Hyperbase Web. (Hyper)Bases, Corpus, Langage

Laurent Vanni

Hyperbase Web
(Hyper)Bases, Corpus, Langage

Laurent Vanni

Univ. Côte d'Azur, BCL, UMR UniCA-CNRS 7320, Nice, France

Résumé

Hyperbase est un logiciel d'Analyse de Données Textuelles (ADT) qui offre une suite d'outils statistiques dédiés à l'étude de corpus. Initialement développé sur ordinateur de bureau, il se décline depuis 2015 en plateforme web offrant une interface à l'ergonomie travaillée pour un usage tourné vers les sciences humaines et sociales. Après un rappel méthodologique de l'ADT, cette contribution présente Hyperbase Web version 2024, à partir d'exemples concrets d'usages, de notes techniques ainsi que des entrées par le menu (manuel d'utilisateur). Cette présentation sert de référence pour la prise en main du logiciel ou l'utilisation avancée des méthodes d'ADT.

Mots-clés : méthodes, outils, logométrie, textométrie, statistiques, bases de données, logiciels

Abstract

Hyperbase is a Textual Data Analysis software offering statistical tools dedicated to the study of corpora. Initially designed for desktop computers, it's now available as a web platform offering an interface dedicated for use in the human and social sciences. After an overview of the underlying methodology, this contribution presents the 2024 version of Hyperbase Web, based on concrete examples of use, technical notes and menu items (user manual). This presentation serves as a reference for getting to grips with the software or developing advanced use of Textual Data Analysis methods.

Key-words : methods, tools, logométrie, text-mining, statistics, data bases, software

1 Introduction

Le choix d'un outil d'analyse est souvent l'occasion de réfléchir à la démarche scientifique adoptée. Pour les Sciences Humaines et Sociales (SHS), l'implémentation d'algorithmes pose aux chercheurs des questions pointues sur leurs objets de recherche. A défaut de formalisation claire, un des principes fondamentaux de l'algorithmique est de diviser un problème complexe en un ensemble de problèmes plus petits et plus simples à résoudre. C'est ce que propose Hyperbase depuis sa création : aborder le langage dans toute sa richesse en comptant simplement des mots. Cette approche d'apparence naïve promet une extraction d'observables dont la combinaison interroge sur les représentations les plus complexes des textes et du langage.

Cette contribution est à la fois méthodologique et pratique. Les fondamentaux de l'Analyse des Données Textuelles (ADT), sur corpus, sont rappelés et adossés au développement de la plateforme d'Hyperbase Web, un outil dédié aux SHS et qui s'inscrit dans la tradition de l'ADT depuis l'essor de l'informatique en France dans les années 80s. En partant des méthodes statistiques classiques, Hyperbase Web propose un paradigme empiriste et un parti pris épistémologique simples dans lesquels la seule hypothèse est le corpus dans son évidence matérielle (les mots, les phrases, les paragraphes, les textes). Après une discussion sur les prérequis de la méthode, un parcours de lecture est proposé avec des exemples concrets utilisant l'ensemble des outils mis à disposition par le logiciel. Des *notes techniques* ainsi que des *manuels d'utilisation* agrémentent la présentation.

2 Prérequis

L'utilisation d'Hyperbase renvoie à une vision théorique du corpus et un traitement pratique des textes. La méthode est exploratoire, elle interroge et fait émerger des connaissances mais les usages sont bornés par des axiomes liés au corpus. Le parcours interprétatif proposé s'inspire du savoir-faire d'une discipline qui s'étend sur plusieurs décennies. L'ergonomie évolue mais reste attachée aux usages historiques de l'Analyse de Données Textuelles.

1.1 Généalogie

À l'origine du logiciel se trouve un linguiste normalien et agrégé de lettre classique, Etienne Brunet. Il fonde dans les années 80 un laboratoire d'analyse de données textuelles après avoir compris qu'une révolution informatique naissante était sur le point de changer les représentations que nous avions des textes. Pour poursuivre ses recherches, il se munit des premiers ordinateurs, prêtés à l'époque par le centre de recherche d'IBM de La Gaude dans le sud de la France, et met en place des chaines de traitements innovantes pour étudier les premiers corpus numérisés en utilisant une méthode partagée par les grands noms de la discipline du moment que sont Charles Muller ou Jean-Paul Benzecri (Beaudouin 2016).

Quelques années plus tard E. Brunet décide de faire profiter la communauté des outils qu'il développe et compile pour ses propres besoins. Il crée alors Hyperbase, un des premiers logiciels proposant une suite d'outils complète pour l'analyse de données textuelles. La technologie qu'il utilise à l'époque, *Hypercard*, lui souffle une partie du nom *Hyper*-base et l'autre partie lui sert de référence aux humanités numériques naissantes qui se construisent autour des *bases* de données numériques. Pour l'anecdote, le laboratoire qu'il fonde en parallèle se nomme par la suite *Bases, Corpus, Langage*, un nom qui réunit définitivement outils, données, méthodes et objet dans une approche interdisciplinaire qui mêle informatique et sciences humaines[1].

Hyperbase connaît de nombreuses mises à jour et opère différents bons technologiques au gré des avancements de la recherche en informatique. Les systèmes évoluent, les moyens progressent et Hyperbase aussi. Deux changements majeurs sont à noter depuis la création du logiciel. Le premier est une réécriture complète du logiciel dans les années 90 pour offrir une vision plus mature et plus souple de l'ADT sur ordinateur de bureau (remplacement d*'Hypercard* par *ToolBook*). Le deuxième est un changement d'échelle. Avec la démocratisation d'internet dans les années 2000, les corpus ont gagné en volume et en accessibilité, Hyperbase s'est approprié ces ressources. À partir de bases vertigineuses comme celles de Google ou plus structurées de FranText et même de la BNF (Brunet 2012, 2023), le logiciel est capable de se connecter automatiquement et d'appliquer les méthodes de l'Analyse de Données Textuelles pour repousser les limites de l'exploration quantitative de macro-corpus.

[1] Outre un logiciel et un laboratoire, Etienne Brunet lègue à la communauté scientifique une centaine d'articles et une douzaine d'ouvrages. On se reportera aux trois derniers volumes parus chez Honoré Champion, *Comptes d'auteurs. Études statistiques de Rabelais à Gracq* (2009), *Ce qui compte. Méthodes statistiques* (2011) et *Tous comptes faits. Questions linguistiques* (2016).

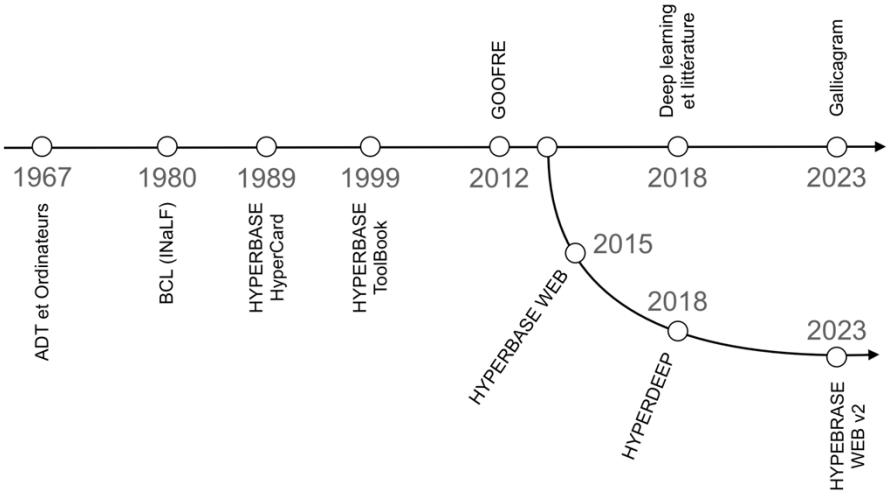


Figure 1 : Hyperbase dans le temps

Au début des années 2010, une autre vision du logiciel voit le jour (figure 1). Une approche orientée web, où la dématérialisation des données s'associe à une dématérialisation du logiciel lui-même. Hyperbase Web est une version du logiciel en ligne accessible *via* un navigateur standard, sans téléchargement ni installation préalable. Avec des bases de données prêtes à l'emploi ou à partir de de corpus nouveaux à charger sur la plateforme, Hyperbase web est livré clé en main sous forme de page internet dédiée. Une des originalités de cette version est l'usage de réseaux de neurones profonds, *deep learning* qui offrent une altérative complémentaire aux outils classiques. Ces méthodes gourmandes en ressources sont permises par la mise en œuvre de serveurs de calcul puissants hébergés par le CNRS, l'Université Côté d'Azur et l'infrastructure ORTOLANG. Ils rendent accessibles ces approches aux utilisateurs néophytes, curieux d'étudier la plus-value de nouveaux modèles probabilistes aux paramètres foisonnants.

1.2 Matérialité textuelle

Avant de prendre en main l'outil, il est suggéré de s'arrêter quelques instants sur le raisonnement qu'il propose. Il existe de nombreuses approches pour étudier le langage, toutes ne partagent pas la même démarche scientifique. Hyperbase adopte une vision matérialiste des textes. Autrement dit, les données textuelles dans leur factualité interrogent le chercheur et pas l'inverse. La littérature anglo-saxonne parle d'une approche *Corpus-driven* en opposition à l'approche *Corpus-Based* (Tognini-Bonelli, 2001). Dans cette démarche, le corpus représente à lui seul le matériel dont la complexité et les interactions composent les phénomènes observables et interprétables du langage. Comme pour la matière physique, le corpus est composé d'atomes. S'il est possible d'en discuter la précision et plus tard de les articuler, le microscope d'Hyperbase se focalise au niveau des *tokens*, c'est-à-dire des chaines de caractères séparés par des espaces blancs ; pour simplifier des mots[2]. En comptant principalement des mots, Hyperbase rend ainsi compte de phénomènes statistiques factuels émergeant des textes numériques, autant d'observables qui permettent de décrire les corpus.

Une des clés de la méthode est l'organisation contrastive du corpus faisant ainsi écho à la sémantique interprétative de F. Rastier : en corpus, « le sens est fait de différences, non de références » (Rastier, 2011 p. 64, cf. aussi p. 24, 29, 51)

[2] En réalité le logiciel permet aussi d'explorer des variations à d'autres niveaux de granularité : caractères, cooccurrences, expressions complexes, … (voir section 3). Ici pour simplifier la démonstration, le mot est considéré comme l'élément atomique de la matière texte.

Loin d'être donné, gratuit ou banal, le corpus représente une hypothèse précise que le chercheur vise à étudier en analysant les caractéristiques d'un sous-ensemble de textes. Il devient la norme par rapport à laquelle les différents textes contrastent. Pour y parvenir Hyperbase a recours aux *métadonnées*. Elles correspondent à un (ou plusieurs) découpage(s) du corpus rendu(s) possible(s) par un étiquetage préalable des textes. Ce découpage permet de délimiter la recherche de différences (ou contrastes) qui marquent les spécificités d'une partie du corpus par rapport à l'ensemble. Pour respecter l'esprit de la méthode, il est préférable que ces informations soient aussi objectives que possible (date, auteur, genre, …). En effet une annotation manuelle et subjective de phénomènes trop précis au sein du corpus est contraire à la démarche. La recherche et le comptage d'objets linguistiques balisés par l'intervention humaine entrainent un focus sur des représentations discutables qui empêchent l'exploration plus large ou plus neutre des données. En d'autres termes, le logiciel se focalise sur la détection d'information pour créer un savoir nouveau et non sur la comptabilisation d'annotations induites pour attester d'un savoir préalable (approche *Corpus-Based*). Les limites de la méthode correspondent aux limites des données collectées et structurées en contraste. L'interprétation des résultats est bornée par le corpus. Pour le chercheur, il constitue le fond endogène de toutes les observations faites. Toute tentative de généralisation des résultats est au mieux hasardeuse au pire érronée.

Note technique : Le corpus est numérique et se limite aux fichiers chargés sur la plateforme qui représentent du texte brut sans stylage ni mise en page particulière. Plusieurs formats sont pris en charge par la plateforme. Le format *.txt* est à privilégier mais d'autres formats plus élaborés comme les documents *.doc*, *.docx* et *.pdf* sont aussi acceptés (dans ce cas ils engendrent une conversion automatique vers du *.txt* par Hyperbase qui peut entrainer dans certains cas une perte d'information). Les métadonnées sont quant à elles soit associées au nom de chaque fichier chargé soit à un format particulier type *Alceste* ou encore répertoriés dans un fichier unique type *csv* contenant l'ensemble du corpus (textes et métadonnées). Une fois ces considérations technico-méthodiques digérées, Hyperbase propose une suite d'outils issue de la longue tradition française de l'Analyse de Données Textuelles qui permet d'interroger et d'explorer le corpus.

1.3 Méthode

Dès l'origine, la *lexicométrie* (puis la *textométrie* et la *logométrie)* a traité des corpus numériques suivant deux approches complémentaires. La première d'ordre lexicale (Muller, 1977) et la seconde d'ordre multidimensionnelle (Benzécri 1973). L'approche lexicométrique considère les textes comme des urnes statistiques et les mots comme le contenu de ces urnes qu'il est possible de dénombrer, comparer et pondérer en fonction de la taille de chaque urne (de chaque texte donc). Cette méthode implémentée dans Hyperbase, appelée **calcul des spécificités** (Lafon 1980), s'appuie sur des lois statistiques robustes (normale ou hypergéométrique) qui n'ont cessé de démontrer leur capacité à extraire de l'information des corpus. Pour dépasser le token (le jeton dans l'urne), l'ADT est passée rapidement de l'occurrence à la cooccurrence statistique de mots (voir à la poly-cooccurrence avec Hyperbase Web). L'analyse des cooccurrences spécifiques permet d'apprécier à la fois des phénomènes de macro distribution (des contrastes liés aux métadonnées) et de micro distribution (des contrastes liées au contexte d'usage des mots, dans les passages, phrases, paragraphes, …). Au-delà du mot seul, les cooccurrences permettent de repérer des objets linguistiques complexes, comme des associations sémantiques ou des choix syntagmatiques particuliers. Ces méthodes sont aussi dotées d'annotations morpho-syntaxiques (automatiques ou manuelles) qui démultiplient le champ des possibles observations (en croisant les mots avec les

catégories grammaticales et les lemmes par exemple). Seul bémol ici, une altération du texte par une interprétation arbitraire sur la nature des mots. « Qui lemmatise, dilemme attise. » prévenait (Brunet 1999), mais Hyperbase propose cette fonctionnalité dont l'usage est laissé libre au choix de l'utilisateur.

La seconde filière de l'ADT à la française considère les mots et les métadonnées comme un seul ensemble ayant autant de dimensions qu'un tableau croisant chacun de ses attributs. Depuis (Benzécri 1973), cette approche a recours à l'Analyse Factorielle des Correspondance (AFC) pour organiser l'espace des données de manière à ne conserver que deux axes (deux dimensions) lisibles par l'humain sur un plan graphique organisé en abscisses et ordonnées. La méthode permet d'appliquer des calculs vectoriels simples et d'observer des regroupements ou des oppositions entre les mots et les textes. Les axes principaux font généralement apparaître des tensions au sein du corpus et favorisent une vue d'ensemble critique des données. Il existe de nombreux usages de l'AFC, Hyperbase propose d'y recourir dès l'instant où l'information est trop dense et semble difficile à interpréter directement (trop de dimensions dans un tableau). Par exemple, la distribution d'une liste de mots convoque automatiquement une AFC si sa taille dépasse un certain seuil (50 mots dans le logiciel). Un parcours exploratoire classique qui utilise l'AFC consiste à demander la distribution des n mots les plus fréquents dans le corpus. Ce type d'exploration fait émerger en général les thèmes, les choix lexicaux ou encore stylistiques qui marquent des similarités ou au contraire des différences entre les textes (métadonnées). L'AFC n'est pas la seule méthode d'analyse multidimensionnelle portée par l'ADT et Hyperbase. Pour compléter l'AFC, le logiciel propose une analyse arborée, sorte de classification hiérarchique des données présentée sous forme d'arbre (et non de dendrogramme comme à l'accoutumé) pour simplifier la visualisation de regroupements et attester graphiquement de proximités ressenties avec l'AFC.

Note technique : **L'analyse arborée** implémentée dans Hyperbase a été originalement proposée par Xuan Luong dans sa thèse ("Méthodes d'analyse arborée. Algorithmes. Applications", 1988, Université de Paris 5). Elle utilise un algorithme de classification hiérarchique qui tranche sur les regroupements possibles dans un espace à n-dimensions fabriqué à partir des données. Hyperbase utilise une variante de l'algorithme original qui s'appelle ***Neighbour Joining.*** Cette méthode issue de la bio-informatique (pour la construction d'arbres phylogénétiques) est associée à un calcul de distance[3] et peut être convoquée pour comparer les métadonnées à partir du Tableau Lexical Entier (TLE) et rendre compte d'une distance intertextuelle (section 2.2) ou dans une distribution générale en croisant une liste de mots avec des métadonnées (section 3.3).

Avec l'AFC et l'Arborée, une troisième approche est introduite dans Hyperbase Web, celle de l'IA. En réalité cela peut ressembler à une extension de l'AFC qui illustre la statistique multivariée comme l'a décrite (Lebart, 1998). En effet, les réseaux de neurones artificiels convergent dans certains cas vers des calculs statistiques connus de l'ADT classique dont l'AFC. Cependant, si on augmente le nombre de couches de neurones et qu'on s'oriente vers des équations non linéaires[4] on obtient par apprentissage des modèles suffisamment profonds (d'où le nom *deep learning*) qui dépassent les capacités de détection de phénomènes statistiques par des modèles classiques. Les cas d'usage précis pour l'ADT interrogent les chercheurs, et l'extraction d'informations de ces modèles reste difficile, mais de *deep learning* est performant (classification automatique, génération,

[3] Le calcul de distance est choisi en fonction de la méthode utilisée. Se reporter aux sections 2.2 et 3.3

[4] Les équations non linéaires font référence ici aux fonctions d'activations du perceptron qui sont aujourd'hui principalement non linéaires en *deep learning*

traduction, …) ce qui signifie que ces méthodes utilisent des représentations pertinentes des corpus qu'il convient d'étudier. Pour ce faire, Hyperbase a recours à des algorithmes de *deep learning* dédiés à la classification automatique de textes. Le logiciel extrait des couches cachées les marqueurs qui permettent au modèle de distinguer les textes (ou métadonnées) entre eux. Le *deep learning* offre ainsi une information complémentaire à l'ADT classique et un parcours de lecture différent. Avec une IA descriptive associée à une approche statistique historique robuste, Hyperbase propose une expérience de pensée unique autour des corpus.

Hyperbase dans sa version web s'inscrit donc dans la tradition de l'ADT à la française tout en s'efforçant de suivre l'évolution technologique qui ne cesse de s'accélérer. La suite logicielle fait le pari du corpus comme objet d'étude clos, et celui du texte comme matériel empirique à interpréter. Si Hyperbase est incapable de modéliser une langue ou de généraliser des résultats, il est au contraire terriblement efficace pour interroger les hypothèses qui se cachent derrière un corpus. Se distinguant de la *linguistique de corpus* (étude de corpus sans *métadonnées)*, Hyperbase permet d'envisager l'exploration de phénomènes macros et micros, qui marquent les variations du langage au sein de corpus contrastifs.

2 Approches empiriques des textes

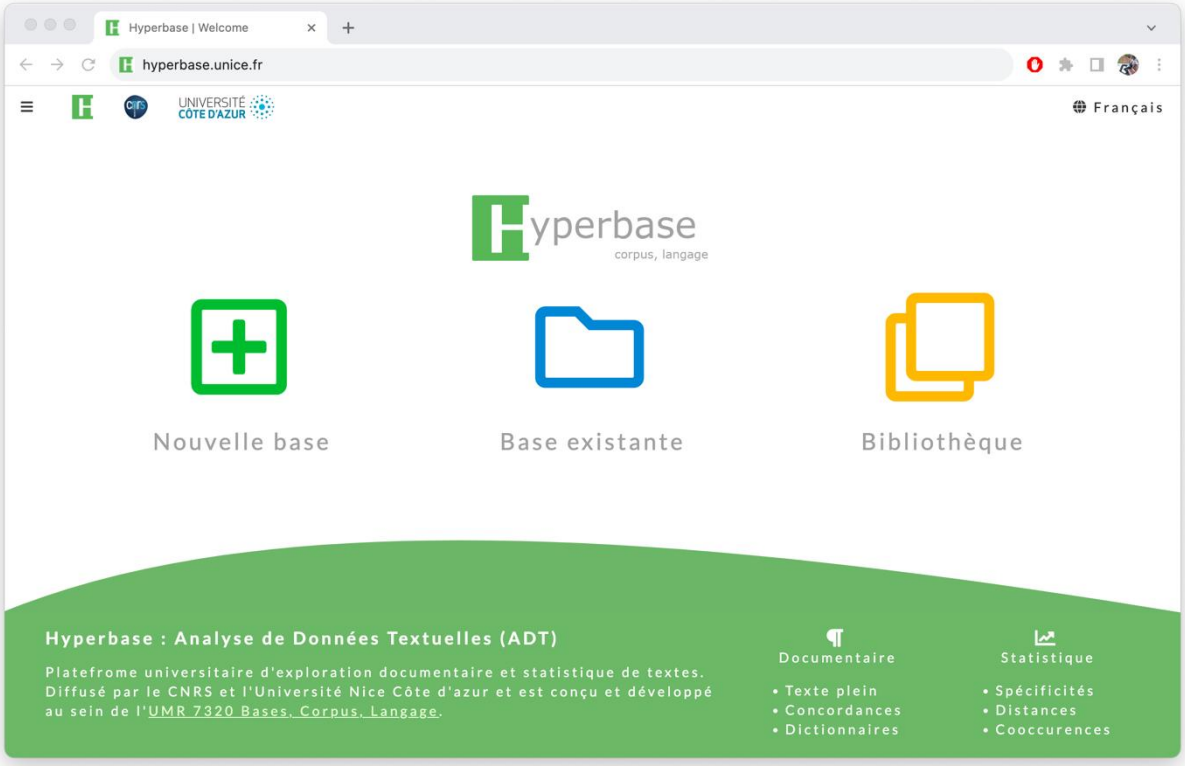


Figure 2 : page d'accueil d'Hyperbase Web

Manuel d'utilisation : La page d'accueil d'Hyperbase propose une « **Bibliothèque** », un panel important de bases (de données textuelles) prêtes à l'emploi. Elles représentent divers exemples d'usages dans des domaines aussi différents que l'analyse de discours politiques, médiatiques ou littéraires. L'usage de ces bases est libre, il est cependant recommandé de se les approprier (lire et comprendre le corpus) avant d'entreprendre une étude y faisant référence. L'interface en ligne permet aussi et surtout de créer une « **Nouvelle base** » ou de charger une « **Base existante** » à partir

de données personnelles et c'est en général le point de départ à emprunter pour démarrer une étude avec Hyperbase (Figure 2). Une fois la base chargée, l'interface proposée se veut aussi épurée et ergonomique que possible en respectant les standards du web actuel (Figure 9). Un moteur de recherche central accueille l'utilisateur et invite à une recherche d'occurrences orientée par l'intuition du chercheur (voir section 3). Si ce choix est assumé d'un point de vue ergonomique, il peut sembler moins évident d'un point de vue méthodologique. En réalité, les menus déroulants renvoient au parti pris émergentiste (et quantitatif) de la méthode, et le moteur de recherche se positionne seulement comme l'outil principal d'interrogation des observations faites. Dans les sections suivantes, les noms de fonction en italique représentent des entrées accessibles directement depuis les menus matérialisés par des icones situées en haut dans les coins gauche et droit de l'interface.

2.1 Profil et étendue du corpus

Le point d'entrée conseillé de la méthode avec Hyperbase est la fonction ***Edition***. Elle affiche le profil du corpus et donne quelques premières mesures effectuées automatiquement par le logiciel. La fenêtre d'édition du corpus contient l'information statistique minimale à connaître avant d'entreprendre une étude avec Hyperbase, à savoir l'étendue du corpus en nombre de mots, taille du vocabulaire et nombre d'hapax selon le contraste (métadonnées) choisi. Cette interface permet aussi d'éditer certaines informations relatives à la description de la base, à savoir son nom, son titre, l'auteur ainsi qu'exporter les données, partager la base ou l'effacer définitivement.

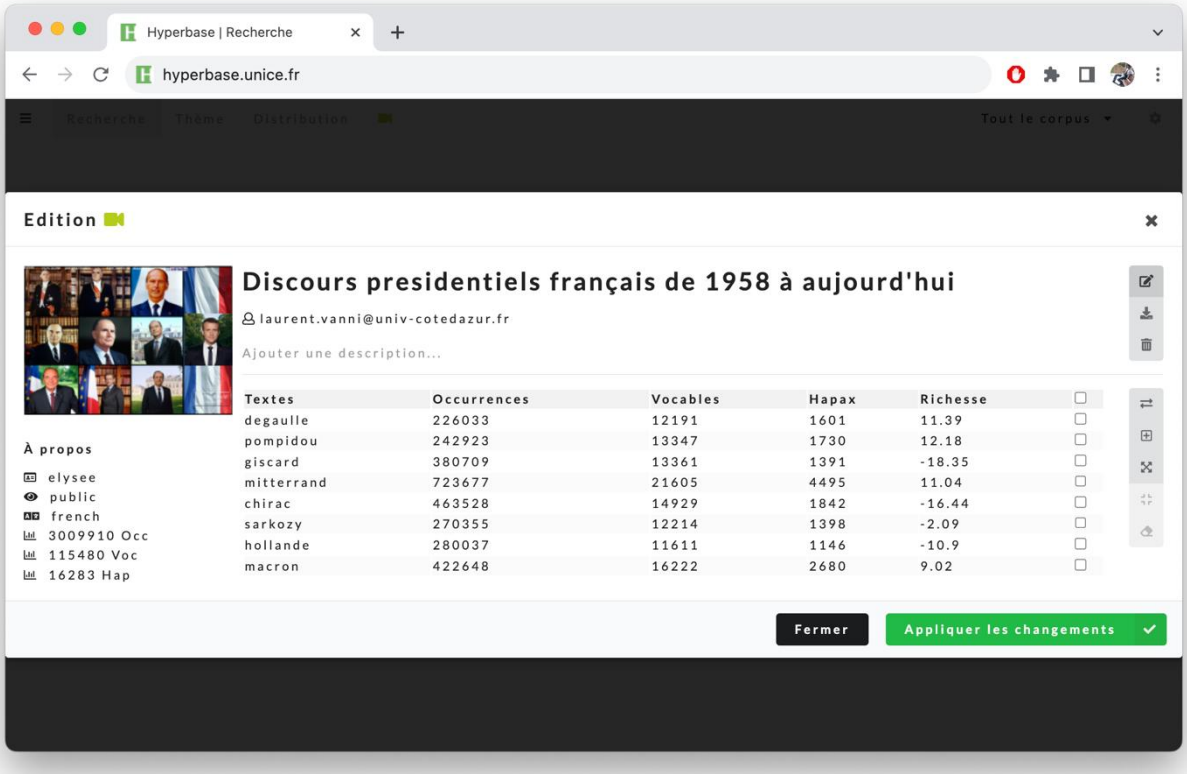


Figure 3 : Menu Edition – Base elysee

La figure 3 représente le profil de la base *elysee* qui a pour objet l'analyse du discours présidentiel français sous la Vème République. Ce corpus contrastif d'auteurs est organisé par défaut selon le nom de chaque président afin de pouvoir explorer les différences statistiques individuelles. Les métadonnées recensées sur chaque discours prononcé permettent de modifier ce point de vue depuis l'interface et de passer par exemple à une analyse de genre en comparant les types de discours (allocution, interview, déclaration, …) ou encore les différents mandats (ou années) pour chaque président. Les outils d'édition proposés par Hyperbase permettent à tout moment de croiser les métadonnées disponibles pour créer de nouvelles représentations du corpus et tester de nouvelles hypothèses. Le tableau central rappelle l'état du corpus observé qu'il faut garder à l'esprit lorsqu'un calcul statistique est convoqué dans l'interface. En effet, la taille générale du corpus et la taille de chaque partie sont des informations essentielles au moment de l'utilisation des méthodes statistiques. En appliquant des changements, toute l'indexation du corpus est modifiée et l'ensemble des résultats est recalculé pour prendre en compte ces changements.

En préambule à toute analyse, le menu ***Edition*** permet donc de vérifier l'état du corpus, c'est-à-dire l'hypothèse de travail, matérialisée par une mise en contraste, chargée dans le logiciel. Certaines analyses sont possibles dès ce menu. La colonne *Richesse*, par exemple, rend compte de la richesse lexicale sur les hapax ; plus précisément, l'écart qu'il y a par rapport à l'effectif théorique du nombre d'hapax dans chaque partie (voir la note technique sur le calcul des spécificités dans la section 2.3).

2.2 Distance intertextuelle

Une autre fonction exploratoire qu'il est possible d'interroger dès le chargement de la base est la fonction ***Distance***. Cette fonction correspond au calcul de la distance intertextuelle proposé par (Brunet, 2003). Ce calcul associé à une analyse arborée rend compte d'une structure dans le corpus organisé suivant les métadonnées sélectionnées. Chaque feuille de l'arbre représente une partie du corpus et chaque nœud intermédiaire de l'arbre constitue une étape dans la distance qui sépare deux parties. Plus ce parcours est long (plus le nombre de nœuds est important et plus les branches s'étendent) entre deux feuilles plus la distance est importante entre deux parties du corpus. L'interprétation de cette distance repose sur le choix du calcul, ici une variante de la distance de Jaccard qui poussent deux feuilles de l'arbre à se rapprocher d'un même nœud si leur proportion de vocabulaire commun augmente.

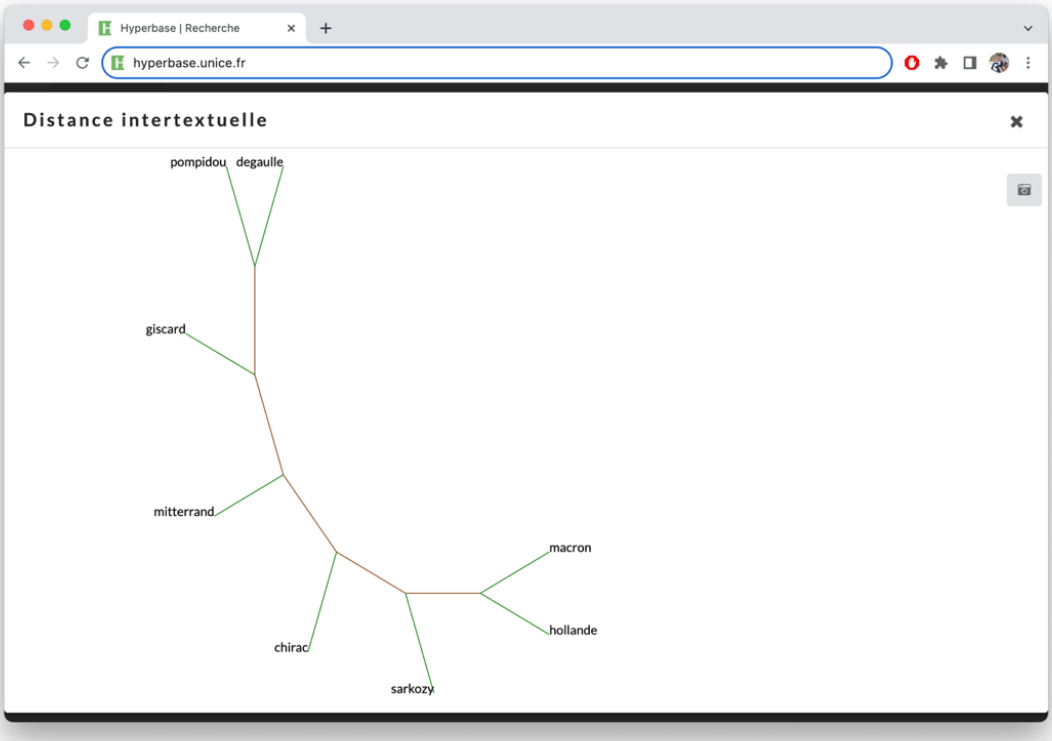


Figure 4 : Distance intertextuelle – Base elysee

La figure 4 montre l'affichage de cette distance sur la base du corpus *elysee.* Cette figure illustre en pareil cas avant tout la chronologie. Le discours semble évoluer principalement en fonction du temps et des événements sociaux-politiques du pays. Cette observation rappelle les précautions à adopter dans l'usage de statistiques multivariées lorsqu'une variable comme le temps (Salem 1991) ou le genre littéraire ont tendance à l'emporter sur tout autre phénomène. Mais la recherche de paternité d'un texte ou la comparaison de styles d'auteur sont des domaines où la distance intertextuelle s'illustre régulièrement et anime des discussions tel que l'éternel débat Molière/Corneille (Brunet, 2021).

Quelle que soit la métadonnée sélectionnée, la distance intertextuelle est un bon moyen d'apprécier l'équilibre du corpus. L'arbre généré permet d'identifier une structure robuste ou au contraire des altérations inattendues ; un moyen simple et efficace d'entamer une étude sur un corpus à partir de sa topologie lexicale (Mellet et Barthélemy 2007). Si cette fonction offre un aperçu rapide de l'organisation des données dans le corpus, elle s'associe généralement à d'autres analyses complémentaires qui dépassent la simple comparaison entre présence ou absence de mots dans les textes.

2.3 Spécificités lexicales

Les spécificités lexicales sont aux origines de la méthode lexicométrique. Elles reposent vulgairement sur un calcul d'écart entre effectif théorique et effectif réel.

Note technique: Sachant la taille d'un texte t et la taille du corpus total T, l' effectif théorique pour un mot dans un texte correspond à : fp où f correspond à la fréquence observée du mot dans le corpus et p correspond la probabilité de trouver un mot dans un texte avec $p = t/T$. L'écart

statistique entre la fréquence théorique et la fréquence réelle est obtenu suivant la formule de l'écart réduit :

$$z = \frac{k - fp}{\sqrt{fpq}}$$

avec $q = 1 - p$. Si ce calcul a déjà fait ses preuves, Hyperbase utilise en réalité une variante plus précise, fondée sur la loi hypergéométrique comme proposée par (Lafon, 1980):

$$p = \frac{f!\,(T-f)!\,t!\,(T-t)!}{k!\,(f-k)!\,(t-k)!\,(T-f-t+k)!\,T!}$$

La probabilité p est ici plus stable lorsque T et t sont petits. Elle converge cependant vers la loi normale (cité avant) lorsque T et t augmentent. Dans tous les cas, le logiciel convertit la probabilité en écart, plus lisible, en utilisant un procédé logarithmique (voir manuel de référence Brunet 2010) qui renvoie au calcul standard de la loi normale.

Les spécificités lexicales sont accessibles dès lors qu'un texte est sélectionné dans l'interface (à partir de la liste déroulante). En effet la suite d'outils statistiques s'adapte en fonction des données sollicitées. La distance intertextuelle (sur l'ensemble du corpus), présente précédemment dans le menu principal, laisse sa place aux spécificités lorsque le chercheur se focalise sur une partie du corpus. Le tableau qui apparaît alors donne la liste des mots spécifiques, c'est-à-dire les mots suremployés par le texte par rapport à une distribution moyenne. Cette liste de vocabulaire représente les premiers véritables contrastes que l'on peut apercevoir dans le corpus. Ces marqueurs linguistiques sont les phénomènes observables que l'hypothèse de travail, c'est-à-dire le corpus partitionné, permet d'étudier. La liste n'est pas filtrée mais est triée par écart décroissant. L'ensemble des *tokens* du corpus est soumis au calcul, ce qui conduit généralement à un mélange de substantifs, verbes, marqueurs de l'énonciation, mots outils, symboles en tout genre, etc. qui partagent la particularité d'être sur-représentés dans le texte par rapport à l'ensemble du corpus.

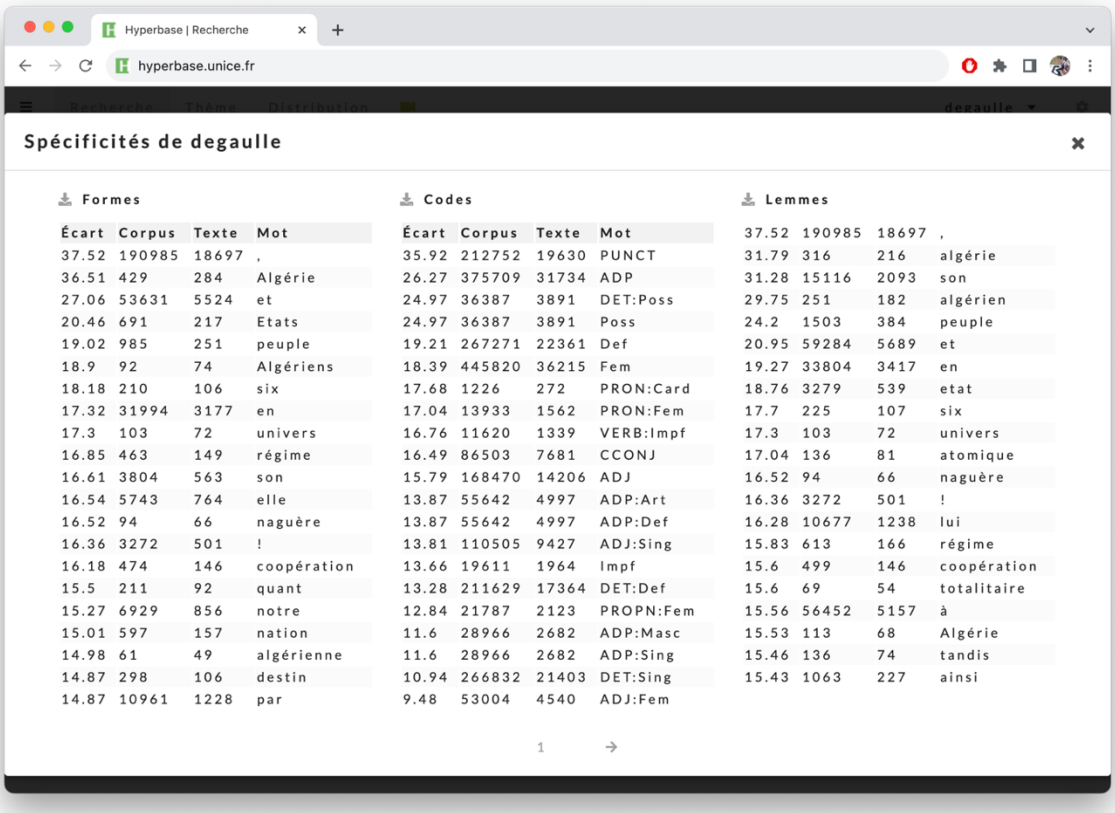


Figure 5 : Distance intertextuelle – Base elysee

La figure 5 illustre cette méthode sur le discours du président de Gaulle. On constate la présence de noms propres comme « Algérie » qui marquent l'époque et les évènements sociaux-politiques, nationaux et internationaux liés. On trouve aussi des signes de ponctuation qu'il faut analyser ici avec prudence : le discours oral est retranscrit à l'écrit et la ponctuation nait d'une tentative de marquer le rythme par la personne qui saisit les textes. D'autres mots en revanche ouvrent sur une interprétation plus élaborée, c'est le cas de « peuple », « univers », « régime » ou encore « destin » qui marquent un discours très solennel du chef de l'Etat. En ajoutant « coopération » ou « nation [5]» le discours semble tourné d'avantage vers la politique internationale, « pré carré » du président de la République. Il y a aussi des choix d'énonciation intéressant comme l'usage du pronom possessif « notre », qui cache ou révèle un usage prononcé de la deuxième personne du pluriel et interroge sur les l'inclusivité/exclusivité du « nous » dans la rhétorique politique (Bouzereau et Mayaffre 2022). Enfin certains mots comme « naguère » illustrent le mouvement de la langue dans le temps, c'est-à-dire ici entre 1958 et aujourd'hui, avec une diminution progressive de certaines expressions au profit d'autres. On le voit, les 20 premières spécificités sont riches en enseignements, et si la méthode est ici avant tout lexicale elle s'étend facilement vers une linguistique plus formelle en ajoutant les catégories grammaticales et les lemmes, qui poussent vers une exploration d'autres contrastes. Par exemple, la figure 5 montre que de Gaulle sur-utilise statistiquement le féminin, signe sans doute d'un discours plus conceptuel que celui de ses successeurs (les concepts, particulièrement les concepts politiques, étant plus souvent féminins que masculin).

[5] Le mot « nation » est à mettre en perspective avec l'usage ou non de la majuscule dans le corpus.

Cette approche robuste de la détection des contrastes du corpus ne se limite pas au simple comptage d'occurrences. Elle s'étend aisément vers la détection de phénomènes linguistiques plus intriqués comme les cooccurrences spécifiques. Il existe de nombreuses approches pour apprécier les phénomènes cooccurrentiels au sein des corpus (Mayaffre et Viprey, 2012) c'est-à-dire les associations particulières entre les mots qui en distinguent l'usage. La première fonction exploratoire des cooccurrences proposée par le logiciel se base sur un repérage généralisé et s'appelle *Corrélats*.

2.4 Corrélats

Corrélats est une des fonctions historiques du logiciel. L'idée promeut une visualisation d'ordre sémantique des mots basée sur leur distribution co(n)textuelle (Viprey 2006). Ainsi un mot comme « classe » peut prendre des sens différents dans les discours lorsque ses proches voisins relèvent du champ de l'éducation (« professeur », « école », « élève », …) ou au contraire de l'organisation sociale (« ouvrier », « travailleur », « patron », …). La cooccurrence offre ainsi des analyses qui révèlent des choix sémantiques, thématiques ou idéologiques au fil du corpus. *Corrélats* est un des outils qui permet une représentation globale de l'organisation des mots, que ce soit au niveau du corpus entier ou de chaque métadonnée prise séparément.

Techniquement, dans un premier temps, l'outil trie les mots par fréquence pour ne garder que les n mots les plus fréquents (n étant un paramètre qu'il est possible d'ajuster dans l'interface d'Hyperbase). Cette sélection permet de focaliser l'analyse sur des mots partagés et éviter des matrices creuses qui donneraient lieu à des distorsions liées à des phénomènes marginaux difficiles à interpréter. À noter que ce nettoyage automatique est entièrement paramétrable et l'étiquetage morphosyntaxique des mots autorise un filtrage par catégories grammaticales (par exemple les 100 premiers verbes).

Dans un deuxième temps une matrice $mots \times mots$ (tableau carré) rend compte de la fréquence d'apparition de chaque paire dans le corpus en considérant une fenêtre contextuelle glissante. La nature et la taille de cette fenêtre fait partie des paramètres généraux d'Hyperbase et utilise 10 mots par défaut (voir la note sur le manuel d'utilisation section 3). Ce choix peut être modifié avec une autre valeur (20, 30…50 mots) ou en utilisant les phrases (ponctuations fortes) ou les paragraphes (retour à la ligne) comme contexte élémentaire.

Enfin, les données générées par une telle extraction d'information sont analysées en utilisant une Analyse Factorielle des Correspondances (AFC) pour compresser l'information et la restituer sur deux axes (Benzécri 1974). Ici c'est une utilisation légèrement détournée de l'usage traditionnel de l'AFC puisqu'il n'y a pas des mots/textes (données rectangulaires) à visualiser, mais seulement des mots/mots (données carrées). En complément de l'AFC, une classification hiérarchique est appliquée sur le tableau (méthode similaire à celle utilisée pour l'analyse arborée). Les classes retenues (fixées par un nombre choisi dans les options de l'interface) sont représentées par des couleurs différentes sur le graphique. La représentation finale obtenue restitue une carte lexicale des textes et offre une visualisation rapide de la cooccurrence générale et des profils cooccurrentiels particuliers au sein des textes sélectionnés (figure 6).

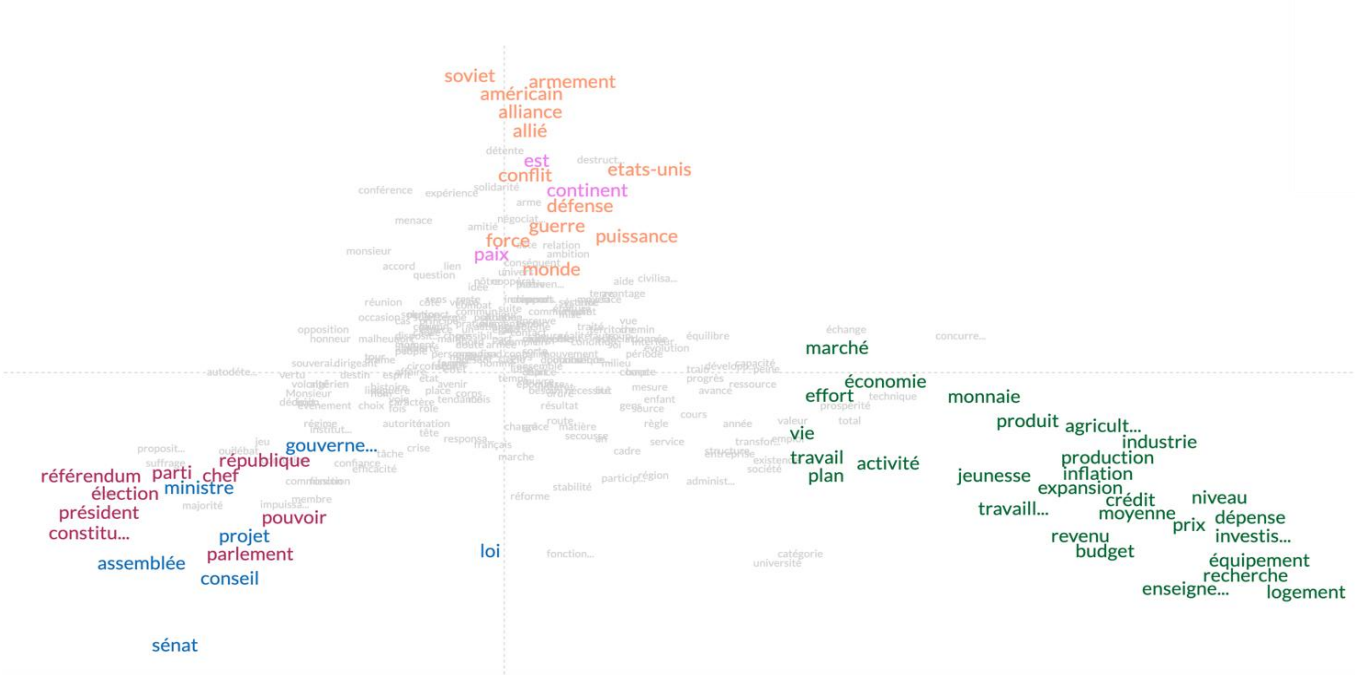


Figure 6 : Corrélats chez De Gaulle – Base elysee

La figure 6 illustre l'application de la méthode aux discours du président de Gaulle. Ici l'analyse se concentre sur les 300 substantifs les plus fréquents. Dans un souci de lisibilité, les contributions aux axes les plus fortes sont mises en valeur visuellement (mots en sur-brillance) ; les autres moins importantes sont masquées (mots grisés). Cet aperçu révèle les principaux thèmes du discours présidentiel. Trois quadrants se distinguent nettement sur le graphique pour distinguer trois nuages de mots qui représentent les sujets dominants : la politique socio-économique (en bas à droite, avec *prix*, *budget*, *logement*, etc.), la politique internationale (en haut au centre, avec *allié*, *Américain*, *armement*, etc.) et le discours lié aux fonctions politiques et institutionnelles du chef de l'Etat (en bas à gauche avec *assemblée*, *parlement*, *sénat*, etc.).

La fonction *Corrélats* donne donc matière à penser de manière global le contenu lexical et son organisation thématico-sémantique. Une autre fonction s'applique à prendre en compte les agencements particuliers des mots dans les textes, c'est la fonction *Associations,* développée dans la section suivante.

2.5 Associations

Contrairement à *Corrélats* qui s'appuie sur des profils cooccurrentiels établis (la matrice $mots \times mots$), la fonction ***Associations*** se base sur la seule hypothèse statistique de la co-présence de deux mots dans le contexte. Proche du calcul statistique des spécificités, cette approche utilise un calcul de cooccurrence théorique entre deux mots qui utilise deux probabilités, celle de l'absence du premier mot dans un segment de texte et celle de l'absence du deuxième mot. La probabilité est obtenue à partir de la même loi que celle utilisée pour le calcul des spécificités, la loi hypergéométrique. À partir de la fréquence du mot dans le corpus f, de la taille du corpus T et de celle du contexte t choisi pour observer la cooccurrence (10 mots par défaut), la probabilité est donnée par la formule :

$$p = \frac{f!\,(T-f)!\,t!\,(T-t)!}{f!\,t!\,(T-f-t)!\,T!}$$

Soit une simplification du calcul présenté dans la section 2.3 (avec $k = 0$). Par symétrie, la probabilité inverse $q = 1 - p$, donne les chances de trouver chaque mot dans le contexte quelles que soient leurs fréquences. Et enfin le produit des deux résultats donne les chances de trouver les deux mots ensemble dans le même contexte :

$$q = q_1 \times q_2$$

L'effectif théorique d'une paire est ensuite obtenu en multipliant ce résultat par le nombre de contexte T/t. Cet effectif est enfin comparé à l'effectif réel au sein d'un texte et un écart est alors obtenu qui rend compte de la force d'attraction entre ces deux mots dans le texte. Ce procédé unique à Hyperbase permet d'obtenir une liste d'associations statistiquement privilégiées relatives à une métadonnée et autorise une autre forme de visualisation qui illustre les liens de cooccurrences forts au sein des textes.

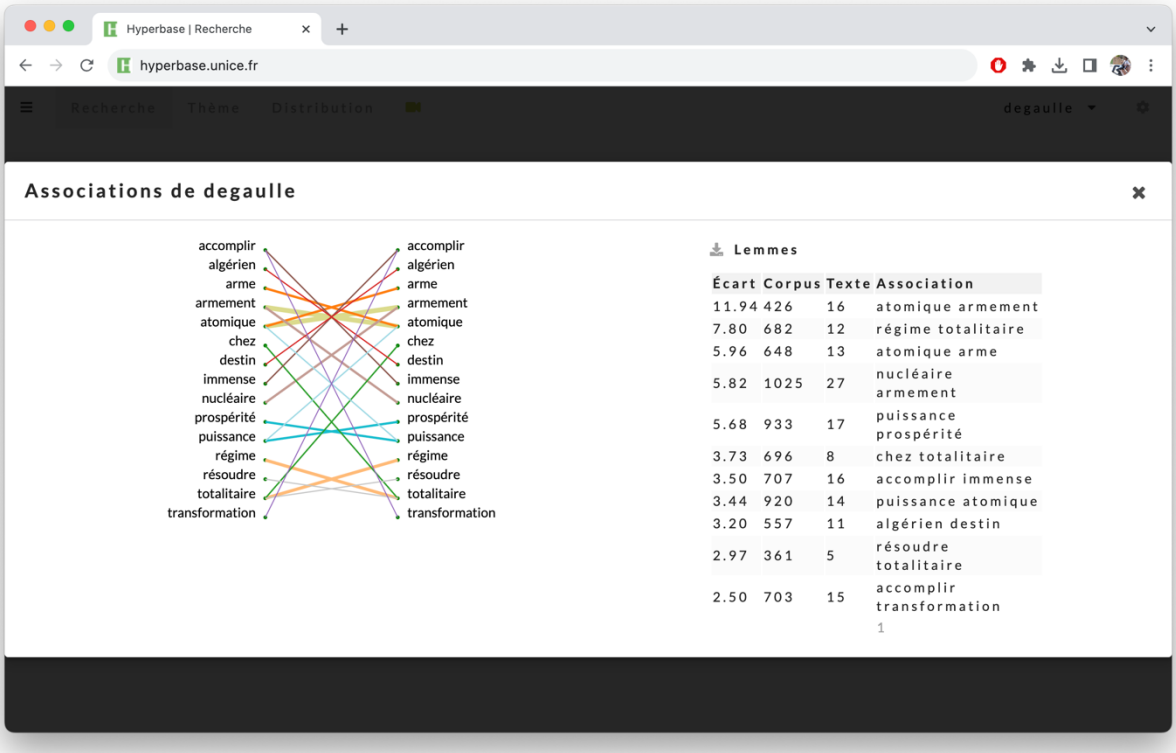


Figure 7 : Associations privilégiées chez de Gaulle

Dans le discours présidentiel gaullien (figure 7), une première interprétation du calcul des associations privilégiées pointe d'abord de manière attendue des thèmes historiques, spécifiques de la période : par exemple, l'armement nucléaire (« armement »/« atomique ») qui devient un enjeu essentiel pour la France après la deuxième guerre mondiale. Mais une deuxième lecture apporte des résultats plus subtiles. C'est le cas de la double association du verbe « accomplir » avec « transformation » et « immense ». L'usage d'un adjectif fort interroge ici sur le poids de la « transformation » à « accomplir » par le président. La « transformation » est d'ailleurs un levier langagier que partage d'autres présidents comme E. Macron qui l'associe volontiers, quant à lui, à des adjectifs comme « profond » ou lié à des sujets émergent comme l'écologie (« transformation écologique ») ou le numérique (« transformation numérique ») (Mayaffre 2021).

On le voit le passage de l'occurrence à la cooccurrence apporte un niveau d'analyse plus fin que celui d'un lexique décontextualisé. En réalité c'est la combinaison de l'ensemble des méthodes

proposées par l'ADT qui offrent au chercheur des parcours interprétatifs riches et des analyses précises. Chaque outil repose sur une représentation particulière des textes qui présente des avantages et des limitations. Et le croisement automatique de l'ensemble de ces représentations reste difficile pour la statistique traditionnelle. Aussi, pour élargir le champ méthodologique et permettre ce croisement, Hyperbase propose d'utiliser les réseaux de neurones profonds dédiés à la description des données.

2.6 Deep learning

Pour les Sciences Humaines et Sociales, l'entrainement de réseaux de neurones profonds ou *deep learning* correspond à des méthodes en boîte noire offrant des outils aux performances souvent intrigantes (Mayaffre et Vanni, eds. 2021). En linguistique, de nombreuses tâches automatiques comme la traduction, la classification ou la génération ont été améliorées significativement. Avec Hyperbase, et appliqué aux textes, le *deep learning* interroge l'intelligibilité de l'IA. En effet, la vocation première de l'ADT, à savoir la description du corpus, n'est pas celle au départ du *deep learning* qui se focalise davantage sur les performances pures. Cependant, considérant la classification automatique comme le pendant prédictif de l'analyse contrastive de corpus, Hyperbase fait le pari d'extraire l'information textuelle encodée dans les couches cachées des réseaux de neurones pour 1) expliquer les performances des modèles 2) découvrir de nouveaux marqueurs linguistiques qui interrogent sur les contrastes qui déterminent chaque partie du corpus.

Note technique : Pour être interprétable, l'architecture utilisée repose sur des couches de neurones standards associées à des méthodes d'extraction de l'information. Plusieurs types de couche sont utilisés pour combiner différents types d'abstraction des textes. Plus précisément, trois approches sont mobilisées. Une couche d'*embedding* responsable de la vectorisation des mots (similaire à une AFC) ; une couche de *Convolution* qui a pour but de modifier cette représentation en fonction du co(n)texte des mots ; et une couche d'*Attention* (Vaswani, 2017) qui se spécialise dans la détection de liens distants entre les mots. Hyperbase offre la particularité de mixer chacune de ces approches dans une seule architecture, appelée *Multichannel Convolutional Transformer* (MCT) (Vanni et al., 2024 sous presse), où *Mulitchannel* signifie que les mots sont considérés à la fois comme forme graphique, catégorie grammaticale et lemme. Ce système garantit l'extraction des marqueurs linguistiques complexes utilisés par le modèle pour accomplir une tâche de classification automatique (prédiction d'un auteur par exemple). Comme pour l'ensemble des méthodes traditionnelles d'ADT, l'applicabilité des réseaux de neurones profonds est dépendante de l'organisation du corpus. Les métadonnées choisies représentent pour l'IA les classes à distinguer. L'apprentissage se fait à partir du corpus étudié, la précision du modèle est dépendante du volume et de l'homogénéité des données.

Manuel d'utilisation : C'est à partir du menu ***Edition*** d'Hyperbase que l'on peut constater la disponibilité et la précision de l'IA sur le corpus chargé[6]. Si le nombre de mots est suffisant et la précision acceptable, une entrée ***Hyperdeep*** est ajoutée dans le menu principal.

Une des fonctions principales du **deep learning** appliqué au texte est la prédiction. À partir d'un nouveau texte, extérieur au corpus d'apprentissage, il est en effet possible de demander à l'IA de prédire la classe, c'est-à-dire de donner la métadonnée la plus probable attachée au segment de texte correspondant. Outre la fonctionnalité finale qui peut trouver un certain sens dans des domaines comme la recherche de *paternité* des textes, la description faite par l'interface révèle un usage différent pour l'ADT. Hyperbase offre au chercheur la possibilité d'explorer les marqueurs du langage qui prévalent dans la détermination de la classe par l'IA. Qu'ils soient d'ordre lexical, sémantique, sélectif ou associatif, ces marqueurs se distinguent en général par leurs combinaisons particulières et ajoutent un niveau d'analyse supplémentaire pour l'ADT. Avec l'architecture MCT,

[6] L'apprentissage d'un modèle de *deep learning* reste une tâche semi-automatique avec de nombreux paramètres à prendre en compte. La plupart des bases publiques bénéficient de ce traitement. Les bases privées sont quant à elles évaluées et entrainées au cas par cas en fonction des besoins et des questions de recherches associées.

Hyperbase embrasse le texte sous toutes ses représentations possibles. La structure du langage est couverte à la fois sur l'axe syntagmatique et l'axe paradigmatique par la *Convolution et* l'*Attention* qui captent des associations particulières en présence (à partir de l'organisation des mots dans le texte à prédire) et en mémoire (à partir de relations apprises sur le corpus d'entrainement).

Un des usages de la prédiction de textes est la détection d'intertextualité (Mayaffre et Vanni 2020). En projetant un auteur-cible dans un corpus contenant un ensemble d'autres auteurs-sources (ayant potentiellement inspiré le premier), il est possible d'apprécier les passages attribués à chacun comme autant de traces d'intertextualité. En utilisant la description des marqueurs donnés par Hyperbase, ces traces se précisent et le travail d'interprétation se spécialise. Chaque mot ayant contribué à la prise de décision s'éclaire dans le texte, et les liens qui existent entre les mots sélectionnés se matérialisent par des lignes colorées.

La figure 8 montre un exemple d'analyse d'intertextualité sur les vœux du président Macron, le 31 décembre 2022, en utilisant l'ensemble de ses prédécesseurs (de de Gaulle à Hollande) comme base d'apprentissage :

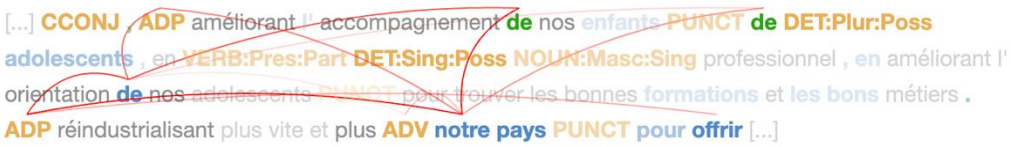


Figure 8 : Passage des vœux 2023 d'E. Macron attribué par l'IA à C. de Gaulle

L'extrait qui est présenté est le passage de Macron le plus fortement attribué au président de Gaulle. Les traces d'intertextualités que l'on peut observer sont riches en interprétation. D'abord, l'usage du déterminant possessif « notre », une spécificité jusqu'ici attribuée à de Gaulle que Macron a repris jusqu'à le détrôner de sa première place au rang des spécificités lexicales (voir section 2.3). Le mot « pays » ensuite, que Macron utilise ici opportunément, porté par le possessif est lui aussi spécifique de de Gaulle. Le détermiant possessif est d'ailleurs en général pointé par l'IA qui reconnaît en lui un marqueur rhétorique de l'actuel président emprunté au fondateur de la Vème république. Surtout, la force de la méthode se situe dans la combinaison de ces marqueurs, qui met le chercheur sur la piste de *motifs* (Mellet et Longrée 2009) linguistiques complexes, comme ici l'enchainement « de + Possessif » utilisé plusieurs fois au sein du même passage. Il révèle une répétition, figure de rhétorique employée en politique pour assener une idée, ici transcendé par la notion de possession comme dans l'extrait affiché: « ... l'accompagnement *de nos* enfants … *de nos* adolescents … l'orientation *de nos* adolescents … ».

Le *deep learning* s'accompagne nécessairement d'un retour au texte pour pouvoir être analysé correctement. Que ce soit manuellement ou à l'aide de calculs statistiques élaborés, la vérification de certaines observations est utile pour attester de certains phénomènes. Pour y parvenir, Hyperbase est articulé à un moteur de recherche, un moyen de cibler des marqueurs linguistiques particuliers pour en apprécier les usages.

3 Moteur de recherche

Manuel d'utilisation : Le moteur principal d'interrogation de la base se matérialise par un champ de recherche central adossé à trois onglets qui correspondent à trois visions de la recherche sur corpus (Figure 9). Le premier onglet par défaut correspond à la ***Recherche*** documentaire, c'est-à-dire à un retour au texte brut présenté sous forme de concordancier. Le deuxième utilise l'analyse des cooccurrences pour afficher le ***Thème*** du mot ou de l'expression choisie. Le dernier onglet affiche la ***Distribution*** statistique de la recherche demandée au fil des métadonnées qui composent le corpus. Toutes ces fonctions reposent sur un système de requêtes qui va du simple mot clé, à l'expression (régulière) complexe. Ce moteur combine aisément des formes graphiques de mots, des

catégories grammaticales ou des lemmes de manière contiguë ou non. Plusieurs expressions peuvent être regroupées dans la même recherche en utilisant des guillemets pour délimiter chacune d'entre elles. Ce système de recherche *ad hoc* se veut suffisamment large pour couvrir la majorité des questions linguistiques posées en ADT sans pour autant verser dans un langage informatique décourageant pour les Humanités. Les détails du fonctionnement du moteur de requêtes sont disponibles depuis l'interface dans une entrée ***Recherche avancée*** du menu principal. Enfin des ***Paramètres*** généraux viennent étendre les possibilités du moteur de recherche. Il est notamment possible de changer la représentation des mots en les regroupant par lemmes ou par étiquettes morpho-syntaxiques. Le corpus peut aussi être filtré par catégories grammaticales pour ajuster un calcul statistique et répondre à une interrogation linguistique particulière. La fenêtre de contexte peut également être manipulée pour le calcul des cooccurrences ou l'utilisation des *Jokers* associés aux requêtes.

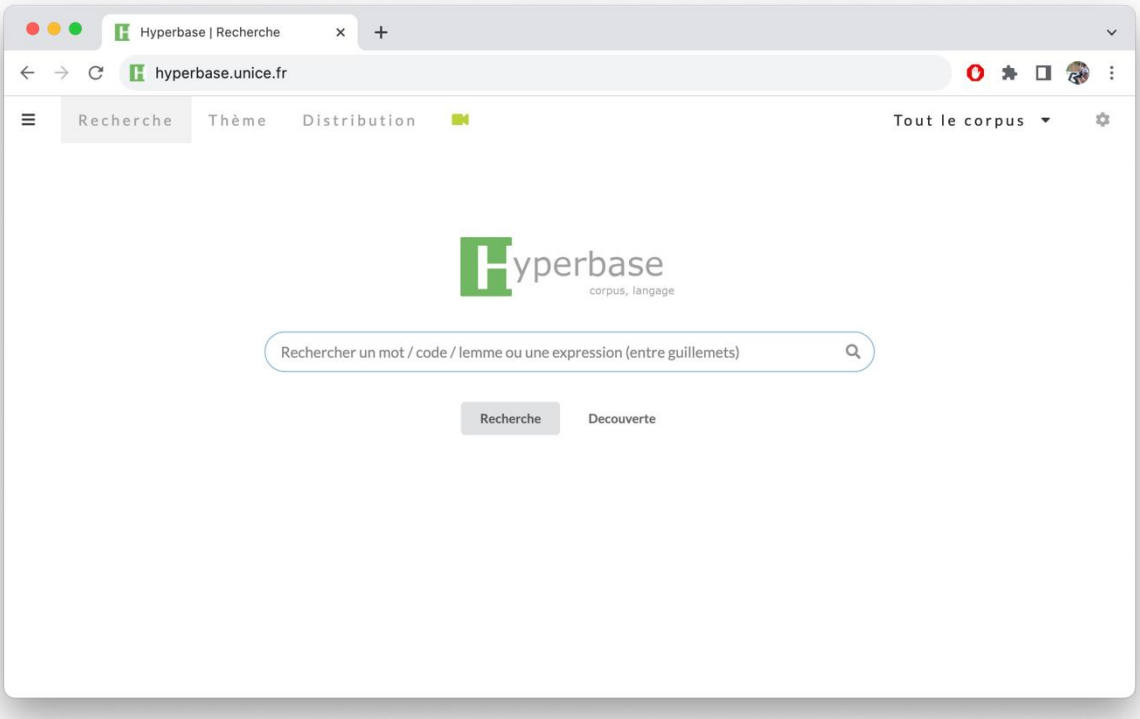
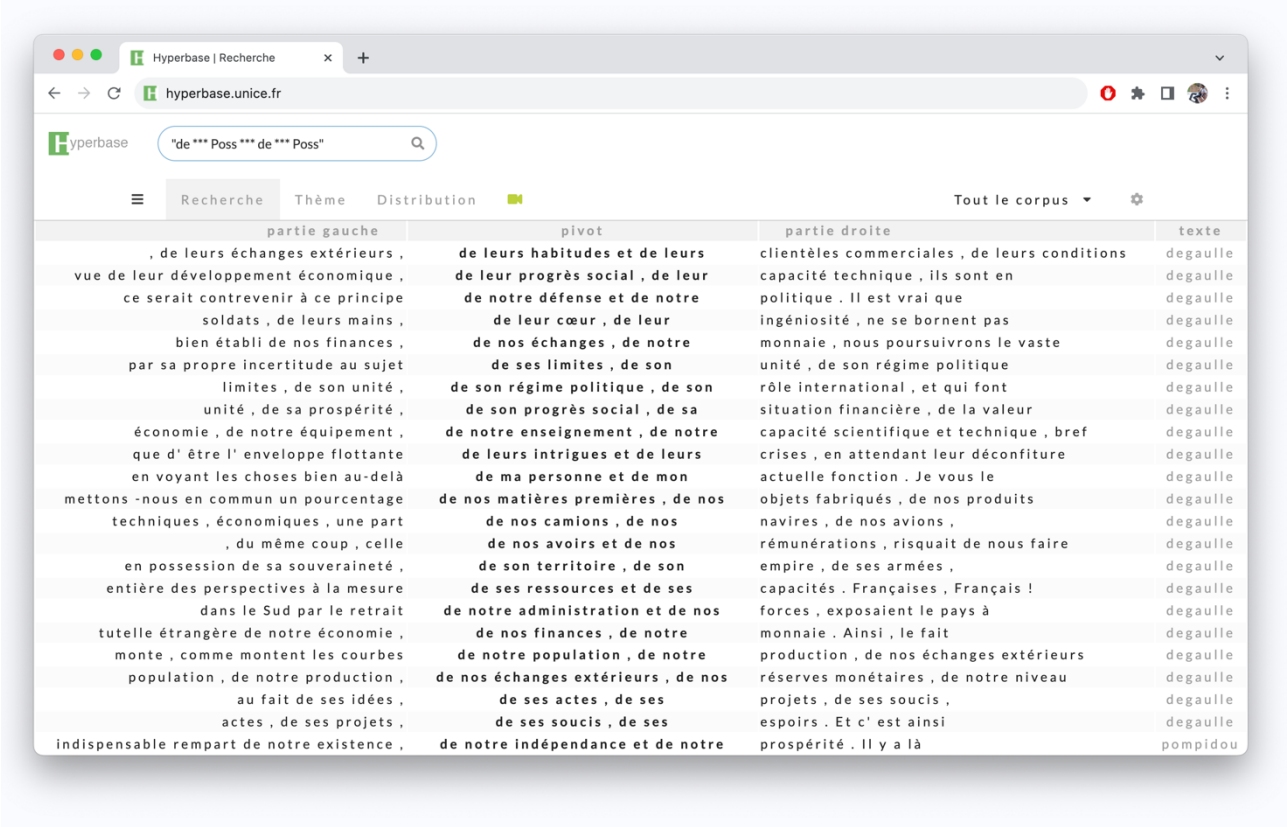


Figure 9 : Moteur de recherche principal

3.1 Recherche de concordance

Le concordancier, outil classique des philologues, est un moyen de peser certaines hypothèses par des sorties machines qualitatives directement interprétables par l'humain (Pincemin 2006). L'outil présente les résultats de la recherche sous forme d'un tableau contenant trois colonnes pour représenter le contexte gauche, le pivot (l'objet de la recherche), et le contexte droit. Cette visualisation peut être triée par ordre alphabétique à droite ou à gauche et permet d'amorcer une analyse (manuelles) des redondances affichées. À cette vue traditionnelle, Hyperbase ajoute une quatrième colonne pour visualiser les métadonnées relatives aux sorties, et un retour au texte plein à partir d'un clic sur la ligne considérée.

Cette vue contextuelle d'un mot ou d'une expression est en général une étape récurrente de l'Analyse de Données Textuelles. Les interprétations statistiques sont ainsi régulièrement valorisées par des exemples concrets dans le texte et *vice versa*. Par exemple le *motif* que nous avions pressenti en utilisant les réseaux de neurones profonds (section 2.6), à savoir la répétition de l'expression « de + Possessif » est identifiée en concordance par une expression utilisant la forme graphique du déterminant « de » plus l'étiquette « Poss » qui couvre la catégorie possessive des déterminants ou pronoms, le tout répété une fois en utilisant le *Jocker* « *** » (Figure 10).



partie gauche	pivot	partie droite	texte
, de leurs échanges extérieurs ,	de leurs habitudes et de leurs	clientèles commerciales , de leurs conditions	degaulle
vue de leur développement économique ,	de leur progrès social , de leur	capacité technique , ils sont en	degaulle
ce serait contrevenir à ce principe	de notre défense et de notre	politique . Il est vrai que	degaulle
soldats , de leurs mains ,	de leur cœur , de leur	ingéniosité , ne se bornent pas	degaulle
bien établi de nos finances ,	de nos échanges , de notre	monnaie , nous poursuivrons le vaste	degaulle
par sa propre incertitude au sujet	de ses limites , de son	unité , de son régime politique	degaulle
limites , de son unité ,	de son régime politique , de son	rôle international , et qui font	degaulle
unité , de sa prospérité ,	de son progrès social , de sa	situation financière , de la valeur	degaulle
économie , de notre équipement ,	de notre enseignement , de notre	capacité scientifique et technique , bref	degaulle
que d' être l' enveloppe flottante	de leurs intrigues et de leurs	crises , en attendant leur déconfiture	degaulle
en voyant les choses bien au-delà	de ma personne et de mon	actuelle fonction . Je vous le	degaulle
mettons -nous en commun un pourcentage	de nos matières premières , de nos	objets fabriqués , de nos produits	degaulle
techniques , économiques , une part	de nos camions , de nos	navires , de nos avions ,	degaulle
, du même coup , celle	de nos avoirs et de nos	rémunérations , risquait de nous faire	degaulle
en possession de sa souveraineté ,	de son territoire , de son	empire , de ses armées ,	degaulle
entière des perspectives à la mesure	de ses ressources et de ses	capacités . Françaises , Français !	degaulle
dans le Sud par le retrait	de notre administration et de nos	forces , exposaient le pays à	degaulle
tutelle étrangère de notre économie ,	de nos finances , de notre	monnaie . Ainsi , le fait	degaulle
monte , comme montent les courbes	de notre population , de notre	production , de nos échanges extérieurs	degaulle
population , de notre production ,	de nos échanges extérieurs , de nos	réserves monétaires , de notre niveau	degaulle
au fait de ses idées ,	de ses actes , de ses	projets , de ses soucis ,	degaulle
actes , de ses projets ,	de ses soucis , de ses	espoirs . Et c' est ainsi	degaulle
indispensable rempart de notre existence ,	de notre indépendance et de notre	prospérité . Il y a là	pompidou

Figure 10: Affichage du motifs "de + Possessif…de + Possessif" dans le discours de C. de Gaulle

Cette lecture particulière du texte est une aide précieuse à l'interprétation du *motif*, une figure de style particulière, en l'occurrence une répétition. Il est facile d'identifier grâce à la concordance la stratégie rhétorique de l'auteur. Le possessif entraine un affect, et la répétition un mécanisme d'axiomes qui tend à valider automatiquement le deuxième concept si le premier est accepté. Par exemple, « progrès social » et « capacité technique », « défense » et « politique », « actes » et « projets » ou encore « indépendance » et « prospérité ». Cette sortie interroge aussi sur l'usage du mot « notre », affiché plusieurs fois dans les résultats et identifié depuis l'analyse des spécificités comme un des leviers du discours du président de Gaulle. Un questionnement qui nécessite de changer de représentation afin d'apprécier d'autres phénomènes statistiques sous-jacents.

Manuel d'utilisation : La navigation par onglets permet pour une même requête de passer de la ***Recherche*** du texte brut (correspondant à la requête) à des représentations d'ordre fréquentiel en cliquant au choix sur ***Thème*** ou ***Distribution***.

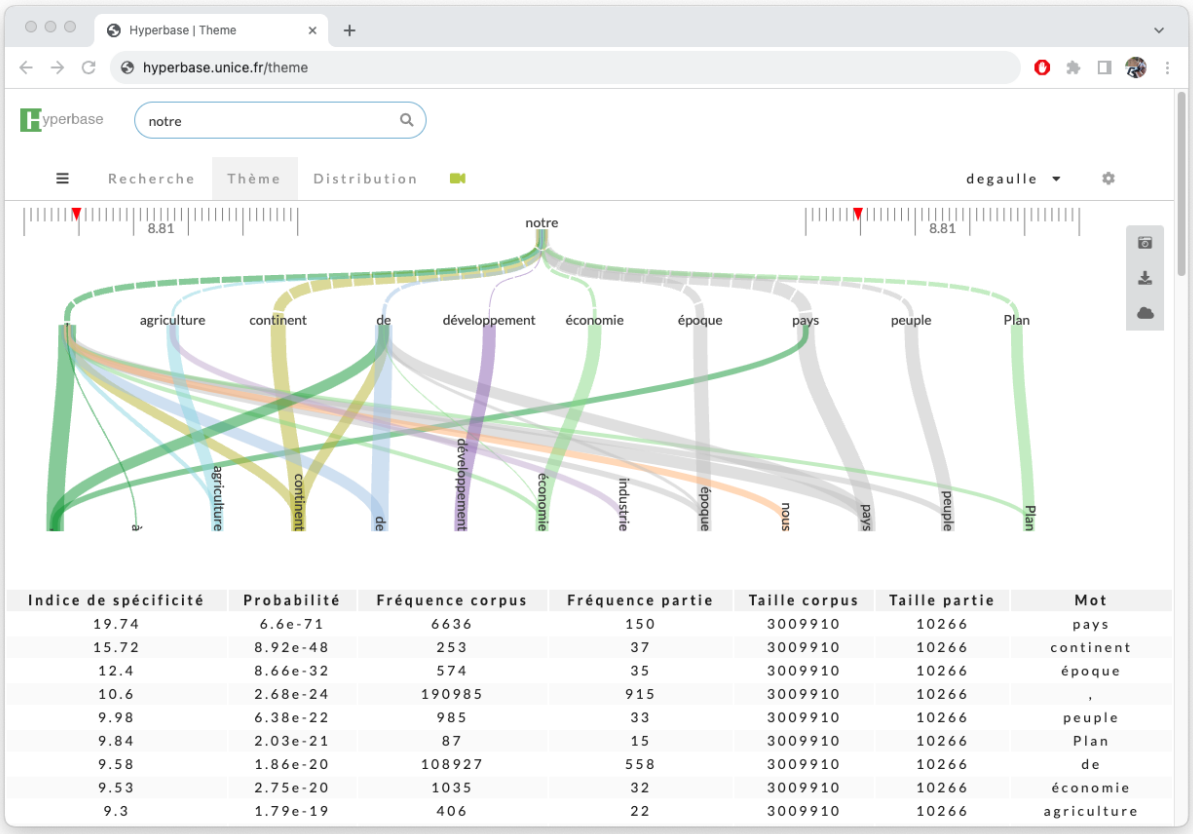
3.2 Recherche thématique

Avec les fonctions ***Corrélats*** et ***Associations***, l'onglet ***Thème*** constitue le troisième type d'analyse de cooccurrences proposé par Hyperbase. Le *Thème* correspond au calcul de cooccurrences

spécifiques, elle permet de lister les mots sur-représentés statistiquement autour du mot ou de l'expression choisie. Le résultat prend la forme d'un graphique associé à la liste exhaustive des termes dont le score de spécificité dépasse 2 (seuil minimal de spécificité). Deux visualisations sont possibles. La première prend une forme hiérarchique sur deux niveaux, il s'agit de la représentation des poly-cooccurrences (Vanni 2016). La deuxième plus classique affiche un nuage de mots représentant l'ensemble des cooccurrents directs. Dans les deux cas, la taille des mots ou des liens est proportionnelle au score de spécificité. Ces visualisations offrent un aperçu global de la cooccurrence qui n'est plus figée. L'interface est dynamique et s'adapte au parcours du chercheur. Certains liens apparaissent en surbrillance et d'autres s'estompent lorsque l'utilisateur pointe un mot en particulier. Pour interpréter les résultats, le retour au texte est accessible en permanence depuis l'interface qui se divise alors entre sorties statistiques et texte brut.

Cette fonction trouve des usages complémentaires en analyse de texte. Elle donne un poids sémantique aux mots recherchés en les associant à des univers lexicaux particuliers. Elle révèle des choix de vocabulaire d'ordre stylistique ou idéologique dans le cas de l'analyse de discours politique ou médiatique. L'utilisation de filtres ou le changement de représentation des mots permet aussi d'aller vers des interrogations linguistiques plus complexes. Il est notamment possible d'étudier la variation de certaines catégories grammaticales ou même d'interroger la syntaxe en jouant sur les ***Paramètres*** pour étudier l'usage de certaines parties du discours autour d'un enchainement syntaxique particulier (avant, après, ou bidirectionnel). Cette fonction poussée de l'analyse des cooccurrences répond aux interrogations liées aux critères associatifs du langage de tous ordres.

Note technique : Le calcul des cooccurrences spécifiques est identique à celui des spécificités lexicales (section 2.3), mais les mesures effectuées sont différentes. La taille du corpus T reste identique (sauf si un filtre est appliqué, le taille est alors ramené à l'ensemble du texte filtré). La taille du texte t quant à elle correspond à la taille représentée par la somme des contextes où le mot/expression apparaît (la taille du contexte étant définie dans les ***Paramètres***. Les deux autres mesures f et k correspondent aux fréquences de l'ensemble des mots qui apparaissent dans le contexte de la recherche. f étant toujours la fréquence du mot dans l'ensemble du corpus et k celle du mot pris uniquement dans le contexte du mot ou de l'expression recherchée. Le calcul des spécificités est répété autant de fois qu'il y a de mots dans le contexte de la recherche considérée. Dans le cas de la poly-cooccurrence, cet algorithme est répété une fois en réduisant le contexte à tous les passages contenant à la fois le mot/expression et un des cooccurrents. Il est important de noter que ce calcul prend en compte de manière implicite la macro distribution des mots (l'organisation par métadonnée du corpus). Les cooccurrences observées sont donc sensibles aux sur-représentations lexicales dépendant de l'approche contrastive de l'analyse du corpus. Pour préférer une analyse dédiée à la micro distribution des mots et ignorer les spécificités lexicales, il est possible de ramener le corpus à un seul texte depuis menu ***Edition***. Il est alors nécessaire de répéter l'opération sur les autres textes pour pouvoir comparer les résultats.



Indice de spécificité	Probabilité	Fréquence corpus	Fréquence partie	Taille corpus	Taille partie	Mot
19.74	6.6e-71	6636	150	3009910	10266	pays
15.72	8.92e-48	253	37	3009910	10266	continent
12.4	8.66e-32	574	35	3009910	10266	époque
10.6	2.68e-24	190985	915	3009910	10266	,
9.98	6.38e-22	985	33	3009910	10266	peuple
9.84	2.03e-21	87	15	3009910	10266	Plan
9.58	1.86e-20	108927	558	3009910	10266	de
9.53	2.75e-20	1035	32	3009910	10266	économie
9.3	1.79e-19	406	22	3009910	10266	agriculture

Figure 11: Thème autour du mot "notre" dans les discours du président de Gaulle

La figure 11 montre un exemple d'usage de cette fonction sur le discours de C. de Gaulle. Le mot « notre », spécifique du président, s'associe à d'autres mots particuliers qui attestent de choix discursifs. Tout d'abord le discours rassemble largement en partageant avec l'allocutaire le temps et l'espace avec « notre... pays », « notre... continent » ou « notre... époque ». A l'analyse, il n'inclut pas nécessairement le locuteur dans le « peuple » comme suggéré par les extraits en concordance : « … en même temps *notre peuple* m'a fait confiance …» [7] ou « … en voulant que *notre peuple* soit un bon compagnon … » [8]. Enfin il fédère sur des questions socio-économiques avec « notre… économie » ou « notre… agriculture ».

Un des usages privilégiés de la fonction *Thème* dans une approche d'analyse contrastive de corpus, est la comparaison des sorties entre les différentes métadonnées. Ainsi, en sectionnant deux présidents il est possible de noter des variations d'usages qui interrogent sur l'usage de la première personne du pluriel, ici entre de Gaulle et Macron (figure 12).

Figure 12: Cooccurrences de "notre" dans le discours de C. de Gaulle à gauche et E. Macron à droite

[7] Extrait de l'allocution radiotélévisée du 10 octobre 1962.
[8] Extrait des vœux du président du 31 décembre 1965.

La figure 12 (représentation de la cooccurrence sous forme de nuage de mots), indique que le discours d'E. Macron reprend certaines cooccurrences de son prédécesseur tout en introduisant de nouveaux marqueurs. Le temps et l'espace sont toujours représentés, mais l'« époque » chez de Gaulle laisse place à l'« histoire » chez Macron, le « pays » reste présent mais le « continent » disparaît et le « territoire » ou la « Nation » prennent sa place. Avec « souveraineté », un mot apparaît et remplace le « peuple » ou le complète.

Manuel d'utilisation : Cette fonction s'accompagne de nouveaux menus qui apparaissent sur le graphique. Ces menus permettent de passer d'un type de graphique à l'autre, de télécharger les données ainsi que les graphiques sous forme d'image. Ces menus sont partagés par l'ensemble des fonctions utilisant des tableaux de données et des visualisations graphiques.

3.3 Distributions statistiques

Le dernier onglet du moteur de recherche d'Hyperbase affiche la ***Distribution*** statistique des mots. Cette fonction croise les mots ou expressions recherchées avec les textes dans un tableau de contingence où les colonnes représentent les métadonnées et les lignes des unités observées. Ce type d'analyses multivariées peut être abordé de différentes manières suivant la densité du tableau. Une recherche simple d'un seul terme s'apprécie en général avec un histogramme classique et permet de visualiser la mesure effectuée. Dans le cas d'Hyperbase il s'agit du calcul des spécificités (section 2.3) et le graphique rend compte des écarts positifs ou négatifs pour chaque métadonnée.

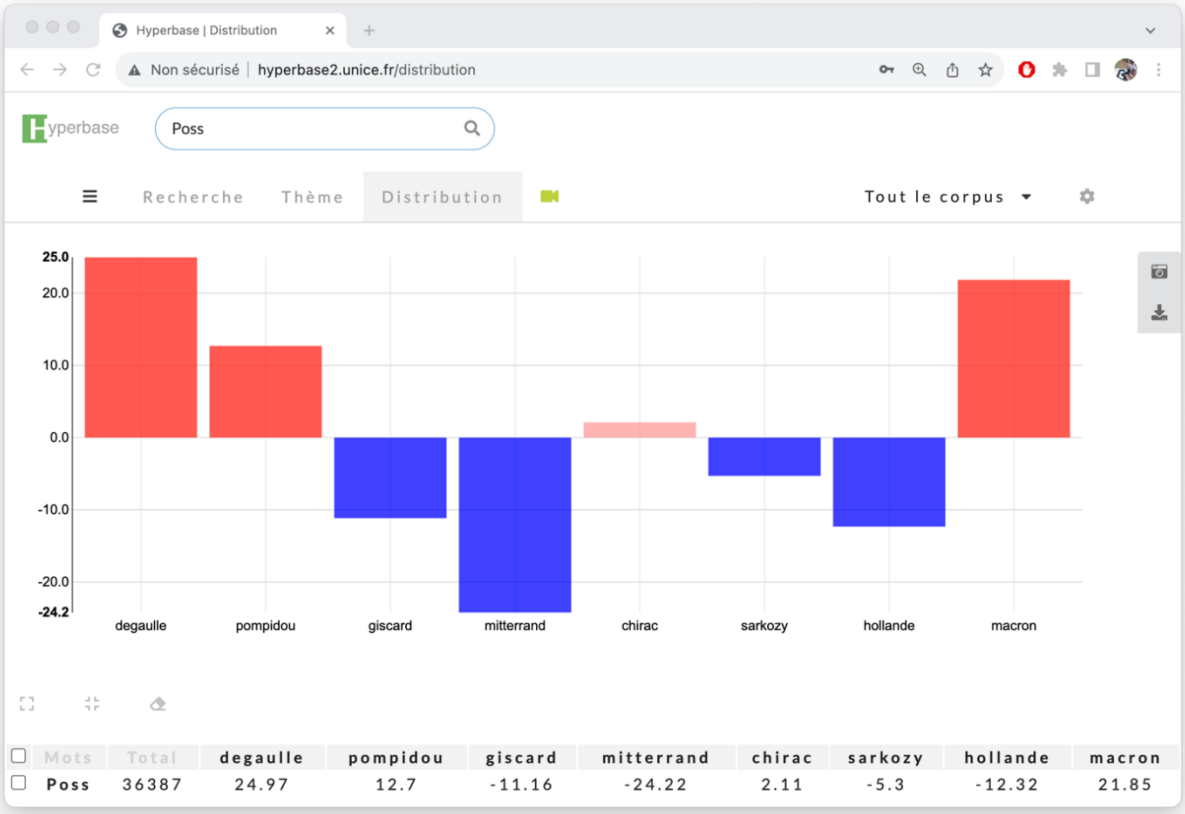


Figure 13: Distribution du mot « notre » dans le corpus « Elysee »

Dans l'exemple de la figure 14 c'est le code grammatical « Poss » qui sert de requête. L'idée est de tester ici l'emploi du possessif dans l'ensemble du corpus. De Gaulle apparaît comme le premier représentant de cette catégorie, mais il est suivi de près par Macron. Cet ordre hiérarchique est

surprenant puisqu'il rompt avec la chronologie du corpus, et interroge donc sur les usages du possessif au début et à la fin de la Vème République.

Manuel d'utilisation: Hyperbase permet, par son moteur de recherche, d'analyser un ou plusieurs mots à la fois, soit en les listant dans le champ de recherche soit à partir d'un code grammatical en demandant le détail via le tableau des données situé sous le graphique. Plusieurs icônes apparaissent au-dessus de l'en-tête du tableau permettant de manipuler les lignes sélectionnées (case à cocher en début de ligne). Parmi les actions, il est possible d'« Afficher les plus fréquents », « Regrouper les mots » ou « Effacer les mots de la liste ». Toutes ces actions modifient les mesures effectuées et changent la visualisation en fonction des besoins. La représentation du tableau de contingence passe automatiquement d'un histogramme à une AFC lorsque le nombre de colonnes ou de lignes dépasse les 50 entrées. Les outils disponibles sur le graphique (menu vertical à droite du graphique) permettent à tout moment de basculer d'une visualisation à l'autre et d'utiliser des fonctions dédiées comme l'affichage d'éléments supplémentaires sur une AFC. À noter la possibilité de convoquer certaines analyses complémentaires comme l'arborée par ces mêmes outils si le nombre de textes le permet (> 4 pour une Arborée comme pour une AFC).

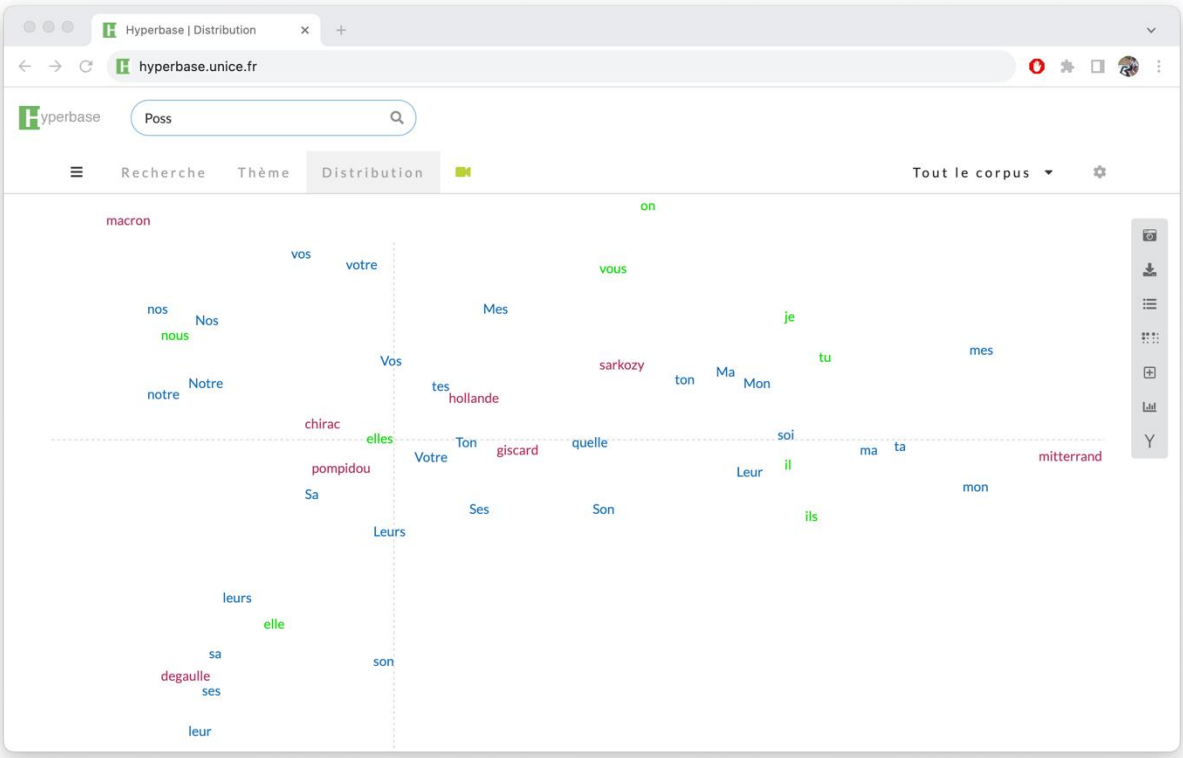


Figure 14: AFC des Possessifs dans la base « Elysee » - Les pronoms personnels en vert sont des éléments supplémentaires

En demandant les possessifs les plus fréquents dans l'ensemble du corpus, l'AFC produite par Hyperbase montre une opposition dans le discours entre pluriel et singulier et entre l'usage de la première, deuxième et la troisième personne (figure 14). Trois pôles se distinguent, celui de Mitterrand, de Macron et de de Gaulle. On le sait depuis (Mayaffre, 2012), Mitterrand introduit l'usage du pronom personnel « je » dans le discours présidentiel sous la Vème République. Les déterminants et pronoms possessifs associés marquent cette singularité dans l'AFC. Le graphique montre ensuite que le « je » deviens « nous » (éléments supplémentaires projetés sur l'AFC) avec Macron plus consensuel et cherchant davantage à représenter le collectif. La distinction avec de Gaulle se fait au niveau de l'usage de la troisième personne (plus particulièrement du féminin, une spécificité statistique du président avec un score de +18.35 ; voir la figure 5). Avec le pronom de la non-personne, le possessif marque une tournure impersonnelle et historique du discours où le chef

de l'Etat est extérieur à la situation, surplombant ainsi les décisions politiques du pays : « … les modalités de *leur* destin seront délibérées dès que l'apaisement sera venu … », « … la France , par *ses* intérêts , *ses* responsabilités , est continuellement impliquée … », …

Hyperbase permet de pousser ce type d'étude suivant plusieurs axes. Par exemple en recherchant des expressions plus complexes suggérées par d'autres analyses. C'est le cas du motif repéré avec le *deep learning*, « de + Possessif », illustré par la figure 10, qui obtient le score maximal de +6.07 pour de Gaulle par rapport à l'ensemble des autres présidents. Une autre approche heuristique avec la fonction ***Distribution*** consiste à convoquer une analyse généralisée des mots les plus fréquents du corpus (utilisation de l'expression « .* » qui couvre l'ensemble des formes graphiques repérées par le moteur). Cette méthode permet l'exploration des traits caractéristiques à visualiser de préférence avec une AFC. La distribution de l'ensemble des parties du discours est aussi possible en utilisant le menu ***Recherche avancée,*** une manière d'étudier le style de chaque auteur en visualisant les proximités entre les catégories grammaticales et les textes. L'analyse arborée enfin reste le moyen le plus sûr de contrôler les regroupements pressentis à partir d'une distribution.

Loin d'être exhaustifs ces exemples montrent un aperçu de l'étendue des outils proposés aux chercheurs par Hyperbase. Le moteur de recherche qui se décline en retour au texte, analyse des cooccurrences et distributions statistiques offre des possibilités d'analyses quasi illimitées sur le texte, les métadonnées et le corpus en général.

4 Conclusion

Hyperbase Web propose aux chercheurs une suite d'outils d'ADT, du plus basique au plus sophistiqué. L'évolution de l'informatique permet des analyses toujours plus poussées et entraine des changements de paradigme permanent. Cependant le logiciel n'a pas vocation à objectiver le langage mais simplement à décrire les textes. Le langage humain reste un phénomène complexe qui nécessite, pour être étudié, d'être matérialisé. Le logiciel fait le choix des mots comme matière première et du corpus comme objet d'étude. L'interface offre un parcours interprétatif complet en maintenant un lien permanent entre mesure quantitative et retour au texte qualitatif. S'articulant autour d'un moteur de recherche avancé, les observables linguistiques extraits des textes sont mis en perspective par des moyens graphiques qui illustrent l'organisation des mots dans le corpus. Ces artefacts du texte expriment surtout des réalités statistiques qui ne cherchent pas à naturaliser l'humain et la culture (ni à humaniser la machine) mais à tendre vers des analyses scientifique du langage en acte, à travers des descriptions de phénomènes idiolectaux ou sociolinguistiques particuliers. En d'autres termes, si le langage est à la source des Humanités, Hyperbase est une tentative pour aider à remonter le cours de la rivière.

Bibliographie

V. Beaudouin (2016). « Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données ». *Proceedings of 13th Internationl Conference on Statistical Analysis of Textual Data* - JADT 2016, vol. 1. pp.17-27

J.-P. Benzecri (1973). *L'Analyse des données*. Tome 2 : L'Analyse des correspondances, Dunod.

C. Bouzereau, D. Mayaffre et V. Montagne (eds.) (2022). « Le roi disait « nous voulons ». Usages et fonctions du nous dans le discours politique ». *Les cahiers de praxématique*.

E. Brunet, L. Lebart, L. Vanni. (2021). « Littérature et intelligence artificielle ». In *L'intelligence artificielle des textes*, , Honoré Champion, pp.73-130.

E. Brunet. (2016). *Tous comptes faits. Questions linguistiques*, textes édités par Bénédicte Pincemin, préface de François Rastier, Paris, Champion, 2016.

E. Brunet. (1999). Qui lemmatise dilemme attise. In *11e Rencontres linguistiques en pays rhénan*, L Kosé, A. Theissen, Nov 1999, Strasbourg, France. pp.7-32.

E. Brunet. (2003). Peut-on mesurer la distance entre deux textes ? In *CORPUS* N°2 - 2003

E. Brunet (2009), *Comptes d'auteurs. Études statistiques de Rabelais à Gracq*, Paris, Champion.

E. Brunet. (2010). HYPERBASE Manuel de référence. 2010.

E. Brunet. (2011). *Ce qui compte. Méthodes statistiques*, textes édités par Céline Poudat, préface de Ludovic Lebart, Paris, Champion, 2011.

B. de Courson, B. Azoulay, C. de Courson, L. Vanni, E. Brunet. (2023) « Gallicagram : les archives de presse sous les rotatives de la statistique textuelle », *Corpus*, 24.

P. Lafon (1980). « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, 1(1) :127–165.

L Lebart, A. Salem and L. Berry (1998). *Exploring Textual Data*. Springer.

D. Mayaffre. (2012). *Le discours présidentiel sous la Ve République. Chirac, Mitterrand, Giscard, Pompidou, de Gaulle.* Honoré Champion. 2021. Presses de Sciences Po.

D. Mayaffre (2021). *Macron ou le mystère du verbe. Ses discours décryptés par la machine*, Editions de l'Aube.

D. Mayaffre, L. Vanni. (2021) *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, 15, Honoré Champion. Lettres Numériques, 2021.

D. Mayaffre, J.-M. Viprey. (2012). La cooccurrence. Du fait statistique au fait textuel. *In CORPUS*, N°11, 2012.

D. Mayaffre, « Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'intelligence artificielle », *7ème Congrès Mondial de Linguistique Française*, 2020. [hal-02520224]

S. Mellet S. et Barthélemy J.-P. (2007), « La topologie textuelle : légitimation d'une notion émergente », Lexicometrica 7. [http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mellet.pdf]

S. Mellet and D. Longrée (2009). Syntactical motifs and textual structures. In *Belgian Journal of Linguistics 23*, pages 161–173.

Pincemin, B. et al. (2006). « Concordanciers : thème et variations », in J.-M. Viprey (éd.), *JADT 2006*, pp. 773-784.

F. Rastier. (2011). *La mesure et le grain. Sémantique de corpus.* Paris : Champion, Collection Lettres numériques, 2011.

A. Salem. (1991) « Les séries textuelles chronologiques », *Histoire & Mesure*, vol. VI, n°1/2, 1991, p.149-175.

Tognini-Bonelli, E. (2001). « Corpus linguistics at work », *Computational Linguistics*, 28:583–583

L. Vanni, M. Ducoffe, C. Aguilar, F. Precioso and D. Mayaffre. (2018). Textual deconvolution saliency (TDS): a deep tool box for linguistic analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 548–557.

L. Vanni, H. Mahmoudi, D. Longrée and D. Mayaffre. (2024, sous presse). « Multi-channel Convolutional Transformer and intertextuality : a Latin case study ». In *JADT2022*, Springer.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones and al.. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Viprey J.-M. (2006), « Structure non-séquentielle des textes », Langages, 163, p. 71-85.

K. Yoon (2014). « Convolutional Neural Networks for Sentence Classification ». In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.* 10.3115/v1/D14-1181.