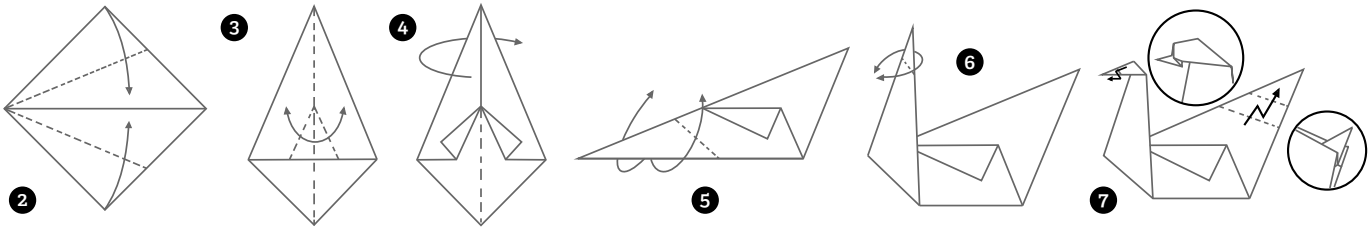
CYGNE NOIR

revue d'exploration sémiotique



L'analyse de texte assistée par ordinateur : introduction à l'un des champs fondamentaux de la sémiotique computationnelle

Davide PULIZZOTTO

Cygne noir, no 7, 2019 : « Algorithmes »

Résumé

La sémiotique computationnelle étudie l'interaction entre les processus d'émergence du sens et les systèmes formels, computables et numériques. En effet, l'une de ses hypothèses est la possibilité de décrire la sémiose à travers des métalangages formels et de la simuler par des procédés algorithmiques. Dans ce contexte, plusieurs pratiques d'analyse sémiotique se sont développées, à l'exemple de l'analyse de texte assistée par ordinateur (ATO). Avec cette dernière, en adoptant des techniques et des méthodes issues de l'informatique et de l'intelligence artificielle, les formes plus classiques de l'analyse de texte se joignent aux champs de recherche des humanités numériques. La sémiotique est ainsi appelée, entre autres, à discuter les enjeux de l'usage de ces techniques dans la recherche en sciences humaines et sociales. L'objectif de cet article est de présenter un survol de la sémiotique computationnelle et d'introduire le lectorat à certains aspects théoriques et méthodologiques de l'assistance informatique à l'analyse de texte. Plus particulièrement, le texte expose les étapes et les hypothèses de la transformation vectorielle du texte que présuppose l'ATO et discute des enjeux sémiotiques de deux procédures : la lemmatisation et la fonction de pondération.

Pour citer cet article

PULIZZOTTO, Davide, « L'analyse de texte assistée par ordinateur : introduction à l'un des champs fondamentaux de la sémiotique computationnelle », *Cygne noir*, no 7, 2019. En ligne : <http://www.revuecygnenoir.org/numero/article/pulizzotto-ato> (consulté le xx/xx/xxxx).

L'ANALYSE DE TEXTE ASSISTÉE PAR ORDINATEUR : INTRODUCTION À L'UN DES CHAMPS FONDAMENTAUX DE LA SÉMIOTIQUE COMPUTATIONNELLE

La « révolution numérique » en train de modifier profondément les sciences humaines et sociales (SHS) est loin d'être ignorée par la sémiotique. L'interaction, de plus en plus grandissante, entre les SHS et les sciences informatiques démontre la pertinence de définir un nouveau domaine d'études, à savoir celui des *humanités numériques*. La sémiotique est partie intégrante de ce nouveau domaine de recherche et y contribue par le développement d'un nouveau champ d'études appelé *sémiotique computationnelle*. Celle-ci regroupe des travaux de nature différente, mais qui partagent une relation, plus ou moins profonde, avec l'informatique. Ainsi, les travaux afférents permettent d'étudier, sous différentes perspectives, la nature computationnelle du *sens*.

Constitué de deux parties, cet article a une visée introductive et didactique. Fournissant des clés de compréhension pour appréhender un champ relativement récent de la sémiotique, ce texte s'adresse aux chercheurs en sciences humaines, plus précisément aux sémioticiens, intéressés à explorer plus avant les liens entre sémiotique, analyse de texte et computation. Dans la première partie, je présenterai un survol de la sémiotique computationnelle, en proposant une classification en trois branches. Il s'agira de dresser un portrait disciplinaire et méthodologique des champs d'études de la sémiotique computationnelle. Par la suite, j'approfondirai la nature de l'un de ses champs d'études les plus importants : l'analyse de texte assistée par ordinateur (ATO). J'ai en effet choisi d'isoler et d'explorer ce champ en vue de répondre à la question suivante : quels sont les enjeux sémiotiques de l'usage des outils numériques et quels éléments permettent d'en juger? L'ATO se prête bien à l'exploration des relations entre sémiotique et usage des outils numériques puisque son cadre épistémologique est compatible avec l'une des pratiques les plus communes de la sémiotique, soit l'analyse des artéfacts sémiotiques. Ainsi, dans la deuxième partie, j'accompagnerai le lectorat à réfléchir sur certains aspects ayant trait aux dynamiques sémiotiques de l'usage des outils de traitement automatique du langage naturel (TALN) et de l'algèbre linéaire. Pour ce faire, je présenterai l'une des étapes qui permet le passage de l'analyse de texte traditionnelle à l'assistance par ordinateur, soit la conversion du texte en un modèle computable, réalisée grâce à l'adoption du *modèle sémantique vectoriel.* Plus particulièrement, je décrirai deux opérations spécifiques de la transformation vectorielle du texte, la *lemmatisation*

et la *fonction de pondération*, puisqu'elles sont fonctionnelles à la compréhension de la transposition mathématique du texte dans le cadre de l'ATO.

Bien que l'article ait une portée pédagogique, les opérations décrites dans la deuxième partie de ce texte me donneront la possibilité de m'arrêter plus longuement sur des questions supplémentaires et de problématiser certains éléments de l'approche concernée qui me semblent être souvent négligés par la communauté de chercheurs. L'opération de lemmatisation, notamment, permet de réfléchir à la nature sémiotique des caractéristiques qui composent les vecteurs du modèle sémantique. Le plus souvent, les auteurs se limitent à présenter ces caractéristiques comme un ensemble de mots. Peter D. Turney et Patrick Pantel, pour citer l'un des textes les plus connus, affirment que le sens d'un texte est décrit par les éléments qui le composent et leurs fréquences[1]. Chaque texte est ainsi conçu comme un *sac-de-mots*, c'est-à-dire comme l'ensemble des signes linguistiques qui apparaissent dans le texte. Cet ensemble constitue le vecteur qui représente un texte dans un espace géométrique. Or, trop peu est dit à propos de la nature linguistique, sémantique ou lexicale de ces unités. Pour cette raison, je m'attarderai sur la lemmatisation. Je mettrai en évidence quelques limites des outils qui sont généralement utilisés en ATO – limites qui, à dire vrai, affectent peu la valeur heuristique du modèle vectoriel, mais qu'il est important de garder à l'esprit.

Dans cette deuxième partie de l'article, j'ai aussi la possibilité de rebondir sur la complémentarité entre l'hypothèse distributionnelle, qui est à la base du modèle sémantique vectoriel, et la linguistique de Saussure. Teun A. van Dijk est, probablement, le premier à signaler cette complémentarité[2]. Il considère identiques le problème que se pose Zellig Harris, un des pères putatif de l'hypothèse distributionnelle, et celui de la sémiotique textuelle des années 1960 et 1970, laquelle se préoccupe de la construction d'outils pour l'analyse des textes. De plus, selon Dijk, l'analyse du discours de Harris propose des solutions qui coïncident avec la vision structuraliste du langage. Ces liens sont décrits de manière explicite par Magnus Sahlgren, qui souligne à plusieurs reprises la nature structuraliste de la sémantique vectorielle[3]. En particulier, l'auteur a mis en évidence le rôle de la « vision différentielle du sens »[4], tirant ses racines de la linguistique structuraliste, pour la définition de l'hypothèse distributionnelle, énoncée la première fois par Harris. Ainsi, il montre que les motivations théoriques de deux types de modèles de la sémantique vectorielle, le syntagmatique et le paradigmatique, se basent sur la notion de *valeur* et sur la définition de l'axe syntagmatique et de l'axe paradigmatique en linguistique. De mon point de vue, la notion saussurienne de valeur peut également être utilisée pour motiver la fonction de pondération, ce qui constitue un prolongement logique des premières réflexions de Sahlgren.

Enfin, l'article s'adresse principalement à ceux qui abordent l'ATO depuis peu. Ces derniers trouveront les notions de base pour comprendre son cadre théorique. Mais il offre également quelques pistes de réflexion pour ceux qui cherchent à clarifier la valeur sémiotique de la transformation vectorielle du texte. Cette dernière discussion est basée sur plusieurs travaux empiriques que j'ai personnellement conduits[5].

L'article est divisé en deux parties : la première présente les trois branches de la sémiotique computationnelle et détaille davantage l'ATO ; la deuxième décrit les procédures et les hypothèses de la transformation vectorielle du texte que l'ATO met en place et discute davantage les fondements sémiotiques de la procédure de lemmatisation et de la fonction de pondération.

1. Un survol de la sémiotique computationnelle

La sémiotique est une discipline intrinsèquement interdisciplinaire et son histoire en témoigne. En effet, en Europe, elle s'est développée avec le concours d'une autre discipline, la linguistique. Pendant plusieurs décennies, les histoires de ces deux disciplines se sont entrecroisées alors que la linguistique englobait la sémiotique en tant que champ d'études. Parallèlement, en Amérique du Nord, la sémiotique se définissait à l'intérieur du pragmatisme américain, par le biais des travaux de Charles S. Peirce. Avec le temps, la sémiotique a redéfini ses approches, ses méthodologies et ses objets d'études spécifiques, mais elle n'a pas cessé de produire des « mélanges disciplinaires » de toutes sortes. La sémiotique computationnelle est le produit de l'un de ces « mélanges ». Celle-ci se caractérise par la rencontre de la sémiotique avec les domaines de l'informatique. La sémiotique computationnelle constitue aujourd'hui un champ des humanités numériques, ces dernières recouvrant « un ensemble de pratiques de recherche à l'intersection des technologies numériques et des différentes disciplines des sciences humaines[6] ».

Les volets de la sémiotique computationnelle sont nombreux et interagissent à différents niveaux avec l'informatique, mais aussi avec les différents champs de l'intelligence artificielle. Pour définir la sémiotique computationnelle, il est nécessaire d'aborder la relation entre la sémiotique et les *propriétés computables* d'un mécanisme signifiant. L'une des perspectives les plus intéressantes pour appréhender cette relation vient de l'intelligence artificielle. Ainsi, je propose de résumer cette relation avec le concept de « simulation » d'opérations cognitives par une machine, ce qui constitue, historiquement, le point de départ du programme de recherche de l'intelligence artificielle. La naissance de cette dernière, comme celle du cognitivisme, est enracinée dans le mouvement cybernétique des années 1940-1950. C'est dans ce contexte que l'intérêt pour les

mécanismes de la pensée humaine et la possibilité de les simuler à l'aide d'un ordinateur a commencé à émerger. Elaine Rich nous fournit probablement une définition de l'intelligence artificielle des plus précises et des plus actuelles[7]. Elle souligne comment le transfert des fonctions cognitives vers l'ordinateur est effectivement le principal but de cette discipline : « L'intelligence artificielle étudie la manière de faire faire aux ordinateurs des choses pour lesquelles, à l'heure actuelle, l'être humain est plus habile[8]. »

Cette définition de l'intelligence artificielle offre une perspective idéale pour appréhender les dynamiques sémiotiques par ses aspects computables. Dans ce cadre, si l'objectif de la sémiotique est de comprendre les mécanismes qui *mettent les « choses » en condition de signifier*[9], la sémiotique computationnelle aurait le but de « simuler » les « conditions de signification ». Le verbe « simuler » devrait être appréhendé dans une acception large, car il doit représenter la sémiotique computationnelle et ses diverses expériences de recherche, dont seule une portion utilise effectivement l'informatique et l'intelligence artificielle pour implémenter des dynamiques sémiotiques.

Compte tenu de leur grand nombre et de leur diversité, il est très difficile de résumer et de classifier les études pouvant faire partie de la sémiotique computationnelle. Pour des raisons de simplicité, je propose de diviser la sémiotique computationnelle en trois grandes branches : (1) l'étude de l'*ordinateur comme machine sémiotique*[10] ; (2) l'étude de l'*ordinateur comme artéfact*[11] ; et (3) l'étude de l'*ordinateur comme outil au service de la recherche sémiotique*[12]. Cette classification, inspirée des travaux de Jean-Guy Meunier[13] et de Kumiko Tanaka-Ishii[14], se caractérise par trois approches différentes du même objet : l'*ordinateur*. Le terme « ordinateur » est ici utilisé de manière métonymique et, selon ces approches, réfère à des éléments de nature différente. Toutefois, ces approches concernent toutes, d'une manière ou d'une autre, l'aspect computable sous-entendu dans la notion générale d'ordinateur. Cette classification ne peut toutefois être considérée comme étant définitive puisque ce champ est en évolution constante. De plus, plusieurs auteurs qui sont inclus dans la classification que je propose ne reconnaissent pas explicitement la nature sémiotique de leurs recherches, ce qui rend la détermination de cette classification encore plus difficile.

1.1 L'ordinateur comme machine sémiotique

La première de ces trois branches de la sémiotique computationnelle considère l'ordinateur comme une machine sémiotique, c'est-à-dire une machine qui *manipule des signes*. Dans cet axe, la sémiose est considérée comme *un fait ou un événement computable* et l'objectif principal des travaux qui y sont menés consiste à simuler des processus de si-

gnification. Ainsi, l'une des questions fondamentales posées dans ce contexte concerne la manière de « modéliser » les processus de construction du sens dans des systèmes artificiels[15]. La sémiotique joue un rôle très important dans ce genre de recherches et plusieurs modèles sémiotiques ont été proposés pour la simulation des systèmes et des processus du sens. Dans les travaux de João Queiroz et Floyd Merrell[16], par exemple, la sémiotique de Peirce est employée comme cadre conceptuel à partir duquel il est possible de comprendre la sémiose et de construire un modèle formel pour sa simulation par l'ordinateur. La sémiose est entendue comme un processus qui s'auto-organise comme le font les organismes biologiques[17].

Toutefois, les études en sémiotique qui se sont concentrées sur le lien entre les dynamiques sémiotiques et la computation sont peu nombreuses. Ces travaux ont surtout essayé de mettre en relation les caractéristiques formelles des modèles sémiotiques avec celles des modèles computationnels[18]. Ainsi, dans ce genre de travaux, le raisonnement logique est constitué par l'identification de fonctions cognitives qui peuvent, en premier lieu, être transformées en un *modèle formel et logique* et, successivement, en un *modèle computable*, pour enfin passer à la véritable phase d'*implémentation informatique*. Ce raisonnement constitue une approche générale pour la sémiotique computationnelle et, de manière plus spécifique, pour la troisième branche de la sémiotique computationnelle. En effet, identifier les fonctions cognitives que, involontairement ou volontairement, certains outils numériques peuvent simuler permet d'entrevoir avec plus d'attention leur potentiel heuristique.

1.2 L'ordinateur comme artéfact

La deuxième branche considère l'ordinateur en tant qu'*artéfact*. Un artéfact doit être appréhendé comme un objet construit par l'humain, qui lui sert pour exécuter des tâches. Dans ce contexte, la sémiose est regardée comme *un évènement qui se génère à partir de l'interaction entre l'humain et la machine*. Dès lors, les concepts de machine et d'artéfact se chevauchent. L'artéfact prend une grande importance en sémiotique pour la définition du concept de culture. Pour François Rastier, par exemple, les artéfacts constituent la composante principale du niveau sémiotique de la culture, et ils regroupent les *objets culturels*[19]. L'ordinateur peut être étudié comme un *instrument* (une des catégories des objets culturels), c'est-à-dire un objet construit par l'humain et qui interagit avec lui dans différentes situations. Ici, la sémiotique est souvent utilisée pour supporter l'étude du design et de la conception des systèmes informatiques afin d'optimiser l'interaction humain-machine. Cette deuxième branche est probablement

la plus vaste. En simplifiant, on peut la subdiviser en trois axes, soit les études sur l'interaction humain-machine (IHM), comme les travaux en ingénierie sémiotique[20], les études sur les systèmes d'information comme systèmes sémiotiques qui sont liés à la sémiotique organisationnelle[21], et enfin les études sémiotiques de la programmation et de ses langages, projet de recherche qui remonte aux années 1960[22].

1.3 L'ordinateur comme outil

La troisième branche appréhende l'ordinateur comme un *outil pour l'analyse sémiotique*. Dans ce contexte, la sémiose qui est prise en considération *émerge d'un artéfact sémiotique qui est indépendant de l'outil informatique*. Les textes écrits ou autres artéfacts sémiotiques sont pris dans leur contexte et convertis en un objet formel sur lequel il est possible d'appliquer des outils pour l'analyse assistée par ordinateur. Le mot « ordinateur » n'est alors qu'un mot-contenant pour indiquer plusieurs éléments de nature mathématique, statistique, formelle, computationnelle et informatique. Pour le dire simplement, les travaux de cette branche continuent à étudier des phénomènes sémiotiques et ses artéfacts, mais au moyen de méthodes hybrides d'analyse sémiotique qui se servent de l'assistance informatique. La troisième branche se distingue des deux précédentes par son objet de recherche, lequel n'est pas la détermination de modèles computationnels pour la simulation sémiotique ni les aspects computables de la signification ou l'ordinateur comme artéfact. Son objectif est de répondre à des questions traditionnelles en sémiotique ou en SHS, à l'exemple de problématiques reliées à l'analyse d'un corpus de textes, et de le faire à l'aide de méthodes assistées par l'informatique.

L'analyse du texte écrit domine les travaux de cette troisième branche, et la majorité des travaux qui y sont développés le sont en ATO. Le traitement automatique du langage naturel est en effet l'une des branches de l'intelligence artificielle qui s'est grandement développée dans les dernières décennies. À l'inverse, en ce domaine, les outils d'analyse de l'image ou du son sont encore en phase de développement. Il est probable que la recherche en apprentissage profond (*deep learning*) pourra faire émerger des méthodes pour l'analyse d'artéfact sémiotique dans lesquels plusieurs systèmes sémiotiques interagissent entre eux, comme le langage verbal écrit, visuel, audio, etc. Pour le moment, il existe très peu de travaux qui se sont penchés sur l'analyse d'artéfacts sémiotiques complexes. Pour toutes ces raisons, je présenterai dans les prochains paragraphes le cadre théorique de l'ATO, lequel se prête bien à une première exploration de l'usage des outils numériques pour l'analyse d'artéfact sémiotique.

1.3.1 L'analyse de texte assistée par ordinateur (ATO)

La troisième branche de la sémiotique computationnelle regroupe plusieurs travaux, parmi lesquels ceux déjà accomplis en ATO, dont l'objet de recherche est à la fois le développement de méthodes d'analyse de textes à l'aide d'outils statistiques, mathématiques ou informatiques ainsi que l'analyse de corpus textuels. En résumé, l'ATO s'applique à des corpus qui sont construits à des fins d'analyse et afin de répondre à des problématiques typiques en SHS. Par exemple, il est possible d'assister l'analyse d'un corpus composé d'articles de journaux, d'œuvres d'un philosophe, de romans d'un courant littéraire particulier, etc. Ce domaine est vaste et regroupe des travaux qui naissent dans différents contextes, comme l'analyse de contenu[23], l'analyse du discours[24], la statistique textuelle et la linguistique de corpus[25], pour ne nommer que cela. Il n'est donc pas simple de définir ce champ de recherche de manière univoque. L'expression « analyse de texte assistée par ordinateur » ne fait pas l'unanimité d'autant plus qu'elle n'est pas la seule utilisée[26]. Il demeure toutefois possible d'esquisser un cadre général des disciplines et des méthodes qui contribuent aux avancées de ce champ de recherche. Sauf erreur de ma part, il n'existe aucune tentative de décrire l'interdisciplinarité de l'ATO. J'en propose ainsi une description schématique (fig. 1) afin de clarifier le plus possible la nature disciplinaire et méthodologique de l'ATO. Pour ce faire, je présente un survol qui va des SHS jusqu'à la science de données. La description de l'ATO que je propose est constituée de six axes de contribution.

D'abord, chaque recherche en ATO se situe à l'intérieur d'un domaine d'application dans différentes disciplines des SHS, comme la sociologie, l'histoire, la communication, les sciences politiques, la philosophie, etc. (premier axe). Par exemple, en histoire de la philosophie les chercheurs étudient, entre autres, l'évolution d'un concept en analysant un recueil de texte d'un philosophe particulier. En science politique, les chercheurs analysent les caractéristiques du discours électoral d'un candidat aux élections à partir des verbatim de ses discours télévisés. En sociologie, l'intérêt de recherche se concentre davantage dans la description de communautés socio-sémantiques à partir de corpus de textes journalistiques, et ainsi de suite. Ce sont donc les intérêts de recherche de ces disciplines qui déterminent le point de départ et les objectifs de recherche de l'ATO. Ainsi, chacune de ces disciplines caractérise de manière spécifique l'approche d'analyse utilisée. Il est donc important de ne pas négliger la contribution qu'elles apportent au développement des applications en ATO.

Ensuite, formant un deuxième axe, des disciplines comme la linguistique et la sémiotique contribuent en fournissant un cadre théorique pour l'analyse du texte. Des concepts relatifs aux paliers textuels (sème, mot, syntagme, phrase, paragraphes,

etc.), les axes syntagmatique et paradigmatique ou la dénotation et la connotation favorisent la compréhension de l'artéfact sémiotique qui est utilisé pour l'extraction des connaissances.

Un troisième axe de contribution à l'ATO est constitué par l'analyse de contenu et l'analyse du discours. Ces pratiques ont certainement fourni à l'ATO des méthodes et des cadres théoriques à la fois qualitatifs et quantitatifs[27] qui apportent des perspectives différentes à l'analyse du texte. Un de ses mérites est, par exemple, d'avoir mis en évidence l'importance d'une approche herméneutique qui se concentre davantage sur l'extraction de connaissances extratextuelles. En analyse du discours, les chercheurs ajoutent le plus souvent à l'analyse du texte une description des relations de son contenu avec son contexte historique, social, politique, etc.

Un quatrième axe regroupe les disciplines qui mettent en lumière plusieurs méthodes quantitatives d'analyse du texte, soit la textométrie, la statistique textuelle et la linguistique de corpus. Depuis le début des années 1960, ces disciplines contribuent à la définition de l'approche statistique à l'étude du langage. Chacune avec ses propres spécificités, elles ont su montrer comment utiliser différents outils de la statistique dans un contexte d'analyse de texte, comme la notion de corrélation[28] ou l'analyse factorielle de correspondances[29].

Un cinquième axe de contribution est celui du traitement automatique du langage naturel (TALN)[30] et de la linguistique computationnelle[31]. À la différence de l'axe précédent, celui-ci met en évidence un autre aspect de l'ATO, soit la description et l'annotation automatique des caractéristiques linguistiques d'un texte. En effet, c'est à l'intérieur de cette discipline que des outils comme les lexicographes automatiques ou les analyseurs morphosyntaxiques[32] se développent et se mettent constamment au point. L'analyse linguistique du texte est, sans doute, une étape fondamentale de l'ATO, et le TALN contribue à améliorer son assistance par ordinateur, mettant ainsi en évidence l'importance de l'informatique et de l'intelligence artificielle. Plusieurs autres applications et outils sont développés dans ce domaine (à l'exemple du résumé automatique[33]), et peuvent, à l'occasion, être intégrés dans une chaîne de traitement en ATO.

Enfin, la science des données met à disposition de l'ATO toute une série de méthodes développées en apprentissage automatique[34] et en apprentissage statistique[35], comme le regroupement automatique de documents[36], les techniques d'analyse thématique[37], etc. Ces outils peuvent être présentés sous des formes et des perspectives différentes, comme celles de la sémantique distributionnelle[38], de la recherche d'information[39] et, plus particulièrement, de la fouille de texte[40].

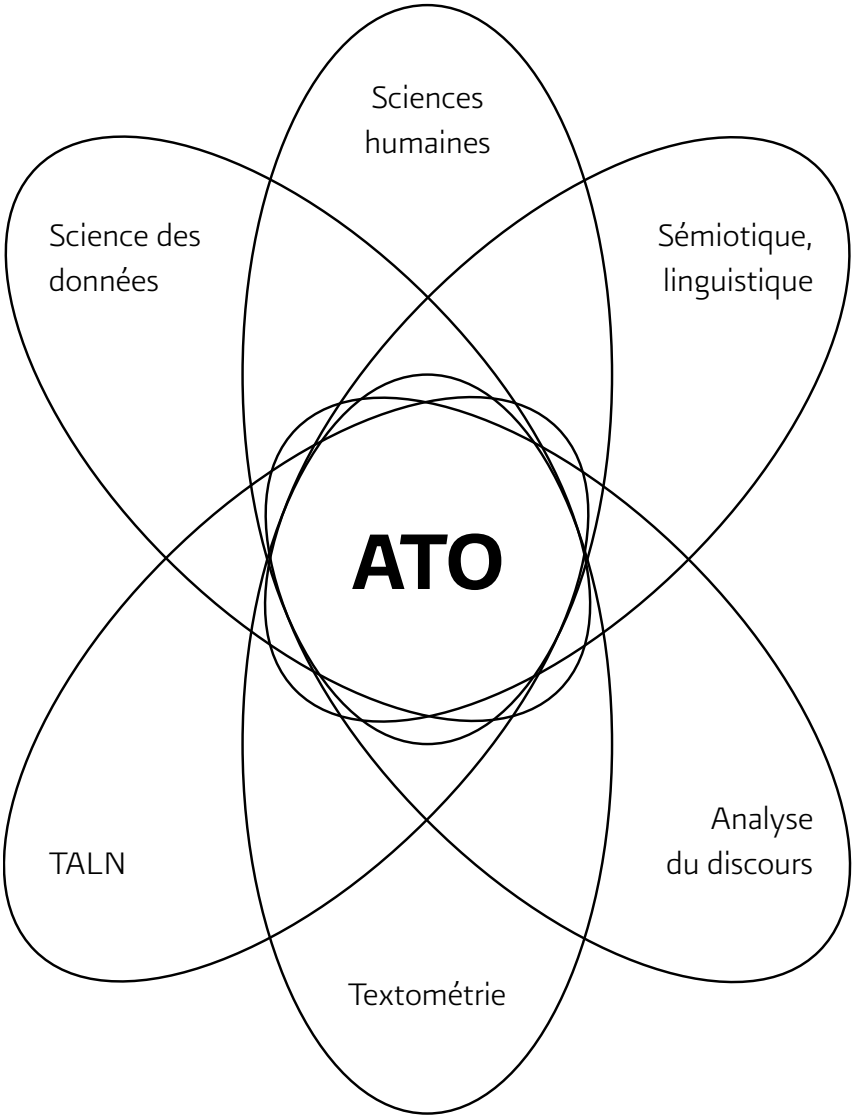


Figure 1. Schéma définissant l'interdisciplinarité de l'analyse de texte assistée par ordinateur.

L'ATO est alors une pratique d'analyse qui, à travers un bricolage de cadres théoriques et de méthodes qualitatives et quantitatives, répond à des questions de recherche typiques en SHS. Une étape particulière de ce bricolage a le mérite d'élargir les possibilités de l'assistance par ordinateur : il s'agit de l'étape de transformation du texte en un modèle formel et computable. C'est seulement lorsque cette transformation est accomplie que la plupart des outils de la statistique, du TALN ou de la science de données peuvent

être impliqués. Cette étape est aussi celle dans laquelle la sémiotique et l'informatique entrent plus facilement en contact, puisque c'est à ce moment que l'artéfact textuel subit des modifications pour s'adapter à un paradigme différent, celui numérique. En d'autres termes, les éléments linguistiques normalement utilisés pour la compréhension d'un texte, comme les mots, les lemmes, les phrases, etc., sont transférés dans un objet mathématique particulier appelé matrice. Il est alors important de comprendre, lors de ce passage du *texte* à la *matrice*, ce qui est véritablement conservé et ce qui est perdu.

Avant de passer à la description de la transformation vectorielle du texte en ATO, je tiens à faire une précision. En fait, il me semble que l'enthousiasme qui accompagne l'apparition des outils de l'apprentissage profond (*deep learning*) porte parfois la communauté qui s'intéresse à l'ATO à confondre différentes approches et applications. L'apprentissage profond est un sous-champ de recherche de l'apprentissage automatique. Les outils que s'y développent sont basés sur un modèle mathématique existant depuis les années 1960[41], à savoir les réseaux de neurones. Sans rentrer dans les détails techniques, je veux ici souligner la différence entre les applications typiques de l'apprentissage profond et l'ATO. En fait, en raison des caractéristiques des réseaux de neurones, l'apprentissage profond est surtout développé pour des applications d'apprentissage de type supervisé. Or, en ATO, il est rare d'utiliser cette approche, car l'extraction de connaissance se fait à partir de matériel non structuré, sans annotation, de taille relativement petite et dans un cadre de type non supervisé. Les outils de l'apprentissage profond ne sont donc pas toujours adaptés à l'ATO. Toutefois, il est vrai qu'il est possible d'accéder à des modèles de langage entraînés par de tels outils pour accomplir des tâches supervisées qui sont fonctionnelles à l'analyse de texte. Par exemple, il est possible d'utiliser des modèles entraînés pour exécuter l'annotation sémantique, l'annotation morphosyntaxique ou pour obtenir des représentations vectorielles de mots ou de documents qui sont généralement appelées *word embedding*. Il est vrai également que des modèles de type non supervisés sont en train d'être développés. Mais, en l'état actuel, il me semble que le cœur des outils de l'ATO ne nécessite pas d'intégrer impérativement l'apprentissage profond.

2. Textes et vecteurs

Le plus grand défi pour l'ATO est la *conversion d'un objet sémiotique comme le texte en un objet de nature computable*. Cette étape est fondamentale, car les différents outils informatiques peuvent être seulement utilisés si l'objet d'étude assume une forme mathématique. La modélisation mathématique de phénomènes humains et sociaux n'est

pas un fait nouveau. Il s'agit d'une pratique nécessaire lorsque la statistique devient un outil d'analyse. Dans ce contexte, le phénomène humain ou social à étudier est transposé en un jeu de données qui est ensuite analysé. Par exemple, en fonction d'hypothèses particulières, il est possible d'étudier l'impact des collaborations entre chercheurs de différentes universités sur la productivité en analysant le nombre de collaborations et le nombre de publications. Dans ce cas, un phénomène social est converti en modèle statistique à travers des faits observables, comme le nombre de publications ou de collaborations. Ce type de pratique est requis également en ATO où, toutefois, les données sont composées à partir des textes. Dans le cadre de cet article, je décrirai le modèle utilisé en ATO pour la transformation du texte : la sémantique vectorielle.

2.1 Sémantique vectorielle

La sémantique vectorielle est un modèle pour la représentation formelle et computable du langage naturel. Son but est d'obtenir une *représentation vectorielle de la sémantique des textes*. Ce modèle partage les mêmes caractéristiques et propriétés que le modèle vectoriel, ce qui permet d'utiliser les outils mathématiques de l'algèbre linéaire. Le modèle sémantique vectoriel a été développé, en Amérique du Nord, par l'équipe de Gerard Salton dans les années 1970[42], et, en Europe, par Jean-Paul Benzécri[43].

La représentation vectorielle du texte consiste dans la transformation de celui-ci en un point dans un espace géométrique, dont les coordonnées dépendent de ses caractéristiques lexicales et sémantiques. En d'autres termes, un document textuel est projeté dans un espace euclidien au moyen de sa conversion en *vecteur*. Ce dernier est constitué de valeurs numériques qui mesurent le poids de chaque caractéristique lexicale ou sémantique du document. Un ensemble de documents est ainsi transformé en une matrice sur laquelle, grâce aux outils classiques de l'algèbre linéaire, plusieurs opérations peuvent être effectuées. Une des premières opérations, qui a permis la naissance du champ d'études de la recherche d'information, est le calcul de la *similarité* ou de la *dissemblance* entre les vecteurs qui représentent les documents dans cet espace. Cette opération se concrétise notamment dans le calcul de la distance entre deux points de l'espace euclidien ou, plus fréquemment, dans le calcul de l'angle entre deux vecteurs. Elle est également l'opération la plus simple à exécuter et à interpréter. Tout simplement, plus les vecteurs sont proches, plus ils sont similaires. Par exemple, dans la figure 2, les points représentent des documents et la distance entre eux constitue leur similarité (ou dissemblance) sémantico-lexicale. Dans ce cas, trois groupes de documents similaires sont représentés.

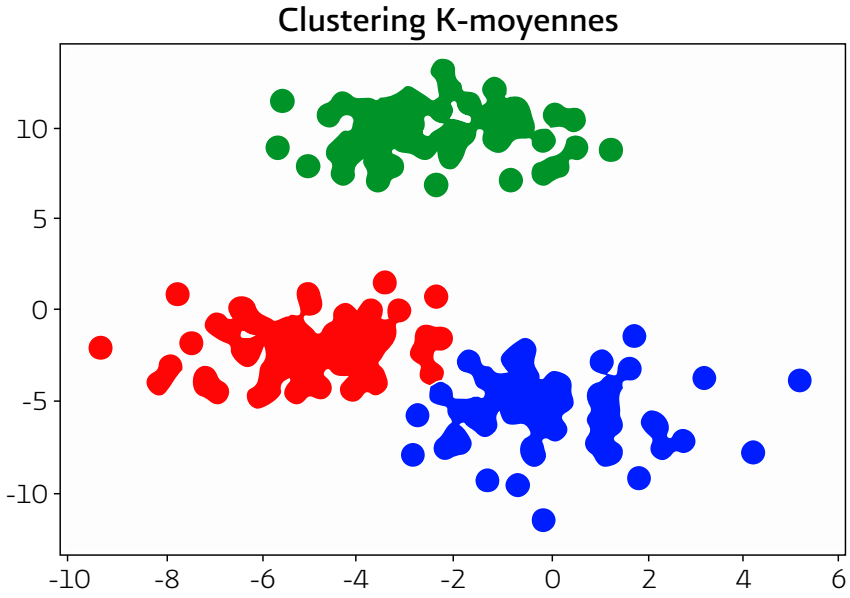


Figure 2. Représentation des similarités entre les documents d'un corpus.

Chaque point représente un document. La distance entre points représente la similarité lexico-sémantique. Les documents ont été organisés en trois ensembles distincts, identifiés par un algorithme de *clustering* (le *K-moyennes*). Chaque ensemble constitue un groupe de documents homogènes du point de vue sémantique. Il est possible d'interpréter ces groupes en termes de thèmes.

Du point de vue strictement mathématique, le modèle vectoriel correspond alors à la construction d'une matrice *U*. Cette modélisation est représentée par la formule 1 (fig. 3). Chaque ligne de cette matrice modélise un document sous la forme d'un vecteur $\vec{s_i} = (V_{i1}, V_{i2}, ..., V_{ij})$ dans lequel V_{ij} correspond à la *valeur de pondération* du $j^{ème}$ mot dans le $i^{ème}$ segment.

$U =$

	mot$_1$	**mot**$_2$	**...**	**mot**$_j$
document$_1$	V_{11}	V_{12}	...	V_{ij}
document$_2$	V_{21}	V_{22}	...	V_{ij}
...	...	...	...	...
document$_i$	V_{i1}	V_{i2}	...	V_{ij}

Figure 3. Formule 1 : matrice Documents-Mots.

Dans ce cadre, chaque document est décrit par son lexique. Par exemple, la transformation vectorielle des phrases « Marc mange une pomme et une fraise par jour » (phrase 1) et « Marie et Julie mangent un ananas et une pomme par semaine » (phrase 2) s'accomplit avec la construction de la matrice suivante :

Phrase/ Lemmes	ananas	et	fraise	jour	Julie	manger	Marc	Marie	par	pomme	semaine	un	une
Phrase 1	0	1	1	1	0	1	1	0	1	1	0	0	2
Phrase 2	1	2	0	0	1	1	0	1	1	1	1	1	1
Phrase n	…	…	…	…	…	…	…	…	…	…	…	…	…

Figure 4. Exemple d'une matrice Documents-Mots.

Dans ces cas, la matrice représente les deux phrases par la fréquence d'apparition de chaque lemme. Pour la phrase 1, l'article « une » a une fréquence égale à 2, tandis que le nom « semaine » la fréquence est de 0.

Ce type de matrice constitue le modèle sémantique de type *sac-de-mots* (de l'anglais, *bag-of-words*). Ce modèle se base sur le postulat qui affirme que la sémantique d'un document peut être représentée par l'ensemble de ses mots et sans tenir compte de leurs relations syntaxiques. Dans ce contexte, alors, les textes qui partagent des entités lexicales similaires tendent à être similaires, ce qui implique que les vecteurs qui les représentent tendent à se rapprocher dans l'espace géométrique construit par le modèle sémantique vectoriel. Ainsi, chaque phrase qui est ajoutée à la matrice avec nos exemples (fig. 4) peut être comparée aux précédentes en calculant la distance sémanti-co-lexicale. Par exemple, le calcul de l'angle entre les vecteurs qui représentent les deux phrases est égal à 0,945, ce qui correspond à une similarité sémantique très grande, considérant que la valeur maximale est 1.

2.2 L'hypothèse distributionnelle

À ce point, le lectorat pourrait s'interroger sur la légitimation sémiotique du modèle vectoriel. Ou, en d'autres termes, quelles sont les raisons sémiotiques qui expliquent, au moins partiellement, pourquoi ce modèle fonctionne? La réponse classique est fournie

par l'*hypothèse distributionnelle*. Formulée par le linguiste John Rupert Firth, l'hypothèse distributionnelle est devenue un classique des travaux en fouille de texte[44]. Cette hypothèse a été utilisée la première fois pour la construction de procédures de manipulation automatique du langage naturel par Zellig Harris, qui a assurément popularisé l'approche distributionnelle[45]. Cette approche se base sur l'idée que *le sens des mots est distribué dans le contexte* dans lequel il apparaît. En d'autres termes, les contextes linguistiques ayant des éléments lexicaux distribués de manière similaire ont des significations similaires. Le paradigme distributionnel se base donc sur la notion selon laquelle la similarité entre entités sémantiques dépend de leurs propriétés distributionnelles. Cela implique que la similarité du contenu entre deux textes est corrélée à la similarité de leurs distributions lexicales. Enfin, deux mots tendent à être similaires s'ils côtoient souvent les mêmes mots et, par conséquent, les entités sémantiques qui partagent des contextes similaires tendent à être similaires.

Cette hypothèse fait entièrement partie du modèle sémantico-vectoriel qui a été présenté dans la section précédente. En effet, chaque vecteur décrit un document sur la base de son lexique, lequel montre des relations avec les autres entités lexicales, c'est-à-dire des *relations de cooccurrence*, ce qui représente, selon l'hypothèse distributionnelle, un indice fiable et suffisant de la sémantique du mot et du texte. Or, l'hypothèse distributionnelle demeure la seule justification linguistique et sémiotique de toute recherche en ATO, mais aussi de toute application qui utilise le cadre théorique de la sémantique vectorielle. Ceci vaut pour la majorité des travaux en fouille de texte, qu'ils adoptent les approches de type supervisé, incluant l'apprentissage profond (par exemple, les représentations de *word embedding*), ou non supervisé. Sa solidité et sa pertinence ne sont pas remises en question dans cet article. Cependant, je veux aborder deux opérations spécifiques de la construction de la matrice, parce qu'elles sont fonctionnelles à la compréhension de la transposition mathématique du texte. La première renvoie à l'opération de *lemmatisation*, la seconde est la *fonction de pondération*. Ces opérations sont fondamentales pour bien construire la matrice et elles peuvent affecter négativement ou positivement le potentiel de l'hypothèse distributionnelle. Les prochaines sections permettent d'explorer si et comment *les caractéristiques textuelles transposées dans la matrice constituent de véritables simulacres des processus de signification qui sont en place dans le texte d'origine*.

2.3 La lemmatisation

La lemmatisation se rapporte à l'opération qui identifie le morphème porteur du signifié pour les unités lexicales des langues flexionnelles[46]. Dans un cadre d'analyse d'ATO, cette opération est accomplie pour *normaliser les formes des mots*, c'est-à-dire pour regrouper les occurrences de chaque mot (*tokens*) en une seule forme normalisée (*type*). Ceci est justifié par le fait que, du point de vue sémantique, ce ne sont pas tous les caractères qui composent l'occurrence d'un mot qui sont pertinents. Par exemple, le mot « mangeait » est composé de deux morphèmes : (1) le radical, qui renvoie à l'unité lexicale abstraite d'un mot et qui est le signifiant qui véhicule le signifié du mot[47] ; et (2) l'affixe, qui est utilisé pour conférer un rôle syntaxique au mot. Par exemple, dans le mot « mangeait », le radical « mange » constitue le signifiant qui est en relation directe avec le signifié du verbe « manger », tandis que l'affixe « ait » spécifie le temps, le nombre et la personne de l'action. Le but principal de ce processus est de mettre en valeur les aspects sémantiques du texte, en simplifiant ainsi les formes et les flexions que chaque mot peut prendre. Ainsi, la matrice *U* sera constituée de lemmes (*type*) plutôt que d'occurrences des mots (*tokens*). Par exemple, le lemme « manger » substituera chaque occurrence du mot « mangeait ». Enfin, cette procédure *confère au modèle vectoriel une nature lexico-sémantique*.

Cette procédure de « normalisation » peut être accomplie avec d'autres méthodes, comme la *racinisation* et le *découpage en n-gram de caractères*[48]. Cependant, surtout dans les applications d'ATO, il est préférable d'utiliser la procédure de lemmatisation puisqu'elle apporte un plus haut degré d'interprétabilité. En fait, les lemmes constituent des mots compréhensibles et font partie du vocabulaire de la langue cible, tandis que la racine ou le n-gram de caractères ne sont pas facilement reconductibles à un mot spécifique. Les deux autres procédures peuvent toutefois être choisies dans d'autres contextes, comme dans des applications de recherche d'information[49]. Cette procédure de sélection des lemmes, qui constitueront les caractéristiques de la représentation vectorielle du texte, peut également porter à la construction d'une matrice composée de n-gramme de lemmes, c'est-à-dire de segments de mots qui se répètent souvent dans le corpus à l'étude et qui peuvent constituer une unité de signification pertinente. Par exemple, de segments répétés[50] comme « agression sexuelle », « gaz à effet de serre », « accommodement raisonnable », etc.

2.3.1 Réflexions sur le processus d'identification des caractéristiques lexico-sémantiques d'un texte

Dans un cadre d'ATO, l'opération de lemmatisation est accomplie pour normaliser les formes des mots, c'est-à-dire pour regrouper les occurrences des mots (*tokens*) en types (forme lemmatisée du mot). La dichotomie *token* vs *type* est fondamentale en ATO. Umberto Eco a associé cette dichotomie au modèle du signe de Louis Hjelmslev[51] :

(matière) *substance* *forme*	CONTENU
forme *substance* (matière)	EXPRESSION

Figure 5. Expression et contenu selon Hjelmslev, reproduit d'après Eco[52].

Ce modèle (fig. 5) implique une structure du signe à deux plans, celui du plan du contenu (signifié) et celui de l'expression (signifiant), lesquels se présupposent réciproquement et sont entièrement réversibles. Le modèle de Hjelmslev complète la théorie du signe proposée par Ferdinand de Saussure en ajoutant à chaque plan trois niveaux de transformation. Suivant Umberto Eco :

> Selon ce modèle, on définit comme matière de l'expression tout continu amorphe auquel un système sémiotique déterminé donne forme en en découpant des éléments pertinents et structurés et en les produisant ensuite comme substance ; et l'on définit comme matière du contenu l'univers en tant que champ de l'expérience auquel une culture déterminée donne forme en en découpant des éléments pertinents et structurés et en les communiquant ensuite comme substance. La différence entre un élément de la forme et un élément de la substance est celle qui intervient entre un type et une occurrence concrète (*token*)[53].

En d'autres termes, la matière de l'expression et la matière du contenu sont un « tout continu amorphe » alors que la substance est la matière structurée par la forme. Ainsi, sur le plan du contenu, la matière amorphe est découpée en unités qui constituent la substance du contenu, ce qui est permis par la forme. Par exemple, le signifié général de « mourir » peut avoir différentes formes : « décéder » ou « partir », etc. Chacune de ces formes est porteuse d'un signifié (plan du contenu) qui se manifeste dans une forme particulière sur le plan de l'expression. Le rapport entre le plan du contenu et le plan

de l'expression est appelé fonction sémiotique. Chacune de ces formes peut avoir une nuance de signifié particulière, qui dépend du contexte d'utilisation de la forme. Ainsi, dans certains contextes, on dira « décédée » pour se référer à une personne qui est morte. Dans d'autres contextes, on utilisera le verbe « partir » pour se référer à la même personne décédée. Ceci implique que la substance du contenu demeure la même, soit la notion générale de « mourir ».

Quel est donc le lien entre forme et contenu, d'un côté, et le processus de lemmatisation et de conversion d'un *token* en *type*, de l'autre? Suivant les réflexions d'Umberto Eco, la dichotomie entre *type* et *token* peut être appréhendée à travers le modèle de Hjelmslev afin de considérer « la différence entre un élément de la forme et un élément de la substance [comme celle] qui intervient entre un type et une occurrence concrète (*token*)[54] ». Dans l'exemple précédent, le *type* « mourir » peut avoir différents *tokens*, comme « partir », « décéder », etc. Mais ces formes du contenu sont découpées à partir de la même substance du contenu, soit la notion générale de « mourir ».

Or, idéalement, la lemmatisation devrait adhérer à une *normalisation sémantique* du *token*, ce qui regrouperait les différentes formes de « mourir » sous le même *type*. Toutefois, la finesse d'une analyse de ce genre n'est pas encore praticable dans un cadre d'ATO avec les outils d'analyse sémantico-syntaxique disponibles aujourd'hui. Les procédures actuelles de lemmatisation permettent de normaliser les différentes occurrences et formes qu'un mot peut prendre, et donc de trouver les *types*, mais seulement dans un cadre linguistique du genre « partir » (*type*) *versus* « partais », « part » ou « partirais » (*token*). Elle ne permet pas de normaliser les formes du contenu et ses occurrences, comme pour le mot « mourir » (*type*) par rapport à « décéder », « partir », etc. Une piste d'amélioration de cette tâche consiste dans l'utilisation des outils d'annotation sémantique automatique, qui sont développés surtout dans le domaine de la recherche d'information et dans le web sémantique. Ces techniques sont utilisées pour détecter des entités sémantiques dans le texte à partir d'une base de connaissance préalablement construite, comme, par exemple, DBpedia[55]. Bien que ces outils semblent être prometteurs, leurs performances ne sont pas encore assez bonnes pour leur faire confiance dans un cadre d'ATO. Selon la plus récente étude comparative, la précision de ces annotateurs ne dépasse pas les 40%[56].

Ceci constitue sûrement une limite des outils en TALN et, par conséquent, de l'ATO. La finesse de la relation sémantique entre *type* et *token* est très difficile à détecter par le biais d'un « lexicographe automatique », car il nécessite des algorithmes de lemmatisation qui, en analysant le contexte d'occurrence d'un mot (*token*) (ex. « partais »), peut trouver son *type lexical* (« partir »), mais aussi son *type sémantique* (« mourir »). Dans ce cadre, l'algorithme de lemmatisation devrait transformer la phrase « il est parti

serein » en la liste suivante de lemmes, « il », « être », « mourir », « serein », ce qu'il est actuellement très difficile d'exécuter avec les outils informatiques.

Malgré ces limites, l'opération de réduction du *token* à son lemme demeure nécessaire pour mettre en valeur les éléments sémantiques des textes. Ceci comporte, toutefois, la perte des informations morphosyntaxiques des mots. Dans un contexte d'ATO, il est préférable de retrancher ces informations, car lorsqu'elles sont conversées, elles réduisent la valeur sémantique du modèle vectoriel. Même si on perd ces informations, la lemmatisation demeure nécessaire dans un contexte computationnel, car chaque réduction lexicale permet d'augmenter la *significativité statistique* du *token*, ce qui est important dans un contexte d'analyse de données et de corpus de grande taille. Ainsi, par exemple, il n'est pas essentiel de garder toutes les différentes variantes du verbe « manger » (il mange, ils mangent, etc.), mais il suffit de conserver son lemme. Il demeure toutefois important de souligner les limites de l'opération de lemmatisation et d'en être conscient, afin d'interpréter correctement les résultats d'une chaîne de traitement en ATO basée sur la sémantique vectorielle et les outils classiques de lemmatisation.

2.4 La fonction de pondération

Le deuxième élément sur lequel je veux porter l'attention du lectorat est la fonction de pondération. Cette dernière est la fonction qui permet de calculer le *poids* que chaque mot possède pour chaque document. Retournons à la matrice *U*. Il s'agit de la représentation mathématique d'un certain nombre de documents. Les chiffres qu'elle contient constituent des valeurs qui résument l'information lexico-sémantique du document sous *forme numérique*. Cette représentation mathématique des textes est fondamentale pour le traitement informatique. Il est alors important de comprendre comment ces valeurs numériques sont calculées et, ensuite, d'analyser ce que ces valeurs « disent » du point de vue sémiotique.

Pour résumer autrement le modèle sémantique vectoriel, on peut dire que la représentation vectorielle d'un groupe de textes est constituée par la matrice *U*, qui est composée des vecteurs d_i et chacun d'entre eux représente un document faisant partie du corpus *D*. Ces documents sont décrits par un certain nombre de caractéristiques lexico-sémantiques v_j (les lemmes), qui font partie du vocabulaire *V*. Le poids de chaque caractéristique est pondéré par une fonction mathématique qui constitue le cœur du modèle sémantique vectoriel. C'est cette fonction qui concrétise la conversion d'un texte dans un modèle computable.

Différentes méthodes de pondération existent et chacune d'elles évalue différemment les poids des caractéristiques dans un document. Les principales fonctions sont : la fonction binaire, la fonction de la fréquence et la fonction TF-IDF (de l'anglais, *Term frequency-Inverse document frequency*). La fonction binaire, comme son nom l'indique, assigne seulement des valeurs binaires (0/1), signalant ainsi la présence ou l'absence d'une caractéristique. La fonction de la fréquence, quant à elle, assigne à chaque caractéristique sa fréquence pour chaque document. Par exemple, si un lemme est répété plus d'une fois, la valeur ne sera pas un, mais un chiffre correspondant au nombre d'apparitions du lemme dans le document.

La dernière fonction, le TF-IDF, est la plus utilisée et il en existe différentes versions. Sa puissance repose sur l'hypothèse que le poids d'une caractéristique (un lemme) est *proportionnel à sa fréquence dans chaque document et inversement proportionnel à sa fréquence documentaire*. C'est-à-dire que la valeur TF-IDF augmente si le mot est très fréquent dans un petit nombre de documents, puisque, si le mot est très répandu dans le corpus, la valeur tend à diminuer. Pour exécuter cette fonction, on calcule d'abord la fréquence documentaire, c'est-à-dire le nombre de documents dans lesquels les caractéristiques apparaissent. Par exemple, si le lemme « étudiant » apparaît dans 232 documents, alors 232 sera la fréquence documentaire du lemme « étudiant ». La fréquence documentaire est ensuite convertie en fréquence documentaire inverse comme suit :

$$Idf = log \frac{n}{f(d_j)} + 1$$

Figure 6. Formule 2 : fonction *Idf*.

Dans cette formule (fig. 6), n est égal au nombre de documents de D et $f(d_j)$ est la fréquence documentaire du lemme j. Cette fonction est complétée par la normalisation logarithmique $log(idf) + 1$. La composante inverse de la fréquence documentaire est donc représentée par la fraction entre le nombre total n de documents du corpus D et la fréquence documentaire $f(d_j)$ du lemme. La fréquence documentaire inverse est ensuite multipliée par la fréquence du lemme, ce qui donne la formule suivante (fig. 7) :

$$TfIdf = f(j) \cdot log \frac{n}{f(d_j)} + 1$$

Figure 7. Formule 3 : fonction *Tf-Idf*.

Ainsi construite, la valeur de la fonction TF-IDF est plus élevée lors de l'augmentation de la fréquence du terme et elle est moins élevée lors de l'augmentation du nombre de documents qui contiennent le lemme. Donc, si un lemme est très fréquent et très répandu dans le corpus, il obtient une valeur TF-IDF plus basse, tandis qu'un lemme peu fréquent en général, mais très fréquent dans un sous-groupe de documents, obtient une valeur TF-IDF plus élevée. L'objectif de cette fonction est d'identifier *les caractéristiques les plus discriminantes dans un groupe de documents* ; en d'autres termes, les caractéristiques qui permettent de distinguer les documents en sous-groupes différents en raison de leur similarité lexico-sémantique.

2.4.1 La *valeur* saussurienne de la pondération

La pondération est la fonction qui assigne une valeur numérique à chaque caractéristique (par exemple, le lemme) de chaque document. Elle constitue la fonction mathématique la plus importante de l'étape de prétraitement du texte dans un contexte sémantique vectoriel. Or, il s'avère que les hypothèses et les postulats linguistiques en cause lors de la transformation d'un corpus en une matrice selon une fonction de pondération sont complémentaires à la conception structuraliste du sens[57]. Tel que je l'ai déjà écrit en introduction, Sahlgren a déjà montré les liens entre Saussure et Harris, et il a souligné l'importance de la notion de valeur linguistique pour la compréhension des modèles sémantiques vectoriels de type syntagmatique et de type paradigmatique. Dans ce contexte, je veux souligner la complémentarité qui, de mon point de vue, existe entre la notion de *poids* présupposée dans la fonction de pondération TF-IDF et la notion de valeur linguistique.

Dans le *Cours de linguistique générale*, Saussure définit le concept de valeur en le distinguant de celui de *signification*. Il donne l'exemple suivant : en anglais le mot « *sheep* » possède la même signification que le mot français « mouton »[58]. Dans la langue anglaise, le mot « *sheep* » ne peut pas être utilisé dans des contextes similaires à ceux dans lesquels le mot « mouton » est employé en français. En effet, pour indiquer une pièce de viande de mouton servie pendant un repas, on utilise en anglais le mot « *mutton* », alors qu'en français le même terme est gardé. Dans ce cas, « *sheep* » n'a donc pas la même valeur que « mouton », car il est remplacé par le mot « *mutton* » dans certains contextes. La valeur est alors une notion qui intègre plus directement le système-langue et les processus qui régissent l'agencement syntagmatique des mots. Comme on l'a vu, le mot « *mutton* » diffère du mot « *sheep* » et cette différence existe dans le système-langue, ce qui confère une valeur différente aux deux mots. Autrement

dit, la valeur indique *la position spécifique de chaque mot dans le système discret de la langue*. Chaque mot organise et structure la pensée d'une communauté linguistique qui, sinon, « n'est qu'une masse amorphe et indistincte[59] ».

> Philosophes et linguistes se sont toujours accordés à reconnaître que, sans le secours des signes, nous serions incapables de distinguer deux idées d'une façon claire et constante. Prise en elle-même, la pensée est comme une nébuleuse où rien n'est nécessairement délimité. Il n'y a pas d'idées préétablies, et rien n'est distinct avant l'apparition de la langue[60].

Pour Saussure, la langue est un système discret de signes qui donnent forme à la pensée amorphe. Dans ce cadre, le signe linguistique se met « en condition de signifier » seulement à l'intérieur d'un *système discret et non analogique*, composé d'éléments qui s'*opposent entre eux*. Saussure propose donc ce type de schéma (fig. 8) :

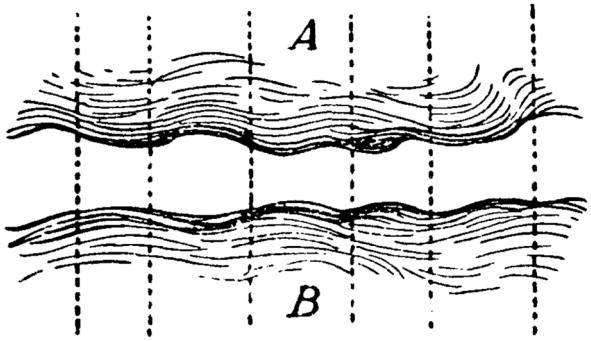


Figure 8. L'exemple de la feuille de papier[61].

La lettre A illustre les « idées confuses » et la lettre B représente le flou continu de tous les sons possibles, alors que les lignes verticales constituent les mots qui donnent forme à la pensée. L'image qui permet de mieux comprendre ce schéma est une feuille de papier. La langue est comparée à une feuille où la pensée est le recto et le signe perceptible est le verso. Lorsqu'on coupe le papier, il est impossible de couper le verso sans couper le recto. De la même manière, la langue coupe la pensée amorphe et construit un système discret.

Par sa proposition selon laquelle « dans la langue, il n'y a que des différences[62] », Saussure accorde de l'importance à *la position que le signe prend dans le système auquel il appartient*. Ce sont les différences entre les mots qui donnent la valeur à chaque forme de pensée constituée et perceptible par la langue. D'autre part, la signification se définit

verticalement à travers la relation arbitraire entre signifiant et signifié, alors que le plan horizontal est dominé par les relations discrètes et intrasystémiques entre les signes.

> En outre, l'idée de valeur, ainsi déterminée, nous montre que c'est une grande illusion de considérer un terme simplement comme l'union d'un certain son avec un certain concept. Le définir ainsi, ce serait l'isoler du système dont il fait partie ; ce serait croire qu'on peut commencer par les termes et construire le système en en faisant la somme, alors qu'au contraire c'est du tout solidaire qu'il faut partir pour obtenir par analyse les éléments qu'il renferme [...]. [...] la langue est un système dont tous les termes sont solidaires et où la valeur de l'un ne résulte que de la présence simultanée des autres [...][63].

Pour revenir à la différence entre signification et valeur, la signification s'organise sur deux plans, soit celui de *l'interaction entre signifiant et signifié* et celui des *relations intrasystémiques*. La *valeur*, de son côté, *est l'élément qui donne du poids aux relations intrasystémiques* et permet l'intégration de ces dernières dans les processus de la signification. En d'autres termes, le processus de signification d'un mot n'est pas limité à la relation arbitraire entre signifiant et signifié, car sa valeur est établie en évaluant les relations que le mot entretient avec les autres. Enfin, pour Saussure, *la valeur d'un mot est le résultat d'une fonction qui permet de le distinguer des autres mots*.

À mon avis, la notion de valeur constitue une base théorique de nature sémiotique pour comprendre la notion de « poids » proposée par la fonction TF-IDF. En effet, de manière similaire à la notion de valeur chez Saussure, l'objectif de la fonction TF-IDF est d'intégrer la globalité du système dans lequel un mot (ou lemme) s'insère afin d'évaluer son « poids » dans un document. En assignant un poids à chaque occurrence d'un mot (lemme), la fonction TF-IDF évalue toujours sa valeur, c'est-à-dire la position qu'il occupe à l'intérieur de l'« univers clos » qui constitue le corpus. Autrement dit, on évalue la valeur sémiotique de la caractéristique qui fait partie du vocabulaire *V* à l'intérieur de l'univers-clos du corpus *D* et on lui assigne une valeur numérique. Si la valeur sémiotique d'un mot se définit à l'intérieur du système-langue, le poids TF-IDF d'un mot (lemme) est défini à l'intérieur du *système-corpus*. Le système dans lequel un mot se positionne, qu'il soit général comme la langue ou relatif comme un corpus, ne peut pas être construit en commençant « par les termes et [...] en en faisant la somme »[64], mais en débutant par le « tout solidaire »[65], c'est-à-dire, par les relations entre les éléments qui le constituent. François Rastier va dans le même sens quand il affirme que :

> [...] sur l'axe syntagmatique, les classes sont définies par observation des cooccurrences : or, la création de listes de cooccurrences est une des opérations les plus simples et les plus éprouvées que permet la linguistique de corpus. Ainsi le concept de valeur étend-il également son efficacité aux groupements syntagmatiques que définissent les

> contextes : il importe donc à présent de redéfinir la valeur comme un rapport du texte à ses unités constituantes d'une part, à son corpus d'autre part[66].

Ce dernier extrait met en évidence le rôle du concept de valeur dans un cadre d'analyse propre à la linguistique de corpus. La valeur a un rôle très important dans la détermination du poids que les unités du texte ont en fonction du corpus, ce qui implique d'identifier la valeur d'un mot à travers le *jeu de cooccurrences* dans les textes et dans le corpus, duquel le mot fait partie.

En résumé, je propose de considérer la fonction TF-IDF comme une tentative « non volontaire » pour l'approximation mathématique du concept de valeur chez Saussure. La fonction TF-IDF évite d'isoler le mot du corpus dont il fait partie et, pour ce faire, la fréquence des mots est combinée à la fréquence documentaire inverse. De cette façon, comme c'est le cas pour la valeur sémiotique, le TF-IDF met en évidence les oppositions et les différences entre les termes au moyen de l'identification de *seuils discriminants* entre les mots, seuils qui coupent le recto et le verso d'une feuille. Enfin, comme le dirait Saussure, *dans le corpus il n'y a que différences*.

Conclusion

Cet article à visée introductive a été réalisé en deux parties. Dans la première partie, un survol des champs d'étude de la sémiotique computationnelle a été présenté. J'ai proposé un portait en trois branches, chacune d'elles caractérisant une approche particulière au regard de l'ordinateur : l'ordinateur comme machine sémiotique, où la sémiose est appréhendée comme un fait computable ; l'ordinateur comme artéfact, où la sémiose dérive de l'interaction entre l'humain et la machine ; et, enfin, l'ordinateur comme outil, où l'ordinateur est au service de l'analyse sémiotique classique. De ces branches, la troisième a fait l'objet d'un approfondissement. Ainsi, la pratique de l'analyse de texte assistée par ordinateur (ATO) a été présentée dans l'optique d'en interroger les enjeux sémiotiques. Poursuivant la démarche didactique adoptée dans cet article, j'ai examiné une étape spécifique de l'ATO, à savoir la transformation vectorielle du texte. Celle-ci est fondamentale pour la compréhension de l'usage de l'ATO et de l'usage d'outils numériques pour l'analyse sémiotique. J'ai ainsi présenté le modèle vectoriel, son cadre théorique et deux de ses opérations. J'ai souligné également l'importance de ces deux opérations pour le transfert des simulacres de la signification d'un modèle complexe et ambigu comme le texte à un modèle formel et computable comme la matrice.

L'une des contributions de cet article est d'avoir fourni une présentation originale de la sémiotique computationnelle, pour laquelle j'ai mis en évidence les différentes typologies de sémiose qu'elle étudie, et de l'ATO, pour laquelle j'ai fourni une description schématique de son interdisciplinarité. Ce texte contribue également à la discussion sur la procédure de lemmatisation et à prolonger les réflexions existantes sur la complémentarité entre la sémantique vectorielle et la linguistique structurale. À travers ces éléments, il met en lumière les fondements sémiotiques dans l'usage de l'ATO. En raison de sa nature introductive, l'article s'adresse principalement aux chercheurs des sciences humaines et sociales qui s'intéressent depuis peu à la sémiotique computationnelle.

En guise d'ouverture, ces réflexions sur les pratiques de l'ATO mènent à un élément important pour la sémiotique computationnelle, soit l'*approche empirique*, qui est favorisée par l'utilisation de corpus de très grande taille. Ceci correspond au contexte épistémologique qui caractérise l'analyse de mégadonnées (*big data*). La locution *big data* désigne le phénomène d'accumulation massive de données. Une grande masse de données sur le même phénomène constitue une valeur ajoutée pour l'analyse, mais elle apporte des défis, et plus particulièrement, l'emploi d'une approche méthodologique qui se caractérise par le concept de « recherche conduite par les données ». Cette approche s'inscrit dans l'approche empirique la plus classique et, en quelque sorte, la revitalise. En effet, cette approche, appelée « approche *data-driven* »[67], inaugure une nouvelle ère de l'empirisme, qui se caractérise par le volume et la diversité des données mises à disposition de l'analyse, ainsi que par une panoplie d'outils appartenant au domaine de la science de données. Dans ce contexte, l'approche empirique des données se radicalise, car l'analyse et l'extraction de connaissances sont effectuées sans les contraintes d'une théorie imposée d'avance[68]. Ainsi, l'analyse de données devient un outil à disposition des SHS, mais au risque de laisser les données parler par elles-mêmes, sans (ou avec peu de) contraintes théoriques.

Avec les changements qu'apporte le numérique, la sémiotique doit affronter de nouveaux défis amenés par les nouvelles techniques offertes par la science des données, l'apprentissage automatique et l'accumulation massive de données. Elle doit également évaluer les implications sémiotiques des outils utilisés, afin d'assister l'interprétation des résultats et la construction d'algorithmes et de chaînes de traitement pour l'analyse d'artéfacts sémiotiques. Cette nouvelle aventure de la sémiotique n'est qu'au début de son parcours, mais le potentiel de recherche qui s'entrevoit est déjà très important.

Bibliographie

ADAM, Jean-Michel, *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Armand Colin, 2011.

AGGARWAL, Charu C. & ChengXiang ZHAI (dir.), *Mining Text Data*, New York, Springer, 2012.

BEAUDOUIN, Valérie, « Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis », *Glottometrics*, no 33, 2016, p. 56-72.

BENZÉCRI, Jean-Paul & Françoise BENZÉCRI, *Analyse des Correspondances : exposé élémentaire*, Paris, Dunod, 1980.

BERNARD, Michel & Baptiste BOHET, *Littérométrie : outils numériques pour l'analyse des textes littéraires*, Paris, Presses Sorbonne nouvelle, 2017.

BISHOP, Christopher M., *Pattern Recognition and Machine Learning*, Singapour, Springer, 2006.

BOYD-GRABER, Jordan, Yuening HU & David MIMNO, « Applications of Topic Models », *Foundations and Trends® in Information Retrieval*, vol. 11, no 2-3, 2017, p. 143-296.

CARLEY, Kathleen, « Content Analysis », dans R. E. Asher *et al.* (dir.), *The Encyclopedia of Language and Linguistics*, vol. 2, Édimbourg, Pergamon Press, 1990, p. 725-730.

CHARTIER, Jean-François, Davide PULIZZOTTO, Louis CHARTRAND & Jean-Guy MEUNIER, « A Data-Driven Computational Semiotics: The Semantic Vector Space of Magritte's Artworks », *Semiotica*, vol. 2019, no 230, 2019, p. 19-69.

CLARK, Alexander, Chris FOX & Shalom LAPPIN, *The Handbook Of Computational Linguistics and Natural Language Processing*, Malden, Wiley-Blackwell, 2010.

COMPAGNO, Dario (dir.), *Quantitative Semiotic Analysis*, New York, Springer, 2018.

DACOS, Marin & Pierre MOUNIER, *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*, rapport de recherche, Institut français, 2015.

DE SOUZA, Clarisse Sieckenius, *The Semiotic Engineering of Human-Computer Interaction*, Cambridge, The MIT Press, 2005.

DIJK, Teun A. van, « Grammaires textuelles et structures narratives », dans S. Alexandrescu *et al.* (dir.), *Sémiotique narrative et textuelle*, Paris, Larousse, 1973, p. 177-207.

ECO, Umberto, « Pour une reformulation du concept de signe iconique », *Communications*, no 29, 1978, p. 141-191.

ERTEL, Wolfgang, *Introduction to Artificial Intelligence*, Londres, Springer, 2011.

ETXEBERRIA, Arantza & Jesus IBÁÑEZ, « Semiotics of the Artificial: The "Self" of Self-Reproducing Systems in Cellular Automata », *Semiotica*, vol. 127, no 1-4, 1999, p. 295-320.

FABBRI, Paolo, *Le tournant sémiotique*, trad. de l'italien par Y. Jeanneret, Paris, Hermès Science publications-Lavoisier, 2008.

FABRE, Cécile & Alessandro LENCI, « Sémantique distributionnelle », *Traitement automatique des langues*, vol. 56, no 2, 2015, 2015, p. 7-23.

FETZER, James H., « Minds and Machines: Limits to Simulations of Thought and Action », *International Journal of Signs and Semiotic Systems*, vol. 1, no 1, 2011, p. 39-48.

FIRTH, John Rupert, *Papers in Linguistics, 1934-1951*, Londres, Oxford University Press, 1957.

GAGNON, Michel, Amal ZOUAQ, Francisco ARANHA, Faezeh ENSAN & Ludovic JEAN-LOUIS, « An Analysis of the Semantic Annotation Task on the Linked Data Cloud », *International Journal of Metadata, Semantics and Ontologies*, vol. 13, no 4, 2019, p. 317-329.

GAMBHIR, Mahak & Vishal GUPTA, « Recent Automatic Text Summarization Techniques: A Survey », *Artificial Intelligence Review*, vol. 47, no 1, 2017, p. 1-66.

GREIMAS, Algirdas J. & Joseph COURTÉS, *Sémiotique : dictionnaire raisonné de la théorie du langage*, Paris, Hachette, 1979.

GUDWIN, Ricardo & Fernando A. C. GOMIDE, « Computational Semiotics: An Approach for the Study of Intelligent Systems-Part I: Foundations », Technical Report RT-DCA 09 - DCA-FEEC-UNICAMP, 1997.

HARRIS, Zellig S., « Distributional Structure », *Word*, vol. 10, no 2-3, 1954, p. 146-162.

HASTIE, Trevor, Robert TIBSHIRANI & Jerome FRIEDMAN, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, Springer, 2013.

HJELMSLEV, Louis, *Prolégomènes à une théorie du langage*, trad. du danois par U. Canger, avec la collab. d'A. Wewer, préface de V. Hjelm, Paris, Minuit, coll. « Arguments », 1968.

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE & Robert TIBSHIRANI, *An Introduction to Statistical Learning: with Applications in R*, New York, Springer, 2013.

KETNER, Kenneth Laine, « Peirce and Turing: Comparisons and conjectures », *Semiotica*, vol. 68, no 1-2, 1988, p. 33-62.

KITCHIN, Rob, « Big Data, New Epistemologies and Paradigm Shifts », *Big Data & Society*, vol. 1, no 1, 2014, p. 1-12.

KRIPPENDORFF, Klauss, *Content Analysis: An Introduction to Its Methodology*, Thousand Oaks, SAGE Publications, 2004.

LEBART, Ludovic, Bénédicte PINCEMIN & Céline POUDAT, *Analyse des données textuelles*, Québec, Presses de l'Université du Québec, 2019.

LEBART, Ludovic & André SALEM, *Statistique textuelle*, Paris, Dunod, 1994.

MAINGUENEAU, Dominique, *L'analyse du discours*, Paris, Hachette Supérieur, 1997.

MALATERRE, Christophe, Jean-François CHARTIER & Davide PULIZZOTTO, « What Is This Thing Called Philosophy of Science? A Computational Topic-Modeling Perspective, 1934-2015 », *HOPOS. The Journal of the International Society for the History of Philosophy of Science*, vol. 9, no 2, 2019, p. 215-249.

MALATERRE, Christophe, Davide PULIZZOTTO & Francis LAREAU, « Revisiting Three Decades of *Biology and Philosophy*: A Computational Topic-Modeling Perspective », *Biology & Philosophy*, vol. 35, no 1, 2020, p. 1-25.

MANI, Inderjeet & Mark T. MAYBURY, *Advances in Automatic Text Summarization*, Cambridge, The MIT Press, 1999.

MANNING, Christopher D., Prabhakar RAGHAVAN & Hinrich SCHÜTZE, *Introduction to Information Retrieval*, Cambridge, Cambridge University Press, 2008.

MANNING, Christopher D. & Hinrich SCHÜTZE, *Foundations of Statistical Natural Language Processing*, Cambridge, The MIT Press, 1999.

MARRONE, Gianfranco, *Corpi sociali: processi comunicativi e semiotica del testo*, Turin, Einaudi, 2001.

MEUNIER, Jean-Guy, « Artificial intelligence and sign theory », *Semiotica*, vol. 77, no 1-3, 1989, p. 43-64.

—, « Humanités numériques et modélisation scientifique », *Questions de communication*, vol. 1, no 31, 2017, p. 19-48.

MORENO, Juan Manuel Torres, *Automatic Text Summarization*, Somerset, Wiley, 2014.

MORETTI, Franco, *Distant reading*, Londres, Verso Books, 2013.

NADIN, Mihai, « Information and Semiotic Processes: The Semiotics of Computation », *Cybernetics & Human Knowing*, vol. 18, no 1-2, 2011, p. 153-175.

NÉE, Émilie (dir.), *Méthodes et outils informatiques pour l'analyse des discours*, Rennes, Presses universitaires de Rennes, coll. « Didact Méthode », 2017.

POLGUÈRE, Alain, *Lexicologie et sémantique lexicale : notions fondamentales*, Montréal, Presses de l'Université de Montréal, 2016.

POUDAT, Céline & Frédéric LANDRAGIN, *Explorer un corpus textuel : méthodes, pratiques, outils*, Paris, De Boeck supérieur, 2017.

PULIZZOTTO, Davide, Jean-François CHARTIER, Francis LAREAU, Jean-Guy MEUNIER & Louis CHARTRAND, « Conceptual Analysis in a Computer-Assisted Framework: Mind in Peirce », *Umanistica Digitale*, vol. 2, no 2, 2018, p. 185-205.

PULIZZOTTO, Davide, Jean-François CHARTIER, Jean-Guy MEUNIER, Louis CHARTAND, Francis LAREAU & Louis HÉBERT, « Vers une sémiotique computationnelle : étude de cas et premières explorations », *Applied Semiotics / Semiotique appliquée*, no 26, 2018, p. 192-208.

QUEIROZ, João & Floyd MERRELL, « On Peirce's Pragmatic Notion of Semiosis—A Contribution for the Design of Meaning Machines », *Minds and Machines*, vol. 19, no 1, 2009, p. 129-143.

RAPAPORT, William J., « Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing », *International Journal of Signs and Semiotic Systems*, vol. 2, no 1, 2012, p. 32-71.

RASTIER, François, *Sémantique interprétative*, Paris, Presses universitaires de France, 2009.

—, « Objets culturels et performances sémiotiques. L'objectivation critique dans les sciences de la culture », dans L. Hébert & L. Guillemette (dir.), *Performances et objets culturels. Nouvelles perspectives*, Sainte-Foy, Presses de l'Université Laval, 2010, p. 15-58.

—, *La mesure et le grain : sémantique de corpus*, Paris, Honoré Champion, 2011.

—, « Computer-Assisted Interpretation of Semiotic Corpora », dans D. Compagno (dir.), *Quantitative Semiotic Analysis*, New York, Springer, 2018, p. 123-139.

RICH, Elaine, Kevin KNIGHT & Shivashankar B. NAIR, *Artificial intelligence*, New Delhi, Tata McGraw-Hill, 2009.

ROSENBLATT, Frank, « The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain », *Psychological review*, vol. 65, no 6, 1958, p. 386-408.

SAHLGREN, Magnus, *The Word-Space Model*, thèse de doctorat, Stockholm, Univerité de Stockholm, 2006.

—, « The Distributional Hypothesis », *Italian Journal of Linguistics*, vol. 20, no 1, 2008, p. 33-53.

SALTON, Gerard & Michael J. McGILL, *Introduction to Modern Information Retrieval*, New York, McGraw-Hill, 1983.

SALTON, Gerard, Amit SINGHAL, Mandar MITRA & Chris BUCKLEY, « Automatic Text Structuring and Summarization », *Information Processing & Management*, vol. 33, no 2, 1997, p. 193-207.

SALTON, Gerard, Andrew WONG & Chung-Shu YANG, « A Vector Space Model for Automatic Indexing », *Communications of the ACM*, vol. 18, no 11, 1975, p. 613-620.

SAUSSURE, Ferdinand de, *Cours de linguistique générale*, publié par C. Bailly & A. Sechehaye avec la collab. d'A. Riedlinger, éd. critique préparée par T. de Mauro, Paris, Payot, 1995 [1916].

SHAH, Neepa & Sunita MAHAJAN, « Document Clustering: A Detailed Review », *International Journal of Applied Information Systems,* vol. 4, no 5, 2012, p. 30-38.

STAMPER, Ronald, *Information in Business and Administrative Systems*, New York, John Wiley & Sons, 1973.

TANAKA-ISHII, Kumiko, *Semiotics of programming*, New York, Cambridge University Press, 2010.

—, « Semiotics of Computing: Filling the Gap Between Humanity and Mechanical Inhumanity », dans P. P. Trifonas (dir.), *International Handbook of Semiotics*, New York, Springer, 2015, p. 981-1002.

TURNEY, Peter D. & Patrick PANTEL, « From Frequency to Meaning: Vector Space Models of Semantics », *Journal of Artificial Intelligence Research*, vol. 37, 2010, p. 141-188.

ZEMANEK, Heinz, « Semiotics and Programming Languages », *Communications of the ACM*, vol. 9, no 3, 1966, p. 139-143.

Notes

1 P. D. TURNEY & P. PANTEL, « From Frequency to Meaning: Vector Space Models of Semantics », *Journal of Artificial Intelligence Research*, vol. 37, 2010, p. 141-188.

2 T. A. van DIJK, « Grammaires textuelles et structures narratives », dans S. Alexandrescu *et al.* (dir.), *Sémiotique narrative et textuelle*, Paris, Larousse, 1973, p. 177-207.

3 M. SAHLGREN, *The Word-Space Model*, thèse de doctorat, Stockholm, Univerité de Stockholm, 2006.

4 M. SAHLGREN, « The Distributional Hypothesis », *Italian Journal of Linguistics*, vol. 20, no 1, 2008, p. 33-54.

5 Voici une sélection des plus récents travaux : J.-F. CHARTIER *et al.*, « A Data-Driven Computational Semiotics: The Semantic Vector Space of Magritte's Artworks », *Semiotica*, no 230, 2019, p. 19-69 ; C. MALATERRE *et al.*, « What Is This Thing Called Philosophy of Science? A Computational Topic-Modeling Perspective, 1934-2015 », *HOPOS. The Journal of the International Society for the History of Philosophy of Science*, vol. 9, no 2, 2019, p. 215-249 ; C. MALATERRE, D. PULIZZOTTO & F. LAREAU, « Revisiting Three Decades of *Biology and*

Philosophy: A Computational Topic-Modeling Perspective », *Biology & Philosophy*, vol. 35, no 1, 2020, p. 1-25 ; D. PULIZZOTTO *et al.*, « Conceptual Analysis in a Computer-Assisted Framework: Mind in Peirce », *Umanistica Digitale*, vol. 2, no 2, 2018, p. 185-205 ; D. PULIZZOTTO *et al.*, « Vers une sémiotique computationnelle : étude de cas et premières explorations », *Applied Semiotics / Semiotique appliquée*, no 26, 2018, p. 192-208.

6 M. DACOS & P. MOUNIER, *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*, rapport de recherche, Institut français, 2015, p. 7.

7 Cf. W. ERTEL, *Introduction to Artificial Intelligence*, Londres, Springer, 2011, p. 2.

8 E. RICH, K. KNIGHT & S. B. NAIR, *Artificial intelligence*, New Delhi, Tata McGraw-Hill, 2009, p. 3. Je traduis : « Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better. »

9 G. MARRONE, *Corpi sociali: processi comunicativi e semiotica del testo*, Turin, Einaudi, 2001.

10 Cf. A. ETXEBERRIA & J. IBÁÑEZ, « Semiotics of the Artificial: The "Self" of Self-Reproducing Systems in Cellular Automata », *Semiotica*, vol. 127, no 1-4, 1999, p. 295-320 ; J. H. FETZER, « Minds and Machines: Limits to Simulations of Thought and Action », *International Journal of Signs and Semiotic Systems*, vol. 1, no 1, 2011, p. 39-48 ; K. L. KETNER, « Peirce and Turing: Comparisons and conjectures », *Semiotica*, vol. 68, no 1-2, 1988, p. 33-62 ; J.-G. MEUNIER, « Artificial Intelligence and Sign Theory », *Semiotica*, vol. 77, no 1-3, 1989, p. 43-64 ; W. J. RAPAPORT, « Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing », *International Journal of Signs and Semiotic Systems*, vol. 2, no 1, 2012, p. 32-71.

11 Cf. C. S. DE SOUZA, *The Semiotic Engineering of Human-Computer Interaction*, Cambridge, The MIT Press, 2005 ; M. NADIN, « Information and Semiotic Processes: The Semiotics of Computation », *Cybernetics & Human Knowing*, vol. 18, no 1-2, 2011, p. 153-175 ; R. K. STAMPER, *Information in Business and Administrative Systems*, New York, John Wiley & Sons, 1973 ; K. TANAKA-ISHII, *Semiotics of programming*, New York, Cambridge University Press, 2010 ; H. ZEMANEK, « Semiotics and Programming Languages », *Communications of the ACM*, vol. 9, no 3, 1966, p. 139-143.

12 Cf. L. LEBART & A. SALEM, *Statistique textuelle*, Paris, Dunod, 1994 ; F. MORETTI, *Distant reading*, Londres, Verso Books, 2013 ; M. BERNARD & B. BOHET, *Littérométrie : outils numériques pour l'analyse des textes littéraires*, Paris, Presses Sorbonne nouvelle, 2017 ; D. COMPAGNO (dir.), *Quantitative Semiotic Analysis*, New York, Springer, 2018 ; F. RASTIER, « Computer-Assisted Interpretation of Semiotic Corpora », dans D. Compagno (dir.), *Quantitative Semiotic Analysis*, New York, Springer, 2018, p. 123-139 ; J.-F. CHARTIER *et al.*, « A Data-Driven Computational Semiotics: The Semantic Vector Space of Magritte's Artworks », *loc. cit*.

13 J.-G. MEUNIER, « Humanités numériques et modélisation scientifique », *Questions de communication*, vol. 1, no 31, 2017, p. 19-48.

14 K. TANAKA-ISHII, « Semiotics of Computing: Filling the Gap Between Humanity and Mechanical Inhumanity », dans P. P. Trifonas (dir.), *International Handbook of Semiotics*, New York, Springer, 2015, p. 981-1002.

15 R. R. GUDWIN & F. A. C. GOMIDE, « Computational Semiotics: An Approach for the Study of Intelligent Systems-Part I: Foundations », Technical Report RT-DCA 09 - DCA-FEEC-UNICAMP, 1997.

16 J. QUEIROZ & F. MERRELL, « On Peirce's Pragmatic Notion of Semiosis—A Contribution for the Design of Meaning Machines », *Minds and Machines*, vol. 19, no 1, 2009, p. 129-143.

17 A. ETXEBERRIA & J. IBÁÑEZ, « Semiotics of the Artificial: The "Self" of Self-Reproducing Systems in Cellular Automata », *loc. cit*.

18 Cf. J.-G. MEUNIER, « Artificial Intelligence and Sign Theory », *loc. cit.* ; J.-G. MEUNIER, « Humanités numériques et modélisation scientifique », *loc. cit*.

19 F. RASTIER, « Objets culturels et performances sémiotiques. L'objectivation critique dans les sciences de la culture », dans L. Hébert & L. Guillemette (dir.), *Performances et objets culturels. Nouvelles perspectives*, Sainte-Foy, Presses de l'Université Laval, 2010, p. 17-19.

20 Cf. C. S. DE SOUZA, *The Semiotic Engineering of Human-Computer Interaction*, *op. cit*.

21 Cf. R. K. STAMPER, *Information in Business and Administrative Systems*, *op. cit*.

22 Cf. L. LEBART & A. SALEM, *Statistique textuelle*, *op. cit*.

23 Cf. K. CARLEY, « Content Analysis », dans R. E. Asher *et al.* (dir.), *The Encyclopedia of Language and Linguistics*, vol. 2, Édimbourg, Pergamon Press, 1990, p. 725-730 ; K. KRIPPENDORFF, *Content Analysis: An Introduction to Its Methodology*, Thousand Oaks, SAGE Publications, 2004.

24 Cf. J.-M. ADAM, *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Armand Colin, 2011 ; D. MAINGUENEAU, *L'analyse du discours*, Paris, Hachette Supérieur, 1997.

25 Cf. C. POUDAT & F. LANDRAGIN, *Explorer un corpus textuel : méthodes, pratiques, outils*, Paris, De Boeck supérieur, 2017 ; F. RASTIER, *Sémantique interprétative*, Paris, Presses universitaires de France, 2009.

26 On parle aussi de « statistique textuelle », de « textométrie », de « sémiométrie », de « *distant reading* », de « linguistique de corpus », etc. Les pratiques auxquelles ces expressions réfèrent, bien qu'avec des différences, sont très comparables à celles de l'ATO.

27 Cf. E. NÉE (dir.), *Méthodes et outils informatiques pour l'analyse des discours*, Rennes, Presses universitaires de Rennes, coll. « Didact Méthode », 2017.

28 Cf. C. POUDAT & F. LANDRAGIN, *Explorer un corpus textuel, op. cit.,* p. 90.

29 Cf. J.-P. BENZÉCRI & F. BENZÉCRI, *Analyse des Correspondances : exposé élémentaire*, Paris, Dunod, 1980.

30 Cf. C. D. MANNING & H. SCHÜTZE, *Foundations of Statistical Natural Language Processing*, Cambridge, The MIT Press, 1999.

31 Cf. A. CLARK, C. FOX & S. LAPPIN, *The Handbook Of Computational Linguistics and Natural Language Processing*, Malden, Wiley-Blackwell, 2010.

32 Avec la locution de lexicographe automatique, je regroupe les techniques qui permettent d'associer une suite de caractères à une entrée de dictionnaire d'une langue, comme la lemmatisation ou, dans une moindre mesure, la racinisation. Les analyseurs morphosyntaxiques, quant à eux, sont des techniques qui permettent d'effectuer une analyse morphosyntaxique automatisée de chaque mot d'un texte.

33 Cf. G. SALTON *et al.*, « Automatic Text Structuring and Summarization », *Information Processing & Management*, vol. 33, no 2, 1997, p. 193-207 ; I. MANI & M. T. MAYBURY, *Advances in Automatic Text Summarization*, Cambridge, The MIT Press, 1999 ; J. M. T. MORENO, *Automatic Text Summarization*, Somerset, Wiley, 2014 ; M. GAMBHIR & V. GUPTA, « Recent Automatic Text Summarization Techniques: A Survey », *Artificial Intelligence Review*, vol. 47, no 1, 2017, p. 1-66.

34 Cf. C. M. BISHOP, *Pattern Recognition and Machine Learning*, Singapour, Springer, 2006 ; T. HASTIE, R. TIBSHIRANI & J. FRIEDMAN, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, Springer, 2013.

35 Cf. G. JAMES *et al.*, *An Introduction to Statistical Learning: with Applications in R*, New York, Springer, 2013.

36 Cf. N. SHAH & S. MAHAJAN, « Document Clustering: A Detailed Review », *International Journal of Applied Information Systems*, vol. 4, no 5, 2012, p. 30-38.

37 Cf. J. BOYD-GRABER, Y. HU & D. MIMNO, « Applications of Topic Models », *Foundations and Trends® in Information Retrieval*, vol. 11, no 2-3, 2017, p. 143-296.

38 Cf. C. FABRE & A. LENCI, « Sémantique distributionnelle », *Traitement automatique des langues*, vol. 56, no 2, 2015, 2015, p. 7-23.

39 La recherche d'information est un domaine de l'informatique qui étudie les manières d'extraire de l'information d'une collection de documents à partir d'une requête précise fournie par l'utilisateur. Voir notamment G. SALTON & M. J. McGILL, *Introduction to Modern Information Retrieval*, New York, McGraw-Hill, 1983.

40 Cf. C. C. AGGARWAL & CX. ZHAI (dir.), *Mining Text Data*, New York, Springer, 2012.

41 F. ROSENBLATT, « The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain », *Psychological review*, vol. 65, no 6, 1958, p. 386-408.

42 G. SALTON, A. WONG & C. S. YANG « A Vector Space Model for Automatic Indexing », *Communications of the ACM*, vol. 18, no 11, 1975, p. 613-620.

43 V. BEAUDOUIN, « Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis », *Glottometrics*, no 33, 2016, p. 56-72.

44 Firth a formulé une expression devenue phare : « Il est possible de connaître la signification d'un mot en regardant avec qui il se tient. » J. R. FIRTH, *Papers in Linguistics, 1934-1951*, Londres, Oxford University Press, 1957, p. 11. Je traduis : « You shall know a word by the company it keeps. »

45 Z. HARRIS, « Distributional Structure », *Word*, vol. 10, no 2-3, 1954, p. 146-162.

46 Les langues flexionnelles, à la différence des langues isolantes, ont une morphologie complexe, ce qui justifie des opérations de lemmatisation.

47 Cf. A. POLGUÈRE, *Lexicologie et sémantique lexicale : notions fondamentales*, Montréal, Presses de l'Université de Montréal, 2016.

48 Cf. G. SALTON & M. J. McGILL, *Introduction to Modern Information Retrieval*, *op. cit.*

49 Cf. C. D. MANNING, P. RAGHAVAN & H. SCHÜTZE, *Introduction to Information Retrieval*, Cambridge, Cambridge University Press, 2008, p. 39.

50 Cf. L. LEBART, B. PINCEMIN & C. POUDAT, *Analyse des données textuelles*, Québec, Presses de l'Université du Québec, 2019.

51 U. ECO, « Pour une reformulation du concept de signe iconique », *Communications*, vol. 29, 1978, p. 141-191.

52 *Ibid.*, p. 141.

53 *Idem*.

54 *Idem.*

55 Relié à Wikipédia, DBpedia forme un « projet universitaire et communautaire d'exploration et extraction automatiques de données dérivées de Wikipédia. Son principe est de proposer une version structurée et normalisée au format du web sémantique des contenus de Wikipedia ». En ligne : <https://fr.wikipedia.org/wiki/DBpedia>. Voir également le site web du projet : <https://wiki.dbpedia.org/>.

56 M. GAGNON *et al.*, « An Analysis of the Semantic Annotation Task on the Linked Data Cloud », *International Journal of Metadata, Semantics and Ontologies*, vol. 13, no 4, 2019, p. 317-329.

57 M. SAHLGREN, « The Distributional Hypothesis », *Italian Journal of Linguistics*, *loc. cit.*

58 F. de SAUSSURE, *Cours de linguistique générale*, publié par C. Bailly & A. Sechehaye avec la collab. d'A. Riedlinger, éd. critique préparée par T. de Mauro, Paris, Payot, 1995 [1916], p. 160.

59 *Ibid.*, p. 155.

60 *Idem.*

61 *Ibid.*, p. 156.

62 *Ibid*., p. 166.
63 *Ibid*., p. 157-159.
64 *Ibid*., p. 157.
65 *Idem*.
66 F. RASTIER, *La mesure et le grain : sémantique de corpus*, Paris, Honoré Champion, 2011, p. 30.
67 R. KITCHIN, « Big Data, New Epistemologies and Paradigm Shifts », *Big Data & Society*, vol. 1, no 1, 2014, p. 1-12.
68 *Ibid.*