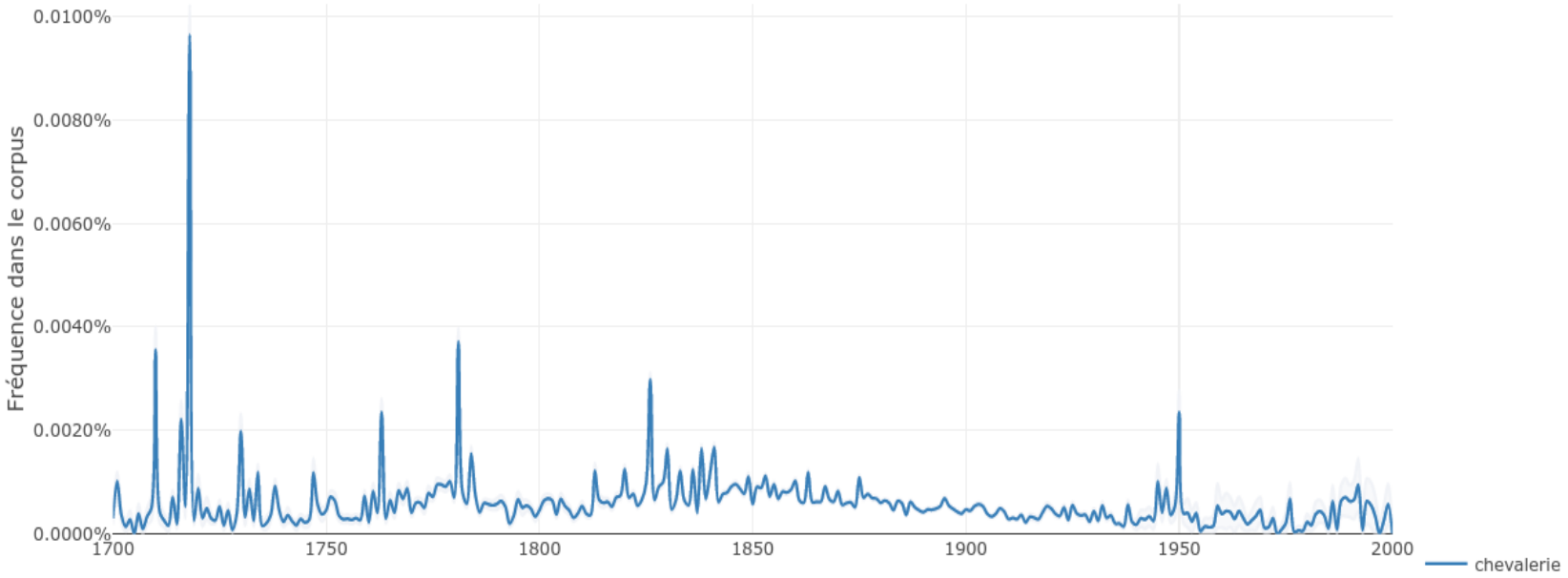
	Word	Trend	Frequency	Sample
35	sampled	↗	2,759	
36	outperforming	↗	697	
37	priors	↗	1,596	
38	impacted	↗	149	
39	treebanks	↗	2,705	
40	lemmatized	↗	1,046	
41	clinical	↗	3,769	
42	lemmatize	↗	173	
43	θ	↗	1,481	
44	α	↗	6,208	
45	revisit	↗	184	
46	φ	↗	1,023	
47	λ	↗	1,311	

Corpus ACL Anthology (2001-

Quelques outils grand public

- Gallicagram

  Evolution de la fréquence relative par année à travers les collections de Gallica (et d'autres bibliothèques françaises)



Méthodes d’analyse diachronique

Traitement lexico-statistique de données diachroniques

- Constat : Le traitement d'une variable chronologique doit tenir compte du fait que celle-ci relève d'une échelle d'intervalle.
- Différentes visées méthodologiques :

1) Regroupement de mots par profils chronologiques
   - Mesures de tendance globale (Herman & Kovář, 2013 ; Hilpert & Gries, 2009, 388-390) ;
   - Spécificités diachroniques (Salem 1988, 126-131 ; Lebart et al. 1998, 155-161)
   - Calcul du barycentre (Salem 1988, 134-139)
   - Curve clustering (Trevisani & Tuzzi, 2016)

2) Diagnostiquer l'influence du facteur « temps » au sein d'une série préconstruite selon des critères chronologiques
   - AFC (Salem 1988, 110-118) ;

Traitement lexico-statistique de données diachroniques

- Différentes visées méthodologiques (suite) :

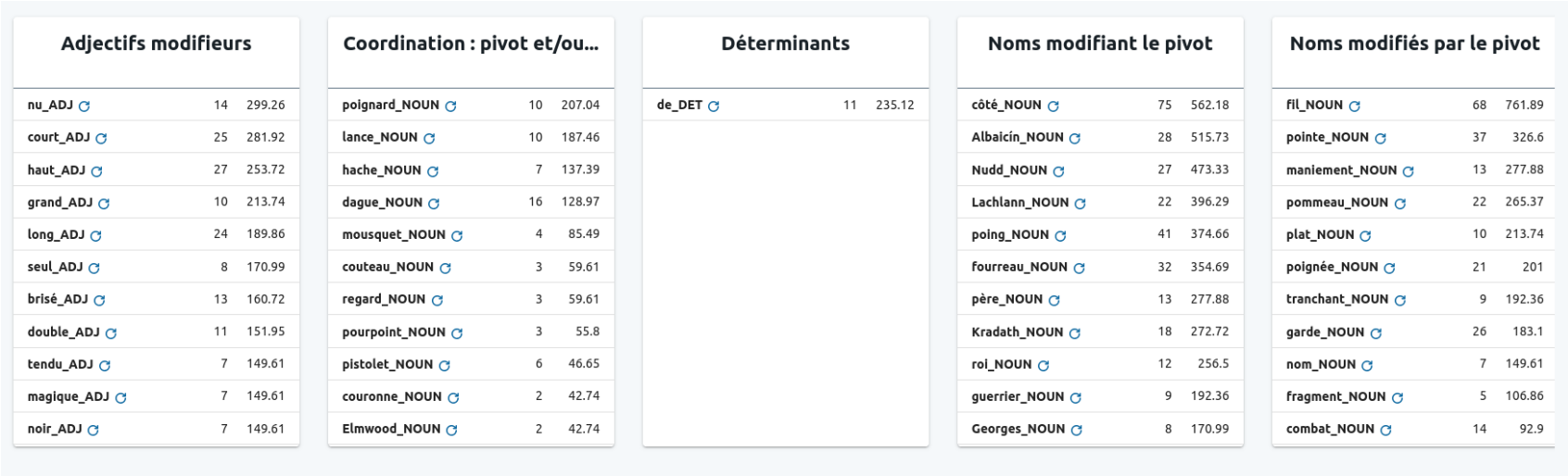
3) Périodisation automatique
   - Variability based neighbor clustering (VNC, Gries & Hilpert 2008)

4) Identification de tournants
   - Modèles additifs généralisés (Gabrielatos et al. 2012)

Présentation du Lexicoscope 2.0

Le Lexicoscope 2.0

- Une conception proche du Sketch Engine
  - Concordances
  - Spécificités (*Keywords*)
  - Index hiérarchiques (*Word lists*)
  - Expressions polylexicales et ALR (*n-grams*)
  - Extraction de cooccurrents syntaxiques et de Wordsketches



Adjectifs modifieurs

nu_ADJ	14	299.26
court_ADJ	25	281.92
haut_ADJ	27	253.72
grand_ADJ	10	213.74
long_ADJ	24	189.86
seul_ADJ	8	170.99
brisé_ADJ	13	160.72
double_ADJ	11	151.95
tendu_ADJ	7	149.61
magique_ADJ	7	149.61
noir_ADJ	7	149.61

Coordination : pivot et/ou...

poignard_NOUN	10	207.04
lance_NOUN	10	187.46
hache_NOUN	7	137.39
dague_NOUN	16	128.97
mousquet_NOUN	4	85.49
couteau_NOUN	3	59.61
regard_NOUN	3	59.61
pourpoint_NOUN	3	55.8
pistolet_NOUN	6	46.65
couronne_NOUN	2	42.74
Elmwood_NOUN	2	42.74

Déterminants

de_DET	11	235.12

Noms modifiant le pivot

côté_NOUN	75	562.18
Albaicín_NOUN	28	515.73
Nudd_NOUN	27	473.33
Lachlann_NOUN	22	396.29
poing_NOUN	41	374.66
fourreau_NOUN	32	354.69
père_NOUN	13	277.88
Kradath_NOUN	18	272.72
roi_NOUN	12	256.5
guerrier_NOUN	9	192.36
Georges_NOUN	8	170.99

Noms modifiés par le pivot

fil_NOUN	68	761.89
pointe_NOUN	37	326.6
maniement_NOUN	13	277.88
pommeau_NOUN	22	265.37
plat_NOUN	10	213.74
poignée_NOUN	21	201
tranchant_NOUN	9	192.36
garde_NOUN	26	183.1
nom_NOUN	7	149.61
fragment_NOUN	5	106.86
combat_NOUN	14	92.9

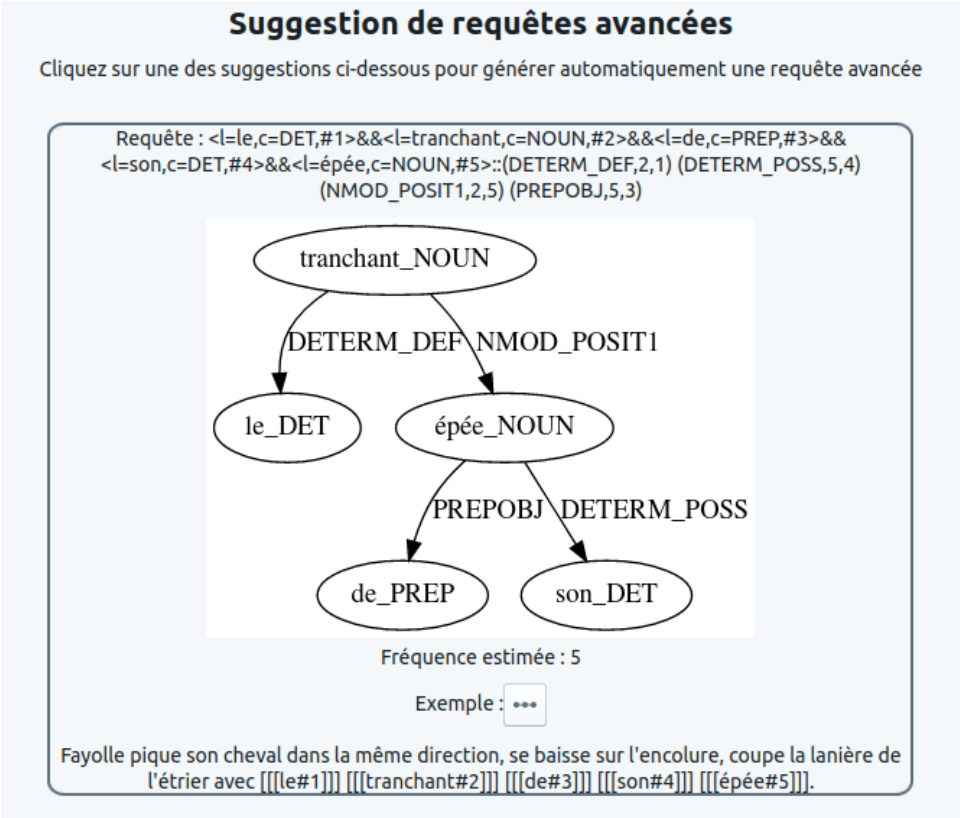
Le Lexicoscope 2.0

- Exemple d'extraction d'ALR

pointe_NOUN	6	<c=DET,l=le,#1>&& <c=DET,l=le,#2>&& <c=NOUN,l=garde|pointe,#3>&& <c=NOUN,l=épée,#4>&& <c=PREP,#5>&& <c=PREP,l=de,#6>:: (DETERM_DEF,3,1) (NMOD_POSIT1,3,4) (PREPOBJ,3,5) (DETERM_DEF,4,2) (PREPOBJ,4,6)	de_PREP le_DET garde|pointe_NOUN de_PREP le_DET épée_NOUN	Il ouvrait la bouche et serrait les dents sur la pointe de l'épée qui les disjoignait.	25
fil_NOUN	6	<c=DET,l=le,#1>&& <c=NOUN,l=fil,#2>&& <c=NOUN,l=épée,#3>&& <c=PREP,l=de,#4>&& <c=PREP,l=à,#5>&& <c=VERB,l=passer,#6>:: (NMOD_POSIT1,2,3) (PREPOBJ,2,5) (DETERM_DEF,3,1) (PREPOBJ,3,4) (VMOD_POSIT1,6,2)	passer_VERB à_PREP fil_NOUN de_PREP le_DET épée_NOUN	Emportant l'odeur de la mort, les cris des femmes que les assassins passent au fil de l' épée, les silhouettes grotesques des enfants pendus au gibet de Bénarès, leur cocarde autour du cou.	25
dégainer_VERB	3	<c=DET,l=le,#1>&& <c=NOUN,l=épée,#2>&& <c=VERB,l=brandir|dégainer,#3>:: (DETERM_DEF,2,1) (NMOD_POSIT1,2,3)	le_DET épée_NOUN brandir|dégainer_VERB	L'escarcelle derechef garnie, je quittai l'hôtel de Joyeuse tout ému de gratitude pour celle qui en était l'âme, regagnant mon logis sur mon Accla galopante, Giacomi et Miroul eux aussi montés et me flanquant, l' épée dégainée, pour ce que Cossolat m'avait avisé de ne plus cheminer à pied nuitamment par les rues de Montpellier, craignant pour moi quelque embûche et revanchement de la truanderie.\n	25
lance_NOUN	4	<c=DET,l=le,#1>&& <c=DET,l=le,#2>&& <c=NOUN,l=dague|lance,#3>&& <c=NOUN,l=épée,#4>:: (DETERM_DEF,3,1) (DETERM_DEF,4,2) (COORDITEMS,4,3)	le_DET épée_NOUN le_DET dague|lance_NOUN	Il montra à Felix l'épée et la lance avec lesquels il avait tué plus de mille taureaux.	24

Le Lexicoscope 2.0

- Mais avec une orientation « recherche »
  - Développement d'expressions complexes et de patterns syntaxiques
  - Étude contrastive de sous-corpus (partitions comme dans TXM)
  - Construction automatique des requêtes complexes (requêtage basé sur l'exemple, Augustinus et al., 2016)
  - Intermédiaire entre Sketch Engine (Kilgarriff et al., 2004) et Grew Match (Guillaume, 2021)



Corpus Imdiachro

Corpus *lmdiachro*

- Compilé par Sascha Diwersy, consultable sur le Lexicoscope 2.0
- Echantillon d'articles du Monde de 1944 à 2015 (2 à 3 éditions par semaine)
- Traitement avec le pipeline Bonsai (Candito et al., 2010)
- Découpé par décades
- lmdiachro V2 : complété jusqu'à 2019. Traité avec Stanza.

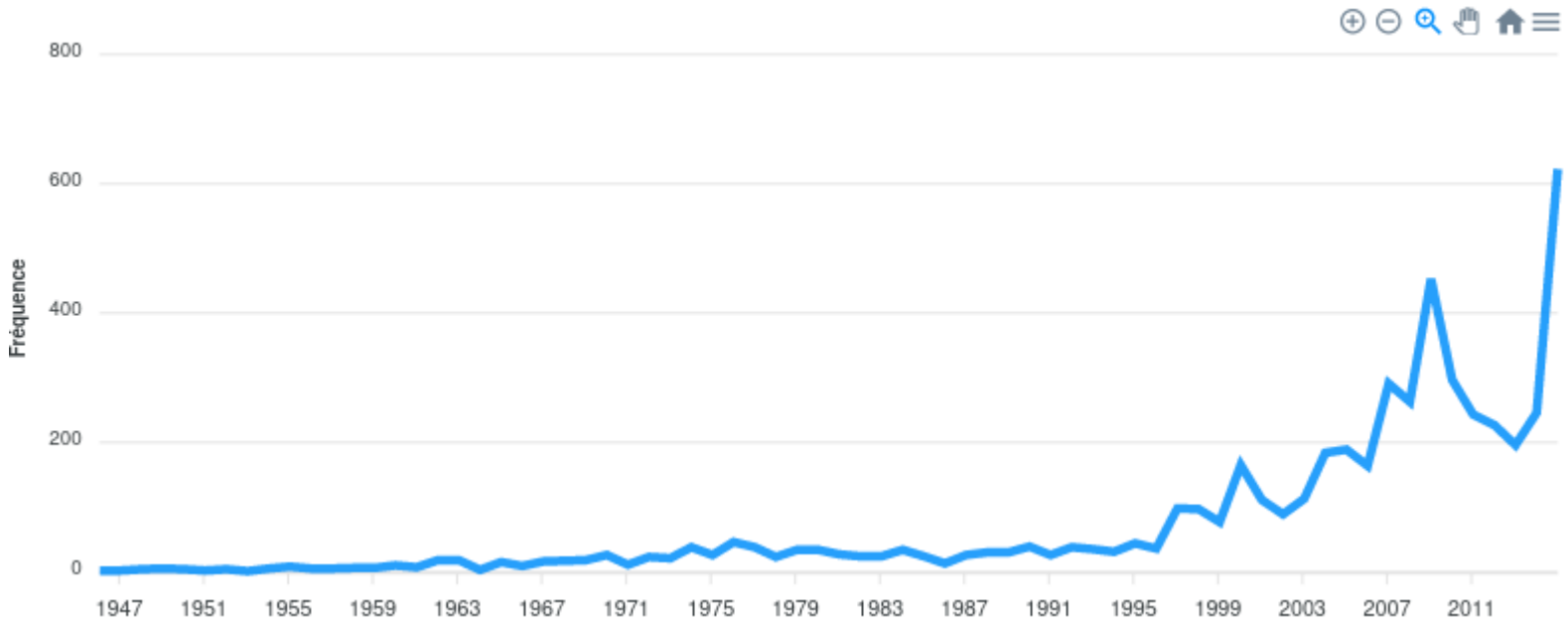
Décade	Editions	Articles	Tokens
1944-49	410	23 089	8 767 430
1950-59	827	71 069	29 048 869
1960-69	785	85 160	37 517 057
1970-79	789	93 415	50 826 934
1980-89	810	91 392	55 097 518
1990-99	830	125 020	66 990 080
2000-09	827	116 726	64 344 690
2010-15	499	66 072	35 905 223
Total	5 777	671 943	348 497 801

Fonctionnalité Chronogramme

étude de cas

Fonctionnalité chronogramme
et étude de cas

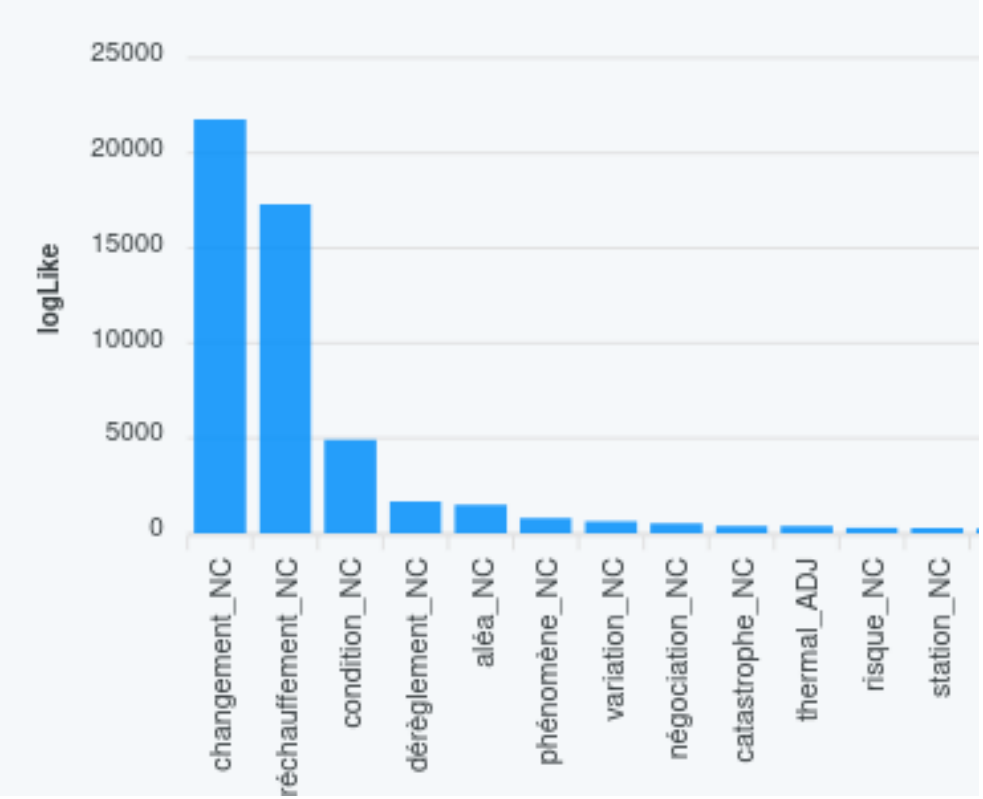
- Étude de l'adjectif *climatique* (5117 occ.)
- Le chronogramme donne l'évolution des fréquences (absolues, relatives et spécificité) pour la taille d'empan choisie (mois, année, décade, etc.)



Fonctionnalité chronogramme

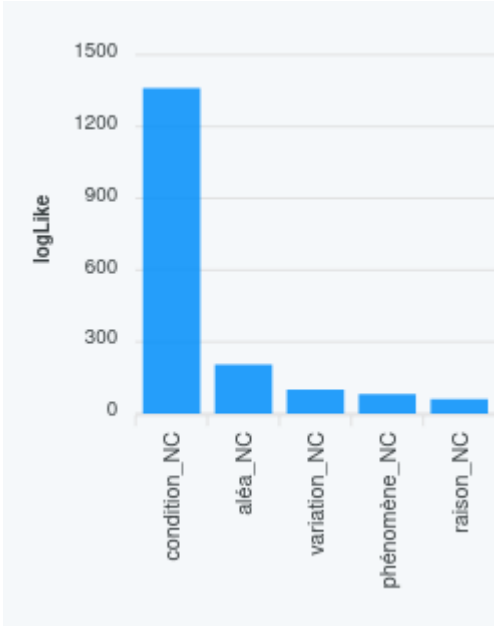
et étude de cas

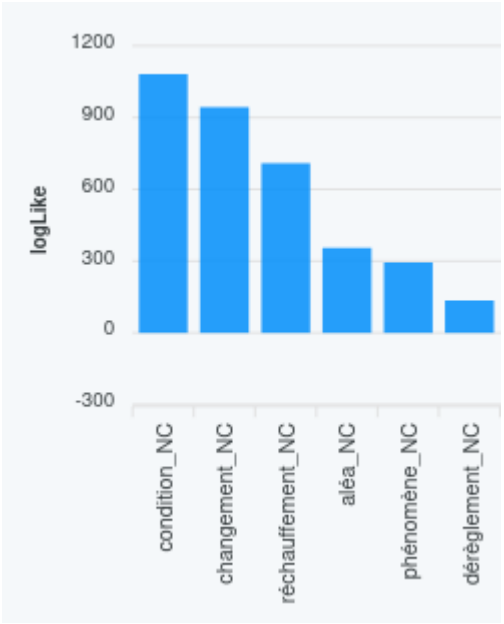
- Cooccurrents de *climatique*



Fonctionnalité chronogramme
et étude de cas

- Deux séries de cooccurrents :
  - *changement*, *réchauffement*, *condition*, *dérèglement*, *aléa*, *phénomène*, *variation*, *négociation*, *catastrophe*, *thermal*, *risque*, *station*
  - une étude par décade montre une bascule entre les années 1980 et 1990

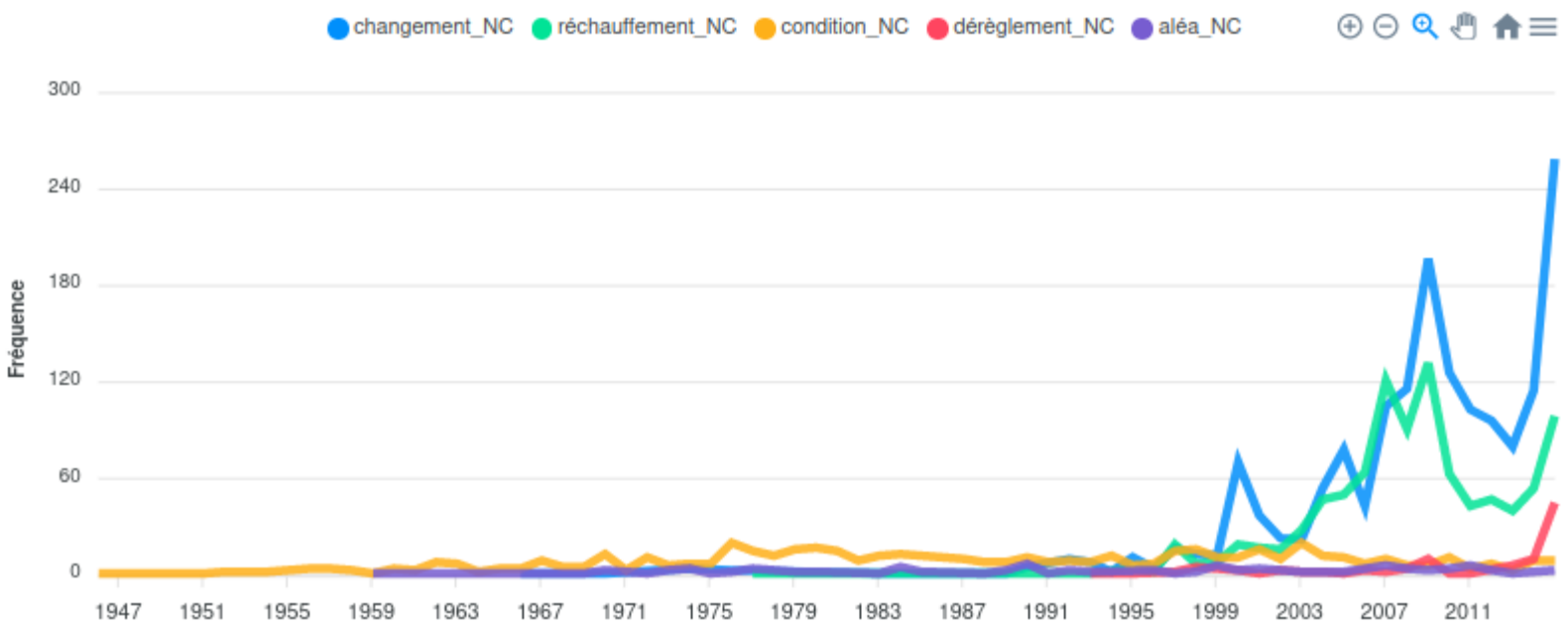




Fonctionnalité chronogramme

et étude de cas

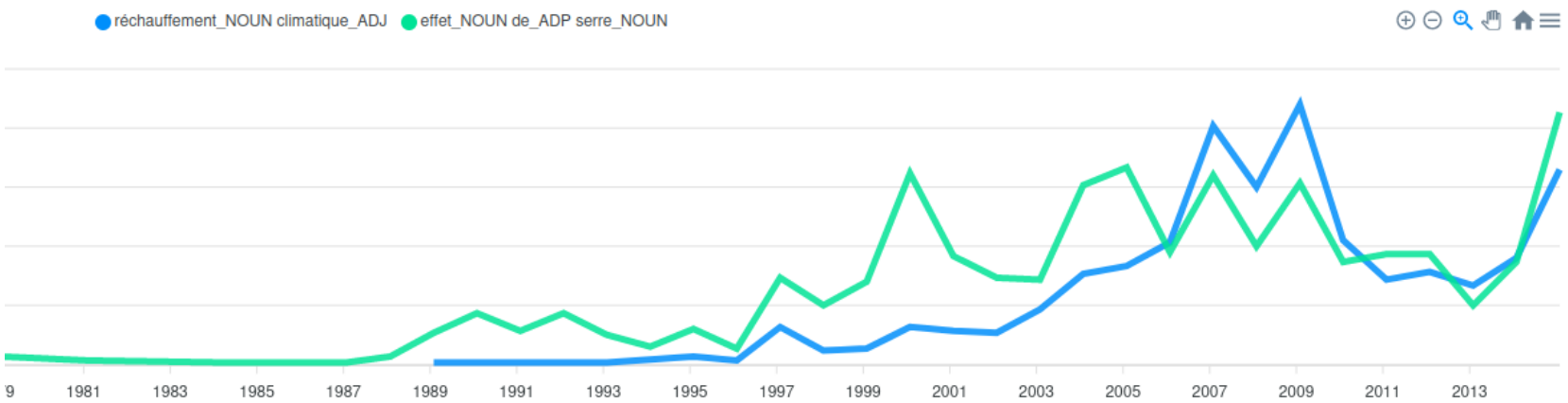
- Possibilité de croiser chronogramme et cooccurrents



Fonctionnalité chronogramme

et étude de cas

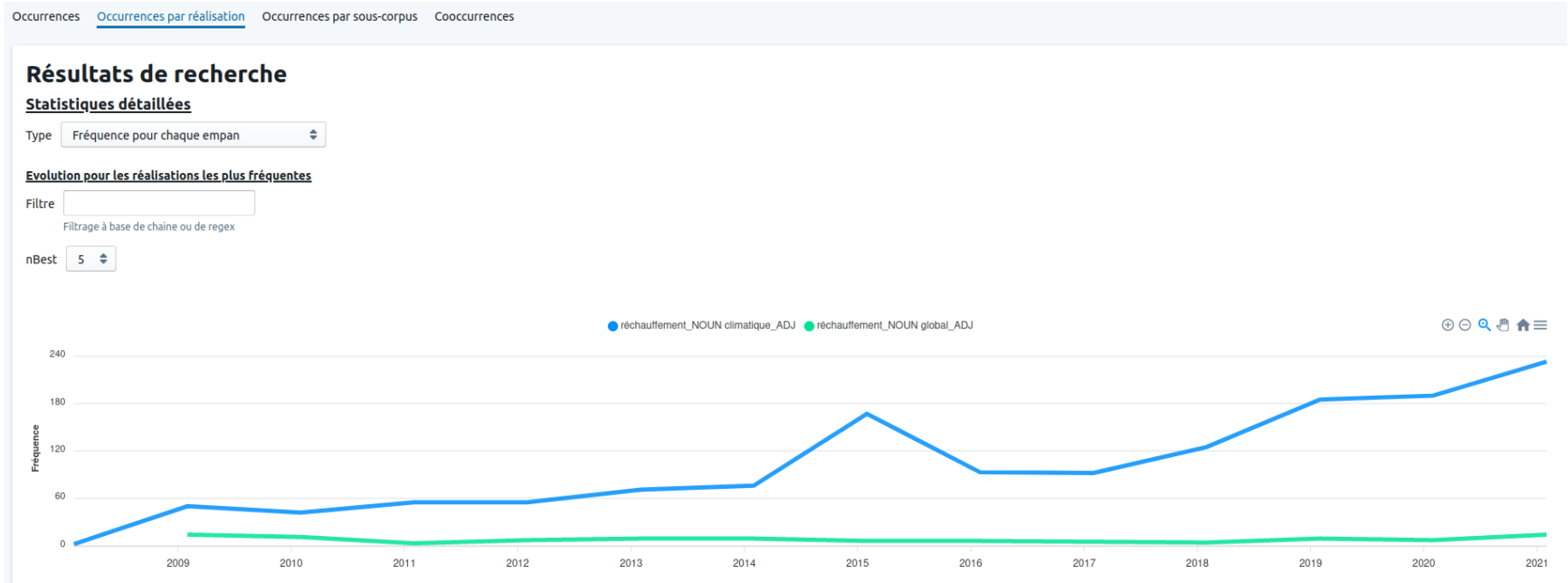
- Possibilité de croiser chronogramme et réalisations



Fonctionnalité chronogramme

et étude de cas

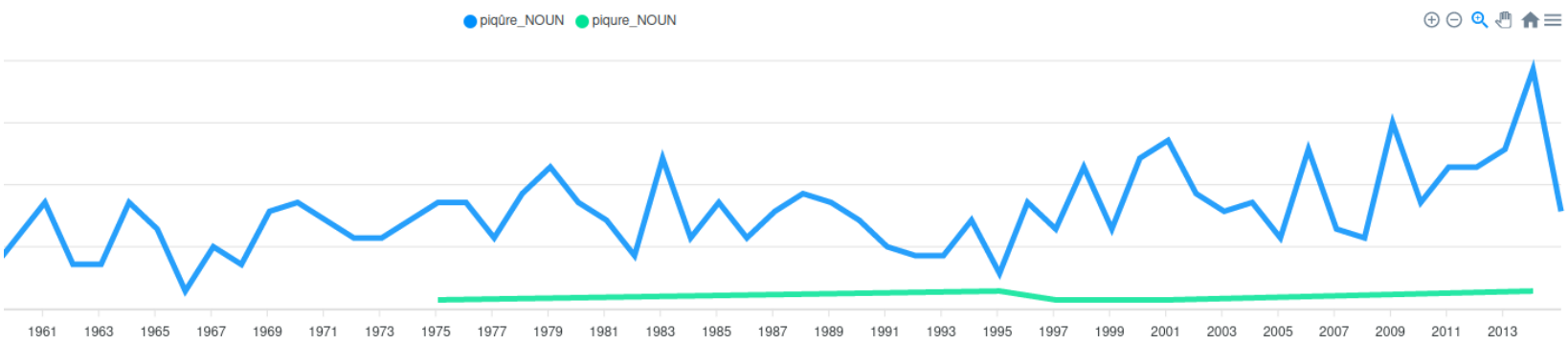
- Possibilité de croiser chronogramme et réalisations



Fonctionnalité chronogramme

et étude de cas

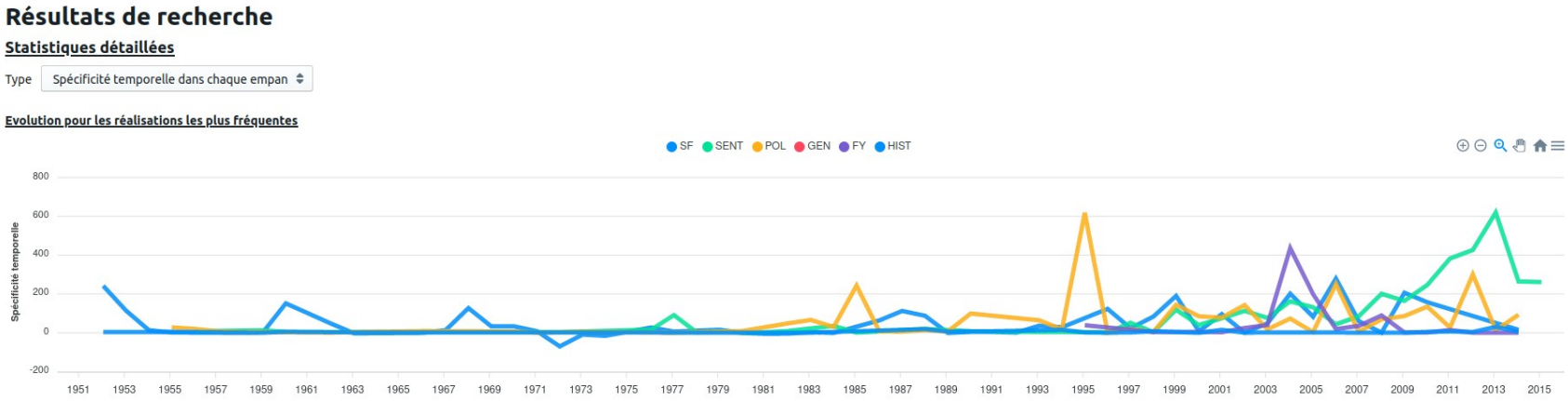
- Possibilité de croiser chronogramme et réalisations : ex. d'évolution orthographique



Fonctionnalité chronogramme

et étude de cas

- Possibilité de croiser chronogramme et sous-corpus



évolution des occurrences de *écran*
(spécificité temporelle)

Fonctionnalité chronogramme
et étude de cas

- Affichage des tendances

Coocurrents en augmentation ↗

- changement_NOUN : Première attestation= 1966 (freq= 1613, coeff= 20.69)
- réchauffement_NOUN : Première attestation= 1989 (freq= 970, coeff= 11.53)
- risque_NOUN : Première attestation= 1955 (freq= 43, coeff= 3.86)
- modèle_NOUN : Première attestation= 1993 (freq= 37, coeff= 2.5)
- archive_NOUN : Première attestation= 1979 (freq= 10, coeff= 2.33)
- aléa_NOUN : Première attestation= 1959 (freq= 112, coeff= 1.88)
- problème_NOUN : Première attestation= 1970 (freq= 22, coeff= 1.83)
- donnée_NOUN : Première attestation= 1974 (freq= 15, coeff= 1.8)
- catastrophe_NOUN : Première attestation= 1973 (freq= 45, coeff= 1.62)
- raison_NOUN : Première attestation= 1961 (freq= 31, coeff= 1.55)

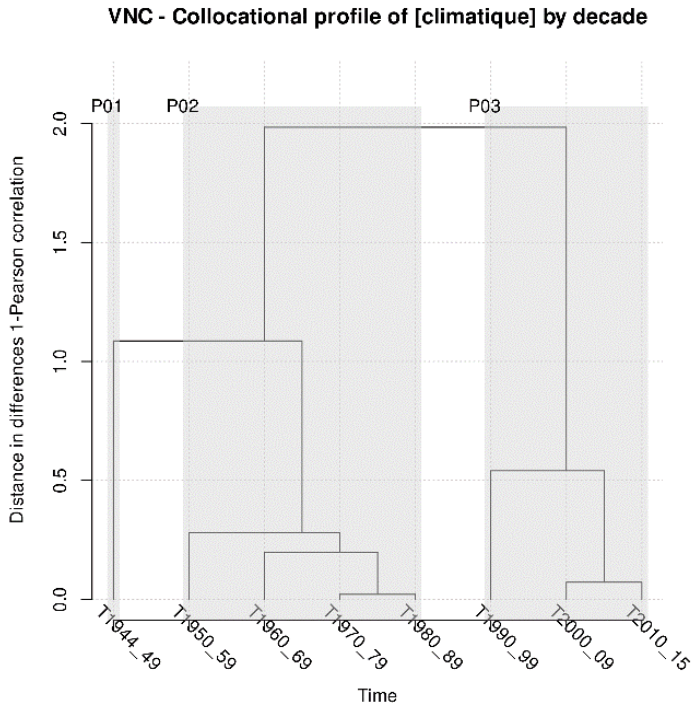
Coocurrents en régression puis en augmentation ↘↗

- conférence_NOUN (freq= 11, coeff= 10)
- désordre_NOUN (freq= 12, coeff= 2)
- politique_NOUN (freq= 43, coeff= 1.82)

Perspectives

Perspectives

- Ajout de mesures de tendance
  - Estimateur Theil Sen (Herman & Kovář, 2013)
  - Correlation des rangs de Kendall (Hilpert & Gries, 2009, 388-390)
    - exemple *climatique* : τ = 0.8571, p-valeur = 0.0044 → tendance à la hausse à travers la série chronologique
- Périodisation automatique
  - Classification ascendante hiérarchique par contiguité - CAHC (Diwersy et al., 2017) ou VNC (Gries & Hilpert, 2008)



Merci de votre attention !

Références

Augustinus, L., Vandeghinste, V., Vanallemeersch, T. (2016). Poly-GrETEL: Cross-Lingual Example-based Querying of Syntactic Constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3549–3554 , Portorož, Slovenia. ELRA.

Candito M., Nivre J., Denis P. and Henestroza Anguiano E. (2010). Benchmarking of Statistical Dependency Parsers for French. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

Coavoux M., Denoyelle, C. Kraif, O., Sorba J. (2022) Phraséologie du roman médiéval en prose. In *Actes du Colloque international DIACHRO X, Le français en diachronie*, Paris, Sorbonne Université, 17-19 janvier 2022.

Diwersy, S., Falaise A., Lay, M.-H., Souvay, G. (2017) Ressources et méthodes pour l'analyse diachronique. *Langages 2017/2* (N° 206), 21-44.

Evert, S. (2007). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, pp. 1212–1248.

Gabrielatos, C., McEnery, T., Diggle, P. J. & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 17(2). 151–175. https://doi.org/10.1075/ijcl.17.2.01gab.

Références

Grobol, L. and Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles,* pages 106–114. Lille, France, June.

Grobol, L., Regnault, M., Oretiz Suarez P., Sagot, B., Romary, L., Crabbé, B. (2022). BERTrade: Using Contextual Embeddings to Parse Old French. LREC 2022, Marseille : 1104-1113.

Gries S. Th. and Hilpert M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. Corpora, 3, 59–81.

Guillaume B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. *EACL (System Demonstrations) 2021*: 168-175.

Herman O. and Kovář V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, 79–85.

Kilgariff A., Rychly P., Smrz P., Tugwell D. (2004). The Sketch Engine". *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116. Lorient, France. Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publisher.

Salem A. (1988). Approches du temps lexical. *Mots*, 17, 105–143. https://doi.org/10.3406/mots.1988.1401

Trevisani, M. & Tuzzi, A. (2016). Analisi di dati testuali cronologici in corpora diacronici: effetti della normalizzazione sul curve clustering. In *JADT 2016: 13ème Journées Internationales d'Analyse Statistique des Données Textuelles*. Nice, France. http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/82630/82630.pdf.