# Théoriser le dynamique, modéliser la variation, et outiller l'herméneutique : le(s) sens en question(s)

Julien Longhi

Critical Hermeneutics, 4(1), 2020
Biannual International Journal of Philosophy
http://ojs.unica.it/index.php/ecch/index
ISSN 2533-1825 (on line); DOI 10.13125/CH/43198

Received: 6/6/2020
Accepted: 28/7/2020
Published: 9/10/2020

# Théoriser le dynamique, modéliser la variation, et outiller l'herméneutique : le(s) sens en question(s)

*(Theorising The Dynamic, Modeling the Variation, and Equipping Hermeneutics: The Meaning(s) in Question)*

**JULIEN LONGHI**

## Abstract

*This paper examines tooled methods used to analyze digital corpora, linked with theoretical orientations of scientific research. Indeed, building corpora with various textual data is not neutral, and the process of constituting corpora gives specific meaning to the data. Currently, the use of Artificial Intelligence (AI), or Deep Learning, questions semioticians, and linguists, on the possible interpretative processes from these treatments which often come from "black boxes". From the concepts of instability, deformability, and chaotic units, this paper considers the hermeneutical dynamic which governs the use of digital corpora and their apprehension by digital tools. A double problem arises: hermeneutics of digital corpora, and hermeneutics of digital tools which allow the analysis of digital corpora. To approach this, the description of the different scientific models (Meunier 2019) makes it possible to propose a tooled analysis method that meets the principles of the Theory of discursive objects (Longhi 2015). This method is illustrated with the analysis of the term "enemy" in a corpus of political interviews, thus posing the different*

*issues from a hermeneutical perspective for the digital analysis of discourse.*

**Keywords**: Digital discourse analysis, Digital humanities, Digital Hermeneutics, Discursive Semantics

**Résumé**

Cet article interroge les méthodes outillées d'analyses des corpus numériques, en lien avec les orientations théoriques propres aux *recherches qui y président. En effet, le rassemblement de données, souvent hétérogènes, dans des corpus, n'est pas neutre, et le processus même de constitution de corpus donne du sens aux données rapprochées ainsi dans ces ensembles. Actuellement, le recours à l'Intelligence Artificielle (IA), au Deep Learning, interroge les sémioticiens, et les linguistes, sur les processus interprétatifs possibles à partir de ces traitements qui sont souvent issus de « boites noires ». A partir des concepts d'instabilité, de déformabilité, et d'unités chaotiques, cet article envisage la dynamique herméneutique qui préside à l'usage des corpus numériques et à leur appréhension par des outils numériques. Une double problématique se pose : herméneutique des corpus numériques, et herméneutique des outils numériques qui permettent l'analyse des corpus numériques. Pour l'aborder, la description des différents modèles scientifiques (Meunier 2019) permet de proposer une méthode d'analyse outillée qui répond aux exigences de la Théorie des objets discursifs (Longhi 2015). Cette méthode est illustrée avec l'analyse du terme « ennemi » dans un corpus d'interviews politiques, posant ainsi les différents enjeux d'une perspective herméneutique pour l'analyse numérique des discours.*

**Mots-clés** : analyse du discours numérique, humanités numériques, herméneutique numérique, sémantique discursive

*Théoriser le dynamique, modéliser la variation, et outiller l'herméneutique : le(s) sens en question(s).* Cet article interroge les méthodes outillées d'analyses des corpus numériques, en lien avec les orientations théoriques propres aux recherches qui y président, dans le large champ que l'on peut appeler « humanités numériques ». En effet, alors que de plus en plus de données (notamment textuelles) deviennent accessibles, que ce soit par des processus de numérisation, ou par des productions nativement numériques, de nombreux projets d'analyses, d'extractions de connaissances, de constitution de banques de données, voient le jour. Or, le rassemblement de ces données, souvent hétérogènes, dans des corpus, n'est pas neutre, et le processus même de constitution de corpus donne du sens aux données (Anquetil, Duteil-Mougel et Llovera 2019) rapprochées ainsi dans ces ensembles. En particulier, le recours à l'Intelligence Artificielle (IA), au Deep Learning, interroge les sémioticiens, et les linguistes, sur les processus interprétatifs possibles à partir de ces traitements qui sont souvent issus de « boites noires ».

Il s'agira donc dans cet article de strates de l'interprétation, et les différents niveaux linguistiques et sémiotiques, qui contribuent à construire et stabiliser le sens. Nous montrerons notamment que les concepts d'instabilité, de déformabilité, et d'unités chaotiques, sont opérants pour considérer la dynamique herméneutique qui préside à l'usage des corpus numériques et à leur appréhension par des outils numériques. Une double problématique se posera à nous : herméneutique des corpus numériques, et herméneutique des outils numériques qui permettent l'analyse des corpus numériques. Plus précisément, à partir de la description des différents modèles scientifiques (Meunier 2019), nous développerons une méthode d'analyse outillée qui répond aux exigences du modèle théorique sur

lequel nous nous appuyons. Cet article vise donc à mettre concrètement en cohérence les différents aspects de la *Théorie des objets discursifs* (Longhi 2015, 2018) en formulant l'articulation entre théorie linguistique, formalisation conceptuelle, et outillage. Nous illustrerons cela avec un exemple d'analyse en lien avec le projet ANR TALAD (Traitement automatique des langues et analyse du discours)[1].

## 1. Herméneutique et analyse du discours numérique

La perspective théorique dans laquelle nous situons notre recherche relève de l'analyse du discours, entendue de manière « pragmatique », puisqu'elle recouvre à la fois une acception du discours comme processus, et du discours comme matérialité (Longhi 2018).

### *1.1. L'analyse du discours comme herméneutique*

L'objectif de notre recherche est d'élaborer une sémantique du discours, qui permette de rendre compte des mécanismes de construction du sens en discours. De ce point de vue, et si l'on considère les définitions données de l'herméneutique par le *TLFI*, il y a une grande cohérence à la convoquer :

1) Science des règles permettant d'interpréter la Bible et les textes sacrés, d'en expliquer le vrai sens.
2) Théorie, science de l'interprétation des signes, de leur valeur symbolique. *Appelons herméneutique l'ensemble des connaissances et des techniques qui permettent de faire parler les signes et de découvrir leur sens* (Foucault 1966: 44).

[1] Les points 2 et 3 de cet article ont été traduits en anglais et étoffés dans Longhi 2020 (paru finalement antérieurement pour des questions de calendriers éditoriaux) dans le cadre d'une réflexion plus spécifique sur l'analyse du discours intégrée aux humanités numériques.

3) *Emploi adj.* Qui concerne, qui a pour objet l'interprétation des textes religieux ou philosophiques, en particulier des Écritures saintes.

Dans les différentes dimensions étymologiques ou historiques, on note l'« art de découvrir le sens exact d'un texte » (1777, Encyclop. Suppl. t. 3), l'« interprétation de ce qui est symbolique » (1890, notamment avec l'herméneutique des couleurs chez Huysmans, et en 1803, en emploi adjectival, du grec « qui concerne l'interprétation, propre à faire comprendre » - comme dérivé d'interpréter ou traduire).

Pour préciser notre point de vue sur les corpus numériques, arrêtons-nous sur cette citation de Michel Foucault : *Appelons herméneutique l'ensemble des connaissances et des techniques qui permettent de faire parler les signes et de découvrir leur sens*. Dans les espaces numériques, informationnels ou sociaux en particulier, les usagers sont confrontés à des données textuelles, à des images, des vidéos, qui fonctionnent comme autant de signes, c'est-à-dire qui restent à interpréter. Ce qui « fait sens », c'est notamment la pratique numérique, identifiée par Milad Doueihi (2015: 711) au regard de « nouvelle sociabilité en ligne, peuplée de textes, animée par des « partages » », soit une « pratique numérique populaire » que doivent penser les travaux en humanités numériques. La question des sociabilités en ligne peut renvoyer à l'appareil conceptuel de l'analyse du discours, et il n'est selon nous pas nécessaire de forger de nouveaux concepts, mais plutôt de les réactualiser au regard de leur opérativité en contexte numérique. Ainsi, des analyses en termes de positionnements énonciatifs, d'interactions verbales, selon des considérations allant de l'argumentation à la sociolinguistique, sont de nature à permettre la prise en compte située des données textuelles, et plus largement

sémiotiques, échangées en ligne. Mais pour prendre en compte la spécificité de la matérialité sémiotique partagée et circulant en ligne nous proposons une articulation conceptuelle de l'herméneutique avec la philologie, et la phénoménologie.

### *1.2. L'herméneutique pour l'analyse du discours numérique : une extension à la philologie et la phénoménologie*

Le lien entre herméneutique et philologie est proposé par François Rastier (2001, chapitre 4 intitulé « herméneutique matérielle ») qui explique qu'« en réunifiant l'herméneutique et la philologie, l'herméneutique matérielle place la problématique de l'interprétation au centre des sciences du langage ». Bien qu'étant tributaire d'un modèle sémique (le sens « est produit dans des parcours qui discréditent et unissent des signifiés entre eux, en passant par des signifiants »), cette conception témoigne d'une vision dynamique de la construction du sens dans les corpus, à travers des parcours d'interprétation. Comme nous l'avons proposé ailleurs (Longhi 2018), nous pouvons considérer le discours, analysable à travers les corpus, comme un champ, à l'intérieur duquel ces parcours de sens se structurent. Ce « champ » est ainsi un observatoire qui permet de considérer les formes linguistiques (et sémiotiques) qui s'y déploient. On rejoint ainsi la phénoménologie, prise de manière générale comme l'« observation et description des phénomènes et de leurs modes d'apparition, considéré indépendamment de tout jugement de valeur », en prenant en compte la matérialité, synonyme de la « facticité » promue par Merleau-Ponty : « la phénoménologie, c'est l'étude des essences [...] c'est aussi une philosophie qui replace les essences dans l'existence et ne pense pas qu'on puisse comprendre l'homme et le monde autrement qu'à partir de leur « facticité ». » (Merleau-Ponty 1945: 1).

Dans *L'archéologie du savoir*, M. Foucault établissait un lien

entre herméneutique et polysémie, puisque « la polysémie – qui autorise l'herméneutique et la découverte d'un autre sens – concerne la phrase, et les champs sémantiques qu'elle met en œuvre : un seul et même ensemble de mots peut donner lieu à plusieurs sens, et à plusieurs constructions possibles ; il peut donc y avoir, entrelacées ou alternant, des significations diverses, mais sur un socle énonciatif qui demeure identique » (144). Du point de vue linguistique, Yves-Marie Visetti a développé une recherche intéressante relativement à l'orientation précédemment évoquée : dans *Formes et théories dynamiques du sens*, il indique :

> L'expérience ne se *présente* pas dans le cadre d'une intuition de type kantien, elle ne se factorise pas, en quelque sorte, à travers un cadre spatio-temporel vide (en particulier vide de tout engagement !). Elle est faite d'anticipations, perçues et opérant comme telles au sein du Présent – constituant même ce présent (36).

Il utilise la phénoménologie « comme un discours objectivant d'un type particulier, qui fait jouer l'Etre-au-Monde, ainsi qu'à certaines structures du champ de conscience (formes et champ thématique), le rôle d'un « modèle » général, partout transposable.

Pour lui il faut investir « le « thème » de l'herméneutique, à condition ensuite de donner aux questions linguistiques et sémiotiques toute leur importance *à ce niveau fondamental*, et le primat de la perception est phénoménologiquement compris comme le primat d'un *sens* perceptif. Dès lors le concept d'énonciation est pertinent pour l'analyse du discours numérique, puisque l'énonciation n'est pas considérée comme une sortie du langage, ni même une « actualisation » de la langue en discours.

### *1.3. Un modèle conceptuel pour les humanités numériques*

La mise en œuvre de cet arrière-plan théorique a notamment été réalisée dans le cadre de la théorie des formes sémantiques (Cadiot et Visetti 2001), présentée de manière synthétique par Cadiot (2009) :

> Nous défendons depuis longtemps l'idée que des propriétés qui n'auraient pas de corrélat intentionnel dans le monde des pratiques et des expériences n'ont aucune place en sémantique lexicale. Les mots doivent d'abord être vus comme des index de discours, donc dans leur texture locale. Les motifs lexicaux (Cadiot & Visetti 2001) ne sont que des supports d'élaboration pour des opérations de profilage (leur mise en syntagme, en phrase, en texte, leurs collocations ou figements) et de thématisation (leur vocation à parler du monde, leur dimension intentionnelle, ou, si l'on veut, référentielle). Ils ne se stabilisent que par l'entremise d'opérations textuelles dont une cristallisation sensible est l'ensemble des effets de figement, semi-figement, délocutivité formulaire

En particulier, dès lors que l'on utilise des outils d'analyse/d'exploration de corpus, l'échantillonnage des textes, les réorganisations textuelles, les différents niveaux de repérage des observables, peuvent poser problème. Or une vision non compositionnelle du sens permet d'appréhender les choses sous un autre angle. Pour Cadiot (2009) :

> Contrairement à une certaine vulgate *bottom / up*, que l'on trouve clairement dans la notion de compositionnalité, la mise en syntagme n'est jamais une simple instanciation

« libre » et nous renvoie donc à cette notion de texture locale.

Cette « texture locale » est intéressante, car des analyses textométriques peuvent permettre de rendre compte de cette mise en syntagme. Rappelons que la textométrie « (ou statistique textuelle, lexicométrie, logométrie) propose une approche instrumentée des corpus, articulant synthèses quantitatives et analyses à même le texte » (Lebart, Salem 1994 cités par Pincemin 2012). En plus de fournir des procédures de tri et de calculs statistiques pour l'étude de corpus numériques, elle « établit une modélisation contextuelle et contrastive : le texte est caractérisé par ses mots par rapport à leur usage dans le corpus, le mot est caractérisé par ses cooccurrents, etc. » (Pincemin 2012 : en ligne).

## 2. Modèles et outils

Ce détour par le modèle théorique de sémantique qui sous-tend l'approche discursive qui est la nôtre est nécessaire, car l'architecture analytique que nous allons mobiliser, à la fois en termes de méthodologie, mais aussi d'outils, doit être ancrée dans une certaine conception du discours.

### *2.1. Les quatre types de modèles et leur application à notre recherche*

Meunier (2019: 23) explicite le lien nécessaire entre quatre types de modèles : le modèle conceptuel, le modèle formel, le modèle computationnel, et le modèle informatique. L'objectif ici est de montrer la cohérence de notre démarche d'analyse, à travers la prise en compte de ces quatre modèles. Si notre modèle conceptuel est la théorie des formes sémantiques, appliquée aux corpus numériques, et mise en œuvre grâce aux ressources textométriques, il convient de

mettre en valeur la manière dont les concepts sont mobilisés dans le modèle formel, dans le modèle computationnel, et comment ils peuvent être traités avec un programme qui relèverait du modèle informatique.

Nous allons donc détailler des différents modèles, et leur application à notre recherche.

(1) Le modèle conceptuel exprime dans un langage naturel (concepts, énoncés, arguments, discours, etc.) les objets, opérations et méthodes qui sont épistémiquement pertinents pour une recherche dans les humanités. Ce modèle est limité : son mode d'expression (le langage naturel) est grevé d'ambiguïtés, de biais, d'impressions. Il n'en demeure pas moins essentiel dans la démarche de recherche, car il est le seul langage qui nous est immédiatement accessible » : ici, le modèle conceptuel de la *Théorie des formes sémantiques* convoque les concepts de motifs, profils, et thèmes, qui, en plus d'être polysémiques, sont utilisés dans plusieurs cadres théoriques. Il convient donc de pouvoir transcrire ces concepts, de manière fidèle à leur définition, et de les rendre fonctionnels pour une analyse outillée ;

(2) « Dans le cadre de projet sur des « humanités » que l'on veut « numériques », le modèle formel aura pour rôle de traduire certains éléments du modèle conceptuel dans un langage formel (mathématique, géométrique, logique, grammatical, etc.).

Cette « traduction » est un point de travail en cours, et ne pourra pas être explicitée dans le cadre de cet article. Nos recherches sur le sujet renvoient notamment aux travaux d'Yves-Marie Visetti qui indique notamment à propos des liens entre modèle conceptuel et modèle formel dans ce cade théorique :

- Notre démarche – encore une fois orientée par la question des formes en sémantique, et non directement par celle du

> continu – a consisté en un retour critique aux écoles historiques de la Gestalt, et en même temps à la philosophie phénoménologique, parcourue le long d'un axe allant de Husserl à Merleau-Ponty en passant par Gurwitsch. Nous avons tenté de développer sur cette base un mode phénoménologique de théorisation, bien distinct des modes formels, même si un certain type de modélisation mathématique (précisément celui évoqué ci-dessus, dans la filiation de R. Thom) nous a servi de tremplin (2004, en ligne);
>
> - l'intérêt d'une mathématisation des théories et des techniques des disciplines cognitives est encore trop largement méconnu, notamment en IA : il ne s'agit pas seulement ici d'obtenir une plus grande généralité et un meilleur contrôle sur les conditions applicatives, mais il s'agit également de permettre une véritable schématisation des concepts théoriques descriptifs par l'intermédiaire des structures à la fois formelles et intuitives des mathématiques. Dans cet esprit, les méthodes topologiques, géométriques, dynamiques, doivent être promues au même titre que les méthodes symboliques ou numériques dans la construction des modèles (2003, en ligne).

Il nous semble donc que des formalisations selon des modèles topologiques, géométriques et dynamiques, dont des pistes pertinentes pour « traduire » les concepts en langage formel. L'écho des modèles 3 et 4 que nous allons présenter peut aussi être une voie d'accès aux enjeux formels de notre proposition, car ils permettent de les rendre compréhensibles.

(3) « Pour sa part, le modèle computationnel est directement lié aux modèles formels. Son rôle est de traduire les énoncés calculables

du modèle formel dans des énoncés d'un langage computationnel, c'est-à-dire en des algorithmes ou des programmes » (Meunier 2019).

Dans le cadre des recherches que nous avons menées en humanités numériques, la transcription des modèles dynamiques et topologiques précédemment évoqués ont trouvé un écho dans les analyses de similitudes d'une part, et dans un certain usage des classifications hiérarchiques descendantes d'autre part (et de leurs représentations graphiques). Cet écho, comme indiqué à la partie précédente, est en cours de caractérisation précise, par une réflexion plus approfondie sur les modèles formels, les conceptions mathématiques qui les sous-tendent, et les enjeux du passage entre modèle conceptuel et modèle formel. Précisions néanmoins les caractéristiques du contenu du modèle computationnel envisagé :

- A propos des analyses de similitude, Loubère (2016) explique que ce modèle est issu de « la théorie des graphes (Flament 1962, 1981; Vergès & Bouriche 2001), et qu'il représente la structure d'un corpus par la schématisation de ces relations, permettant ainsi de faire ressortir les liens des formes dans les segments de textes (Marchand & Ratinaud 2012) ». Plus précisément, Marchand et Ratinaud (2012) expliquent qu'« après segmentation, reconnaissance et lemmatisation des formes, puis partition en UCE, la matrice du corpus global peut être représentée de diverses façons (arbres linéaires ou circulaires; taille des formes proportionnelle à la fréquence ou à la liaison statistique…). On représente ici l'arbre des liaisons lexicales du corpus (calcul de cooccurrence et algorithme de Fruchterman-Reingold ». Le calcul de cooccurrence, et l'algorithme de

Fruchterman-Reingold[2], sont conçus ici comme le moyen de prendre en compte les « profils », car ils rendent compte à la fois de la proximité syntaxique et la fréquence des associations, et la force de la relation entre les unités ;

- A propos de la classification hiérarchique descendante, nous suivons Loubère (2016) dans le choix d'une classification de type Reinert proposée par le logiciel Iramuteq : « cette classification implémentée pour la première fois dans le logiciel Alceste® » (Reinert 1983) permet de mettre en avant les mondes lexicaux. Ces structures du discours partent du principe que l'énoncé est un point de vue dépendant du sujet mais aussi de son activité et son contexte, où « le vocabulaire d'un énoncé particulier [est considéré] comme une trace pertinente de ce « point de vue » il est à la fois la trace d'un lieu référentiel et d'une activité cohérente du sujet-énonciateur. Nous appelons mondes lexicaux, les traces les plus prégnantes de ces activités dans le lexique » (Reinert 1993) ». Au niveau de la méthodologie, comme décrit par Loubère, « après lemmatisation les textes sont segmentés, puis la ponctuation est supprimée ; à partir de ce matériau est construit un tableau à double entrée répertoriant la présence ou absence dans les segments des formes pleines retenues ; sur ce tableau est effectuée une série de bi-partitions reposant sur une analyse factorielle des correspondances » ;
- En termes de visualisation et de représentation des résultats, l'analyse factorielle des correspondances peut être utilisée : c'est

[2] https://github.com/gephi/gephi/wiki/Fruchterman-Reingold: The Fruchterman-Reingold Algorithm is a force-directed layout algorithm. The idea of a force directed layout algorithm is to consider a force between any two nodes. In this algorithm, the nodes are represented by steel rings and the edges are springs between them. The attractive force is analogous to the spring force and the repulsive force is analogous to the electrical force. The basic idea is to minimize the energy of the system by moving the nodes and changing the forces between them. For more details refer to the Force Directed algorithm.

> une méthode statistique « qui s'applique aux tableaux de contingence, tels par exemple les tableaux résultant du décompte de différents types de vocabulaire (lignes du tableau) dans les différentes parties (colonnes du tableau) d'un corpus de textes » (Salem, Tutoriel du logiciel Lexico3). On commence par « calculer une distance (dite distance du χ2) entre chacune des paires de textes qui constituent le corpus. On décompose ensuite ces distances sur une succession hiérarchisée d'axes factoriels. [...] Cette méthode permet d'obtenir des représentations synthétiques portant à la fois sur les distances calculées entre les textes et celles que l'on peut calculer entre les unités textuelles qui les composent ». Il est néanmoins important de noter que si « l'intérêt principal de l'AFC réside dans sa capacité à extraire à partir de vastes tableaux de données difficilement appréhendables des structures simples qui rendent compte approximativement des grandes oppositions sous-jacentes dans un corpus de textes », il s'agit d'une « approximation », et que les résultats des fonctions précédentes (calculs, tableaux de chiffres) doivent être considérés précisément.

Ici, par la pratique que nous avons, le logiciel *Iramuteq* permet de rassembler ces différents algorithmes et fonctionnalités. Le recours au logiciel Lexico (3 puis 5) est également fréquent, notamment pour les fonctions de segments répétés, et d'AFC, qui peuvent permettre de rendre compte des dynamiques temporelles d'un corpus.

Ce choix est motivé par notre double intérêt conceptuel pour les liens entre formes et profils d'un côté, et formes et thèmes de l'autre. On notera que ces algorithmes ne répondent pas directement aux modèles dynamiques/topologiques évoqués à propos du modèle formel. L'articulation proposée réside dans le fait que les travaux que

nous menons en analyse du discours sont fondés sur la variation et la comparaison/différentialité : aussi, c'est l'application variationnelle des algorithmes d'Iramuteq (ou de Lexico) qui peut permettre de rendre compte des dynamiques des corpus (comparer des états de discours, avec des « instantanés » de différents stades, compris comme un tout qui varie en différents moments discursifs). Les illustrations de la partie 3 rendront ces considérations plus claires. Un exemple pour saisir les dynamiques en corpus, d'un point de vue chronologique, est d'avoir recours aux séries textuelles chronologiques. Comme l'explique André Salem dans un tutoriel pour Lexico 3 (mais ces fonctions sont généralisables à d'autres outils de textométrie), les séries textuelles chronologiques « sont des corpus constitués par la réunion de textes similaires produits par une même source textuelle au cours d'une période de temps » : la prise en compte de la dimension chronologique de tels corpus permet « de mettre en évidence des variations qui surviennent au cours du temps dans l'emploi du vocabulaire, de mettre en évidence des moments importants dans l'évolution de celui-ci ». On s'inscrit donc pleinement dans l'analyse des dynamiques du sens propres à un corpus, ici en lien avec la dimension temporelle.

(4) « Enfin, le modèle informatique traduit dans des formes mécaniques de types électroniques les algorithmes créés dans le modèle computationnel. Il offre une architecture matérielle qui permet d'effectuer concrètement les calculs ou la computation dans les modèles computationnels » (Meunier 2019).

Selon le modèle théorique que nous avons présenté, le modèle informatique qui offre selon nous la meilleure « architecture matérielle » est l'ordinateur (permettant l'installation du logiciel) ou la plateforme, qui intégrerait dans son fonctionnement les algorithmes décrits (c'était le cas pour la plateforme #Idéo2017). Il reste à expliciter la manière dont les algorithmes décrits peuvent

rendre compte des concepts de la théorie des formes sémantiques.

### *2.2. Profilage et analyse de similitudes, thématiques et classification lexicale*

Pour Cadiot et Yves-Marie Visetti (2001: 130), les dynamiques de profilage « renvoient pour une part à des frayages déjà enregistrés en lexique et, sous une forme bien plus générique, en grammaire. Mais elles se font aussi par inscription dans des thématiques inédites, qui les reprennent dans leur propre grille, possiblement extrinsèque, soit aux motifs donnés en langue, soit aux normes de profilage lexical déjà attestées ». Aussi, l'Analyse de similitudes (ADS), qui établit un calcul sur la base d'« un indice de co-occurrence (combien de fois les éléments vont apparaitre en même temps) » pour donner un résultat visuel « où la taille des mots est proportionnelle à la fréquence et où la taille des arêtes et proportionnelle à la force », permet de rendre compte de ces opérations de profilage, leurs frayages, leur stabilité, etc. Ceci permet de rendre compte visuellement de la fréquence des mots en lien avec les associations spécifiques. Cette fonctionnalité représente d'une certaine manière le « profilage » des unités, c'est-à-dire qu'elle rend compte de leur stabilisation dans le corpus à travers des associations fréquentes qui « profilent » les usages des formes dans tel ou tel domaine ou pratique.

Concernant les thématiques, Cadiot & Visetti (2001) expliquent que le thème « traduit la stabilisation et l'actualisation dans et par un domaine "référentiel", voire aussi "conceptuel" ». Aussi, la classification lexicale implémentée dans Iramuteq qui permet de faire ressortir les thématiques propres à un corpus et de regrouper des « mondes lexicaux », peut correspondre à cet enjeu de caractérisation des thématiques (elle vise à « rendre compte de l'ordre interne d'un discours, à mettre en évidence ses mondes

lexicaux »[3].

Pour illustrer ces choix et cette mise en cohérence des différents modèles, nous allons proposer une illustration à partir d'un exemple.

## 3. Illustration du questionnement numérique du sens dans un corpus numérique

L'illustration de ces questionnements et choix méthodologique est ici appliquée, de manière illustrative, au corpus du projet ANR TALAD. Pour contextualiser, ce choix, le projet TALAD vise à montrer comment le Traitement automatique des langues (TAL) permet à l'Analyse du discours (AD) d'aller plus loin dans ses explorations, d'éprouver son appareil théorique et de renforcer son outillage méthodologique. Il s'agit d'adapter des techniques TAL pour fournir à l'AD des jeux de descripteurs plus complexes, relatifs à différents niveaux d'organisation discursive. En retour, l'AD offrira un éventail de phénomènes complexes à étudier qui seront autant de défis à soumettre aux dernières avancées en TAL.

### *3.1. L'« ennemi » dans la dynamique discursive politique*

Nous nous sommes focalisés sur les transcriptions automatiques d'interviews « matinales » corrigées manuellement : un jeu de données qui contient actuellement 3166 interviews correspondant à 561 personnalités politiques interviewées, entre le 10 juin 2016 et le 4 décembre 2017 (soit environ 10 millions de mots). Pour faire écho à un autre travail réalisé avec André Salem sur le corpus du *Père Duchène* (Longhi et Salem 2018), l'analyse est ici centrée sur le terme « ennemi ».

Pour s'intéresser à ce terme (ici considéré par le lemme /ennemi/), nous avons fait le relevé de tous les lemmes, afin de

[3] https://datahist.hypotheses.org/11

pouvoir établir ensuite le concordancier de tous les segments de textes contenant ce lemme :

| | | |
|---|---|---|
| heureusement | 137 | adv |
| inquiétude | 137 | nom |
| tendance | 137 | nom |
| électorat | 137 | nom |
| compétence | 136 | nom |
| exercice | 136 | nom |
| immobilier | 136 | adj |
| naturellement | 136 | adv |
| pollution | 136 | nom |
| présence | 136 | nom |
| soumettre | 136 | ver |
| unique | 136 | adj |
| confronter | 135 | ver |
| correspondre | 135 | ver |
| ennemi | 135 | nom |
| ficher | 135 | ver |
| former | 135 | ver |
| moitié | 135 | nom |
| oser | 135 | ver |
| évoluer | 135 | ver |
| banlieue | 134 | nom |
| carte | 134 | nom |
| coalition | 134 | nom |
| cohérent | 134 | adj |

1. Fréquence des lemmes dans le corpus

Et nous avons procédé à l'extraction d'un sous-corpus qui rassemble tous ces segments de texte :

****** *theme_J_Chirac *theme_Patron_du_parti *theme_Extreme_droite *theme_General_de_Gaule *theme_Gaullisme *theme_Synthese *theme_L_Wauquiez *theme_Constructifs *theme_Veut_rassembler_et_exclut *theme_Plus_de_discours_a_droite *theme_Que_la_France_se_remette_a_marcher *theme_E_Macron *theme_P_Seguin *theme_N_Sarkozy *theme_JP_Raffarin *theme_Europe**

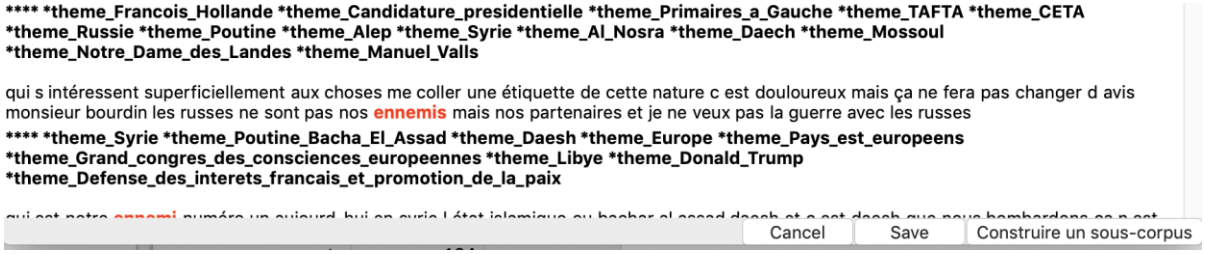
c est toujours un espoir d aller comme disait la gauche jamais d **ennemi** à gauche à droite on dit

****** *theme_J_Chirac *theme_Patron_du_parti *theme_Extreme_droite *theme_General_de_Gaule *theme_Gaullisme *theme_Synthese *theme_L_Wauquiez *theme_Constructifs *theme_Veut_rassembler_et_exclut *theme_Plus_de_discours_a_droite *theme_Que_la_France_se_remette_a_marcher *theme_E_Macron *theme_P_Seguin *theme_N_Sarkozy *theme_JP_Raffarin *theme_Europe**

jamais d **ennemi** droite hé bien on retrouve les mêmes choses mais encore une fois à l époque il y avait un patron qui fixait et ce patron il était parfois contesté mais il était admis par tout le monde

****** *theme_Proces_Merah *theme_Justice_laxiste *theme_Bracelet_electronique *theme_D_Trump *theme_Attentat_New_York *theme_Peine_de_mort *theme_M_Le_Pen *theme_L_Wauquiez *theme_Liberation_534_elus_FN_de_2014_ne_siegent_plus *theme_Une_France_village_d_Asterix *theme_Viol_a_Calais *theme_Refondation_du_FN *theme_Alliance_LR_FN *theme_J_L_Melenchon *theme_A_Berge *theme_Sens_Commun *theme_Europe *theme_Opposition_a_LREM *theme_E_Macron *theme_Sondage_IFOP**

écoutez donald trump réagit à la mesure de ces attentats qui touchent le peuple américain c_est_à_dire qu il a pris conscience de qui était l **ennemi** d où il venait au nom de quoi il combattait




2. Extraction d'un sous-corpus

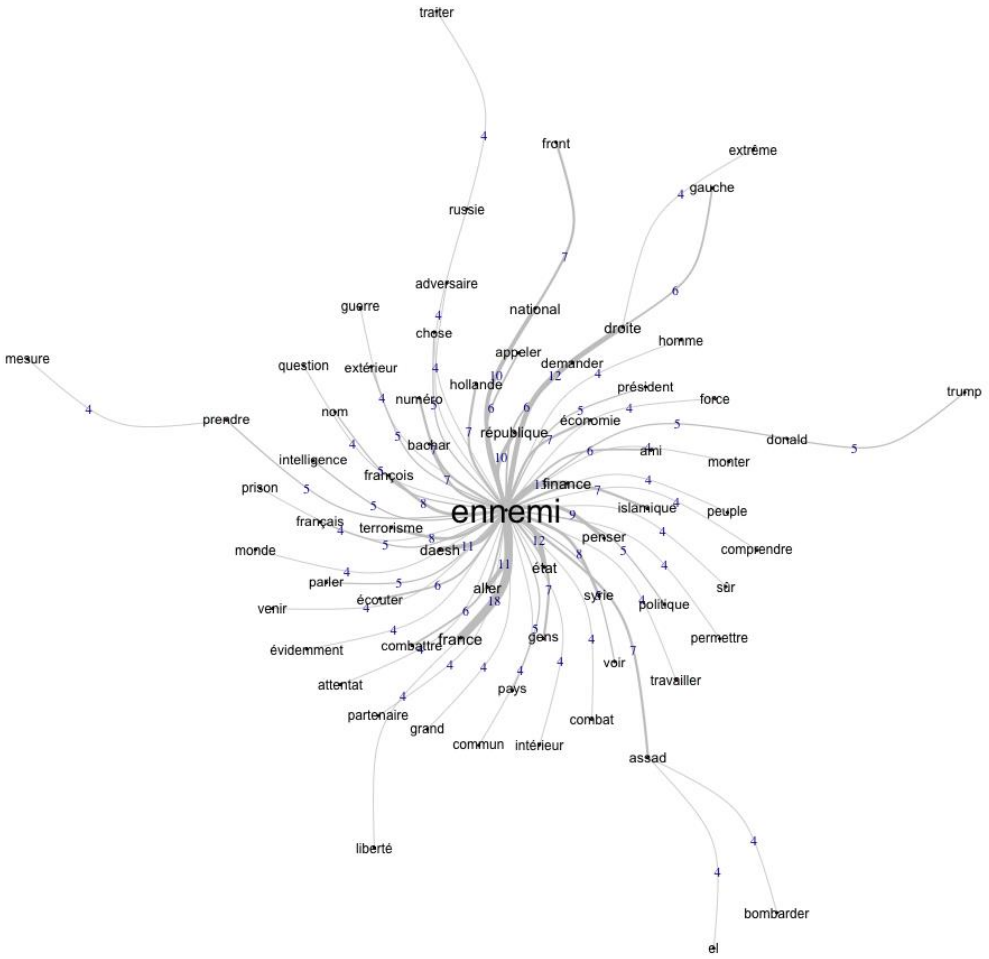
Cela nous donne un sous-corpus spécifique, centré sur les segments qui contiennent le lemme /ennemi/ :

| Forme | Freq. | Types |
|---|---|---|
| ennemi | 135 | nom |
| finance | 19 | nom |
| france | 19 | nr |
| droite | 17 | nom |
| daesh | 16 | nr |
| guerre | 13 | nom |
| état | 12 | nom |
| aller | 11 | ver |
| national | 10 | adj |
| penser | 10 | ver |
| république | 10 | nom |

3. Fréquences des lemmes dans le sous-corpus

Bien sûr, cette procédure induit la perte de certains contextes larges des interviews, mais néanmoins les contextes gauche et droits étant d'envergure convenable, cette étape nous permet de réduire l'attention sur /ennemi/ lui-même.

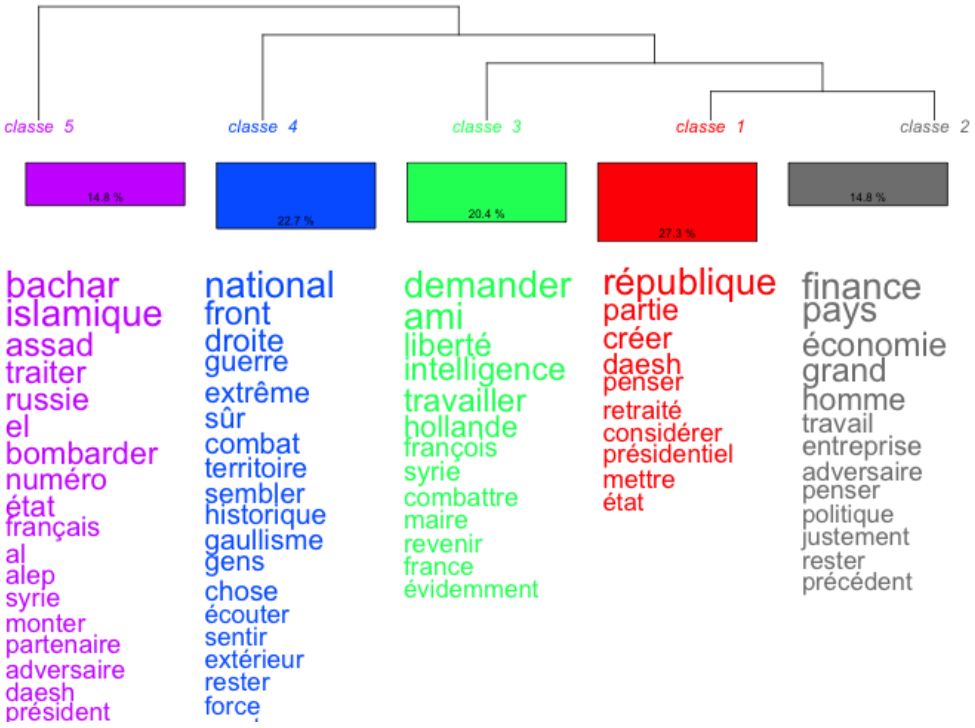
Pour décrire ses profilages, on peut utiliser l'analyse de similitude, définie dans la partie 2. Le résultat est le suivant :




4. Analyse de similitudes dans le sous-corpus

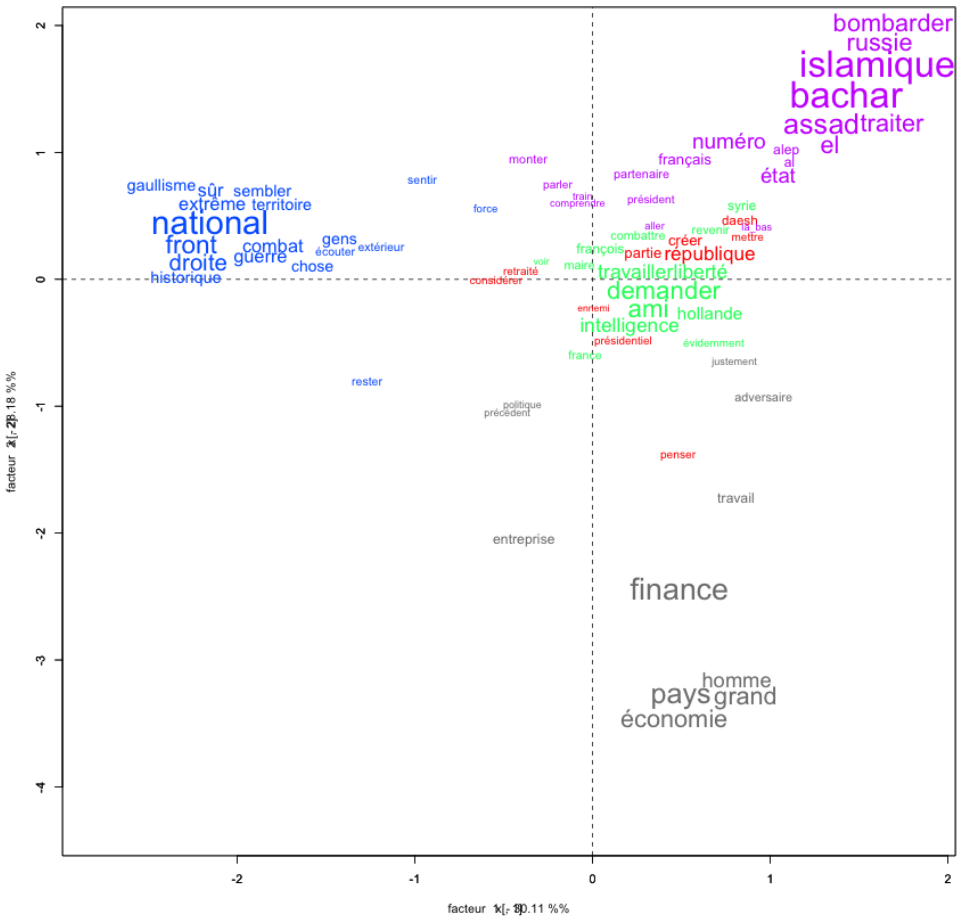
On repère des cooccurrences fréquentes avec les lemmes /France/, /daesh/, /république/, /bachar/, ou encore /islamique/. Ici le propos n'est pas l'analyse exhaustive des cooccurrences autour de ce lemme, mais plutôt de décrire la manière dont cette analyse de similitude peut donner accès aux « frayages déjà enregistrés en lexique » ou à l'« inscription dans des thématiques inédites ». On trouve en effet différents types de caractérisation de l'ennemi dans ce corpus de matinales politiques : un ennemi des valeurs (République), un ennemi politique (France, bachar), un ennemi religieux (islamique) ou politico-religieux (daech). Ces profilages ouvrent la voie aux

thématiques dans lesquelles ce lemme s'intègre, et ceci s'illustre dans la classification thématique que l'on peut produire sur le même sous-corpus :




5. Classification hiérarchique descendante dans le sous-corpus

5 classes principales sont identifiées ici, que l'on peut aussi représenter sur une AFC :




6. AFC dans le sous-corpus

Une caractérisation de l'ennemi est son ancrage dans l'extrême-droite et le front national ; une autre est en lien avec la finance et l'économie ; une autre concerne l'Etat Islamique et Bachar El Assad ; une sur la République et Daech ; et enfin une qui contient les lemmes ami et liberté et qui mérite contextualisation.

On peut alors retourner au corpus, avec les segments de textes caractéristiques des différentes classes, pour comprendre le profilage et la thématisation de /ennemi/. Voici quelques exemples de la classe 3, qui aideront à mieux la comprendre (les termes surlignés sont ceux caractéristiques de la classe) :

- c est à dire que des gens qui *reviennent* de *syrie* qui sont allés égorger qui sont nos ennemis *reviennent* en *france* en *liberté* je suis là aussi un des rares à *demander* leur arrestation immédiate auprès du tribunal pour *intelligence* avec l ennemi
- donc vous *demandez* vous aussi aux *maires* de *france* de ne pas accorder aux ennemis islamistes de la *liberté* la moindre parcelle de *liberté* mais *évidemment*
- alors pour vous qui sont les rebelles d alep ce sont des *amis* ou des ennemis de la *france* on aurait dû les aider ou pas on les a aidés malheureusement

Dans la classe 4, on relève notamment les exemples suivants :

- et que pour moi comme tous ceux qui étaient attachés au *gaullisme historique* où comme ceux qui avaient une histoire centrée ou comme les humanistes de *droite* le *front national* l *extrême droite* était et *reste* un ennemi
- moi mon ennemi c est le *front national* d_abord parce_que à paris vous ne le *sentez* pas mais en province ça monte et vous savez pourquoi ça monte parce_que les *gens* se *sentent* complètement abandonnés
- notre ennemi à nous c est bien l *extrême droite* avant toute *chose* et bien *sûr* la *droite* représentée par françois fillon c est bien l *extrême droite* qui a le projet le plus dangereux pour la france et c est bien la *droite* qui a le projet le plus inégalitaire

On peut donc observer les dynamiques sémantiques, au sens de parcours de sens, tels qu'ils sont illustrés dans l'analyse de similitude, étayés par l'analyse thématique, et spécifiés avec le retour aux

exemples du corpus.

*3.2. Du lexique à l'herméneutique : théoriser les objets discursifs*

En adoptant le principe de variation préconisé, on peut comparer l'analyse de ce corpus, contemporain et politique, aux résultats d'une autre analyse (Longhi et Salem 2018) à propos du corpus Père Duchesne (constitué par la réunion d'un ensemble de livraisons du journal Le Père Duchesne de Jacques-René Hébert, parues entre 1793 et 1794 : voir aussi Salem 1988). A partir de l'étude des segments répétés liés à /ennemi/, l'analyse avait montré qu'aux *plus cruels ennemis, plus mortels ennemis, ennemis du dehors* (les puissances étrangères, les expatriés), des périodes du début, succédaient *les ennemis du dedans et du dehors,* (les *ennemis du dehors* ne constituent pas le seul danger), puis la mention des *ennemis de l'intérieur* qui complétait la notion d'*ennemis du dedans*. Progressivement, *nos ennemis*, devenait *vos ennemis*, puis *les ennemis*. Dans la dernière période les ennemis, désormais désignés, de manière préférentielle, au pluriel, n'étaient plus qualifiés par leur localisation ou par leur rapport aux destinataires du message (*nos/vos ennemis*) mais par des valeurs supposées communes auxquelles ils sont censés s'opposer : *ennemis du peuple, ennemis de la république, ennemis de la révolution, ennemis de la liberté, ennemis de l'égalité.*

On peut donc concevoir des convergences dans ces analyses de corpus, distincts par leur ancrage temporel comme leur genre : une distinction intérieur/extérieur ; la question des valeurs ; la prise en compte du point de vue dans la nomination de l'ennemi. Aussi, ces études nous renseignent à la fois sur le sens et le contexte des corpus spécifiques, mais aussi sur les régimes d'interprétation du sens, et les processus de stabilisation à l'intérieur des thématiques. On accède ainsi à la dimension herméneutique par le biais de la saisie

de la « microgénétique des profils » dont les dynamiques « renvoient pour une part à des frayages déjà enregistrés en lexique et, sous une forme bien plus générique, en grammaire », mais « se font aussi par inscription dans des thématiques inédites, qui les reprennent dans leur propre grille, possiblement extrinsèque, soit aux motifs donnés en langue, soit aux normes de profilage lexical déjà attestées » (Cadiot et Visetti 2001: 130). C'est ce qu'illustre l'analyse de « ennemi », à la fois dans le cadre restreint du discours politique, puis en comparaison avec un autre corpus, d'un autre genre et d'une autre période. Du point de vue de l'analyse du discours, mise en pratique à travers l'analyse de corpus, il est intéressant de mettre à profit la différentialité, par le principe de variation, proche des travaux et discussions issus de la Gestalttheorie :

> comme la psychologie de la Gestalt a hérité du paradigme de la perception les problèmes de la discrétisation et de la description des invariances, elle peut d'autant mieux inspirer une description de la perception sémantique que le langage n'est pas par nature, malgré le postulat de la sémantique cognitive, l'expression de la perception : il en est un objet. Partie de notre monde, il est un moyen essentiel du couplage avec ce monde (Rastier 2006: 103).

Si « les fonds sémantiques semblent des suites de points réguliers et comme les formes sont discrétisées par leurs points singuliers, le parcours productif ou interprétatif de ces formes et de ces fonds suppose un rythme, cellule de base de toute action » (106), au niveau de l'analyse du discours, celui-ci opère comme un champ , à l'intérieur duquel les formes se disposent et se constituent d'un sens. Cette constitution s'opère de manière dynamique, c'est-à-dire par phases qui mettent à profit les différents paliers de la sémantique

du discours, et finalement « l'activité énonciative et interprétative consiste à élaborer des formes, établir des fonds et faire varier les rapports entre fond et forme. […] Ces variations permettent à l'énonciateur de concilier autant qu'il le peut ou le veut les contraintes de la langue, du discours, du genre, de la situation et les rémanences de ce qu'il a déjà dit ou écrit » (112). L'enjeu est alors de pouvoir intégrer de manière constitutive à cette notion de forme la dimension sociodiscursive, qui devient en dernière instance la prise en compte du fond sociodiscursif propre à un discours. Cette prise en compte du « fond sociodiscursif » relève en fait de ce que Georges-Elia Sarfati (1996) a théorisé dans sa reprise du concept de sens commun d'un point de vue linguistique et discursif. Ce travail sur la constitution des formes permet en outre de saisir d'une autre manière cette question du système du sens commun, non plus du point de vue de son instanciation, mais de sa saisie perceptive par les sujets parlants lors de la production des formes. Il s'agit ainsi de prendre en considération cette théorie du sens comme une théorie de l'action des sujets parlants sur leur environnement sémiotique.

## Conclusion

En considérant les liens possibles, dans notre recherche, entre herméneutique et numérique, nous avons pu affiner la méthodologie et l'ancrage théorique de la Théorie des objets discursifs, qui s'appuie notamment sur l'appareillage conceptuel de la Théorie des formes sémantiques. La prise en compte des besoins en termes d'outillage, et des ambitions en termes d'analyse, nous ont permis de mettre en cohérence la dimension théorique et la dimension informatisée de la recherche. L'outil n'est pas considéré comme un simple moyen d'accéder à des observations, mais bien un *ensemble des connaissances et des techniques qui permettent de faire parler les signes et de découvrir leur sens* tel que défini par Foucault. Le

numérique est donc ici au service de l'herméneutique, dans la mesure où il donne au chercheur des moyens d'observer statistiquement et/ou visuellement des résultats cohérents avec la théorie linguistique convoquée (en particulier les profilages et les thématiques).

Dans le contexte d'une production croissante de discours en particulier dans les environnements numériques, et au regard des enjeux critiques de l'analyse du discours, une sensibilisation aux liens entre herméneutique nous semble fondamentale : prise en compte de la spécificité des productions natives du web ; importance des outils numériques pour leur analyse, et précautions nécessaires ; éducation aux médias numériques, et aux processus interprétatifs qui leur sont propres.

**Références**

Anquetil, S., Duteil-Mougel, C. & Llovera, V. (2019, éds.). Le sens des données. Le statut du corpus et herméneutique à l'aune des humanités numériques. Paris : l'Harmattan.

Cadiot, P. (2009). Couleur des mots ou synonymie. *Pratiques* [En ligne], 141-142, mis en ligne le 19 juin 2014. URL : http://journals.openedition.org/pratiques/1273 ; DOI : 10.4000/pratiques.1273 (10 mai 2019).

Cadiot, P. (2002). La métaphore, ou l'entrelacs des motifs et des thèmes. *Semen* [En ligne], 15 | 2002, mis en ligne le 29 avril 2007. URL : http://journals.openedition.org/semen/2374 (20 avril 2016).

Cadiot, P. & Visetti, Y.-M. (2001). *Pour une théorie des formes sémantiques. Motifs, profils, thèmes.* Paris : PUF.

Foucault, M. (1966). *Les Mots et les choses.* Paris : Gallimard.

Longhi, J. (2015). *La Théorie des objets discursifs: concepts, méthodes, contributions*. Mémoire d'HDR, université de Cergy-Pontoise.

Longhi, J. (2018). *Du discours comme champ au corpus comme terrain. Contribution méthodologique à l'analyse sémantique du discours*. Paris : l'Harmattan.

Longhi, J. & Salem, A. (2018). Approche textométrique des variations du sens. *Actes des JADT 2018*, 452–458.

Longhi, J. (2020). Proposals for a Discourse Analysis Practice Integrated into Digital Humanities: Theoretical Issues, Practical Applications, and Methodological Consequences. *Languages*, *5*(1): 5.

Loubère, L. (2016). L'analyse de similitude pour modéliser les CHD. *Actes des JADT 2016*, http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83440/83440.pdf (13 novembre 2017).

Marchand, P. & Ratinaud, P. (2016). L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des JADT 2012*, http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Marchand,%20Pascal%20et%20al.%20-%20L'analyse%20de%20similitude%20appliquee%20aux%20corpus%20textuels.pdf

Meunier, J.-G. (2019). Le paradoxe des humanités numériques. *Quaderni,* 98: 19–31.

Rastier, F. (2001). Art et science du texte. Paris : PUF.

Rastier, F. (2006). Formes sémantiques et textualité. *Langages*, *163*(3): 99–114.

Salem, A. *Tutoriels pour l'analyse textométrique*, http://lexicometrica.univ-paris3.fr/numspeciaux/special8/tutoriel1.pdf

Salem, A. Séries textuelles chronologiques, http://lexicometrica.univ-paris3.fr/numspeciaux/special8/tutoriel2.pdf

Visetti, Y.-M. (2003). Formes et théories dynamiques du

sens », *Texto ! mars 2003* [en ligne], http://www.revue-texto.net/Inedits/Visetti/Visetti_Formes1.html (12 juin 2005).
Visetti, Y.-M. (2004). Le Continu en sémantique : une question de formes », *Texto ! juin 2004* [en ligne], http://www.revue-texto.net/Inedits/Visetti/Visetti_Continu.html (12 juin 2005).
TLFI : http://atilf.atilf.fr/