CORPUS

Corpus

27 | 2025
Motifs, textométrie et stylistique des textes littéraires

Identification des motifs textuels. Entre statistique et *deep learning*

Text pattern identification. Between statistics and deep learning

Dominique Longrée and Laurent Vanni

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/10326
DOI: 10.4000/13woj
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Dominique Longrée and Laurent Vanni, "Identification des motifs textuels. Entre statistique et *deep learning*", *Corpus* [Online], 27 | 2025, Online since 13 May 2025, connection on 09 June 2025. URL: http://journals.openedition.org/corpus/10326 ; DOI: https://doi.org/10.4000/13woj

Identification des motifs textuels. Entre statistique et *deep learning*

Text pattern identification. Between statistics and deep learning

Dominique Longrée and Laurent Vanni

1. Introduction

1 Dès son apparition, la notion de “motif” a été développée dans l’optique de fournir un nouvel observable linguistique susceptible de faire l’objet d’identification et d’extraction automatiques. Tel était bien l’objectif du travail pionnier de Ganascia (2001) visant au repérage des “motif syntaxiques”. Ce souci d’une formalisation d’un objet de recherche détectable automatiquement et susceptible de se prêter à diverses analyses statistiques était également bien présent dans la définition que Longrée, Luong et Mellet (2008) ont proposée pour le “motif” : celui-ci est décrit comme un sous-ensemble ordonné de l’ensemble que constitue un texte, sous-ensemble formé par l’association récurrente de *n* éléments de cet ensemble muni de sa structure linéaire. Cette définition permet de voir dans le “motif” non seulement un objet topologique, au sens mathématique du terme (Mellet, Luong, Longrée Barthélémy, 2009), mais aussi (Mellet & Longrée, 2012), « comme un moyen d’intégrer la multidimensionalité (ou le caractère multi-niveau) de certaines formes récurrentes », fournissant « une base solide à des traitements automatiques (Evert 2008), notamment à la détection non supervisée des motifs dans un texte » (Longrée & Mellet 2013; 2018).

2 A ses origines, le “motif” apparaît donc comme un cadre conceptuel susceptible de recouvrir des patrons linguistiques récurrents de natures très variées, d’où la possibilité de voir dans le “motif” une notion permettant de subsumer divers types d’unités phraséologiques (Longrée & Mellet 2013). A ce titre, tous les patrons dont M. Bendinelli (2017) a dressé la liste dans un article publié dans cette même revue – à savoir cadres collocationnels, colligations, collocations, constructions, constructions préformées, *formulaic expressions, lexical bunddles, multi-word expressions,* n-grammes, *phrase frames, patterns*, poly-cooccurrents, quasi-segments répétés, routines discursives,

segments répétés, séquences textuelles – peuvent être décrits comme des formes particulières de “motifs”, qu’il s’agisse de patrons identifiés par la phraséologie ou par la textométrie.

3 En parallèle avec cette définition large du “motif” comme cadre collocationnel s’est développé un emploi plus restrictif du terme “motifs” pour désigner des associations récurrentes multi-dimensionnelles pourvues d’une fonction textuelle (Legallois,2006; Mellet & Longrée 2012; Longrée & Mellet 2018), le plus souvent structurante ou caractérisante. Une expérience de psychologie cognitive (Lavigne, Longrée & Mellet, 2016, 2019) a montré que la reconnaissance de ces “motifs textuels” reposait sur des associations sémantiques enregistrées dans la mémoire des lecteurs, la récurrence de ces “motifs textuels” étant essentielle pour leur stabilisation, leur mémorisation et leur identification. Ces processus cognitifs sont d’autant plus complexes que ces “motifs textuels” peuvent varier dans leur forme sans que cela ne nuise à leur reconnaissance : contrairement aux segments répétés (Salem, 1986), aux séquences de temps verbaux (Longree & Luong, 2003) ou aux motifs syntaxiques (Mellet & Longrée, 2009), les “motifs textuels” multidimensionnels, constitués d’items pouvant être de plusieurs natures (formes, lemmes, catégories grammaticales, etc.), autorisent de multiples variations en leur sein : variations lexicales ou morphologiques, permutations, présence ou absence d’un élément, espaces entre les éléments du “motif”...

4 La détection automatique de motifs que les inventeurs de la notion appelaient de leur vœux s’est avérée toutefois très rapidement être beaucoup moins aisée qu’il ne pouvait paraître de prime abord. Chacune des variations que l’on vient d’évoquer constitue un défi de taille pour une identification automatisée reposant sur des critères d’ordre statistique. Une autre difficulté majeure réside dans la récurrence qui représente un paramètre constitutif de la notion même de “motif” : le “motif” doit être suffisamment fréquent pour être reconnaissable et mémorisable, on l’a dit, mais les patrons linguistiques les plus récurrents, à savoir les simples patrons syntaxiques qui sous-tendent les structures de base de la langue, ne constituent évidemment pas des “motifs”, alors que par ailleurs certains patrons syntaxiques particuliers (Mellet & Longrée, 2009) peuvent eux caractériser des styles d’auteurs. De même, une forte récurrence d’un patron multi-dimensionnel correspond généralement à des unités phraséologiques, mais ces mêmes unités peuvent soit correspondre à de simples patrons figés en langue, soit constituer de véritables “motifs textuels”. Si l’observable recherché n’est pas seulement un simple patron récurrent, mais bien un véritable “motif” doté d’une fonction textuelle, le critère de l’intensité de la fréquence n’est pas toujours pertinent, sauf si le “motif” est caractérisant et qu’un calcul de spécificité permet de l’identifier comme tel. Par ailleurs, peu de critères, -à l’exception sans doute de la position dans la phrase-, permettent à une détection automatique d’isoler des “motifs” présentant une fonction textuelle structurante.

5 Le présent article aura pour objectif de faire le point, -sans prétendre à l’exhaustivité-, sur les différentes techniques de détection des motifs existantes, en partant essentiellement des possibilités offertes par le Logiciel Hyperbase Web Edition (Vanni 2024) pour le confronter à d’autres méthodes et logiciels. Il s’agira de lister à chaque fois les paramètres linguistiques que la détection automatique peut ou non prendre en compte : intensité de la récurrence, longueur et multi-dimensionnalité des “motifs”; variations morphosyntaxiques; contiguïté ou non des éléments qui le composent. On pourra ainsi évaluer, pour chaque méthode, ses performances, ses avantages et ses

inconvénients, présenter quelques nouvelles pistes possibles reposant sur des méthodes statistiques et enfin explorer dans quelle mesure le recours à l'intelligence artificielle peut s'avérer utile, voire indispensable en la matière.

6 Notre enquête partira d'un exemple précis, à savoir le *De Bello Gallico* de César, dans le format numérique qui est celui des fichiers du Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) de l'Université de Liège. Ce choix s'explique par plusieurs raisons. En tout premier lieu, la tradition philologique s'est intéressée de longue date aux structures récurrentes que recèle le texte césarien; nous avons voulu préciser dans quelle mesure la liste des "motifs" détectés automatiquement correspondait ou non à celle de ces structures bien connues des latinistes. Ensuite, les fichiers du LASLA contiennent des textes lemmatisés et annotés morphosyntaxiquement : à chaque forme du texte correspond un lemme, pourvu d'un indice de désambiguïsation, ainsi qu'une analyse morphologique complète, complétée pour les verbes d'une indication du type de la proposition dont ils sont le prédicat; la granularité fine des fichiers du LASLA permet donc une excellente prise en compte de la multi-dimensionalité des "motifs". Enfin, le logiciel Hyperbase Web Edition a été adapté de longue date pour pouvoir traiter l'ensemble de ces informations. En outre, les fichiers du LASLA sont facilement adaptables à d'autres logiciels de détection, comme nous le verrons dans la suite de cet article.

7 Dans cette enquête, on classera les méthodes de détection selon trois critères essentiels. Le premier repose sur le caractère contrastif ou non contrastif de l'approche : les approches non-contrastives se contentent de repérer des patrons récurrents et donc de détecter des répétitions qui peuvent correspondre autant à de simples phrasèmes qu'à de réels "motifs textuels"; les approches contrastives, elles, s'appuient sur des spécificités, qui permettent de mettre en évidence des patrons linguistiques qui, du fait de leur spécificité, auront *a minima* une fonction caractérisante et dès lors pourront être considérés comme de vrais "motifs textuels" propres à un texte, un auteur, un genre, un mode d'expression, etc. Le deuxième critère correspond à la séquentialité ou la non séquentialité de la recherche : une détection séquentielle s'appuiera sur la linéarité du texte, tandis qu'une détection non-séquentielle reposera essentiellement sur sa réticularité, et donc sur des collocations. Le troisième et dernier, mais qui n'est pas le moins important, est celui connu de l'opposition entre *corpus-based* et *corpus-driven*. C'est autour de cette opposition que se structurera d'abord notre exposé : il s'agira de préciser dans quelle mesure cette opposition se superpose ou non à une distinction entre méthode supervisée ou non-supervisée. Dans chacun des deux cas, on distinguera ensuite approche linéaire et approche non linéaire et enfin, pour chacune de ces approches, on passera en revue les méthodes faisant appel à la seule fréquence de celles calculant des spécificités.

2. Les approches Corpus-Based

8 Les approches *corpus-based* ont toutes en commun la particularité de partir d'un "pivot" que celui-ci soit un token simple (forme, lemme ou code grammatical) ou d'un ensemble de tokens (pouvant combiner formes, lemmes et codes). Elles se distribuent essentiellement en deux types de méthodes : soit la recherche est séquentielle et s'appuie sur la linéarité du texte, soit l'approche est associative, s'appuyant sur les réseaux collocationnels des textes et donc sur leur réticularité.

2.1. Séquentielles

9 Quand elles sont *corpus-based*, les approches séquentielles basées sur la linéarité du texte sont essentiellement qualitatives : elles partent d'un pivot que le chercheur a déjà reconnu comme susceptible d'être inclus dans un "motif textuel" ou d'en constituer le schème sous-jacent. Il peut s'agir de "motifs" déjà reconnus de longue date par la tradition philologique : il en va ainsi des "motifs textuels" caractéristiques des textes historiques latins, et en particulier du style césarien, que J.P. Chausserie-Laprée (1969) a mis en évidence et qui ont donné lieu à plusieurs études textométriques (Longrée & Mellet 2012 ; 2018).

2.1.1. Approche qualitative : concordancier avec tri gauche et droite

10 Un de ces "motifs" particulièrement caractéristique de l'œuvre césarienne a pour représentant emblématique la proposition participiale Ablatif Absolu *his rebus cognitis* "ces affaires ayant été connues / ayant eu connaissance de ces affaires/événements", dont diverses variantes ont été listées dans les études précitées (par ex. *his rebus gestis*, "ces affaires/événements ayant eu lieu / sur ces entrefaites"). Comme le montre la figure 1, une recherche avec Hyperbase Web Edition permet de mieux cerner la liste des variantes de ce "motif" dont la fonction (structurante) est d'assurer la transition entre deux épisodes narratifs.

Figure 1. Concordance de *his rebus* avec tri à droite chez César

partie gauche	pivot	partie droite	référence
sua Romanis satisfacere seu uiuum tradere uelint . mittuntur de	his rebus	ad Caesarem legati . iubet arma tradi principes produci .	cesbgall_7,89,2_49353
milia passuum CCXL in latitudinem CLXXX patebant .	his rebus	adducti et auctoritate Orgetorigis permoti constituerunt ea quae ad proficiscendum	cesbgall_1,3,1_348
et quod fere libenter homines id quod uolunt credunt .	his rebus	adducti non Viridouicem reliquos que duces ex concilio dimittunt quam	cesbgall_3,18,7_16050
quae ad eruptionem praeparauerant proferunt priores fossas explent diutius in	his rebus	administrandis morati suos discessisse cognouerunt quam munitionibus adpropinquarent . ita	cesbgall_7,82,4_48588
belli gloriam libertatem que quam a maioribus acceperint recuperare .	his rebus	agitatis profitentur Carnutes se nullum periculum communis salutis causa recusare	cesbgall_7,2,1_36984
regionibus ueniant quas que ibi res cognouerint pronuntiare cogat .	his rebus	atque auditionibus permoti de summis saepe rebus consilia ineunt quorum	cesbgall_4,5,3_18125
remiges ex prouincia institui nautas gubernatores que comparari iubet .	his rebus	celeriter administratis ipse cum primum per anni tempus potuit ad	cesbgall_3,9,2_14563
in Gallia numerus conquiri et ad se mitti iubet .	his rebus	celeriter id quod Auarici deperierat expletur . interim Teutomatus Ollouiconis	cesbgall_7,31,4_41051
sui liberandi facultas daretur si Romanos castris expulissent demonstrauerunt .	his rebus	celeriter magna multitudine peditatus equitatus que coacta ad castra uenerunt	cesbgall_4,34,6_22053
uti ea quae apud eos gerantur cognoscant se que de	his rebus	certiorem faciant . hi constanter omnes nuntiauerunt manus cogi exercitum	cesbgall_2,2,3_9091
naues subduci et cum castris una munitione coniungi . in	his rebus	circiter dies X consumit ne nocturnis quidem temporibus ad laborem	cesbgall_5,11,6_24122
eiusdem que generis sub aqua defixae sudes flumine tegebantur .	his rebus	cognitis a perfugis captiuis que Caesar praemisso equitatu confestim legiones	cesbgall_5,18,4_25062
omnia in potestate eius essent omnes cruciatus essent perferendi .	his rebus	cognitis Caesar Gallorum animos uerbis confirmauit pollicitus que est sibi	cesbgall_1,33,1_5134
Romani contegit nec iam arma nostris nec uires suppetunt .	his rebus	cognitis Caesar Labienum cum cohortibus sex subsidio laborantibus mittit imperat	cesbgall_7,86,1_48965

11 L'utilisation d'une recherche en concordancier sur la base du pivot *his rebus*, suivie d'un tri sur le contexte droit, permet de capter non seulement *his rebus cognitis*, mais aussi des formes telles que *his rebus agitatis*, "ces affaires ayant été traitées", ayant parfois échappé à la recherche philologique, ou encore une variante du type *his rebus adducti*

“poussés par ces évènements” dont la nature syntaxique diffère (un participe accordé à la place d’un Ablatif absolu), mais qui peut être considérée comme une instance du même “motif”, non seulement en raison de l’inclusion du pivot *his rebus*, mais surtout d’une identité de fonction syntaxique et de fonction structurante. Un tri à droite montre en outre que toutes ces structures apparaissent systématiquement en tête de phrase.

12 Pour optimiser la recherche supervisée d’un motif, on peut également utiliser comme pivot le schème abstrait sous-tendant le motif. Ainsi, toujours pour le même “motif” du type *his rebus cognitis*, on peut faire appel à la requête “Fin de Phrase - Pronom démonstratif à l’ablatif pluriel - substantif à l’ablatif pluriel - participe à l’ablatif pluriel”. Comme illustré par la figure 2, cette recherche, accompagnée d’un tri alphabétique sur le pivot, permet non seulement de relever toutes les instanciations du schème abstrait, mais aussi d’évaluer la variété de ces instanciations. Le choix du schème abstrait a toutefois pour conséquence d’exclure des variantes du “motif” qui s’en écarteraient quelque peu : il peut s’agir de variantes liées à l’insertion d’un élément dans la structure (ex. *His rebus celeriter administratis*, “ces affaires ayant été rapidement réglées”, Caes., *B.G.*, 3, 9), mais ces variantes peuvent être aisément repérées par l’insertion d’un élément vide (dans le recherche avec Hyperbase, le symbole ???) dans le schème recherché; plus gênantes sont les variations d’ordre lexical ou morphologique du type *Quibus rebus cognitis* où le démonstratif est remplacé par un relatif dit de liaison ayant à la fois la valeur d’un coordonnant et d’un démonstratif, mais étant bien codé comme relatif et ne pouvant donc pas être récupéré avec la même requête.

Figure 2. Concordance de ". DemPro:Abl:Plur Subs:Abl:Plur Part:Abl:Plur:" chez César

partie gauche	pivot	partie droite	texte	référence
properarct ad se cum exercitu uenire omnia quc post haberet	. his litteris acceptis	quos aduocauerat dimittit ipse iter in Macedoniam pararc incipit paucis	cacsar	cacsbc3_3_33_1_4798
conari et ea loca finitimae prouinciae adiungere sibi persuasum habebant	. his nuntiis acceptis	Galba cum neque opus hibernorum munitiones quc plene essent perfectae	caesar	caesbg3_3_2_5_300
uctercm belli gloriam libertatem quc quam a maioribus acceperint recuperare	. his rebus agitatis	profitentur Carnutes se nullum periculum communis salutis causa recusare principes	caesar	caesbg7_7_1_8_192
oppida omnia in potestate eius essent omnes cruciatus essent perferendi	. his rebus cognitis	Caesar Gallorum animos uerbis confirmauit pollicitus quc est sibi eam	caesar	caesbg1_1_32_5_5132
eum locum coniecisse quo propter paludes exercitui aditus non esset	. his rebus cognitis	exploratores centuriones quc praemittit qui locum castris idoneum deligant .	caesar	caesbg2_2_16_5_2182
addere et se in posterum diem similem ad casum parare	. his rebus cognitis	Caesar summo studio militum ante ortum solis in castra peruenit	caesar	caesbg7_7_41_4_5844
occultauerant Romani contegit nec iam arma nostris nec uires suppetunt	. his rebus cognitis	Caesar Labienum cum cohortibus sex subsidio laborantibus mittit imperat si	caesar	caesbg7_7_85_6_12180
incolumes in castra conferunt . milites ad unum omnes interficiuntur	. his rebus cognitis	Marcius Rufus quaestor in castris relictus a Curione cohortatur suos	caesar	caesbc2_2_42_5_6024
possent itaque ex eo concursu nauium magnum esse incommodum acceptum	. his rebus cognitis	Caesar legiones equitatum que reuocari atque in itinere resistere iubet	caesar	caesbg5_5_10_3_1572
munita eiusdem que generis sub aqua defixae sudes flumine tegebantur	. his rebus cognitis	a perfugis captiuis que Caesar praemisso equitatu confestim legiones subsequi	caesar	caesbg5_5_18_3_2607
Italia adduxerat in Heluios qui fines Aruernorum contingunt conuenire iubet	. his rebus comparatis	represso iam Lucterio et remoto quod intrare intra praesidia periculosum	caesar	caesbg7_7_7_5_903
in terra patitur scd sublatis ancoris excedere eo loco cogit	. his rebus confectis	Caesar ut reliquum tempus a labore intermitteretur militcs in proxima	caesar	caesbc1_1_31_3_4645
enim erat annus quo per leges ei consulem fieri liceret	. his rebus confectis	cum fides tota Italia esset angustior neque creditae pecuniae soluerentur	caesar	caesbc3_3_1_1_23
ex reliquis captiuis toti exercitui capita singula praedae nomine distribuit	. his rebus confectis	in Haeduos proficiscitur ciuitatem recipit . eo legati ab Aruernis	caesar	caesbg7_7_89_5_12620

13 Une telle recherche permet de se faire une idée de la variété des différentes instanciations d’un schème soutenant un motif, mais le format “concordancier” oblige le chercheur à un travail qui peut devenir très vite fastidieux s’il veut récolter des données quantitatives sur cette base. Hyperbase lui offre en la matière une autre manière de procéder.

2.1.2. Approche quantitative : distribution des occurrences les plus fréquentes d’un patron de codes choisi

14 Le logiciel Hyperbase Web Edition permet de procéder à une analyse quantitative des variantes les plus fréquentes d’un schème multidimensionnel au sein d’un corpus.

L'approche peut être purement quantitative, sans dimension contrastive. Ainsi, avec la fonction distribution, dans un corpus réduit aux deux œuvres conservées de César, le *De bello Gallico* et le *De bello ciuili*, les résultats de la même requête "Fin de Phrase - Pronom démonstratif à l'ablatif pluriel - substantif à l'ablatif pluriel - participe à l'ablatif pluriel" montrent (figure 3) que les deux œuvres contiennent 27 occurrences de la structure et la fonction "afficher les plus fréquents" permet de constater que 2 variantes, *-his rebus gestis* et *his rebus constitutis*, "ces affaires ayant été établies"-, représentent un tiers des occurrences (9/27), toutes les autres variantes ne présentant qu'une seule occurrence dans le corpus.

Figure 3. Distribution de ". DemPro:Abl:Plur Subs:Abl:Plur Part:Abl:Plur:" chez César

Mots	Total	bellumcivile	bellumgallicum
. DemPro:Abl:Plur Subs:Abl:Plur Part:Abl:Plur	27	8	19
._his_rebus_gestis	6	2	4
._his_rebus_constitutis	3	1	2
._his_rebus_agitatis	1	0	1
._his_rebus_conpletis	1	1	0
._his_rebus_expositis	1	0	1

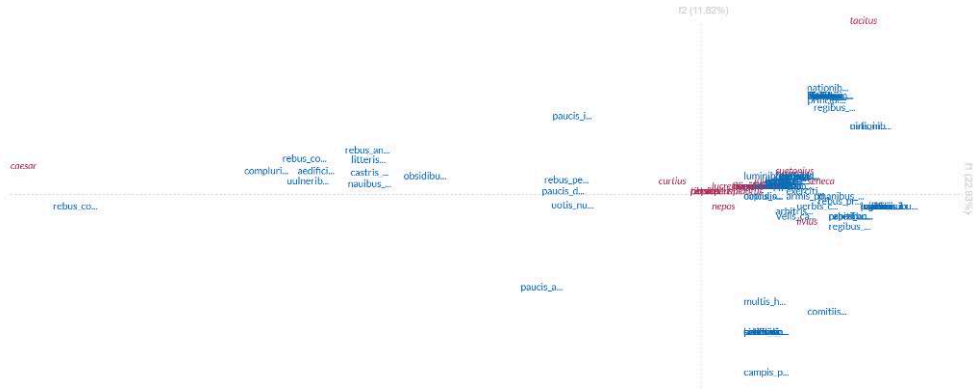
15 La réduction de la requête à "Subs:Abl:Plur Part:Abl:Plur:" (figure 4) montrent que, parmi les 9 variantes de cette structure qui sont attestées plus d'une fois, 3 incluent *rebus* et relèvent du "motif" : *rebus gestis, rebus constitutis, rebus animaduersis* (l'augmentation du nombre d'occurrences se justifie par le remplacement de *his* par *quibus*). Cette recherche permet également de relever un autre "motif" dont le pivot est le participe *acceptis* : *uulneribus / obsidibus / litteris acceptis*, "des blessures / des otages / une lettre ayant été reçu.e.s".

Figure 4. Distribution de "Subs:Abl:Plur Part:Abl:Plur:" chez César

Mots	Total	bellumcivile	bellumgallicum
Subs:Abl:Plur Part:Abl:Plur	206	67	139
rebus_gestis	8	3	5
uulneribus_acceptis	6	0	6
obsidibus_acceptis	3	0	3
rebus_constitutis	3	1	2
armis_traditis	2	0	2
litteris_acceptis	2	2	0
nauibus_iunctis	2	1	1
proeliis_factis	2	1	1
rebus_animaduersis	2	2	0
Atrebatibus_superatis	1	0	1
Germanis_uictis	1	0	1
Trinouantibus_defensis	1	0	1
aedificiis_occupatis	1	0	1
agris_uastatis	1	0	1

16 Cette approche quantitative peut également prendre une dimension contrastive s'appuyant sur un calcul de spécificité. La recherche du même schème dans la base "Latin" d'Hyperbase Web Edition[1] permet de mettre en évidence les variantes de la structure spécifiques de chaque auteur contenu dans la base. Une AFC basée sur la spécificité (figure 5) des 100 variantes les plus fréquentes montre que César s'oppose clairement à gauche à l'ensemble des autres auteurs de la base et que cette opposition repose essentiellement sur la spécificité des variantes des deux motifs relevés précédemment *rebus gestis / constitutis / animaduersis* et *uulneribus / obsidibus / litteris acceptis.* La liste de variantes spécifiques contient également des expressions telles *compluribus interfectis* "un assez grand nombre ayant été tué" ou *aedificiis incensis* "les bâtiments ayant été incendiés": celles-ci n'entrent pas dans des séries, n'apparaissent pas en tête de phrase et ne présentent pas de fonction textuelle structurante comme les deux motifs précités, mais peuvent être considérée également comme des "motifs" caractérisant le style de César, en l'opposant à celui des autres auteurs de la base.

Figure 5. AFC des 100 variantes les plus fréquentes de “Subs:Abl:Plur Part:Abl:Plur:" (Base “Latin”)



17 Une variante de ce type de recherche est proposée par MotiveR qui permet de rechercher des séquences spécifiques d’un auteur, séquences combinant parties du discours et formes (par ex. “si ADJ et si ADJ que”, D. Legallois & A. Silvestre de Sacy, (2021), A. Silvestre de Sacy & D. Legallois, ce numéro).

18 Une limitation majeure de ce type de recherche est qu’il est difficile d’y intégrer des positions “vides” du type ??? comme nous l’avons fait en concordancier avec Hyperbase Web Edition. Pour contourner ce problème, Hyperbase Web Edition, comme d’autres logiciels, permet de faire appel à des recherches non-séquentielles.

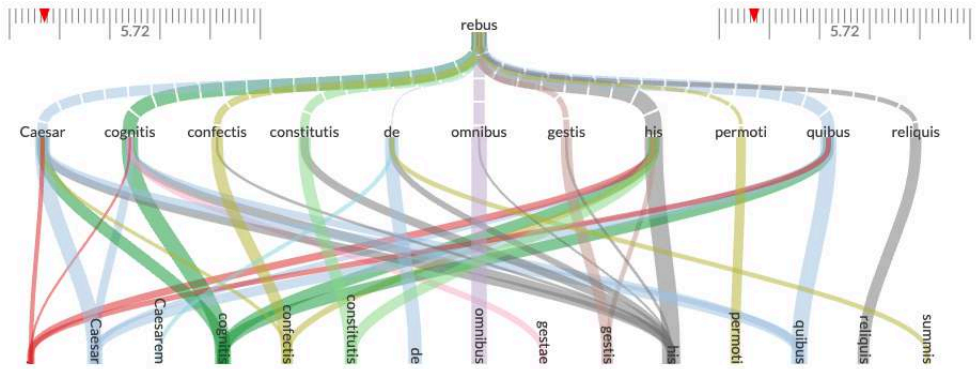
2.2. Non séquentielles

19 Les approches non séquentielles se fondent sur l’existence de réseaux associatifs au sein du texte, et donc en quelque sorte, sur sa réticularité. Celle-ci s’exprime à plusieurs niveaux : associations lexicales, sémantiques, syntaxiques. Leur détection peut reposer soit uniquement sur un calcul d’associations spécifiques, soit sur une combinaison de parsing syntaxique et de calcul de spécificités

2.2.1. Cooccurrences spécifiques

20 La fonction “Thème” du logiciel Hyperbase Web Edition permet de déterminer les cooccurrents spécifiques d’un pivot qui peut être soit une forme, soit un lemme, soit un code grammatical. L’empan textuel utilisé pour calculer les écarts réduits peut varier : empan textuel calculé en nombre de mots, à droite et/ou à gauche du pivot; phrase; paragraphe. Le choix de l’empan dépend des objectifs que le philologue assigne à la recherche. La figure 6 montre les résultats obtenus pour le pivot *rebus* dans un empan textuel de 10 mots à gauche et 10 mots à droite, pouvant capturer des associations tant à courte qu’à plus longue distance.

Figure 6. Cooccurrents spécifiques de *rebus* chez César



21 On ne s'étonne pas de trouver parmi les cooccurrents spécifiques de *rebus* les formes *his* (premier cooccurrents), *cognitis, confectis, constitutis, gestis* mais aussi *quibus*, que la recherche en concordancier avait déjà révélées, toutes formes relevant du même "motif". La présence de *permoti* renvoie pour sa part à la variante du "motif" avec un participe épithète détachée *his rebus permoti* "ébranlés par ces faits". Le concurrent *omnibus* permet de découvrir des extensions du "motif", du type *quibus omnibus rebus permoti* (3 occurrences, *B.G.*, 2, 24; 2, 37; 5, 51). La présence de la préposition *de* s'explique par un autre "motif" césarien, *de rebus certiorem facere* "informer quelqu'un des faits", cette expression n'étant attestée dans la base "Latin" que chez César (*B.G.*, 3, 9; 4, 5; 5, 37). Dans chacun de ces cas, le "motif textuel" correspond à un syntagme formé d'un verbe, éventuellement de son sujet (dans le cas de l'Ablatif absolu) et de ses compléments. Comme le laissait pressentir l'œuvre pionnière de J.G. Ganascia (2001), la dimension syntaxique est donc essentielle dans la définition et le repérage des motifs, ce qui explique la qualité des résultats obtenus par les méthodes qui combinent cette dimension avec des spécificités lexicales.

2.2.2. Arbres lexico-syntaxiques spécifiques - ALRs

22 Ainsi, la méthode que le Lidilem (Université Grenoble Alpes) a implémentée dans le Lexicoscope[2], prend, elle, en compte cette dimension syntaxique. Visant à extraire des structures hiérarchiques et non plus séquentielles, elle s'appuie sur des corpus syntaxiquement arborés pour en extraire des "Arbres lexico-syntaxiques récurrents » (ALRs), reposant sur la combinatoire d'unités lexicales et de structures syntaxiques (Kraif 2016; 2019). L'extraction des arbres repose sur le choix d'un pivot lexical, -nominal ou verbal, simple ou complexe-, et sur une mesure statistique de la saillance de l'arbre (Bertels & Speelman 2013). Grâce à ce calcul de spécificité, les arbres extraits sont interprétables comme des "motifs" que l'on peut considérer comme caractéristiques d'un texte, d'un auteur, d'un genre ou d'un sous-genre littéraire (Novakova & Siepmann, 2020, Kraif, Novakova, Sorba, ce numéro). Un des avantages majeurs de la méthode est de s'affranchir de la séquentialité et donc de dépasser la difficulté que représente les permutations ou les places vides que peuvent autoriser certains "motifs". L'intégration d'une dimension syntaxique aux "motifs" permet de s'affranchir des variations lexicales ou morphologiques.

23 O. Kraif a très aimablement accepté d'appliquer aux fichiers du LASLA une analyse par les "arbres lexico-syntaxiques" sur base du pivot *res*, lemme de la forme *rebus*. La

figure 7 fournit un exemple d'arbre lexico-syntaxique obtenu à partir des fichiers repris dans la base "Historia" d'Hyperbase Web Edition.

Figure 7. Exemple de ALR du lemme *res* dans la base "Historia"

first colloc	gero_VERB
size	7
query	<c=ADV,#1>&&<c=ADV,#2>&&<c=NOUN,#3>&&<c=NOUN,#4>&&<c=NOUN,l=res,#5>&&<c=VERB,#6>&&<c=VERB,l=gero,#7>::(advmod,7,1) (advmod,7,2) (obl,7,3) (obl,7,4) (nsubj:pass,7,5) (advcl,7,6)
realization	(.*)_ADV (.*)_ADV (.*)_NOUN (.*)_NOUN gero_VERB res_NOUN (.*)_VERB
example	[[[primo#2]]] [[[utrimque#1]]] aequis [[[uiribus#4]]] eodem [[[ardore#3]]] animorum [[[gerebatur#7]]] [[[res#5]]] deinde ab laeuo cornu hastati Romani non [[[ferentes#6]]] impressionem Latinorum se ad principes recepere.
frequency	16
dispersion	1

24 La structure mise en évidence ici ne peut être aisément repérée par les méthodes décrites précédemment, à la fois en raison de la distance entre les éléments composant l'ALR et du fait de la taille de la structure composée de 7 éléments. A cet égard, la méthode présente un avantage indéniable par rapport aux méthodes *corpus-based*. L'outil suppose néanmoins de disposer d'arbres syntaxiques suffisamment fiables pour pouvoir en extraire les informations recherchées : dans ce cas, c'est le modèle Stanza qui a été utilisé. En ce qui concerne le latin, on doit malheureusement constater que les parsers, faute de fichiers d'entraînement suffisants, ne peuvent fournir des résultats totalement satisfaisants.

25 Les méthodes de recherche *corpus-based* décrites ici présentent de larges possibilités de détection des motifs pour autant toutefois que le chercheur ait fourni des données de départ pertinentes. Celles-ci ne garantissent pas que des "motifs" caractéristiques d'un genre, d'un auteur ou d'une œuvre puissent échapper aux investigations. Diverses méthodes *corpus-driven* permettent de résoudre partiellement cette difficulté. Parmi les outils offrant des possibilités d'approches *corpus-driven,* se retrouvent certains des outils que nous venons de décrire, selon que le chercheur privilégie ou non une démarche partant d'un pivot donné ou une recherche sans pivot initial. Ainsi, les ALRs peuvent être construits automatiquement sans supervision à partir des occurrences ou cooccurrences spécifiques d'un texte, sans que le chercheur ne doive fixer un pivot en particulier. Il en va de même avec les motifs spécifiques émergents de MotiveR (D. Legallois & A. Silvestre de Sacy, (2021), A. Silvestre de Sacy & D. Legallois, ce numéro).

26 Plusieurs approches sont par ailleurs conçues spécifiquement, dans une perspective *corpus-driven* pour limiter la supervision du chercheur.

3. Les approches Corpus-Driven

27 Parmi les approches *corpus-driven*, nous distinguerons celles qui s'appuient sur l'axe syntagmatique du texte et celles qui se fondent sur des associations en dehors de la linéarité textuelle.

3.1. Séquentielles

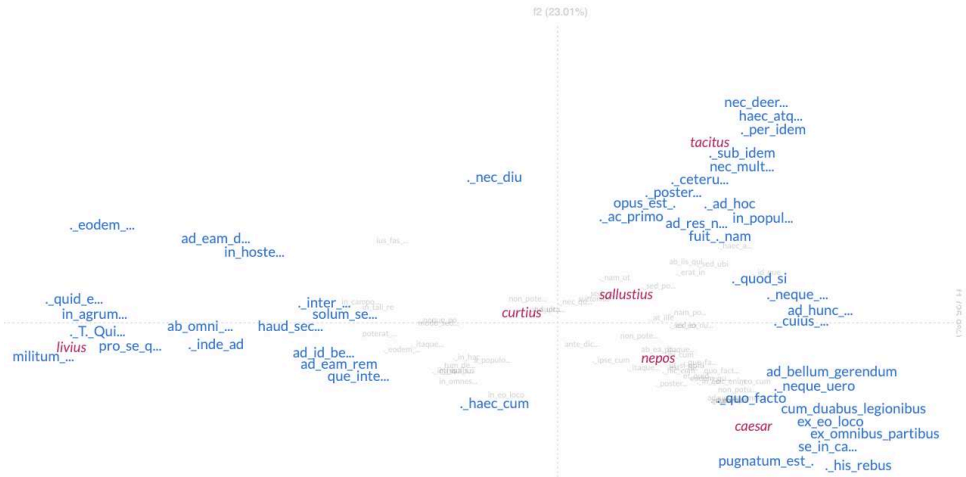
28 L'approche séquentielle la plus basique consiste en une recherche de segments répétés. Hyperbase Web Edition offre à cet égard diverses possibilités.

3.1.1. Segments répétés monodimensionnels ou multidimensionnels

29 La recherche de segments répétés peut s'envisager sous une forme "mono-channel", en ne s'appuyant que sur un seul niveau du texte, par exemple les formes ou les lemmes, ou sous une forme "multi-channel" en autorisant la combinaison multidimensionnelle de divers niveaux, formes, lemmes et codes grammaticaux. La requête peut également contraindre une position par un code grammatical en fonction des objectifs de la recherche. S'il s'agit bien d'une recherche *corpus-driven*, celle-ci n'est donc pas nécessairement totalement non-supervisée.

30 Un exemple d'une recherche de ce type est fourni par la figure 8. Ont été relevés dans la base "Historia" d'Hyperbase Web Edition[3] les 100 premiers segments répétés de longueur 3 sur base des formes et sans aucune contrainte. Un calcul de spécificité et donc une analyse contrastive permet de mettre en évidence les segments caractéristiques de chaque auteur.

Figure 8. AFC des segments répétés de longueur trois dans la base "Historia"



31 Cette AFC met en évidence, dans le cadran inférieur droit, diverses expressions caractéristiques de César : *ex eo loco* "à partir de ce lieu", *ex omnibus partibus* "de tous les côtés", *ad bellum gerendum* "pour mener la guerre", *cum duabus legionibus* "avec deux légions", *se in castra (recipere)* "se replier dans le camp". Ces expressions peuvent être considérées comme des "motifs" caractérisants du fait de leur spécificité. Les autres segments répétés caractéristiques de César se rencontrent eux en début de phrase, la

ponctuation tenant lieu du premier élément du segment (*his rebus*, *quo facto* “et par ce fait”, *neque uero* “et non pas à vrai dire”), ou en fin de phrase, la ponctuation fonctionnant comme dernier élément du segment (*pugnatum est*, “on combattit”). Parmi ces quatre segments, on reconnaît aisément le “motif” *his rebus cognitis*, mais les 3 autres segments sont également constitutifs de “motifs” structurants, ce qui explique leur position en début ou fin de phrase.

32 Cette recherche “standard” avec Hyperbase Web Edition ne permet pas de prendre en compte la multi-dimensionnalité du motif. Une programmation spécifique a donc été mise au point pour détecter des segments multi-dimensionnels, -formes, lemmes, codes grammaticaux-, de 3 éléments présentant une spécificité élevée. Pour les codes grammaticaux, l’ensemble de l’analyse morphosyntaxique encodée par le LASLA a été conservée. La spécificité de ces segments a été calculée selon deux méthodes : soit en sommant les spécificités propres à chacun des 3 éléments composant le segment, soit en calculant la spécificité de l’ensemble du segment.

Figure 9. 15 premiers segments répétés (3 éléments) multidimensionnels du *De Bello Gallico* sur base de la somme des spécificités des éléments

```
"Subs:2Decl:Acc:Plur: LEM:CAESAR_N LEM:OMNIS" ; 97.36
"Subs:3Decl:Abl:Plur: atque LEM:HIC_1" ; 93.36000000000001
"Subs:2Decl:Nom:Plur: atque LEM:HIC_1" ; 93.36000000000001
"Subs:3Decl:Abl:Plur: LEM:ATQVE_1 LEM:HIC_1" ; 93.07
"Subs:2Decl:Nom:Plur: LEM:ATQVE_1 LEM:HIC_1" ; 93.07
"Subs:5Decl:Abl:Plur: LEM:COGNOSCO LEM:CAESAR_N" ; 92.93
"Subs:2Decl:Abl:Plur: LEM:CAESAR_N LEM:COGNOSCO" ; 92.93
"Verb:1Conj:Abl:Sing:Part:Perf:Pas:CodeSubAD:Feminin: LEM:CAESAR_N LEM:OMNIS" ; 92.72
"Subs:2Decl:Nom:Plur: LEM:AD_2 LEM:CAESAR_N" ; 92.53
"RelPro:Acc:Plur:Masculin: LEM:CAESAR_N LEM:AD_2" ; 92.53
"Subs:3Decl:Abl:Plur: LEM:AD_2 LEM:CAESAR_N" ; 92.53
"Subs:2Decl:Acc:Plur: LEM:AD_2 LEM:CAESAR_N" ; 92.53
"Subs:1Decl:Acc:Plur: LEM:AD_2 LEM:CAESAR_N" ; 92.53
"Verb:3Conj:Abl:Plur:Part:Perf:Pas:CodeSubAD:Commun: LEM:AD_2 LEM:CAESAR_N" ; 92.53
"Subs:2Decl:Abl:Plur: LEM:AD_2 LEM:CAESAR_N" ; 92.53
```

33 La figure 9 présente les résultats de la première des deux méthodes. En 6e ligne, la structure "Subs:5Decl:Abl:Plur: LEM:COGNOSCO LEM:CAESAR_N" correspond effectivement au segment *rebus cognitis Caesar*, inclus dans le “motif” *his/quibus rebus cognitis.* Cette donnée permet de préciser le “motif” : en retournant au texte, on constate que celui-ci peut être étendu à un élément au nominatif venant après *cognitis*, la plupart du temps *Caesar* et fonctionnant comme sujet de la proposition principale. En revanche le segment "Subs:2Decl:Abl:Plur: LEM:CAESAR_N LEM:COGNOSCO" qui suit dans la liste ne correspond qu’à une seule occurrence dans le *De Bello Gallico* qui ne constitue en aucun cas un “motif” textuel. La spécificité du segment repose ici sur la spécificité très élevée du lemme *Caesar* qui se retrouve d’ailleurs 11 fois dans les 15 premiers segments présentés ici. Les 4 autres segments relevés ne constituent pas plus des motifs, mais leur présence dans cette liste repose essentiellement sur les spécificités du coordonnant *atque* et du démonstratif *hic*. La méthode donne donc trop de poids à des spécificités pouvant relever d’un seul des éléments du triplet. Pour éviter cet écueil, une méthode pourrait consister à combiner spécificité et seuil de fréquence minimale du segment. Une autre façon de résoudre ce problème est de prendre en compte non la spécificité de chaque élément, mais la spécificité de l’association des 3 éléments.

Figure 10. 15 premiers segments répétés (3 éléments) multidimensionnels du *De Bello Gallico* sur base de la spécificité du segment (indice de spécificité et probabilité)

```
"LEM:MILLE Subs:4Decl:Gen:Plur: Num:Card:Indecl" ; 10.23, 6.95e-23
"LEM:MILLE LEM:PASSVS_1 Num:Card:Indecl" ; 10.23, 6.95e-23
"LEM:MILLE passuum Num:Card:Indecl" ; 10.01, 4.94e-22
"LEM:QVI_1 LEM:SVM_1 Prep" ; 10.01, 4.89e-22
"Prep: LEM:LOCVS LEM:SVPERVS" ; 9.06, 1.22e-18
"Prep: LEM:IS LEM:LOCVS" ; 8.99, 2.11e-18
"Prep: LEM:IS Subs:2Decl:Abl:Sing" ; 8.46, 1.31e-16
"Subs:2Decl:Acc:Sing: Subs:2Decl:Acc:Sing: LEM:LEGATVS" ; 8.3, 4.35e-16
"Prep: omnibus Subs:3Decl:Abl:Plur" ; 8.24, 7.18e-16
"Prep: LEM:OMNIS Subs:3Decl:Abl:Plur" ; 8.24, 7.18e-16
"ad Caesarem LEM:MITTO" ; 8.19, 1.04e-15
"ad LEM:CAESAR_N LEM:MITTO" ; 8.19, 1.04e-15
"Prep: Caesarem LEM:MITTO" ; 8.19, 1.04e-15
"Prep: LEM:CAESAR_N LEM:MITTO" ; 8.19, 1.04e-15
"LEM:AD 2 Caesarem LEM:MITTO" ; 8.19, 1.04e-15
```

34 Dans ce cas (figure 10), les segments les plus spécifiques correspondent soit à des expressions de type phraséologique telles que *milia passuum X*, “à x milliers de pas”, *legatos ad Caesarem mittere* “envoyer des députés à César” ou *prénom nom legatus* “le légat x x”, soit des structures syntaxiques supposant une contrainte lexicale du type “relatif + verbe être + complément prépositionnel de lieu” ou complément prépositionnel de lieu avec un substantif déterminé par un adjectif (*ex loco superiore* “d’un lieu superieur”, *ex omnibus partibus* “de tous les côtés”) ou un démonstratif (*in eo oppido* “dans cette place forte”). Dans la mesure où ces segments sont spécifiques au *De Bello Gallico*, on peut les considérer comme des “motifs textuels” à fonction caractérisante, mais il faut attendre la 26e position dans la liste pour voir apparaître le “motif” *his rebus cognitis.* Ce “motif” apparaît sous la forme d’un démonstratif à l’Ablatif pluriel, de la forme *rebus* et d’un participe à l’Ablatif parfait pluriel prédicat d’un Ablatif Absolu. Si la méthode a l’inconvénient de fournir des structures phraséologiques figées comme premiers segments repérés, elle permet donc en revanche de prendre en compte dans la détection la variabilité lexicale du démonstratif (*his / iis / tantis*) et du participe (*animaduersis / cognitis / constitutis / exploratis / gestis*) .

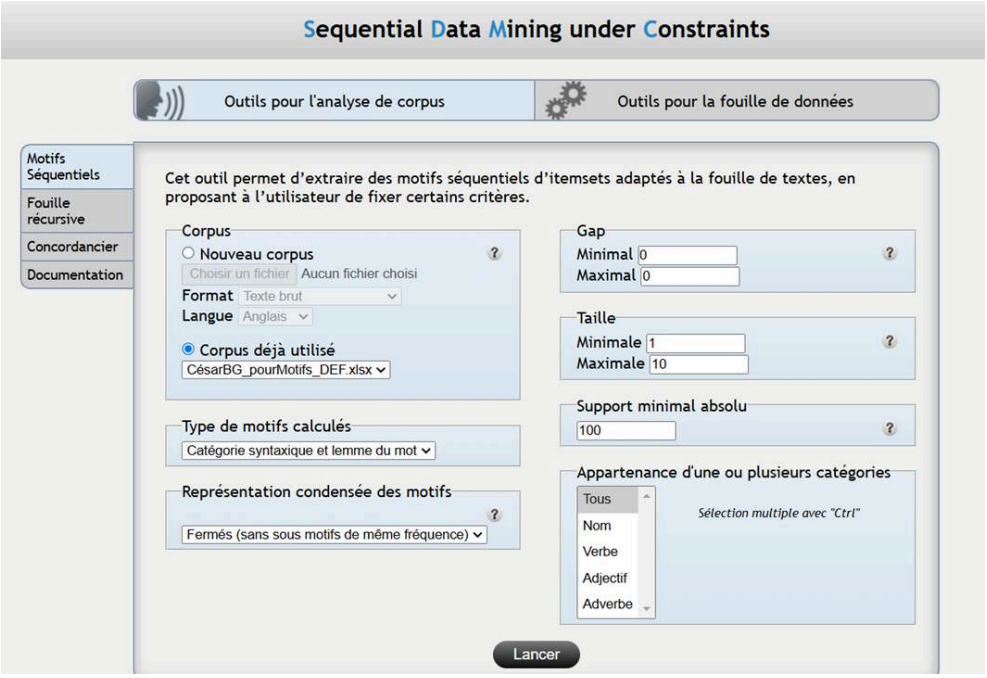
35 Cette dernière méthode contrastive permet d’affiner l’approche des segments répétés, et donc des “motifs textuels”, qui leur correspondent par une prise en compte et de la multi-dimensionnalité, et de la granularité fine des annotations grammaticales du fichier du LASLA. Elle présente toutefois deux inconvénients : elle suppose que la longueur du segment soit prédéterminée et elle n’offre aucune possibilité d’insertion d’éléments complémentaires, de positions en quelque sorte “vides” au sein des “motifs”, contrairement à ce que peut autoriser un logiciel de fouille de données.

3.1.2. Fouille de données séquentielle

36 Parmi les logiciels de ce type, le logiciel SDMC (Sequential Data Mining under Constraint ; https://sdmc.greyc.fr/index.php) présente l’avantage d’avoir pu être aisément adapté au traitement des fichiers du LASLA. Ce logiciel (Quiniou., Cellier, Charnois & Legallois 2012) permet une recherche séquentielle simple (forme, lemme, partie du discours) ou une recherche combinée sur les lemmes et parties du discours, avec la possibilité de calibrer le nombre de positions vides (“gaps”) entre deux éléments du motif, la longueur en nombre de mots de la séquence ou sa fréquence d’apparition. En outre, SDMC peut prendre en compte l’inclusion de motifs les uns dans

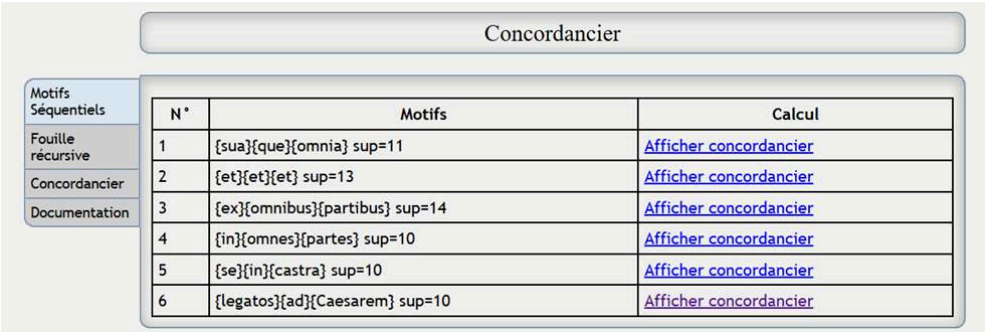
les autres. La figure 11 présente la fenêtre de requête de SDMC permettant de modifier les différentes contraintes portant sur la recherche.

Figure 11. Fenêtre de requête de SDMC



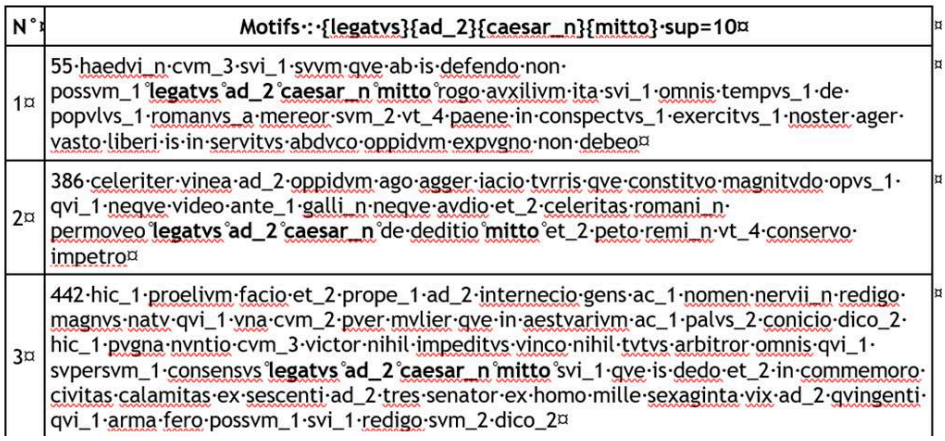
37 A partir de cette fenêtre de recherche, nous avons procédé au sein du *De Bello Gallico* à une requête contrainte par les paramètres suivants : schème en lemmes, gaps de 0 à 2 positions, longueur du “motif” entre 3 et 5 éléments, fréquence minimale de la séquence en nombre d’occurrences. La figure 10 présente les 6 séquences repérées par SDMC correspondant à ces critères.

Figure 12. Présentation par SDMC des 6 motifs correspondant aux critères [0,2], [3,5], [10]



38 Pour chaque “motif séquentiel”, SDMC propose un concordancier. La figure 11 présente les trois premières occurrences du concordancier pour le 6e “motif séquentiel”.

Figure 13. Présentation des 3 premières lignes du concordancier



N°	Motifs·:·{legatvs}{ad_2}{caesar_n}{mitto}·sup=10
1	55·haedvi_n·cvm_3·svi_1·svvm·qve·ab·is·defendo·non· possvm_1 °legatvs °ad_2 °caesar_n °mitto °rogo·avxilivm·ita·svi_1·omnis·tempvs_1·de· popvlvs_1·romanvs_a·mereor·svm_2·vt_4·paene·in·conspectvs_1·exercitvs_1·noster·ager· vasto·liberi·is·in·servitvs·abdvco·oppidvm·expvgno·non·debeo
2	386·celeriter·vinea·ad_2·oppidvm·ago·agger·iacio·tvrris·qve·constitvo·magnitvdo·opvs_1· qvi_1·neqve·video·ante_1·galli_n·neqve·avdio·et_2·celeritas·romani_n· permoveo °legatvs °ad_2 °caesar_n °de·deditio °mitto °et_2·peto·remi_n·vt_4·conservo· impetro
3	442·hic_1·proelivm·facio·et_2·prope_1·ad_2·internecio·gens·ac_1·nomen·nervii_n·redigo· magnvs·natv·qvi_1·vna·cvm_2·pver·mvlier·qve·in·aestvarivm·ac_1·palvs_2·conicio·dico_2· hic_1·pvgna·nvntio·cvm_3·victor·nihil·impeditvs·vinco·nihil·tvtvs·arbitror·omnis·qvi_1· svpersvm_1·consensvs °legatvs °ad_2 °caesar_n °mitto °svi_1·qve·is·dedo·et_2·in·commemoro· civitas·calamitas·ex·sescenti·ad_2·tres·senator·ex·homo·mille·sexaginta·vix·ad_2·qvingenti· qvi_1·arma·fero·possvm_1·svi_1·redigo·svm_2·dico_2

39 Les données fournies par le logiciel permettent de repérer non seulement de simples patrons linguistiques récurrents, mais également de véritables “motifs textuels” dotés d’une fonction textuelle. Ainsi, une recherche dans Hyperbase Web Edition de la séquence "LEM:LEGATVS ad *** LEM:MITTO", où le symbole *** correspond au “gap” de SDMC, indique qu’au sein de la base “Latin”, cette tournure présente une spécificité de 8,9 pour le *De Bello Gallico*, ce qui permet de la considérer comme un “motif textuel” caractérisant.

40 SDMC constitue donc un instrument précieux dans la recherche des motifs, mais la méthode présente néanmoins quelques limitations : la recherche combinée ne peut pas porter sur les formes; la notation grammaticale se limite à la partie du discours et ne permet pas de prendre en compte une granularité plus fine d’annotations morphosyntaxiques ; les résultats produits sont souvent surabondants, rendant difficile le classement des patrons extraits, et enfin le logiciel n’inclut aucune dimension statistique. Le recours à l’IA semble pouvoir dès lors offrir un moyen de conjuguer approche séquentielle et spécificité des séquences détectées.

3.1.3. Convolution

41 Une autre approche qu’explore le laboratoire BCL de l’Université Côte d’Azur repose sur des outils logiciels permettant d’analyser les paramètres sur lesquels s’appuient le *Deep Learning* - réseaux neuronaux convolutionnels et Transformers - pour prédire l’appartenance d’un texte à un auteur ou à un genre (Magri, Mayaffre & Vanni, ce numéro). *Hyperdeep*, implémenté dans l’interface logométrique Hyperbase (https://hyperbase.unice.fr/), permet ainsi d’identifier les traits saillants sur lesquels repose cette prédiction, traits saillants parmi lesquels apparaissent des “motifs” (Vanni, Mayaffre & Longrée 2018 ; Vanni, Corneli Mayaffre & Precioso 2023).

42 Les réseaux neuronaux convolutionnels s’appuient sur la linéarité du texte pour procéder à des tâches de classification. Hyperdeep peut pratiquer une convolution sur chaque niveau du texte, -formes, lemmes et codes grammaticaux-, et peut combiner la convolution sur chacun de ces trois niveaux. Un logiciel de déconvolution mis au point par L. Vanni peut mettre en évidence, -dans les passages attribués à César-, les marqueurs convolutionnels les plus saillants, ceux présentant les taux d’activation les plus hauts des réseaux neuronaux. Les figures 14, 15 et 16 présentent les résultats

obtenus pour le passage qu'Hyperdeep a identifié comme le plus représentatif du *De Bello Gallico*.

Figure 14. Marqueurs convolutionnels sur les formes graphiques d'un passage attribué à César

[...] occupato non longius mille passibus a **nostris munitionibus** considunt . postero die **equitatu ex** castris educto omnem eam planitiem quam in longitudinem **milia** passuum III **patere** demonstrauimus **complent** pedestres que copias paulum ab eo loco abditas in locis superioribus constituunt . erat **ex oppido Alesia** despectus in campum [...]

Figure 15. Marqueurs convolutionnels sur les lemmes d'un passage attribué à César

[...] occupato non longius **MILLE PASSVS_1 AB NOSTER MVNITIO CONSIDO** . postero **DIES EQVITATVS_1 EX** castris educto omnem eam planitiem quam in longitudinem milia passuum III patere **DEMONSTRO COMPLEO** pedestres que copias paulum ab eo loco abditas in locis superioribus constituunt . erat ex oppido **ALESIA_N DESPECTVS_1** in campum [...]

Figure 16. Marqueurs convolutionnels sur les codes grammaticaux d'un passage attribué à César

[...] occupato non longius mille passibus a **PosPro:Abl:Plur: Subs:3Decl:Abl:Plur:** considunt . postero die equitatu ex castris educto omnem eam planitiem quam in **Subs:3Decl:Acc:Sing:** milia passuum III **Verb:2Conj:Inf:Pres:Act:CodeSubAG: Verb:1Conj:Plur:Ind:Perf:Act:1Pers:CodeSubLN: Verb:2Conj:Plur:Ind:Pres:Act:3Pers:CodeSub00:** pedestres que copias paulum ab eo loco abditas in locis **Adj:1Cl:Abl:Plur:Comp:** constituunt . erat ex oppido Alesia despectus in campum [...]

43 Les items en couleur sont ceux qui présentent les plus hauts pics d'activation à chaque niveau. L'examen conjoint des trois niveaux révèle que les réseaux ont été sensibles à ce que l'on peut identifier comme des "motifs" : l'association des lemmes *mille* et *passus* est hautement spécifique du *De Bello Gallico*, comme nous l'avons vu précédemment, mais est en outre associé ici à des lemmes, des codes et des formes qui, sur la base des données de l'Hyperbase Web "Historia" présentent la plus haute spécificité dans cette œuvre : les formes *nostris* (z=9,2) ou *munitionibus* (z=6,7), les lemmes *munitio* (z=20,4) et *consido* (z=6,7), le code grammatical substantif de la troisième déclinaison ablatif pluriel (z=22,1). Il ne faut toutefois pas oublier que l'activation du réseau neuronal ne se fait pas sur un mot seul, mais sur un empan textuel (ici de longueur 3) : chacun des items en couleur correspond au centre de l'empan textuel. Si l'on considère ainsi les empans textuels autour des items colorés, on repère le segment *quam in longitudinem millia passuum quattuor patere demonstrauimus*, qui forme une variante plus développée du "motif" *ut supra demonstrauimus*, caractéristique de l'œuvre césarienne (Mellet & Longrée 2012; Longrée & Mellet 2013).

3.2. Non séquentielles

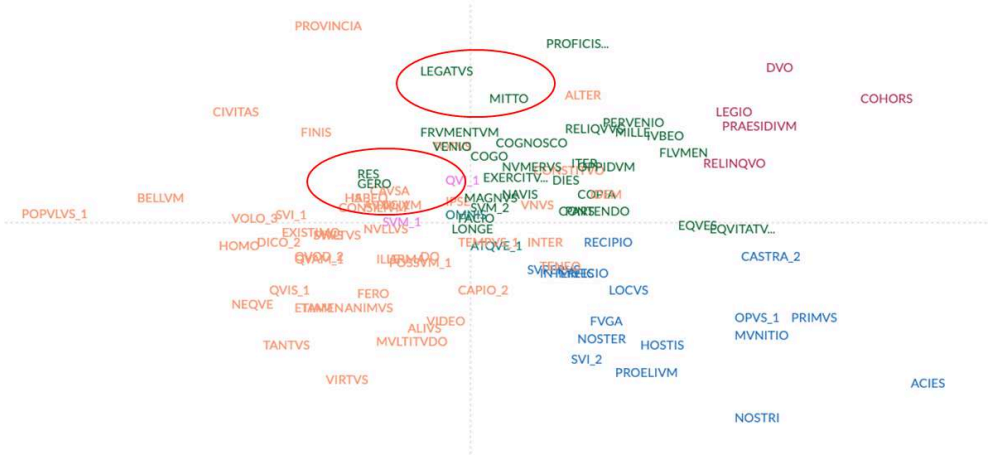
44 La recherche *corpus-driven* peut également adopter une approche non séquentielle qui sera elle uniquement contrastive.

3.2.1. Corrélats

45 Hyperbase Web Edition offre la possibilité d'étudier par le biais d'une AFC les proximités et distances entre les listes de cooccurrents des 100 lemmes les plus

fréquents du corpus considéré. La figure 17 montre le résultat de cette recherche pour un corpus réduit à l'œuvre césarienne.

Figure 17. AFC des profils cooccurentiels des 100 lemmes les plus fréquents chez César

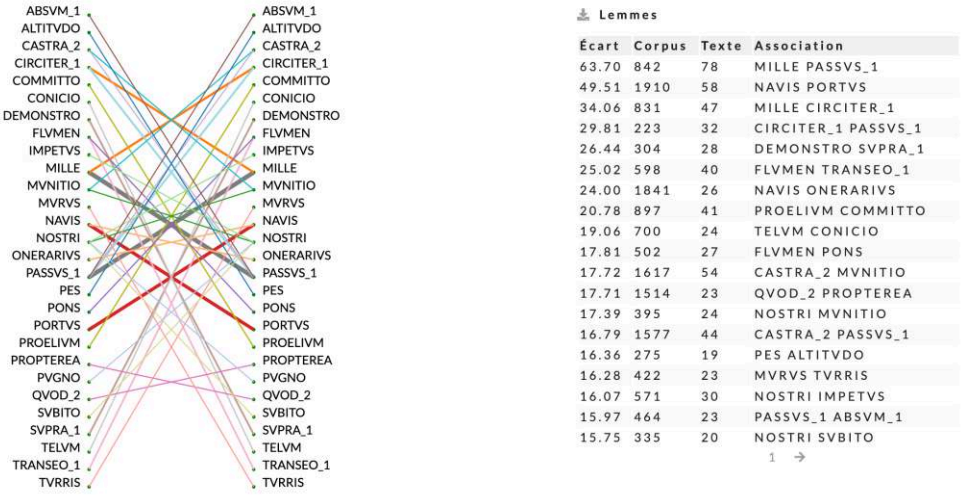


46 On ne s'étonne pas de voir les lemmes *legatus* et *mitto* présenter des profils cooccurentiels proches, sachant que nous avons vu précédemment que ces deux mots sont constitutifs du motif *legatos ad Caesarem mittere*. De même pour les lemmes *res* et *gero* que nous avons vu associés dans le "motif" *Quibus rebus gestis.*

3.2.2. Associations

47 Hyperbase offre également une fonction permettant de relever dans un texte l'ensemble des associations spécifiques de lemmes deux à deux. La figure 18 présente le résultat obtenu pour le corpus césarien. Bon nombre de ces associations relèvent simplement de phraséologismes (ex.: *proelium committere*, "engager le combat"), dont certains ont été également repérés par d'autres méthodes (ex.: *mille* et *passus*) ou apparaissent dans des "motifs" reconnus par ailleurs (ex. : *supra* et *demonstro,* constitutif du "motif" *ut supra demonstrauimus*). Pour rechercher des motifs à plus longue distance, l'IA peut ici à nouveau venir au secours du linguiste.

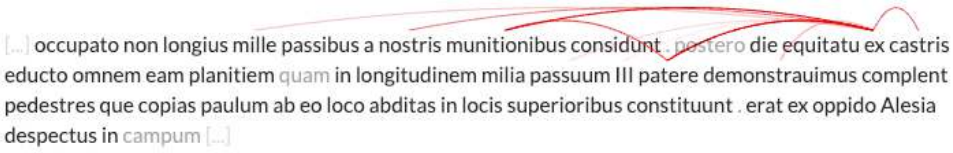
Figure 18. Associations spécifiques de lemmes chez César



3.2.3. Transformers

48 Le logiciel Hyperdeep ne s'appuie pas seulement sur des réseaux convolutionnels, mais également sur les Transformers et sur la "Self-attention" qui se fonde sur des associations à courte ou à longue distance. Hyperdeep permet de visualiser ces associations. Dans le passage déjà repéré comme caractéristique de César par la convolution (figure 16), les Transformers repèrent des associations visualisées par des liens en rouge dans la figure suivante. Ces associations doivent être significatives, dans la mesure où les tâches d'attribution réalisées par les Transformers sont généralement d'une qualité équivalente ou supérieure à celle des réseaux convolutionnels, mais elles ne sont pas toujours aisées à interpréter (comment expliquer, par exemple, dans la figure 19, le lien établi entre *ex* "hors de" et *patere* "ouvrir", qui suit ?) si on les considère seules. Combinées avec les informations des réseaux convolutionnels, ces associations peuvent en revanche devenir très instructives.

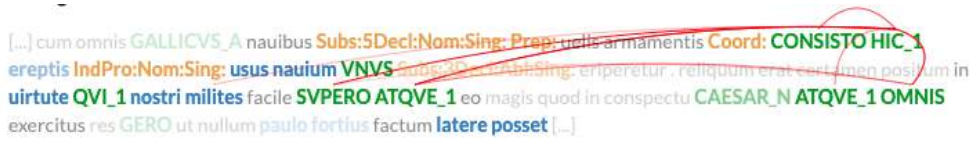
Figure 19. Marqueurs de Self-Attention dans un passage attribué à César



3.3. Hybride : approches séquentielle et non séquentielle combinées

49 Hyperdeep offre la possibilité de combiner deux approches, l'approche séquentielle des réseaux convolutionnels et l'approche non séquentielle des Transformers, grâce à un Multichannels Convolutional Transformer (MCT). Ici le Transformer ne s'appuie pas sur les formes seules du texte, mais bien sur les données de sortie des réseaux neuronaux aux trois niveaux du texte : formes, lemmes et codes grammaticaux. La figure 20 présente le résultat de cette méthode hybride sur un passage attribué à César par MCT.

Figure 20. Marqueurs hybrides (convolution multichannels + self-attention) sur un passage attribué à César



50 L'approche contrastive par MCT permet de mettre ici en évidence un nouveau type de "motif" ayant échappé à toutes les détections précédentes : la séquence de lemmes *atque* x mots *atque omnis* "et ... et tout" ne se rencontre que 5 fois dans le corpus et toutes les occurrences sont propres à César. En revanche, l'association à longue distance entre le lemme *hic* (démonstratif) et les lemmes *supero* ("prendre le dessus sur") et *atque* (coordonnant) n'est que peu interprétable, même en prenant en compte l'environnement de ces différents lemmes. L'examen d'un plus grand nombre de cas permettra sans doute de mieux comprendre à terme à quoi correspondent ces associations qui ne peuvent se résumer ni à la spécificité d'une séquence, ni à une cooccurrence spécifique.

4. Conclusion

51 L'inventaire que nous venons de proposer des méthodes de détection de patrons linguistiques multidimensionnels ne se veut en aucun cas exhaustif : d'autres méthodes existent et de nouvelles méthodes seront très certainement encore mises au point dans les années qui viennent grâce à l'amélioration des outils d'étiquetage et des possibilités offertes par l'IA. Notre objectif était avant tout d'établir une typologie des différentes types d'approches qui se laissent aisément répartir en quelques catégories. Mais quel que soit le type d'approche, - *corpus-driven* ou *corpus-based*, totalement ou partiellement supervisée, séquentielle ou non séquentielle, mono- ou multi-dimensionnelle, fondée sur la fréquence ou sur la spécificité-, l'élément déterminant est en fait la nature du patron recherché : la détection d'une unité phraséologique ou d'une routine discursive reposera sans aucune difficulté sur la répétition du patron linguistique; un "motif textuel" à fonction structurante n'aura pas nécessairement un caractère spécifique à un genre ou un auteur et la répétition pourra là aussi éventuellement suffire, mais s'il s'agit d'identifier des "motifs textuels" caractéristiques d'un genre ou d'un sous-genre littéraire, d'un auteur ou d'une œuvre, le recours à des calculs statistique de spécificités ou à l'IA s'imposera pour garantir objectivement la nature "caractérisante" du "motif".

52 Les moyens de détection des "motifs" dont nous disposons sont très efficaces, -nous l'avons vu-, quand, dans un enquête *corpus-based*, le chercheur a déjà une idée de ce qu'il recherche soit grâce à des études philologiques ou linguistiques antérieures, soit grâce à sa propre connaissance des textes, mais le rêve de quiconque s'intéresse aux "motifs textuels" est évidemment, dans une perspective *corpus-driven*, de pouvoir disposer d'outils permettant d'une part, une recherche de tous les "motifs" d'un corpus, quelle que soit leur fonction textuelle, d'autre part la mise en retrait du descripteur en rendant la détection aussi peu supervisée qu'elle soit. Le *Deep Learning* offre à cet égard des pistes particulièrement prometteuses, avec des résultats dont l'interprétabilité peut varier selon la langue, le type de texte ou la granularité des annotations. Toutefois, même si comme nous l'avons montré sur un exemple latin, la compréhension de certains des marqueurs demandera encore très certainement des

investigations plus poussées, le *Deep Learning*, qu'il s'agisse de convolution, de *Transformers* ou de leur combinaison dans MCT, a permis des avancées considérables dans l'étude des relations entre "motifs textuels" et intertextualité (Vanni, Hadi, Longrée & Mayaffre, 2024).

53 La recherche sur les "motifs" ne pourra à coup sûr dans le futur faire l'économie de l'utilisation de l'IA, et du *Deep Learning* mais son avenir se trouve également dans l'amélioration des annotations syntaxiques dont nous pouvons disposer. Une des composantes des "motifs" multi-dimensionnels est à l'évidence de nature syntaxique : ainsi, dans un "motif" structurant et caractérisant du type *dum haec in Gallia geruntur* "tandis que ces évènements se déroulaient en Gaule" (Longrée & Mellet, 2013, 2018), la composante *in Gallia* pourra commuter avec des compléments de nature très diverses (*Romae*, "à Rome"), dont le point commun sera d'exercer la même fonction de complément dit "locatif" (lieu-situation). Dans cette optique, le LASLA a entrepris de développer un parser LASLA-SynthIA lui permettant de produire des arbres syntaxiques (Treebanks) latins performants pour pouvoir intégrer une dimension syntaxique aux paramètres de recherche des "motifs textuels" latins. On peut espérer que l'introduction plus large de cette nouvelle dimension syntaxique, -présente actuellement essentiellement dans les ALRs-, combinée avec les diverses méthodes de détection relevant de la fouille de données, de l'analyse statistique ou encore et peut-être surtout du *Deep Learning* ne pourra faire qu'avancer notre perception de ce qu'est ce phénomène complexe du "motif textuel".

BIBLIOGRAPHY

Bendinelli M. (2017), "Segments phraséologiques et séquences textuelles", Corpus 17, 51. URL : http://journals.openedition.org/corpus/2844 ; DOI : https://doi.org/10.4000/corpus.2844

Bertels, A., & Speelmann, D. (2013). ‘"Keywords Method" versus "Calcul des Spécificités": A comparison of tools and methods.*" International Journal of Corpus Linguistics*, 18(4), 536–560.

Cellier, P., Quiniou, S., Charnois, Th., & Legallois, D. (2012). What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In *Lecture Notes in Computer Science*, Springer Vol.7181, 166-177.

Chausserie-Laprée J.P. (1969). *L'expression narrative chez les historiens latins, Histoire d'un style*, Paris.

Ganascia, J. G. (2001). "Extraction automatique de motifs syntaxiques". In J. Véronis, L. Danlos, P. Zweigenbaum, N. Gasiglia, and P. l Amsili (éds), *Actes de la 8ème Conférence sur le Traitement Automatique des Langues Naturelles* (TALN'2001). Tours http://talnarchives.atala.org/TALN/TALN-2001/taln-2001-long-017.pdf.

Evert S. (2008), "Corpora and collocations".In A. Lüdeling & M. Kytö (eds), *Corpus Linguistics. An International Handbook*, article 58, vol. 2, Berlin: Mouton de Gruyter, 1212-1248.

Kraif, O. (2016). "Le lexicoscope: un outil d'extraction des séquences phraséologiques basé sur des corpus arborés". *Cahiers de lexicologie*, 108, 91-106.

Kraif, O. (2019). “Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope”, *Langue française*, 203, 67-83.

Kraif O., Novakova I. & Sorba J. (2025). « Périmètres des motifs phraséologiques : réflexion conceptuelle et méthodologique sur corpus littéraire contemporain », ce numéro.

Lavigne, F., Longrée, D., Mellet, S., & Mayaffre, D. (2016). “Semantic Integration by Pattern Priming:Experiment and Cortical Network Model.” *Cognitive Neurodynamics,* DOI :1007/s11571-016-9410-4

Legallois, D. (2006), “Des phrases entre elles à l’unité réticulaire de textes”, *Langages* 163, 56-70.

Legallois, D. et Silvestre de Sacy, A. (2021). “MotiveR : un programme pour la stylistique”. In I. Galleron et I. Fatiha (éds), *Livre des résumés : dix ans avec CAHIER : des corpus d’auteurs pour les humanités à leur exploitation numérique*, 82–4.

Longrée, D. & Luong, X. (2003). Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés. Corpus, 2URL: http://journals.openedition.org/corpus/33; DOI: https://doi.org/10.4000/corpus.33

Longrée, D., Luong, X., & Mellet, S. (2008). *Les motifs : un outil pour la caractérisation topologique des textes*. In S. Heiden & B. Pincemin (éds), *Actes des JADT 2008, 9èmes Journées internationales d’Analyse statistique des DonnéesTextuelles, Lyon, 12-14 mars 2008*, Lyon, 733-744.

Longrée, D. & Mellet, S. (2012). “Continuité et ruptures dans l'expression narrative des historiens latins.” In M. Biraud (Ed.), *(Dis)continuité en linguistique grecque et latine, Hommage à Chantal Kircher-Durand*, Paris, 323-338

Longrée, D. & Mellet, S. (2013). “Le motif: une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours”. *Langages* 189, 68-80.

Longrée, D. & Mellet, S. (2018). Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach. In The Grammar of Genres and Styles: From Discrete to Non-Discrete Units, edited by Dominique Legallois, Thierry Charnois, and Meri Larjavaara, 140–163. Berlin/Boston: de Gruyter.

Longrée, D., Mellet, S., & Lavigne, F. (2019). Construction cognitive d’un motif : cooccurrences textuelles et associations mémorielles. CogniTextes. doi:10.4000/cognitextes.1202

Magri V., Mayaffre D. & Vanni L. (2025). « Enregistrer les genres littéraires. Repérage et apprentissage de marqueurs de registres discursifs », ce numéro.

Mellet, S. & Longrée, D. (2009). Syntactical Motifs and Textual Structures. Considerations based on the Study of a Latin historical Corpus. Belgian Journal of Linguistics, 23. doi:10.1075/bjl.23.13mel

Mellet, S. & Longrée, D. (2012). Légitimité d'une unité textométrique : le motif. In A. Dister, D. Longrée, G. Purnelle (éds.), *Actes des Journée d'analyse des données textuelles 2012*, 715-728.

Mellet, S., Luong, X., Longrée, D., & Barthélémy, J.-P. (2009). “Représentations du texte pour la classification arborée et l’analyse automatique de corpus : application à un corpus d’historiens latins.” *Mathématiques et Sciences Humaines*, 187 (3), 107-121.

Novakova, I & Siepmann, D. (éds), (2020). *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*, Cham.

Quiniou, S., Cellier, P., Charnois, Th., & Legallois, D. (2012). “Fouille de données pour la stylistique: l’exemple des motifs émergents”. In A. Dister, D. Longrée, G. Purnelle (éds.), *Actes des Journée d'analyse des données textuelles 2012*, 821-833.

Salem, A. (1986), “Segments répétés et analyse statistique des données textuelles.” *Histoire & Mesure*,1986, 1-2, 5-28

Silvestre de Sacy A., Legallois D. (2025). Les librairies *MotiveR* et *Pymotifs* pour identifier les patrons lexico-grammaticaux significatifs (ou motifs séquentiels), ce numéro.

Vanni L., “Hyperbase Web. (Hyper)Bases, Corpus, Langage”. *Corpus,* 2024, 25, ⟨10.4000/corpus.8770⟩. ⟨hal-04523479⟩

Vanni L., Corneli M., Mayaffre D., & Precioso F (2023). “From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture”, *Corpus*, 24 https://journals.openedition.org/corpus/7667

Vanni L., Hadi M., Longrée D & Mayaffre D. (2024), “Multi-channel Convolutional Transformer and intertextuality : a Latin case study.” In Misuraca M. & Giordano G., *New Frontiers in Textual Data Analysis*, Springer, 197-207.

Vanni, L., Mayaffre, D. & Longrée, D. (2018). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. 14es Journées internationales d’Analyse statistique des Données Textuelles. JADT 2018, Rome, p. 459-466.

NOTES

1. La base publique “Latin” diffusée sur Hyperbase est conçue en vue d’une exploitation statistique à partir d’une sélection d’œuvres parmi les textes latins classiques traités par le LASLA.
2. http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/
3. La base publique “Historia” diffusée sur Hyperbase et conçue en vue d’une exploitation statistique et organisée selon un ordre chronologique, comprend l’ensemble des œuvres latines historiques classiques traités par le LASLA.

ABSTRACTS

Based on a specific example, Caesar's *De Bello Gallico*, the present article aims to review, without claiming to be exhaustive, the various techniques allowing today to detect “textual motifs”. The study relies on the Latin corpus of the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège, The LASLA files offering for each word form a verified lemmatization and a morphosyntactic analysis. The study will mainly compare the possibilities offered by the Hyperbase Web Edition software (Vanni 2024) with those of other open-source software. It will explore various methods of data mining and statistical analysis, and will look at the contribution that artificial intelligence can make in this area.

A partir d’un exemple précis, à savoir le *De Bello Gallico* de César, le présent article vise à faire le point, sans prétendre à l’exhaustivité, sur les différentes techniques de détection des “motifs textuels” existant aujourd’hui. Le corpus latin utilisé sera celui des fichiers du Laboratoire d’Analyse Statistique des Langues Anciennes (LASLA) de l’Université de Liège qui offrent une

lemmatisation et une analyse morphosyntaxique vérifiées de chaque forme des textes. L'étude confrontera principalement les possibilités que le Logiciel Hyperbase Web Edition (Vanni 2024) offre en matière des motifs à celles d'autres logiciels en libre accès. Elle explorera diverses méthodes de fouilles de données ou d'analyses statistiques et s'intéressera également à l'apport que l'intelligence artificielle peut avoir en la matière.

INDEX

Keywords: textual motifs, data mining, textometry, statistics, deep learning, corpus-based, corpus-driven
Mots-clés: motifs textuels, data mining, textométrie, statistique, deep learning, corpus-based, corpus-driven

AUTHORS

DOMINIQUE LONGRÉE

Université de Liège, UR Mondes Anciens, LASLA

LAURENT VANNI

Université Côte d'Azur, UMR 7320, Bases, corpus, langage