Eleonora Marzi

LANGUE, LITTÉRATURE ET INFORMATIQUE : L'INTERCULTURALITÉ EN PERSPECTIVE

Études de traitement automatique des langues et analyse du discours

I LIBRI DI EMIL

ISBN : 978-88-6680-423-9

I libri di Emil
via Carlo Marx 21– 06012 Città di Castello (PG)
www.ilibridiemil.it

Table de matières

SECONDE PARTIE

Introduction au volume

Appliquer des techniques de traitement automatique des langues à des typologies de discours, voilà un des enjeux les plus importants du domaine de la recherche en humanités numériques. C'est dans ce cadre que nous avons cherché à indiquer des pistes de travail qui, tout en se focalisant sur le discours institutionnel et le discours littéraire, visent à identifier les stratégies linguistiques qui y sont à l'œuvre. Notre recherche s'est centrée sur le discours interculturel pour en décerner les stratégies linguistiques ainsi que ses possibles sens latents.

La nature interdisciplinaire de cette étude motive la structure bipartite du volume qui cherche à fournir des points de repères théoriques et à démontrer son application dans des cas d'étude. La première partie –consacrée aux aspects théoriques des Humanités Numériques- prête une attention particulière à la naissance de cette discipline et à son panorama actuel. Bien que les Humanités Numériques touchent à l'ensemble des disciplines humanistes (art, histoire, etc…), nous nous intéresserons plus particulièrement à l'extraction d'information du texte, où, d'un point de vue technologique, prévaut le rapport entre sens et parole numérisée. Les méthodes et les outils informatiques du traitement automatique des langues seront illustrés sous un angle permettant leur accessibi-

lité sans que des compétences informatiques spécifiques soient nécessaires.

Cette première partie théorique propose également une réflexion sur l'interculturalité et sur le dialogue interculturel. Les stratégies se mettant en place pour parler à l'autre et pour parler *de* l'autre se fondent souvent sur l'implicite linguistique : pour la plupart, porteur de préjugés ou malentendus, l'implicite linguistique est de fait fondamentale pour assurer une coexistence entre les différentes cultures.

La deuxième partie est consacrée à deux cas d'études illustrant une application concrète des techniques pour des thématiques humanistes, au sein des Humanités Numériques.

Le premier cas d'étude prend en considération le discours institutionnel tenu par le *Ministère de l'Immigration, de la Francisation et de l'Intégration du Québec* (MIFI). Le corpus a été collecté à partir du matériel présent dans le site web institutionnel. Les stratégies linguistiques employées pour décrire la figure du migrant sont analysées à travers l'approche ACDAC – analyse critique du discours assistée par corpus.

Le deuxième cas d'étude est consacré au discours littéraire, à travers l'analyse d'un corpus de littérature française contemporaine, construit à partir des romans ayant obtenu le *Prix Goncourt* au cours des vingt dernières années. Par l'utilisation de la perspective des *Digital Literary Studies* et en particulier en s'inspirant au *Distant Reading* ainsi que par le recours auxtechniques du géo-référencement, il sera

possible d'analyser comment les cultures sont mises en fiction, quel est le lexique attribué aux diverses entités culturelles et quels sont les éventuels jugements de valeur qui y sont associés.

Les diverses typologies de discours (dans notre cas institutionnel et littéraire) supposent des prémisses de contexte nécessaires à l'introduction des deux cas d'études. Le commun dénominateur des deux cas d'études s'inscrit, toutefois, dans la thématique de l'interculturalité, à savoir comment le rapport à *l'Autre* et à sa culture se révèle dans le discours au niveau lexical, syntactique, ainsi que dans l'implicite présent dans la communication.

La méthodologie adoptée se propose de systématiser des étapes de la recherche au sein des Humanités Numériques pour l'extraction d'information des textes.

En effet, dans une démarche interdisciplinaire, la transposition d'une problématique donnée d'un domaine à l'autre représente une difficulté méthodologie majeure et avérée. Nous nous interrogerons donc sur comment comprendre l'expression « dialogue interculturel » ? S'il s'agit d'un discours sur l'autre, il peut être analysé à travers des ressources linguistiques, à la recherche de structures syntactiques ou d'un lexique précis. Ainsi, pour cette phase, la connaissance des instruments et des techniques du traitement automatique des langues apparait comme fondamentale, car elle permet de formuler la « bonne » question pour la recherche. L'adjectif « bonne » ne renvoie pas ici à un jugement de qualité ; il se réfère plutôt à une idée de pertinence par rapport aux instruments disponibles : nous ver-

rons tout au long du volume comment, dans la perspective des Humanités Numériques, la formulation des questions de recherche doit tenir compte des possibilités offertes par les outils d'élaboration. Cela est dû au caractère circulaire du rapport s'établissant entre les deux faces de la discipline – les humanités et le numérique – une fois la question posée, il convient d'identifier les techniques et les instruments d'analyse automatique adéquats et vice-versa : la question de la recherche peut être formulée sur la base de la connaissance préalable des instruments disponibles. En effet la recherche actuelle en traitement automatique des langues s'articule généralement autour de tâches bien identifiées et plus ou moins complexes là où les HN utilisent des techniques et des méthodes de TAL comme outils pour décrire des scénarios de recherche complexes sur des thématiques variées. De même, les progrès en TAL sont censés favoriser des retombées positives pour les recherches dans le secteur des humanités sachant quels défis ultimes dans la perspective des Humanités Numériques ne visent pas uniquement à améliorer les performances des outils de TAL en eux-mêmes mais bien leur utilisation en vue d'une recherche innovante capable de faire véritablement avancer la connaissance disciplinaire dans les différents champs des Humanités. Appliquer les techniques de traitement automatique ne signifie pas pour autant effacer les méthodes traditionnelles, mais bien les compléter avec d'autres typologies de recherches, sur les mêmes thématiques sous une perspective différente, influencée à la fois par la taille et la

nature des données disponibles, et par les outils développés pour les manipuler.

Le but de ce travail est donc de montrer des parcours possibles pour l'extraction automatique des informations du texte tout en posant une réflexion sur la nature de l'union entre informatique et sciences humaines.

PREMIÈRE PARTIE

Les humanités numériques : une chimère d'aujourd'hui

1.1. LE CADRE THÉORIQUE

1.1.1. Les enjeux d'une définition

Les Humanités Numériques, ou *digital humanities*, sont un domaine d'études qui dérive de l'intégration de procédures informatiques dans les sciences humaines et sociales, notamment en relation avec la représentation des données, la formalisation des phases de recherche et les techniques de diffusion des résultats. La question d'une définition unanime est, depuis toujours, une question épineuse : si, sur le fond, il ne subsiste aucune difficulté majeure pour s'accorder sur les disciplines inclues dans les Humanités Numériques selon les diverses traditions géoculturelles (il y a souvent débat entre le monde anglo-saxon et français quant au statut des sciences sociales faisant part des humanités)[1], le point critique réside, néanmoins, dans le débat sur la spécificité et l'apport de l'union entre le numérique et les Humanités. Affirmer simplement que l'on parle d'Humanités Numériques lorsqu'on se sert d'un ordinateur pour traiter les savoirs humanistes n'est pas suffisant, étant donné le rôle transversal que ce dispositif occupe aujourd'hui dans notre quotidien : il est donc urgent de

[1] Levi Strauss C., « L'apport des sciences sociales à l'humanisation de la civilisation technique », *Courrier de l'UNESCO*, 2008 (Archives inédits UNESCO 8 août 1956). [https ://fr.unesco.org/courier/2008-5/aportacion-ciencias-sociales-humanizacion-civilizacion-tecnica, [dernier accès 1/1/2021].

poser une limite à partir de laquelle l'usage du numérique est déterminant pour les résultats de la recherche.

La partie littéraire des Humanités Numériques, connue également comme *Digital Literary Studies*, a accordé, dès ses débuts, une grande importance à la parole, unité minimale de sens. Deux approches principales la composent : d'un côté la conservation du matériel et l'enrichissement d'éditions critiques qui visualisent un réseau d'informations : sur la biographie de l'auteur, ou bien sur les possibles éditions critiques, ou encore sur les versions des traductions. Les résultats sont des éditions critiques « augmentées » qui se rapprochent des représentations de réalité virtuelle. De l'autre, une approche visant à extraire des informations du texte, ce qui implique la capacité de formuler des modèles de langage et des modèles de connaissance.

L'expansion des *Big Data* au cours des dernières décennies a considérablement élargi les outils disponibles pour élaborer un texte et en extraire des informations, obligeant ainsi les chercheurs à redéfinir le statut épistémologique des Humanités Numériques en général et des *Digital Literary Studies* en particulier. Cette nécessité renvoie à la théorie sur les structures des révolutions scientifiques élaborée par Thomas Khun[2], théorie selon laquelle la science procède par mouvements cycliques de remise en cause du paradigme scientifique de référence devenu inadapté pour constituer le cadre théorique nécessaire à l'explication des

[2] Khun, T. S., *Structure of Scientific Revolution*, University of Chicago Press, Chicago, 1962.

phénomènes scientifiques. Le fonctionnement est toujours le même, indépendamment des découvertes : il est un état où les nouvelles théories et le paradigme sont alignés, ensuite les nouvelles données ne sont plus explicables par les théories de référence existantes ; un changement de paradigme devient donc nécessaire. Khun s'attarde sur l'état dans lequel se trouve la communauté scientifique à ce moment précis : une communauté scientifique perdue, sans repères, à la recherche d'un nouveau cadre théorique mais qui, grâce à l'effervescence de ses membres va engendrer une grande créativité, et faire surgir de nouvelles théories.

Dans la définition des Humanités Numériques, on peut certainement observer cette nécessité de trouver un nouveau cadre de référence pouvant interpréter les nouvelles données auxquelles on accède grâce à l'application des techniques numériques. Cette approche est complétée par Jean-Gabriel Ganascia[3] qui propose en outre une distinction entre les sciences naturelles – caractérisées par une approche inductive – et les sciences de la culture – visant à comprendre des phénomènes isolés, bien que choisis selon un critère de représentativité. Ganascia considère les Humanités Numériques et en particulier les *Digital Literary Studies*, comme étant à l'intersection des deux approches :

> even if inductive inferences play a role in the digital humanists' investigations, their main modalities of reasoning are essential-

[3] Ganascia J.-G.,« The Logic of the Big Data Turn in Digital Literary Studies », *Frontiers in Digital Humanities*, 2 (7), 2015.

> ly abductive, which means that digital humanists as humanists are looking for explanations, i.e., they are seeking facts that strengthen new hypotheses within a theoretical framework.[4]

Cette observation lui permet de dresser une perspective dans laquelle de nouveaux interprétatifs hybrides surgissent, qui tiennent compte de la l'extrémité opposée à l'autre bout du fil le long duquel se développe le raisonnement, dans les deux domaines, avant de les dépasser dans une approche mixte et nouvelle.

L'exemple de la sculpture la « Porte de l'Enfer » d'Auguste Rodin, tel que le raconte Edward Vanhoutte[5], représente de manière claire cette mixité et sa valeur : l'auteur en y voit une métaphore pour définir les Humanités Numériques.

En 1879, le Secrétaire d'Etat Edmond Turquet commanda une porte décorative pour l'entrée du Musée des Arts Décoratifs à Paris. Malheureusement, l'échec de la construction du musée arrêta le projet, laissant cependant au sculpteur la possibilité de terminer l'œuvre à son goût. Rodin, qui était déjà à la moitié de son travail, fut donc libre de décorer cette surface. L'œuvre qui en résulta était une porte munie de toutes ses composantes : deux volets, le tympan et les deux piliers qui encadrent l'ensemble. Cette porte avait toutefois perdu sa fonction puisqu'aucun méca-

---

4 Ganascia, *op. cit*. p. 3-4.

5 Terras, M., Vanhoutte, E., Nyhan, J., *Defining Digital Humanities : A Reader*, Routledge, London/New York, 2013.

nisme d'ouverture n'était prévu : les figures sculptées sur la surface se superposant aux bords des éléments architecturels, empêchant, de fait, l'ouverture. Une métaphore, selon Vanhoutte pour mettre en relief que :

> Just as Rodin's 'door', Humanities Computing consisted of two clearly separated leaves with their own history and understanding behind them but, when put together, they became so heavily interlinked that they could not be separated without any loss of meaning.[6]

Les deux domaines sont comme les deux volets, conçus pour être séparés mais que, une fois unis, il est impossible de diviser sans perdre du contenu. En d'autres mots, l'union de deux domaines constitue non seulement un enrichissement, mutuel mais aussi la création d'une troisième discipline à l'intérieur de laquelle les deux se fondent, perdant leurs limites pour acquérir un nouveau statut du fait même de leur union. C'est exactement cette impossibilité à réduire les HN à une simple opération mathématique qui fait de sa définition un domaine de recherche autonome.

Pour pouvoir répondre à la question quant au seuil au-delà duquel l'usage de l'ordinateur dans les sciences humaines se traduit en *humanités numériques*, il faut opérer

---

[6] *Ivi*, p. 120, M., Vanhoutte, « The Gates of Hell : History and Definition of Digital | Humanities | Computing », in Terras, M., Vanhoutte, E. and Nyhan, J., Defining Digital Humanities : A Reader. London/New York : Routledge, 2013.

un changement de perspective épistémologique. En raison de leur puissance, les instruments utilisés autorisent la récolte des données différentes par rapport à celles qui sont collectées manuellement. Il est nécessaire de disposer à la fois d'un cadre théorique de référence en mesure d'interpréter ce nouveau matériau, et d'une connaissance approfondie des potentialités des instruments numériques, et ce pour pouvoir formuler des questions de recherches qui – issues des théories et des perspectives humanistes- soient autres par rapport à la tradition car réalisables avec les moyens des HN. La disponibilité des grandes bases de données modifie le rapport au questionnement à la base de la recherche ; il est possible, par exemple, de mettre en relation les modifications qu'un mot – et donc un concept – subit au cours d'un siècle, en se basant sur les documents de l'époque, dont le volume est tel qu'il est impossible de les analyser à l'œil nu. Au moment même où l'usage de l'ordinateur cesse de n'être qu'un instrument, la manière de questionner et d'interpréter les données et de les interpréter se modifie : on se trouve dès lors face aux Humanités Numériques.

1.1.2. A l'aube des Humanités Numériques

L'intuition qu'une machine aurait pu faire autre chose que des calculs mathématiques remonte au XVIIème siècle. En 1840, le mathématicien anglais Charles Babbage présente, à l'Académie des Sciences de Turin, à l'occasion du deuxième Congrès des *Scienziati italiani*, la « Analytical Engine ». Sa machine possédait une mémoire à 1000

chiffres et 50 digits et elle employait des fiches trouées pour l'input, l'output et les instructions. Fasciné par cet engin, l'ingénieur et diplomate italien, Luigi Federico Menabrea, publia en 1842 un travail en langue française intitulé « Notions sur la Machine Analytique de M. Charles Babbage »[7]. La traduction anglaise par Ada Lovelace a paru en 1943 ; il n'est pas exagéré d'affirmer que, par cette traduction, c'est une femme qui a apporté une contribution fondamentale à l'histoire naissante de l'informatique. Dans ses notes, en effet, elle livre une intuition qui est à la base des HN :

> might act upon other things besides *number*, were objects found whose mutual fundamental relations could be expressed by those of the abstract science of operations, and which should be also susceptible of adaptations to the action of the operating notation and mechanism of the engine. Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent.[8]

---

[7] Menabrea, L.F., « Sketch of the Analytical Engine Invented by Charles Babbage. Translated by Ada Augusta, Countess of Lovelace », in Morrison, E. (dir.), *Charles Babbage and his Calculating Engines. Selected Writings by Charles Babbage and Others*, Dover Publications, New York, pp. 225-245, 1961; [Premiere publication : *Bibliothèque Universelle de Genève*, 82 (Oct. 1842); trad. Lovelace, A.A. in Taylor, R., (dir.), *Scientific Memoirs, Selections from The Transactions of Foreign Academies and Learned Societies and from Foreign Journals*, 1843.

[8] *Ibidem*.

C'est après la deuxième guerre mondiale que s'affirme, avec plus de vigueur, l'idée qu'un ordinateur peut effectuer d'autres opérations que des calculs et que ses fonctions peuvent s'appliquer à des domaines autres que les mathématiques et la physique. Warren Weaver[9] – qui avait travaillé dans la balistique pendant la guerre avant de devenir le Directeur de la Rockfeller Foundation – commence à proposer des pistes d'applications *pacifiques* de la machine qu'il appelle « mathematics and machine translation »[10].

La traduction – qui nait d'une nécessité de communication interculturelle – fut le premier domaine des humanités où les ordinateurs furent utilisés, ce qui souligne l'importance et l'intérêt que toutes les applications à venir auront envers l'analyse linguistique. C'était l'époque de la guerre froide et le décryptage des documents russes de la part des Américains représentait une nécessité vitale. Weaver collaborait avec Andrew Booth [11] pour des échanges scientifiques sur la traduction automatique. Ils publient ensemble un *memorandum*, qui suscita l'intérêt des milieux scientifiques, des étudiants de linguistique, de mathématique

---

9 Weaver, W. (1965). « Translation', in W.N. Locke and A.D. Booth (eds) *Machine Translation of Languages. Fourteen Essays*, Cambridge, MA : The MIT Press, pp. 15-23. Original publication 1949.

10 Weaver, W. *Scene of Change. A Lifetime in American Science*, Scribner, New York, 1970.

11 Booth, A.D. (dir.), *Machine Translation*, North-Holland Publishing Company, Amsterdam, pp. 173-94, 1967 ; Booth, A.D., Booth, K.H.V., « The beginnings of MT », in Hutchins, W.J., (dir.), *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 253–61, 2000.

et de logique, et des experts de différents domaines. Une série de réflexions donneront ensuite vie aux premières études sur la traduction automatique, études touchant aussi bien l'ambigüité des mots, la fonction sémantique de la syntaxe,que l'ordre des mots dans des langages différents. L'intérêt croissant envers cette discipline prend forme et des occasions d'échanges à caractère scientifique sont créées : en 1952, 18 experts se réunissent lors de la première conférence sur la « Machine Translation » au MIT. Deux ans plus tard, une démonstration de traduction automatique – qui fera date auprès des quartiers généraux de IBM – propose 49 phrases en langue russe, un vocabulaire de 250 mots russes et ses équivalents (un vocabulaire aligné) et six règles de syntaxe. Le communiqué de presse de l'époque de IBM déclare :

> A girl who didn't understand a word of the language of the Soviets punched out the Russian messages on IBM cards. The "brain" dashed off its English translations on an automatic printer at the breakneck speed of two and a half lines per second.[12]

La même année, la première revue consacrée au thème est publiée : « Mechanical Translation ». Durant les années 50, on observe un intérêt croissant pour la traduction automatique; des séminaires, des conférences, des moments de recherche et de divulgation sont organisés grâce aussi à

[12] IBM press release, 8 January 1954, http ://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html .

la fondation de *l'Association for Machine Translation and Computational Linguistics* (AMTCL). Il est important de relever ceci pour deux raisons : montrer de manière systématique comment l'ordinateur peut être employé dans les sciences humaines, développer une attention pour la parole, élément minimal de sens, et par conséquent pour la linguistique. L'intérêt ne se limita pas aux Etats-Unis :il se répandit partout dans le monde[13] . Les Etats-Unis toutefois consacraient des ressources économiques importantes à la recherche et n'avaient pas subi les dégâts de la guerre. Néanmoins, six ans plus tard, un rapport émanant du comité indépendant ALPAC (Automatic Language Processing Advisory Committee) constitué sous requête des investisseurs, mit fin au mouvement d'optimisme dont la recherche sur la traduction automatique avait pu bénéficier jusque-là. Ce rapport, intitulé « Languages and Machines : Computers in Translation and Linguistics »[14] et paru en 1966, critique les coûts et les efforts engagés dans ce domaine de recherche et déclare qu'une traduction parfaite est impossible, que les coûts ne sont pas soutenables, et

---

13 Hutchins W.J. (dir.), *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins Publishing Company, Amsterdam/ Philadelphia, 2000.

14 Aa. Vv. ALPAC, « Languages and Machines : Computers in Translation and Linguistics », A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC : National Academy of Sciences, National Research Council, 1966; http ://www.nap. edu/books/ARC000005/html.

qu'il vaudrait mieux former de bons traducteurs. Cet échec est provisoire comme on peut le constater aujourd'hui puisque grâce aussi à un développement de techniques comme celles des réseaux neuronaux, le recours à la traduction automatique est de plus en plus massif. Ce même échec, par ailleurs réoriente également l'intérêt scientifique des linguistes qui se sont penchés sur la traduction. Ainsi, les pistes nouvelles de recherche se développent avec l'objectif de fournir des outils aux traducteurs et aux linguistes computationnels. La recherche prend (temporairement) une autre direction : l'idée de pouvoir créer un outil pour une traduction autonome et sans fautes ayant été abandonnée, le choix retombe sur une un outil de support.

Même si les travaux sur la traduction automatique enregistrent, à l'époque du rapport, un coup d'arrêt en raison d'un scepticisme et d'un manque de financements, les techniques employées (qui renvoient à l'analyse lexicale du texte) continuent. La statistique appliquée aux textes, qui permet le calcul des termes, la création d'index et la liste des fréquences se perfectionne : les programmes pour l'analyse statistique d'un texte sont développés dès le début des années 60.[15]

> Le présupposé de cette ligne de recherche est que le mot est un élément numériquement formalisé : la parole est numérisée, quantifiée, augmentée. Une fois cette transformation faite, les applications peuvent faire partie de la traduction

[15] Levison, M. « The Mechanical Analysis of Language », *NPL*, pp. 562 -574, 1962.

> automatique et des études littéraires, comme le chapitre intitulé « The Computer in Literary Studies »[16] in *Machine Translation* le démontre. Dans son texte, Booth identifie les problèmes littéraires qui auraient pu être résolus par le recours à des ordinateurs : concordances, glossaires, attribution d'auteur, études de stylistique, chronologie des textes et édition des textes critiques et leur conservation.[17]

Durant les mêmes années, commence l'un des projets les plus importants pour les Humanités Numériques. Son application aux textes est souvent considérée comme étant à la base de la naissance des *Digital Literary Studies* : il s'agit du travail du jésuite Roberto Busa[18]. Il se proposait, déjà en 1949, d'étudier le concept de « présence » et d' « incarnation » dans la totalité de l'œuvre de Thomas d'Aquin. Il avait tout de suite saisi qu'une recherche manuelle du mot « présence » n'aurait pas été efficace pour deux raisons : la première, concrète, due au volume de l'œuvre de Thomas D'Aquin ; la deuxième, méthodologique, consistait dans le fait que le mot seul« présence » n'avait pas d'utilité ; il fallait chercher la combinaison « en » « présence »[19], une

---

16 Booth, A.D. (dir.), *Machine Translation*, *op. cit*.

17 Levison, M., « The Computer in Literary Studies », Booth, A.D. (dir.), *Machine Translation*, North-Holland Publishing Company, Amsterdam, pp. 173–94, 1967.

18 Busa, R., « The Annals of Humanities Computing : The Index Thomisticus », *Computers and the Humanities*, 14, pp. 83-90, 1980; Jones, S.E., *Roberto Busa, and the emergence of Humanities Computing : The Priest and the Punched Cards*, Routeledge, London, 2016.

19 traduction du latin : « in » « praesentia ».

concordance donc, porteuse d'un sens ponctuel. Busa s'adressa à l'entreprise IBM qui accepta le défi. Après deux millions d'heures de travail, en 1980, il parvint à compléter l'index sur des fiches trouées : ce fut ainsi que l'*Opera Omnia of Thomistic Linguistics*[20], mère de toutes les éditions numériques et hypertextuelles à venir, vit le jour.

La concordance peut donc être la solution à une recherche thématique au sein d'un texte : le calcul quantitatif (le nombre de mots) commence à être relié à une appréciation de type qualitatif. La parole numériquement formalisée couplée aux développements ultérieurs et puissants du calcul des ordinateurs, c'est de là que cette intuition puise toute sa force.[21] Même si le projet de Busa ne comprenait pas des calculs à proprement parler, il démontrait néanmoins que la mécanique pouvait résoudre des problèmes de recherche pour les humanistes, à savoir ces disciplines qui, par définition, nécessitent d'un degré interprétatif différent de l'automatisme, qui, à son tour se caractérise par la répétition qui ne tienne pas compte du contexte.

Entre 1970 et 1990, on peut lire dans une série de conférences, d'études et de publications les qualificatifs de « Computing Humanities » ou « Digital Humanities », comme pour témoigner de la recherche d'une stabilisation lexicale qui représente en même temps la stabilisation d'une nouvelle discipline.

---

20 http ://www.corpusthomisticum.org/ [dernier accès le 1/10/2021].

21 Mounier, P., *Les humanités numériques. Une histoire critique*, Éditions de la Maison des sciences de l'homme, Paris, 2018.

1.1.3. Le manifeste d'aujourd'hui

Deux manifestes ont paru dernièrement et, bien que différents par leur style et par leur aire de provenance (l'un est américain et l'autre francophone), ils nous donnent à voir des traits communs et une orientation future assez stable. Il s'agit du « Manifeste des Digital Humanities »[22], écrit par Pierre Mounier, en 2010, suite à la Conférence THATCamps[23] (The Humanities and Technology Camp) qui s'est tenue à Paris les 18 et 19 mai 2010 et à « The Digital Humanities Manifesto 2.0 », écrit en 2015 par Jeffrey Schnapp et Todd Presner. Bien que signés par des auteurs bien définis, ces travaux n'en sont pas moins le résultat des réflexions d'une communauté plurielle. Le manifeste français, de style « académique » relate :

> de ce bouillonnement est née l'idée, très française, de proposer un « manifeste » pour les *digital humanities*, rédigé de manière collaborative et validé au cours d'une assemblée plénière pour le moins animée... [24]

---

[22] Mounier, P., « Manifeste des *Digital Humanities* », *Journal des anthropologues* [En ligne], pp. 122-123, 2010, mis en ligne 01/12/2012, URL : http ://journals.openedition.org/jda/3652 ; DOI : https ://doi.org/10.4000/jda.3652.

[23] Rapport de l'ACLS : http ://www.acls.org/programs/Default.aspx?id=648 ; Conférences Digital Humanities : http ://www.digitalhumanities.org/ ; Le blog du ThatCamp Paris : http ://tcp. hypotheses.org/ ; Le Manifeste en 20 langues : http ://www.digitalhumanities.cnrs.fr/wikis/tcp/index.php?title=Traduisez_le_Manifeste La liste de discussion Digital Humanities : https ://listes.cru.fr/sympa/info/dh.

[24] Mounier P., « Manifeste des *Digital Humanities* », *op. cit.* p. 449.

Lui fait écho, dans un style plus informel, le manifeste américain : « if you are wondering who is reaching out here, the answer is plural ».[25]

Le manifeste français définit par ailleurs une communauté « pluridisciplinaire et internationale »[26] dans un projet de traduction en 20 langues du monde entier. Parfois considéré comme étant un excès d'autoréférence, « l'idée très française » est, de fait, une déclaration d'intention programmatique : le manifeste français coïncide avec le fondement d'une communauté possédant des orientations et des caractéristiques bien précises : « une communauté sans frontières. Nous sommes une communauté multilingue et multidisciplinaire. »[27]

Les domaines d'appartenance de celle qui est définie comme une « transdiscipline » sont ainsi « l'ensemble des Sciences humaines et sociales, des Arts et des Lettres ». En particulier :

> multiples communautés particulières issues de l'intérêt pour des pratiques, des outils ou des objets transversaux divers (encodage de sources textuelles, systèmes d'information géographique, lexicométrie, numérisation du patrimoine culturel, scientifique et technique, cartographie du web, fouille de données, 3D, archives orales, arts et littératures numériques et hypermédiatiques, etc.), ces communautés étant en train de converger pour former le champ

---

25 Schnapp J., Lunenfeld P., Presner T., « Manifesto 2.0 », https ://www.toddpresner.com/?p=7 [dernier accès : 1/10/2021].

26 Mounier P., « Manifeste des *Digital Humanities* », p. 448.

27 *Ivi*, p. 451.

des digital humanities.[28]

Les humanités numériques, définies comme « une intégration intense et à plusieurs niveaux des technologies numériques dans tous les processus de recherche, depuis la collecte de données jusqu'à la publication »[29] se fondent sur la prémisse d'une révolution en cours, celle du tournant numérique, qui « modifie et interroge les conditions de production et de diffusion des savoirs ». L'importance du passé est mise en valeur : une grande attention sera donc accordée au fait de ne perdre ni les bases posées ni les expériences accumulées au fil du temps. Les traits fondateurs sont ceux des disciplines humanistes, l'équilibre entre le numérique et les humanités penchant en faveur des humanités.

Le concept de « cyberinfrastructure » se fraie alors un chemin quand le manifeste intègre cette notion. C'est le Council of American Learned Societies qui le définit dans un rapport intitulé « Our Cultural Commonwealth ») et qui est repris dans le Manifeste :

> Un ensemble d'informations, d'expertise, de standards, de stratégies, d'outils et de services qui sont partagés largement entre les communautés mais développés spécifiquement pour des usages savants. Une cyberinfrastructure est quelque chose de plus précis que le réseau lui-même, mais de plus général qu'un outil ou une ressource développés

[28] *Ivi*, p. 448.
[29] *Ivi*, p. 447.

> pour un projet particulier, ou même, plus largement, pour une discipline particulière.[30]

L'orientation, c'est-à-dire les actions que les membres vont entreprendre, vise à la diffusion d'un savoir commun avec un vocabulaire partagé, avec libre accès aux données et aux métadonnées, avec une institutionnalisation des Humanités Numériques grâce à la création de diplômes académiques.

Le manifeste de l'aire américaine part de la même nécessité de définition du domaine :

> Digital Humanities is not a unified field but an array of convergent practices that explore a universe in which : a) print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated; instead, print finds itself absorbed into new, multimedia configurations; and b) digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and social sciences.[31]

L'idée que le Humanités Numériques sont un ensemble hétérogène de disciplines est soutenue tout comme est soutenue l'idée que la nécessité de définition nait d'un tournant révolutionnaire. Le support matériel déterminant cette révolution prend ainsi toute sa valeur : le monde n'est plus celui de l'impression sur papier, et ce constat entraine une diffusion du savoir visuel. Les arts, les sciences sociales

---

30 *Ibidem*.

31 Schnapp J., Lunenfeld P., Presner T., « Manifesto 2.0 », *op. cit*.

et les sciences humaines sont les trois domaines pris en considération. Une fracture, toutefois, commence à émerger : si, pour l'aire francophone, il fallait éviter la rupture avec le passé, l'aire anglophone envisage deux « vagues ». La première : « Like all media revolutions, the first wave of the digital revolution looked backward as it moved forward ».[32] La première prend en considération que toute révolution s'appuie au début sur le passé (le premier cinéma avait des techniques théâtrales, la presse de Gutenberg reproduisait les techniques de la culture des manuscrits du Moyen-âge) ; par la suite, il y a un passage – nécessaire – où l'instrument modifie des idées. La deuxième « vague » des HN est celle qui accueille cette transformation :

> The first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is qualitative, interpretive, experiential, emotive, generative in character.[33]

Un changement dans la manière de penser est donc nécessaire : « interdisciplinarity/transdisciplinarity/multidisciplinarity are empty words unless they imply changes in language, practice, method, and output » pour créer une collaboration entre les disciplines : « It is not about the emergence of a new general culture, Renaissance humanism/Humanities, or universal literacy. On the con-

[32] *Ibidem*.
[33] *Ibidem*.

trary, it promotes collaboration and creation across domains of expertise »[34].

Les deux composantes doivent fonctionner ensemble :

> The revolution is not about transforming literary scholars into engineers or programmers. Rather, it is about :
> - expanding the compass and quality of knowledge in the human sciences
> - expanding the reach and impact of knowledge in the Humanities disciplines
> - direct engagement in design and development processes that give rise to richer, multidirectional models, genres, iterations of scholarly communication and practice.

Le « Manifesto 2.0 » se clôture avec une déclaration sur la valeur de l'union entre humanités et numérique. Il refuse toute définition provenant d'une modification du passé, mais insiste sur l'idée de quelque chose de nouveau :

> We reject the phrase to whatever degree it implies a digital turn that might somehow leave the Humanities intact : as operating within same stable disciplinary boundaries with respect to society or to the social and natural sciences that have prevailed over the past century. We further reject the phrase to the degree that it suggests that the humanities are being modified by the digital, as it were, "from the outside" with the digital leading and the Humanities following. On the contrary, our vision is of a world of fusions and frictions, in which the development and deployment of tech-

[34] *Ibidem*.

> nologies, and the sorts of research questions, demands, and imaginative work that characterize the arts and Humanities merge.[35]

Bien que des différences culturelles imprègnent les deux manifestes, nous pouvons en dégager des éléments communs :

- situation révolutionnaire (numérique) qui comporte une urgence de définition de la discipline et en même temps de la communauté ;
- une dialectique entre les deux composantes : numérique et humanité ;
- l'*open access* ;
- la cyberstructure ;
- l'esprit ouvert, fluide, en mouvements, pluriel, collaboratif.

1.1.4. La (bonne) question de recherche : une capacité de transposition

L'interdisciplinaire se caractérise par un mouvement intellectuel qui va d'un domaine à l'autre : le concept de passage et de transposition, devient fondamental. L'on s'approprie les méthodes d'un domaine pour les appliquer aux thématiques de l'autre : une bonne question de recherche doit tenir compte de cet aspect multidimensionnel. Lorsqu'on utilise l'adjectif « bonne » cela ne tient nullement à un jugement de valeur, il se réfère plutôt à un aspect métho-

[35] *Ibidem*.

dologique. Poser la bonne question signifie savoir traduire en modèles linguistiques les thématiques que l'on veut explorer. La bonne question présuppose une connaissance suffisante des deux disciplines pour les mettre en dialogue.

Prenons l'exemple de Roberto Busa s'intéressant au concept de présence physique au sein de l'œuvre d'un philosophe et théologien. Or, rechercher simplement le mot « présence » n'aurait pas abouti à des résultats dignes d'intérêt ; au contraire le mot « présence » avec le mot « in » prenait une autre valeur car il considérait qu'une présence charnelle aurait pu être liée à la préposition « in ». Une fois posée cette hypothèse de modèle linguistique, l'étape suivante consiste à identifier les techniques et les outils qui permettent d'extraire les données linguistiques et de les manipuler. L'interprétation de la thématique en termes linguistiques, c'est-à-dire chercher des patterns conceptuels à travers la forme morphologique ou le comportement syntactique des mots, représente le passage le plus compliqué, celui qui exige une capacité de réflexion de la part du spécialiste.

Il n'est pas exclu que le mépris exprimé à l'encontre de ce genre d'application pour la critique littéraire et l'analyse du discours dérive de la complexité inhérente à la formulation de la « bonne » question en veillant d'une part à ne pas perdre de vue la thématique à étudier et d'autre part au fonctionnement du langage et, enfin, au fonctionnement des instruments numériques disponibles.

Un spécialiste de littérature du XVII siècle peut être sceptique vis-à-vis d'une méthode qui travaille sur l'ana-

lyse des sentiments et qui cherche à évaluer la polarité émotive (négative, positive, neutre) d'un texte grâce au recours à un lexique annoté. Il existe dans le panorama du TAL une technique et des instruments qui autorisent ce qu'on appelle « *l'analyse de sentiments* » ou « *sentiment analysis* », point sur lequel nous reviendrons plus loin, de manière plus approfondie. Le principe est le suivant : pouvoir analyser automatiquement un texte en extrayant les polarités émotives de certains fragments de texte (positive, négative, neutre) à travers une comparaison avec un lexique annoté. L'objection la plus courante est liée au fait que la machine ne peut pas reconnaitre, par exemple, les sentiments de Dante Alighieri. Toutefois, cette objection ne prend pas en compte le fait que dans le cadre des Humanités Numériques il ne faut pas poser à la machine la question sous la forme qu'elle prendrait si on la posait à un être humain. Si on ne peut pas connaitre les sentiments de Dante, en élaborant l'ensemble de son œuvre il est possible de détecter d'éventuelles correspondances entre son style et les événements de sa vie personnelle. Est-il possible, par exemple que pendant son long exil – et selon les endroits où il a vécu – ses écrits puissent se caractériser par des traits textuels spécifiques ? Il s'agit d'informations latentes, non visibles à l'œil nu ou avec une lecture détaillée (*close reading*), mais détectables lorsqu'on compare par exemple l'œuvre de Dante avec d'autres éléments tels que des événements historiques (de sa vie privée ou de l'histoire en général), la variété et le style linguistique utilisé à l'époque contemporain à Dante. La bonne question prend donc en

considération une typologie de données pouvant être analysées à travers les techniques du traitement automatique des langues et dont les résultats seront interprétés avec une sensibilité interdisciplinaire.

Pour poser la bonne question de recherche, il est d'abord nécessaire de procéder par abstraction en prenant en compte le fonctionnement du langage. Pour analyser le concept de beauté, la recherche du seul lemme « beauté » n'est pas suffisante. Par abstraction, la question à poser devient : comment le concept de beauté est représenté au niveau morphologique, lexical et syntactique? Est-il nécessaire de l'analyser dans une perspective diachronique qui mette en valeur le changement de sens qu'il pourrait subir au fil du temps ?

Est-il nécessaire de chercher d'autres événements liés à la beauté qui ne sont pas relatés dans le texte ? Ces choix convoquent la sensibilité du chercheur, de l'humaniste qui doit connaitre les possibilités techniques du numérique. Dans ce processus d'abstraction, il est tout aussi nécessaire de connaitre les instruments à notre disposition : il s'agit à la fois d'un procès *top-down* (on décide d'abord la thématique et ensuite les outils), et de *bottom-up* (cela implique la connaissance du panorama des technologies disponibles). Ce processus met en jeu l'interdisciplinaire. La métaphore de la « Porte de l'enfer » de Rodin (où les deux volets ne peuvent pas être divisés) prend donc ici toute sa valeur.

1.2. LES OUTILS

1.2.1. Le Traitement Automatique des Langues

La discipline consistant à concrétiser les manipulations nécessaires aux humanités numériques et à permettre l'extraction des informations du texte est le TAL[1], acronyme de Traitement automatique des Langues (NLP – *Natural Language Processing* en anglais) ou linguistique computationnelle, discipline qui étudie et produit des modèles du langage pouvant être élaborés par une machine.

Bien que souvent l'on fasse remonter l'institutionnalisation de la linguistique computationnelle au rapport Bar-Hillel de l'ALPAC[2] en 1966, c'est bien avant cette date que des études et une activité scientifique autour de cette discipline l'ont structurée. En 1962, nait l'Association « Machine Translation and Computational Linguistics (AMTCL) » sous la présidence de Victor Yngve et la vice-présidence de Hays[3]. Le premier colloque, intitulé « International Conference on Computational Linguistics », tenu en 1965, est organisé par plusieurs associations (l'AMCTL, l'association française

---

1 Chaumartin F. R., Lemberger, P., *Le traitement automatique des langues : comprendre les textes grâce à l'intelligence artificielle*, Dunod, Paris, 2021 ; Kurdi Z., *Traitement automatique des langues et linguistique informatique*, ISTE éditions, Paris, 2017.

2 Aa. Vv. ALPAC, *op. cit*.

3 Léon, J., « De la TA à la linguistique computationnelle et au TAL », *Histoire de l'automatisation des sciences du langage,* ENS Éditions, Lyon, 2015.

ATALA, des associations scandinave, japonaise et sud-américaine). Cependant L'ALPAC joue un rôle capital dans la reconfiguration de la linguistique computationnelle et du traitement automatique des langues, à l'époque en plein essor.

La linguistique computationnelle trouve ses sources dans deux différents paradigmes de recherche : le premier est lié au dépouillement des textes, à la création d'index (Roberto Busa fut un pionnier de cette approche), et au calcul des fréquences des mots. Le deuxième s'inspire de la grammaire générative de Chomsky[4] qui postule que le langage est géré par un ensemble de règles syntaxiques. La machine, pour l'élaboration d'un texte et pour reconnaitre ou produire un bon exemple, dispose de deux méthodes : l'analyse du comportement d'un mot grâce à la fréquence d'apparition dans des cotextes, l'utilisation d'un ensemble de règles syntaxiques et d'un vocabulaire étiqueté. L'arrivée du *machine learning* a ajouté une troisième approche qui consiste dans l'apprentissage structuré ou non structuré. Grâce au premier, la machine est entrainée avec une quantité de bons et de mauvais exemples ; avec le deuxième, la machine est entrainée avec des données non-structurées, laissant à l'algorithme la tâche de trouver des comportements récurrents et des règles.

Ces techniques définies dans le cadre de l' « Intelligence Artificielle » nous permettent de fournir un texte à la machine et de le recevoir analysé, selon les instruments et les techniques que nous avons choisis.[5]

---

4 Chomsky, N., *Syntactic Structures*, Mouton & Co, The Hague, 1957.

5 Lenci, A., Montemagni, S., Pirrelli, V., *Testo e Computer. Elementi di linguistica computazionale*, Carocci, Roma, 2016.

1.2.2. Les techniques fondamentales du TAL

La question de l'entrée des données est importante au sens que les ordinateurs élaborent les données de manière binaire, travaillant avec les chiffres 1 et 0, et tout doit passer par cette opération d'encodage. Traduire les mots, unité minimale de signifié, les rendre représentables numériquement, devient ainsi un passage fondamental.

La première opération nécessaire lors de l'élaboration d'un texte est la tokenisation des textes, c'est-à-dire qu'il faut savoir reconnaitre où le mot commence et où il se termine. Cela parait une tache relativement simple pour l'œil humain, mais la machine doit distinguer entre les signes de ponctuation, les apostrophes et les accents. La première tâche nécessaire, à savoir la compréhension des limites physiques des mots, permet d'identifier une unité, qui peut ensuite être traitée. Dans l'exemple du tableau 1, nous observons comment le « s' » est de fait considéré comme un mot :

Elle se demanda alors s'il se souvenait vraiment de lui avoir écrit pour lui demander de venir.
Elle | se | demanda | alors | s'| il | se | souvenait | vraiment | de | lui | avoir | écrit | pour | lui | demander | de | venir

Tableau 1. Exemple d'un texte tokenisé. Fragment pris du roman « Trois femmes puissantes » de Marie Ndiaye.

Ensuite grâce à la lemmatisation, nous pouvons réduire chaque mot à sa forme du lemme : généralement le masculin singulier pour les noms et les adjectifs et l'infinitif pour les verbes. La possibilité de réduire aux lemmes les formes

des mots est fondamentale pour pouvoir ensuite les regrouper : chercher, par exemple, la fréquence du verbe « demander » sans faire de distinctions entre les temps peut être utile pour les finalités de notre analyse. La machine est capable de reconnaitre la même racine lexicale.

Elle se demanda alors s'il se souvenait vraiment de lui avoir écrit pour lui demander de venir.	
forme dans le fragment	*forme lemmatisé*
demanda	demander
demander	demander
souvenait	souvenir
elle	il

Tableau 2. Exemple de lemmatisation. Fragment pris du roman *Trois femmes puissantes* de Marie Ndiaye.

L'annotation consiste dans l'ajout d'informations aux mots par le biais d'un métalangage, notamment le XML. Il y a plusieurs niveaux d'information qui peuvent être ajoutés à un terme : morphologique, syntactique, sémantique. Le premier niveau consiste à ajouter à chaque mot sa catégorie morphologique, c'est-à-dire le nombre, le genre, le mode, le temps et la personne. Le deuxième niveau consiste à attribuer une catégorie de parties du discours (nom, adjectif, verbe, adverbe) ou une catégorie syntaxique (sujet, complément direct, complément indirect, etc..).

Elle se demanda alors s'il se souvenait vraiment de lui avoir écrit pour lui demander de venir.	
Elle	[frpos="PRO :PER"]
se	[frpos="PRO :PER"]
demanda	[frpos="VER :simp"]
alors	[frpos="ADV"]
s'	[frpos="KON"]
il	[frpos="PRO :PER"]
se	[frpos="PRO :PER"]
souvenait	[frpos="VER :impf"]
vraiment	[frpos="ADV"]
de	[frpos="PRP"]
lui	[frpos="PRO :PER"]
avoir	[frpos="VER :infi"]
écrit	[frpos="VER :pper"]
pour	[frpos="PRP"]
lui	[frpos="PRO :PER"]
demander	[frpos="VER :infi"]
de	[frpos="PRP"]
venir	[frpos="VER :infi"]

Tableau 3. exemple d'annotation morphosyntactique. Fragment pris du roman *Trois femmes puissantes* de Marie Ndiaye, élaboré avec Treetagger[6].

[6] Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994; Schmid, H., « Improvements in Part-

La légende est la suivante :
[frpos="PRO :PER"] (*french part of speech = pron. personnel)
[frpos="VER :infi"] (*french part of speech = verbe infinitif)
[frpos="VER :pper"] (*french part of speech = verbe plus que parfait)
[frpos="VER :impf"] (*french part of speech = verbe imparfait)
[frpos="ADV"] (*french part of speech = adverbe)
[frpos="PRP"] (*french part of speech = préposition)
[frpos="VER :simp"] (*french part of speech = passé simple)

Tableau 4. Legende

L'annotation syntactique s'est servie en particulier de la théorie des « Dépendances universelles » qui se fonde sur une approche élaborée par Lucien Tesnière. Comme on peut l'observer (tableau 4), l'élément fondamental est la structure arborescente. Grâce à la légende pour l'étiquetage (indiqué au tableau 4.a) nous pouvons interpréter le graphe – ou *treebank* – comme une structure munie d'une racine qui est le verbe « demanda », le verbe de la préposition principale qui a un sujet « elle » et un complément d'objet direct « s' ».

---

of-Speech Tagging with an Application to German », *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995.

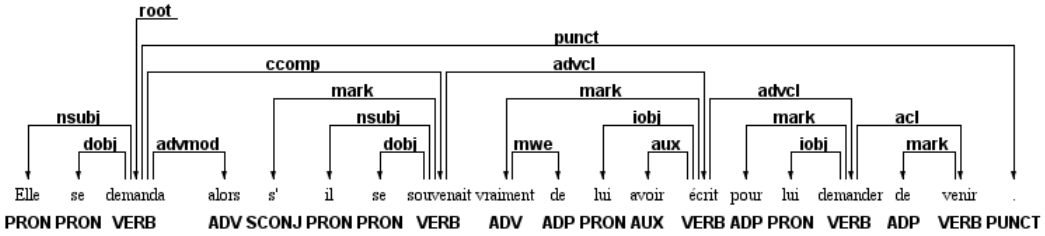

Tableau 4. Exemple d'annotation syntactique. Fragment pris du roman « Trois femmes puissantes » de Ndiaye, élaboré avec la librairie Spacy de Python et visualisé par le biais du logiciel DAG annotator.

étiquette	**catégorie UD**
nsubj	*sujet*
dobj	*complément objet direct*
root	*racine*
aux	*auxiliaire*
punct	*signe de ponctuation*
advmod	*adverbe de mode*
iobj	*complément d'objet indirect*

Tableau 4.a. Légende des Dépendances Universelles.

Les annotations sémantiques représentent les catégories de sens qui répondent aux nécessités de recherche. Elles peuvent être faites en phase de prétraitement du texte, de manière automatique grâce au *machine learning*, et servent alors à trier et regrouper les résultats. Elles peuvent être également ajoutées aussi en phase de traitement du texte, avec une syntaxe XML de sorte que se crée un corpus annoté réutilisable pour d'autres chercheurs.

1.2.3. La parole formalisée / La parole manipulable

Ces techniques de base permettent de traiter le mot et de le rendre liquide. Nous pouvons parler de trois caractéristiques du mot informatisé : il est représentable dans l'espace (par le biais de la sémantique distributionnelle), il

est calculable (grâce à la statistique), augmentable (grâce à l'annotation).

1.2.3.1. Le mot représentable

Quand le mot devient vecteur, chiffre, il est quantifiable. Transformer le mot en vecteur équivaut ainsi à le positionner dans un point de l'espace. Si cet espace est sémantique, le résultat est une carte de sens. En 1957, le linguiste John Rupert Firth[7] résume, avec la phrase célèbre « You shall know a word by the company it keeps », le principe de l'hypothèse de distribution : si deux mots paraissent dans des contextes similaires, ils sont eux-mêmes similaires. Cette hypothèse est à la base de la sémantique distributionnelle selon laquelle la valeur sémantique d'un mot est le résultat d'une interprétation contextuelle. L'attribution de vecteurs à des mots en fonction de leur distribution permet d'identifier des *topics*, c'est-à-dire des groupes de mots qui ont une valeur sémantique et qui sont reliés à des portions plus ou moins grandes du corpus. À l'idée de relier la sémantique à une dimension spatiale s'ajoute la possibilité de traduire les mots en vecteurs, permettant ainsi de construire des espaces distributifs sémantiques à l'aide de représentations géométriques : il est donc possible d'effectuer des opérations algébriques avec la sémantique des mots[8].

---

7 Firth J.R., *Papers in Linguistics*, Longman, London, 1957.

8 Lenci, A., « Distributional semantics in linguistic and cognitive research », *Italian Journal of Linguistics*, 20, 2008.

La sémantique distributive est le fondement des techniques de « topic modeling » : observer le comportement des mots et leurs regroupements permet de détecter les topics présents dans un texte. La sémantique distributive permet une analyse sémantique latente, aide à reconnaitre des structures non visibles à la lecture rapprochée.

Il est toutefois important d'avoir une interprétation correcte des mots-clés du topic (qui doivent être interprétés dans leur ensemble) et de la manière dont ils construisent la signification en fonction de leur relation. Par exemple, dans la tradition des *Digital Literary Sudies*, Lisa Rhody a suivi cette intuition dans son article « Topic Modeling and Figurative Language », articledans lequel la chercheuse examine 4500 poèmes anglais et identifie quatre topics. Si les deux premiers ne sont pas pertinents, les deux autres – que Rhody appelle « semantically evident topics » et « semantically opaque topics » -montrent 'd'évidents indices sémantiques. Dans certains cas, les topics contiennent des mots utilisés pour indiquer des métaphores : night, light, moon, stars qui, bien qu'associés au thème de la nuit, indiquent dans le corpus l'expression de la folie. Le topic, s'il est correctement interprété, peut donner la mesure des métaphores présentes dans le texte. Dans d'autres cas, les topics incluent des mots utilisés dans le même genre littéraire, donnant ainsi des informations extratextuelles. Sharon Block rejoint la même ligne de recherche avec son analyse des articles publiés dans le Pennsylvania Gazette.

1.2.3.2. Le mot calculable

Le mot est aussi calculable au sens statistique. L'on peut calculer les fréquences, extraire les mots-clés d'un texte, calculer la concentration de certains mots dans des sous-parties du corpus. Bien qu'une approche uniquement quantitative ait été la cible de maintes critiques, il est important de savoir que la quantité peut être interprétée à travers des filtres quantitatifs. Grâce à la statistique, nous pouvons calculer les concordances, c'est-à-dire les mots qui apparaissent en contexte. Les concordances permettent de revenir au contexte immédiat et d'identifier des structures syntaxiques récurrentes car il est possible de classer les résultats selon les éléments de la fenêtre du contexte droit ou du contexte gauche. Un terme pivot ayant été donné il est possible d'obtenir chaque occurrence dans un contexte défini en fonction des besoins de recherche. Les résultats sont classés selon des termes sémantiquement discriminants ou selon les structures syntaxiques similaires, permettant ainsi des lectures détaillées. Le regroupement sur la base de structures syntaxiques récurrentes renseigne, entre autres, sur les dénominations attribuées et sur les relations sujet-objet qui sous-tendent les actions.

La mesure des cooccurrences ou *collocate* est plus complexe Les cooccurrences mesurent la fréquence relative de deux mots : pour un mot-pivot donné, ses co-occurrentes sont tous les mots qui, par rapport à la valeur absolue de leur fréquence, apparaissent le plus grand nombre de fois. La co-occurrence est caractérisée par une valeur de fré-

quence du mot qui co-occurre au pivot, ensuite par une valeur qui indique la co-fréquence des deux mots ensemble, l'indice de spécificité qui nous fournit une mesure relative à l'ensemble du corpus et nous permet de déchiffrer la « significativité » de l'association, et enfin la valeur de la distance entre les deux mots. Damon Mayaffre en parle comme de :

> la co-présence ou présence simultanée de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d'un même contexte linguistique : les paragraphes ou la phrase par exemple, ou encore une fenêtre arbitraire.[9]

Les cooccurrences informent sur le comportement sémantique visible grâce aux autres mots qui se regroupent autour du premier, en ayant établi une longueur de contexte (Cheng, 2013). Hunston souligne comment cette technique montre des corrélations latentes et « often unavailable to intuition or conscious awareness » (Hunston, 2002), capables de transmettre des messages implicites. Les cooccurrences se traduisent par des mesures associées à des termes permettant ainsi d'identifier les noyaux sémantiques. La cooccurrence généralisée ou polyoccurrence considère les cooccurrences non pas en couple, deux par deux, mais comme une matrice où les mots sont en rela-

[9] Mayaffre, D., « L'entrelacement lexical des textes. Cooccurrences et lexicométrie », *Journées de Linguistique de Corpus*, Lorient, pp. 91-102, 2007.

tion les uns avec les autres. Victorri & Fuchs affirment que « cette influence n'est pas à sens unique mais réciproque entre les différents éléments »[10]. Cette intuition est suivie par Viprey qui considère le texte (au niveau statistique) comme une entité réticulaire où tous les mots sont corrélés et où le sens de chacun est le résultat d'une négociation de sens mesurée par la distance, la fréquence et la cooccurrence.

Ces mesures font partie des techniques de la textométrie, discipline qui propose une approche et des outils pour analyser les corpus en format numériques, articulant synthèses quantitatives et analyses à même le texte. La disponibilité de grandes quantités de données a permis une systématisation des savoirs. Les analyses texto-métriques permettent de décrire des caractéristiques lexicales, syntaxiques, morphologiques, etc. d'un texte. Il est possible donc de confronter des hypothèses linguistiques à une grande quantité d'occurrences et/ou de créer des hypothèses à partir des données observées. Il est aussi possible d'aborder le texte non seulement par une lecture linéaire, mais aussi par l'observation ciblée de certains éléments et caractéristiques du texte même que l'on sélectionne en fonction des questions que l'on se pose.

---

[10] Victorri B., Fuchs, C., *La polysémie : construction dynamique du sens*, Hermès, Paris, 1996.

1.2.3.3. Le mot augmentable

Le mot est manipulable au sens d'augmentable. Le langage de codage permet d'ajouter des annotations à plusieurs niveaux : morphologique, syntaxique, sémantique et cela permet de traiter un texte avec toutes les informations qui y sont associées. Les recherches se font sur plusieurs niveaux. Le mot numérisé change sa structure, devient liquide. A côté d'une annotation morphologique et syntaxique, il est possible d'annoter des catégories sémantiques définies par les nécessités de la recherche.

Une technique qui s'est développée à partir des années 2000 est l'analyse des sentiments ou la fouille d'opinion, principalement liée aux études de sociologie quantitative. L'analyse des sentiments permet ainsi de retracer, par exemple, les tendances de la communication au sein des discours publics. Dans le domaine du *Digital Literary Criticism,* des études cherchent à définir la forme de l'intrigue grâce à la courbe émotive du texte, ou bien à analyser les personnages selon la valeur émotive contenue dans leurs discours. Les nombreuses applications qui ont été développées utilisent les lexiques de polarité, créés manuellement ou automatiquement, qui permettent de comparer les données et de procéder à l'analyse.

L'analyse des sentiments, combinée dans certains cas à la fouille d'opinion avec laquelle elle partage des paramètres et une méthodologie, identifie automatiquement la polarité émotionnelle des portions de texte. Sa structure conceptuelle est quintuple : 1) la cible vers laquelle le sen-

timent ou l'opinion sont orientés, 2) l'entité de la cible sur laquelle le sentiment ou l'opinion sont exprimés, 3) le sentiment ou l'opinion exprimés, 4) la source de l'opinion ou du sentiment, 5) un temps donné pendant lequel l'événement se produit. À son tour, le sentiment est déterminé par une triple structure : 1) la typologie : rationnelle ou émotionnelle, 2) une orientation : positive, négative ou neutre, 3) une intensité qui augmente ou diminue la valeur initiale. Le résultat est une valeur (positive ou négative) et une valence marquée par des intensificateurs ou des diminuants.[11]

L'articulation de la recherche en TAL vise à perfectionner les techniques et les instruments dans une direction de plus en plus complexe et performante, à un niveau technique, lié au calcul. Il revient aux HN de savoir appliquer ces instruments et techniques de fouille d'information à des domaines variées et à des recherches ponctuelles, suivant les problématiques humanistes.

1.2.4. La linguistique du corpus

L'essor du TAL a été possible grâce à la disponibilité du matériel à élaborer, des données réelles et non des données issues d'un laboratoire. L'organisation de ces données appelle la notion de corpus, qui devient linguistique de corpus. À partir du moment où des technologies du TAL sont appliquées à un texte, la typologie de texte a été déjà

[11] Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.

choisie en raison de notre question de recherche. Ce choix est donc à la fois une collecte et une réponse à des critères déterminés.

La linguistique de corpus est une discipline qui a pris son essor lorsque le matériel en format numérique est devenu grandement disponible ainsi que grâce à la diffusion des méthodes statistiques. La possibilité de scanner, enregistrer et transformer en numérique un texte, de même que la diffusion et la présence toujours plus croissante du matériel à la disposition de la communauté scientifique, ont permis à cette discipline de se renforcer. Le *machine learning,* qui a des performances meilleures de manière directement proportionnelle à la qualité des données, contribue à cette diffusion.

Le corpus est considéré comme l'exemplaire d'une variété linguistique et il est le résultat d'une œuvre de sélection. Il doit posséder les caractéristiques suivantes :

1. le corpus doit être *représentatif* en variété et en proportion selon le phénomène à analyser. Il doit être équilibré en terme de quantités des sous-corpus et de proportion de genre textuel;
2. le corpus doit être *fini,* il doit avoir des dimensions définies;
3. le corpus est considéré en format *numérique* : manipulable et consultable de manière automatique, ce qui permet les annotations.
4. Il doit devenir un *standard de référence* : une fois

construit, un corpus représente une référence pour les experts.

Le corpus peut être *général* lorsque le chercheur vise une variété linguistique non caractérisée de manière ponctuelle ou *spécifique* lorsque le chercheur veut représenter un phénomène précis (le langage politique du XXème siècle) ou un domaine particulier (musical, artistique, etc…). Il peut être composé par des textes (différence entre écrits, oraux ou mixtes) et/ou par des images, ce qui lui confère l'attribut de « multimodal ». L'aspect temporel joue également un rôle important : le corpus peut être considéré s*ynchronique* lorsque l'on ne prête pas attention au changement dans le temps, et *diachronique* lorsque les données sont capturées sur un échelon temporel considérable pour les exigences de la recherche. Enfin la langue aussi joue un rôle fondamental dans la collecte d'un corpus qui peut être monolingue, bilingue ou multilingue.

1. 3. LA THÉMATIQUE

1.3.1.L'interculturalité multiculturalisme : enjeux et perspectives

Les recherches en Humanités Numériques qui se focalisent sur l'extraction d'information des textes peuvent porter sur une variété de thématiques presque infinie : notre travail se situe à l'intérieur de la thématique de l'interculturalité, vu l'importance que celle-ci revêt dans les sociétés contemporaines de plus en plus composées par des mélanges de cultures. Nous voulons trouver les traces du discours interculturel à travers les instruments de traitement automatique des langues. Une telle affirmation reste néanmoins trop vague, aussi lui préférerons-nous la question suivante : qu'est-ce que le discours interculturel ? Pour y répondre, nous introduirons le concept d'interculturalité, et en particulier celui de communication interculturelle, avant de passer au traitement de l'implicite linguistique qui est, à notre avis, une pratique pouvant souvent accueillir le discours interculturel.

L'Unesco définit ainsi la culture :

> La culture est l'ensemble des traits distinctifs, spirituels, matériels, intellectuels et psychologiques, d'une société ou d'un groupe social et englobe la totalité des manières d'être existant au sein d'une société ; elle comprend, au minimum, l'art et la littérature, les modes de vie, les manières

de vivre ensemble, les systèmes de valeurs, les traditions et les croyances.[1]

Les êtres humains appartenant à différentes cultures ont des échanges constants, en tant qu'individus ou en tant que communauté, dans des situations où les caractéristiques individuelles déterminent des rapports de forces posés dans des relations qui ne sont pas toujours équitables. Quand deux cultures entrent en contact, deux visions du monde, deux langages différents doivent trouver un accord interprétatif dans un cadre commun, de sorte que le message puisse passer de l'une à l'autre de manière correcte. La coexistence entre cultures exige une réflexion : c'est dans ce cadre que plusieurs théories ont pu naitre. Les deux principales théories, bien qu'elles soient constellées de variantes, sont l'interculturalité et le multiculturalisme. Le multiculturel est traditionnellement attribué aux pays anglo-saxons alors que l'interculturel est attribué à l'aire francophone : il s'agit, en réalité, de deux manières de questionner le rapport à l'autre, chacune attribuant une importance différente à la définition de groupe culturel, aux relations et au concept d'autonomie.

Si l'interculturalité se réfère à un échange réciproque

---

[1] *Déclaration de Mexico sur les politiques culturelles*. Paris, UNESCO, 1982.http://portal.unesco.org/culture/fr/files/35197/11919407161mexico_fr.pdf/mexico_fr.pdf ; UNESCO. (3 novembre 2001). *Déclaration universelle de l'UNESCO sur la diversité culturelle*. http ://portal.unesco.org/fr/ev.php-URL_ID=13179&URL_DO=DO_TOPIC&URL_ SECTION=201.html [dernier accès 25/9/2021].

entre normes et visions de valeurs, le multiculturalisme[2] privilégie le statut des cultures et leur autonomie.

Le multiculturalisme est un modèle dans lequel les cultures sont juxtaposées, composant une mosaïque où les échanges ne sont pas nécessaires. A l'origine du concept du multiculturalisme, il y a à la fois une réalité sociologique (les mouvements migratoires ont toujours été très présents dans l'histoire du monde et ils continuent sans cesse à mélanger les cultures les plus différentes) et une école philosophique qui valorise la pluralité, en établissant des droits spécifiques pour les minorités culturelles. Le multiculturalisme canadien peut fournir, à cet égard, un bon exemple. Il fonctionne, en effet, sur des « accommodements raisonnables », sur des instruments de droits qui permettent de lutter contre les discriminations en respectant des exigences ponctuelles formulées par les minorités ethniques. Le multiculturalisme canadien prône la coexistence de différentes cultures au sein du pays, par opposition à l'intégration et à la constitution d'une identité commune. En ce sens, le terme « multiculturalisme canadien » est souvent utilisé par opposition à celui d'« interculturalisme québécois » qui préconise, pour les minorités, une intégration et un dialogue avec la culture majoritaire.

L'interculturalité vise au bien-être de la culture majoritaire, tout en alimentant un dialogue avec les minorités culturelles, en visant un cadre commun des droits ainsi que

---

[2] May, P., *Philosophies du multiculturalisme*, Les Presses de Sciences Po, Paris, 2016.

le respect du vivre ensemble. L'interculturalité fait sien la compétence culturelle, la capacité de voir les choses selon la perspective de l'autre, en adoptant le postulat que chaque culture n'est qu'une option parmi de nombreuses possibilités.[3]

Sans trop s'attarder sur le vif débat en cours entre les deux théories, il suffira de relever la manière dont l'interculturalité appelle l'élément du dialogue, sans pour autant forcer les nouveaux arrivés dans la société à une intégration invasive. La dimension du dialogue est propre à l'interculturel ; celui-ci doit, en effet, s'efforcer de mettre en place une action de compréhension mutuelle à travers des compétences spécifiques. L'Unesco définit les compétences culturelles comme une série d'instruments pour comprendre l'autre, où le terme « compétence » serait de l'ordre du mesurable :

> Cette diversité croissante des cultures, avec sa fluidité, son dynamisme et son pouvoir de transformation, implique pour les individus comme pour les sociétés des compétences*(note dans le texte : Le terme est employé au pluriel puisqu'il recouvre beaucoup de qualifications, d'attitudes connexes et de nombreux types de savoir qui doivent exister simultanément) et des capacités spécifiques à apprendre, réapprendre et désapprendre pour parvenir à l'épanouissement personnel et à l'harmonie sociale. L'aptitude à déchiffrer d'autres cultures de manière à la fois

[3] Aa.Vv., *Compétences interculturelles, Cadre conceptuel et opérationnel*, Organisation des Nations Unies pour l'éducation, la science et la culture, Paris, pg. 11, 2013.

> impartiale et significative ne suppose pas seulement un esprit ouvert et pluraliste, il y faut aussi une conscience de sa propre identité culturelle. Une culture qui mesure sans indulgence ses points forts et ses limites est à même d'élargir ses horizons et d'enrichir ses ressources intellectuelles et spirituelles en tirant des enseignements de conceptions épistémologiques, éthiques et esthétiques et de visions du monde autres que les siennes.[4]

Le dialogue interculturel fait partie de ces compétences : la capacité de comprendre passe par la capacité de changer de perspective et de créer un cadre de référence commun. Le dialogue interculturel désigne ainsi :

> spécifiquement le dialogue qui a lieu entre des personnes appartenant à des groupes culturels différents. Celui-ci présuppose que les participants acceptent d'écouter et de comprendre de nombreux points de vue différents, y compris ceux de groupes ou d'individus avec lesquels ils sont en désaccord.[5]

Il ne s'agit pas de définir un échange direct, mais de s'intéresser à la manière dont l'autre est présenté, au moment où l'énonciateur met en scène une représentation de l'autre. Le dialogue interculturel concerne également la représentation de l'identité culturelle de l'autre.

---

4 *Ibidem*.

5 *Ivi*, p. 14.

1.3.2. Le dialogue interculturel : l'importance de la parole

C'est dans ce cadre que l'UNESCO a mis en place un projet, *Intercultural Dialogue*[6], qui est un hub international et collaboratif consacré à la thématique du dialogue interculturel et ce dans le but d'augmenter la compréhension mutuelle et le respect de la diversité. Concepts clairs et simples, mais difficiles à concrétiser lorsqu'ils sont mis en pratique. D'un point de vue du droit par exemple, une expression telle que « respecter la diversité » a une grande ampleur : quelle diversité ? Et surtout à quel niveau de respect ? Les concepts clés qui s'affichent dans le home page (fig. 1) nous renvoient immédiatement à l'idée qu'il s'agit de repenser les catégories conceptuelles, d'amener des perspectives fonctionnelles différentes.

Fig. 1. Homepage du site Dialogue Interculturel de l'UNESCO (https ://fr.unesco.org/interculturaldialogue/)

6 https ://fr.unesco.org/interculturaldialogue/ [dernier accès 15/09/2021].

L'interculturalité passe ainsi par une action d'éducation, par une sensibilisation à des thématiques nouvelles. Le parcours est collectif et en construction. L'UNESCO a également rédigé un document de compétences interculturelles à l'intérieur duquel trouve sa place le rapport « vocabulaire interculturel » dont la métaphore visuelle est celle d'un arbre comme on peut le voir ci-dessous (Fig. 2).

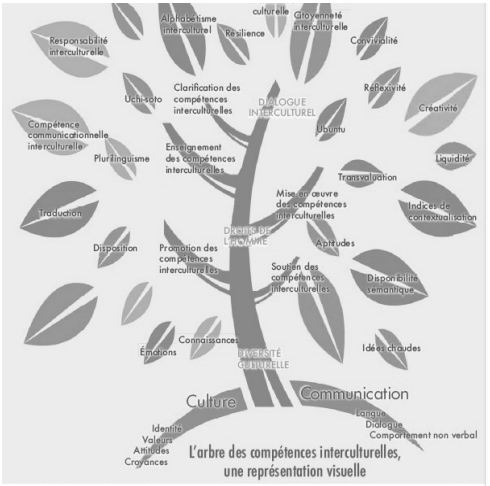

Fig. 2. Source : Les compétences interculturelles, rapport UNESCO

Il est présenté comme un arbre, où la différence entre branches, tronc etfeuilles est conceptuelle. Les racines sont la culture (identité, valeurs, attitudes et croyance) et la communication (langue, dialogue, comportement non verbal), les bases permettant à tout le système d'exister. Ensuite le

tronc s'élance, comprenant la diversité culturelle, les droits de l'homme et le dialogue interculturel.

Les branches sont, de fait, des étapes opérationnelles : clarification, enseignement, promotion, soutien et mise en œuvre des compétences interculturelles.

Les feuilles sont les concepts qui enrobent la structure. Certaines feuilles sont laissées vides pour signifier dans quelle mesure la collaboration commune et l'ouverture sont un signe distinctif de l'opération visant à repenser des catégories.

Parmi les feuilles, nous retiendrons l'exemple du mot *Ubuntu* : pour s'entrainer à comprendre et à respecter les autres cultures, il faut apprendre leur *manière de dire*.

Le mot « Ubuntu » se définit de la manière suivante :

> Ubuntu, mot africain renvoyant à toute une conception philosophique de l'interdépendance dans les relations humaines, est à la fois un idéal éthique et un aspect de l'identité australo-africaine. Le proverbe « xhosa ubuntu ngumuntu ngabantu », que l'on peut traduire à peu près par « une personne n'est une personne que par l'intermédiaire d'autres personnes », exprime le sens profond du mot ubuntu. Ubuntu ne désigne pas seulement une manière d'être louable chez un être humain ou un ensemble de valeurs et de pratiques, mais l'essence même de l'humanité en tant que reconnaissance de l'humanité d'autrui. Ce qui fonde le statut de personne, ce sont les relations éthiques avec autrui.[7]

[7] Aa.Vv., *Compétences interculturelles, op. cit.,* pg. 21.

Le mot est à la fois signe graphique appartenant à une autre culture et porteur de concept d'une autre culture. Les mots sont des outils puissants pour véhiculer des concepts culturels, : les étudier selon cette perspective revient aussi à reconnaitre où se cachent les écueils auxquels le dialogue interculturel peut être exposé.

1.3.3. L'implicite linguistique : quel rapport avec le dialogue interculturel ?

D'où viennent les malentendus ? D'où viennent les idées reçues et, sous une forme plus grave, d'où viennent les stéréotypes ?

Le langage – par sa nature économique – se sert de l'implicite. Cet implicite se définit comme du contenu qui n'est pas le véritable objet de l'énoncé, mais qui peut être déduit ou reconstruit grâce à d'autres éléments présents dans l'énoncé. L'importance de l'implicite se révèle lors du « tournant contextuel »[8] du XXIème siècle, tournant où la nouvelle dignité attribuée à la notion de contexte amène à une analyse du discours qui considère aussi le système locuteur-destinataire.

Ce système est composé d' un cadre de connaissances considérées comme partagées et auxquelles on se réfère pour véhiculer le message. Lorsqu'il y a un échange communicatif, l'énonciateur ne fournit pas tous les détails sont

---

[8] Sbisà, M., « Presupposizioni e contesti », *La svolta contestuale*, C. Penco, (dir.) Milano-New York 2002, pp. 221-239.

au destinataire, car certaines normes ou connaissances sont considérées comme partagées, sans qu'il soit donc nécessaire de les expliciter. Ici, ce n'est pas l'éventuelle intentionnalité qui nous occupe : nous nous limitons à remarquer l'existence de ce dispositif linguistique (fig. 3)

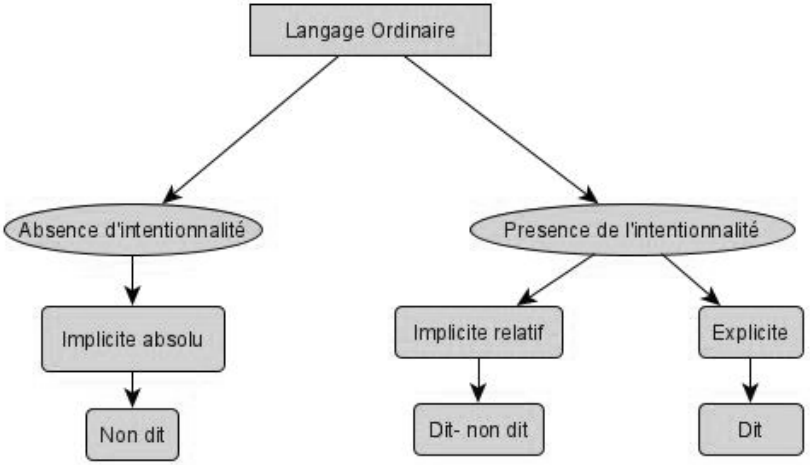

Fig. 3. L'impplicite dans la communication.

Un aspect important de l'implicite est son caractère « collaboratif » dans la construction du sens lors de la communication.

Le premier à avoir systématisé l'idée selon laquelle la construction de sens est une pratique non univoque et qui va dans les deux sens d'une communication fut le philo-

sophe Paul Griece[9] grâce au Principe de Coopération structuré dans les quatre maximes suivantes :

1) maxime de la qualité : chacun fournit une contribution réelle à la conversation ;

2) maxime de la quantité : chacun fournit une contribution qui ne soit ni excédante ni lacunaire par rapport au caractère nécessaire de l'information ;

3) maxime de relation : chacun fournit une contribution à la conversation qui soit pertinente, en relation avec le but de la conversation ;

4) maxime de manière : chacun fournit à la conversation une contribution qui en facilite la jouissance, en particulier : être clair, être concis.

Ces maximes règlent la conversation, c'est-à-dire que les participants déduisent les informations manquantes véhiculées par l'implicite de manière à ne jamais transgresser ces quatre lois.

Si, par exemple un journaliste, dans le titre d'un article écrit : « Une fille a été attaquée par un jeune marocain », il ne respecte pas deux maximes : celle de la relation, car dans l'information relative à une agression la provenance géographique du criminel n'est pas pertinente, et de quantité, car la provenance est une information en plus par rapport à ce que l'on sait sur la fille. Le destinataire, pour reconstruire logiquement le message est amené à penser qu'il doit y avoir une relation entre le fait d'être un criminel et

---

[9] Griece, P., *Logic and conversation* , Cole, P., Morgan, J. L. (dir.), *Syntax and Semantics, III : Speech Acts*, Academic Press, New York, 1975.

le fait d'être marocain. Il s'agit d'un exemple de persuasion au négatif, mais il montre bien toute la puissance dont l'implicite peut se revêtir dans le discours interculturel. Souvent, le stéréotype se cache derrière l'implicite : certaines idées peuvent passer même si elles ne sont pas dites ouvertement. En effet,« souvent l'implicite est utilisé pour nous laisser entendre quelque chose qui – dit ouvertement – ne serait jamais accepté ».[10] L'implicite se distingue en *presupposition* et *implicature*, ou selon la tradition héritée de Ducrot[11] en *présupposition* et *sous-entendu*. Au-delà de la terminologie, la différence entre les deux concepts est que la *présupposition* est inscrite dans l'énoncé, tandis que l'implicature (ou implicite) doit être déduit par l'interlocuteur selon un raisonnement plus ou moins spontané[12]. Le présupposé fait donc partie intégrante du sens de la phrase ; condition d'activation du processus de sens, c'est la connaissance requise pour pouvoir procéder à une première interprétation des indices linguistiques, afin de reconnaître et d'insérer les références fournies dans un cadre sémantique correct.

Le sous-entendu ou implicature, d'autre part, est le résultat non pas tant d'une opération logique d'inférence, mais d'une opération de systématisation des indices four-

---

10 Sbisà, M., *Detto non detto. Le forme della comunicazione implicita,* Laterza, Bari, 2010.

11 Ducrot, O., *Dire et ne pas dire. Principes de sémantique linguistique*, Hermann, Paris, 1972 ; Ducrot, O., Todorov, T., *Dictionnaire encyclopédique des sciences du langage*, Seuil, Paris, 1972.

12 Griece, P., *Logic and conversation, op. cit.*

nis directement par le texte et des éléments présents dans le contexte. Ce dernier, fondamental dans ce type d'opération, comprend la connaissance des autres acteurs de la communication et par conséquent du rôle qu'ils jouent, mais aussi la maîtrise des techniques qui règlent la communication, la rendent complète et pertinente.

Dans cette possibilité de déduire ou de reconstruire, et qui laisse de fait l'interlocuteur compléter le message grâce aux informations qu'il détient, se cache un danger pour la compréhension interculturelle. Il est possible que, dans la communication interculturelle, la représentation de l'autre et de sa culture de l'autre soit déformée. Mais il est également possible que, dans cette reconstruction, se formalisent des traits typiques de certaines cultures, qui peuvent à leur tour contribuer à bâtir un imaginaire collectif : nous avons là le stéréotype.

S'il est vrai qu'une transmission correcte du message se fonde sur l'équivalence des cadres de normes et connaissances que les deux parties considèrent comme communes, un décalage entre ces deux cadres peut amener à un échec dans la transmission du message. Un malentendu est tout simplement analysable comme la réception par le co énonciateur d'un concept de manière différente par rapport aux intentions de l'énonciateur. L'implicite crée un court-circuit : de moyen qui sert à l'économie linguistique il devient moyen qui empêche une transmission correcte du message.

Si l'implicite est transversal à tout type de discours, il est cependant typique de certaines stratégies, comme celle de la persuasion. Par exemple, le discours publicitaire, ou

le discours politique, utilise principalement des techniques persuasives. La présupposition équivaut à considérer comme acquis certains concepts de départ pour pouvoir bien interpréter le message, en les soustrayant de fait à la discussion. Ce qui est indiqué ou suggéré comme présupposé est donné comme accepté et valable pour le système de valeurs dans lequel se trouvent le locuteur et le destinataire. Dans le cas du discours publicitaire, la persuasion débute par un mouvement quasi involontaire ; poser des bases indiscutables pousse le destinataire à se concentrer davantage sur le contenu du message. Une autre fonction du présupposé d'un point de vue persuasif est de renforcer les liens de groupe entre le locuteur et le destinataire : en effet, tenir l'information pour acquise permet de souligner la perception du destinataire de partager des valeurs communes, pour lesquelles certaines explications sont superflues. Dans le cas du discours politique, appartenir au même groupe revient à renforcer la notion de cible dans la perception du destinataire, rendant ainsi plus efficace la fonction persuasive.

Le dialogue culturel doit tenir compte de ces enjeux : l'accès au langage grâce à des modèles interprétables par des machines permet une analyse approfondie de ces dynamiques : ce sera l'objet de la deuxième partie de cet ouvrage.

SECONDE PARTIE

La fouille de textes : deux cas d'étude

2. POUR ENTRER EN MATIÈRE

La théorie nous a montré la complexité d'une démarche touchant à la linguistique, à la littérature et à l'informatique. Pour mieux saisir les procédures, deux cas d'études vont tenter de mettre en pratique les idées proposées. Il s'agit de deux cas différents, l'un touchant au discours institutionnel, à travers un corpus construit à partir du discours officiel contenu sur le site web du Ministère de l'Immigration, de la Francisation et de l'Intégration du Québec (MIFI), l'autre touchant au discours littéraire, à travers un corpus construit à partir des lauréats au Prix Goncourt. En effet, chaque cas d'étude suit une procédure qui peut être généralisée. Même si le contenu spécifique de la recherche, les choix thématiques faits et les problématiques abordées sont différents, les étapes à suivre sont toujours les mêmes :

1. La construction d'un corpus représentatif de la variété linguistique qui correspond à la thématique d'étude choisie.
2. L'identification de modèles linguistiques correspondant à la problématique selon des hypothèses linguistiques à la recherche de l'extraction du sens latent.
3. L'application des techniques d'extraction automatique d'informations à travers l'usage d'outils de traitement automatique des langues.
4. Interpréter les résultats en utilisant les perspectives issues des Humanités Numériques : dans notre cas,

il s'agira du *distant reading* pour le discours littéraire et de l'analyse du discours assistée par ordinateur pour le discours institutionnel.

Nous allons donc appliquer cette procédure pour rechercher les éléments soulignant l'interculturel dans les deux typologies de discours annoncés : le discours institutionnel et le discours littéraire.

2.1. LE DISCOURS INSTITUTIONNEL

2.1.1. La reconfiguration identitaire dans la communication du Ministère

Le premier cas d'étude est une analyse ayant pour base la Communication du Ministère de l'Immigration, de la Francisation et de l'Intégration du Québec (MIFI) à la recherche des stratégies linguistiques et discursives mises en place pour définir les contours identitaires des migrants d'un point de vue institutionnel et politique. À travers les outils du TAL et particulier de la textométrie et en adoptant la perspective de l'analyse du discours assistée par l'ordinateur, nous montrerons une représentation polarisée, où l'énonciateur (le Ministère) se montre accueillant envers le migrant tout en conservant un pouvoir décisionnel sur le statut du migrant qui est considéré comme une ressource positive pour la société d'accueil à condition de pouvoir s'intégrer linguistiquement.

Les migrations caractérisent l'histoire de l'humanité depuis ses origines, contribuant à modeler le monde en termes sociaux, culturels et géopolitiques. Toutefois, même si l'acte migratoire appartient à la nature de l'être humain, le terme *migrant* n'a pas de définition universellement reconnue sur le plan juridique. L'Organisation des Nations Unies (ONU) définit le migrant comme :

> toute personne qui a résidé dans un pays étranger pendant plus d'une année, quelles que soient les causes, volontaires ou involontaires, du mouvement, et quels que

soient les moyens, réguliers ou irréguliers, utilisés pour migrer.[1]

Indépendamment des raisons qui poussent les peuples et les individus à migrer, et des moyens utilisés pour le faire, le migrant se caractérise par un déplacement qu'il effectue et qui est le début d'un processus de reconfiguration identitaire[2] engendré par la différence perçue entre les valeurs de la communauté d'origine et celles de la communauté d'accueil. Au sein de ce processus continu, les deux communautés négocient leurs normes de coexistence : la communauté d'arriver identifie dans le nouvel espace les normes sociales compatibles avec son identité, celles qui son indispensables et celle qui peuvent être négociées. Egalement la communauté d'arrivé se pose les mêmes interrogatifs et doit procéder au même mouvement de recherche d'équilibre entre les normes qu'elle peut accepter et celles qui mettraient en danger sa culture.

Chacune des parties impliquées doit se traduire face à l'autre, à la recherche d'un nouvel équilibre social, culturel et juridique.[3] Ces formes de rencontre et de dialogue font

---

1 Site web ONU – Réfugiés et migrants : https ://refugeesmigrants.un.org/fr/d%C3%A9finitions, et aussi https ://www.iom.int/fr/termes-cles-de-la-migration [dernière consultation 2/10/2021].

2 Lacrampe-Camus, I., « Reconfiguration des ancrages et construction des origines dans un contexte de double migration. Jeunes d'origine équatorienne entre l'Espagne et Londres », *Cahiers des Amériques latines*, 91, pp. 153-170, 2019.

3 Abdessadek, M., « Identité et migration : le modèle des orientations identitaires », *L'Autre* 3 (13) : pp. 306-317, 2012 ; Fargues, P., « Migration and Identity : The Paradox of Reciprocal Influences », *Esprit*, 1 : pp. 6-16, 2010. DOI : 10.3917/espri.1001.0006.10.3917/lautr.039.0306.

l'objet d'activités législatives de la part des gouvernements du monde entier qui travaillent en permanence entre la compréhension de l'implicite socioculturel caractérisant une communauté et l'imposition de normes juridiques équitables pour tous. Les politiques de gestion des flux migratoires sont très variées au niveau mondial, influencées par de nombreux facteurs géopolitiques, économiques et sociaux. L'exemple du Québec est intéressant en raison de son histoire : d'abord colonie française (XVIème-XVIIIème siècles) puis anglaise (XVIIIème-XXème siècles), le Québec se caractérise aujourd'hui par une diversité culturelle de plus en plus grande (Durand 2003) et fait de l'intégration et de l'interculturalité ses principes de gestion des politiques. La population qui l'habite aujourd'hui est culturellement stratifiée, composée d'individus d'origine française, anglaise, autochtone ainsi que d'autres nationalités attirées par la dimension ouvertement multiculturelle du Québec. Dans un tel espace, où se déroule constamment un processus de reconfiguration identitaire, le discours public et politique joue un rôle fondamental, mettant en scène la dynamique de déconstruction et de reconstruction d'un espace partagé entre la communauté d'accueil et la communauté d'origine de l'immigrant ou nouvel arrivant. Les traces de ce genre *de dialogue reconfiguratif* peuvent être repérées en plusieurs lieux, tels le discours de presse, le discours institutionnel ou le discours publicitaire, chacun porteur d'une différente scène d'énonciation définies par des rôles et des limites de l'espace énonciatif et conceptuel.[4]

[4] Maingueneau, D., *Les termes clés de l'analyse du discours*, éditions du

Une première étape présentera la discipline de l'analyse du discours assistée par corpus et la prosodie sémantique ; suivra l'examen de l'état de l'art, un panorama des études qui se sont servies de la même approche pour étudier le discours autour de l'immigration pour revenir ensuite à la description détaillée du corpus construit à partir du site web du Ministère de l'Immigration, de la Francisation et de l'Intégration du Québec. Nous présenterons ensuite les résultats de la recherche en nous concentrant en particulier sur la représentation qui est faite de l'individu et du concept de migrant, et sur la prosodie discursive qui en résulte et la représentation de l'énonciateur et du co-énonciateur.

2. 1. 2. L'analyse du discours assistée par corpus et la prosodie sémantique

Au cours des dernières années, le cadre conceptuel de l'analyse critique du discours[5] s'est unie à celui de la lin-

---

Seuil, Paris, 2009.

[5] van Dijk, T.A., *Handbook of Discourse Analysis : Discourse analysis in society*, Academic Press, London, 1985 ; Van Dijk, T.A. (dir.), *Discourse studies : A multidisciplinary introduction*, SAGE Publications, London, 2011. DOI : 10.4135/9781446289068 ; Fairclough, N. *Critical discourse analysis : the critical study of language*, Longman, London/ New-York, 1995 ; Fairclough, N., *Language and Power,* Longman, London, 2001 ; Fairclough, I., Fairclough, N. *Political Discourse Analysis : A Methods for Advanced Students*, Routledge, London, 2012 ; Maingueneau, D., Charaudeau, P., *Dictionnaire d'analyse du discours,* Seuil, Paris, 2002.

guistique de corpus[6] engendrant ce qu'on définit « analyse critique de discours assistée par corpus » ou ACDAC[7]. Elle a été saluée comme une « useful methodological synergy »[8] ce qui nous indique comment l'association entre analyse qualitative et quantitative peut être heureuse si l'on parvient à définir une modalité paritaire de faire dialoguer les deux approches entre elles.

L'analyse critique du discours s'oriente vers une analyse qualitative, par une approche interdisciplinaire qui étudie les structures linguistiques et les stratégies discursives mises en œuvre pour construire des représentations sociales, politiques et culturelles, et pour établir des relations de pouvoir entre les communautés. Par ailleurs, la linguistique de corpus se sert principalement des techniques quantitatives, qui exploitent des mesures statistiques, telles que des listes de mots clés, des collocations et des concordances, afin d'analyser le niveau de régularité linguistique et les struc-

---

6 Condamines, A., « Linguistique de corpus et terminologie », *Langages, La terminologie : nature et enjeux*, 157, p. 36-47, 2005.

7 Baker, P., *Using Corpora in Discourse Analysis*. Continuum, New-York / London, 2006 ; Cheng, W., « Corpus-Based Linguistic Approaches to Critical Discourse Analysis », in Chapelle, C.A. (dir.), *The encyclopedia of applied linguistics*, Wiley-Blackwell Publishing, Oxford, pp. 1-8, 2013.

8 Baker, P., Gabrielatos, C., Khosravinik, M., McEnery, T., Wodak, R., « A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press », *Discourse and Society,* 19(3) : pp. 273-306, 2008.

tures récurrentes qui émergent des textes[9]. L'ACDAC insiste sur la nécessité de relier des modèles syntaxiques définis à grande échelle avec des éléments issus d'une lecture détaillée des concordances et d'aligner ces mêmes modèles syntaxiques aux dispositifs rhétoriques, aux processus de production, distribution et utilisation du texte qui fait de l'acte linguistique une pratique socioculturelle à interpréter par rapport à un contexte spécifique.

Les études se servant de l'approche ACDAC sont particulièrement attentives au concept de prosodie sémantique, appelée aussi prosodie du discours[10]. Le premier à avoir introduit ce concept, lors d'une communication orale, a été John Sinclair, s'inspirant du concept de prosodie phonologique introduite par Firth[11]. Le premier à l'avoir utilisé dans une communication écrite fut Louw, qui le définit comme « a consistent aura of meaning with which a for is imbued by its collocates »[12]. Il s'agit donc d'évaluer le comportement d'un terme par rapport aux mots qui l'accompagnent et de déterminer ainsi une tendance sémantique dépassant la définition qui en est faite dans le dictionnaire et qui s'ancre dans la compétence linguistique. La prosodie

---

9 Freddi, M., *Linguistica dei corpora*, Carocci, Roma, 2014.

10 Stubbs, M., *Words and phrases : Corpus studies of lexical semantics*. Blackwell, London, 2001.

11 Firth, J.R., *Papers in Linguistics*, *op. cit*.

12 Louw, B., « Irony in the text or insincerity of the writer? The diagnostic potential of semantic prosodies », Baker, M., Francis, G., Tognini-Bonelli, T., (dir.), *Text and Technology : In Honour of John Sinclair*, John Benjamins, Amsterdam, 1993, pp. 30-50.

sémantique ou discursive d'un terme peut être négative, positive ou neutre ; elle veut aussi définir des ensembles de mots qui autorisent – autrement qu'en termes de polarité –l'attribution de sphères sémantiques au terme en question. La prosodie sémantique peut produire aussi des effets de style comme l'ironie qui dérive alors de l'utilisation d un terme avec prosodie négative, pour identifier un concept positif et *vice-versa*. Au-delà de l'ironie, cette manipulation peut indiquer un contenu caché, ou inconscient de l'énonciateur ou locuteur. Un enrichissement ultérieur est apporté par l'application de la prosodie discursive à l'interprétation des comportements linguistiques des acteurs d'une scène d'énonciation et de leur rôle dans la création d'un discours cohérent[13].

La mise à disposition d'amples *corpora* numériques a été fondamentale pour la découverte de phénomènes qui seraient « much less evident to the naked eye »[14] et comme Louw le confirme :« One consequence of the advent of large corpora has been their increased potential for revealing consistencies in the influence of collocation on the behaviour of particular linguistic forms ».

L'ACDAC utilise, pour mener ses analyses, à côté du cadre théorique, les techniques et les mesures suivantes : l'annotation, les concordances et les cooccurrences.

---

13 Stubbs, M., *Words and phrases : op. cit.*

14 Partington, A. « Utterly content in each other's company : Semantic prosody and semantic preference », *International Journal of Corpus Linguistics,* 9(1), pp. 131-56, 2004.

L'annotation d'un corpus enrichit les possibilités de recherche grâce à l'ajout d'informations complémentaires : il est possible d'annoter le texte en suivant un niveau morphologique, syntaxique et sémantique. L'annotation est évidemment un processus qui peut se faire avant ou après le traitement d'un corpus : dans le premier cas, on utilise le *machine learning* pour étiqueter de manière automatique avec des catégories morpho-syntactiques. Dans le deuxième cas, on ajoute des informations aux textes en langage XML[15] pour l'enrichir d'informations sémantiques pouvant être utilisées dans les recherches. Notre corpus a été annoté avec les catégories des parties du discours, de manière automatique dans un premier temps, puis révisé manuellement. De cette manière, nous avons pu reconstruire des structures syntaxiques récurrentes autour des termes de l'immigration.

Les concordances donnent comme résultat le contexte proche d'un terme, une fois qu' a été choisie la fenêtre de mots à droite et à gauche du mot-pivot en fonction des besoins de recherche et de la typologie linguistique du corpus. Une lecture approfondie permet d'identifier les structures syntaxiques présentes et de déterminer la prosodie du discours.

À un autre niveau, se situent les cooccurrences qui mesurent la fréquence relative d'un terme pivot par rapport à ses co-occurrents : cette mesure trouve des mots qui, bien

[15] Pour plus d'informations voir la TEI – Text Encoding Initiative : <https ://tei-c.org/> [dernier accès : 1/10/2021].

que fréquemment associés au terme pivot, ne sont pas forcément dans leur contexte séquentiel. Hunston souligne comment cette technique montre des corrélations latentes et « often unavailable to intuition or conscious awareness »[16] capables de transmettre des messages implicites.

Dans l'analyse du discours sur l'immigration, le genre textuel le plus étudié avec une approche ACDAC a été celui de la presse, tendance sans doute motivée par l'importance que ce type de discours revêt dans la construction de l'opinion publique.[17] Le rôle majeur que joue la représentation de l'Autre dans le discours public est bien mis bien en évidence par Baker[18] qui, analysant un corpus construit à partir de la presse anglaise divisé par parution (quotidienne, hebdomadaire), orientation politique, style (*broadsheet*, *tabloid* ou *middle-market*) et couverture (nationale, régionale), fait ressortir les stratégies de construction linguistique, et donc identitaire, du *nous* (*we*) qui est chrétien et *vous (you)* qui est musulman.

---

16 Hunston, S., *Corpora in Applied Linguistics*, Cambridge University Press, Cambridge, 2002 ; Pincemin, B., « Concordances et concordanciers : de l'art du bon KWAC », *XVIIe colloque d'Albi Lagages et signification – Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, pp. 33-42, Albi, Jul. 2006.

17 van Dijk, T.A., *Communicating Racism : Ethnic Prejudice in Thought and Talk*, Sage, London, 1987 ; van Dijk, T.A., *Racism and the Press*, Routledge, London, 1991.

18 Baker, P., Gabrielatos, C., McEnery, T., *Discourse analysis and media attitudes : the representation of Islam in the British press*, Cambridge University Press, Cambridge, 2013.

Al Fajri[19] vise à analyser le comportement discursif du terme *immigrants* dans le corpus ukWac construit à partir du web en comptant 1.127.056.026 mots, en le comparant à un corpus plus général comme le British National Corpus. À travers l'analyse de cooccurrences et concordances Al Fajri identifie un « hegemonic discourse » qui se développe autour du migrant. Rares sont les exemples d'une représentation positive des migrants (sauf dans les cas où ils sont représentés comme une solution aux problèmes économiques du pays d'accueil) alors que le discours tend à se polariser sur une idée d'illégalité et à représenter les immigrants de manière indifférenciée passant par sa quantification (nombres, chiffres) et sa déshumanisation, en quelque sorte dans le processus. L'auteur identifie une prosodie sémantique négative : l'immigrant est associé à une incapacité de s'intégrer dans le pays d'accueil et, par conséquent, il est étiqueté comme un être dangereux. Ce genre de représentations semble contribuer à renforcer l'hostilité de l' opinion publique envers les immigrants.

Alcaraz-Mármol[20] se concentre directement sur la pro-

---

[19] Al Fajri, M. S., « Hegemonic and minority discourse around immigrants : a corpus-based critical discourse analysis », *Indonesian Journal of Applied Linguistics* 7(2) : 381-390, 2017.

[20] Alcaraz-Mármol, G., Soto Almela J., « The semantic prosody of the words inmigración and inmigrante in the Spanish written media : A corpus-based study of two national newspapers », *Revista Signos Estudios de Linguistica*, 49(91) pp. 145-167, 2016. DOI : 10.4067/ S0718-09342016000200001.

sodie sémantique des mots *immigracion* (en tant que nom) et *immigrante* (en tant que personne) à l'intérieur d'un corpus construit à partir de la presse nationale espagnole, en particulier des deux quotidiens (*El Pais*, *El Mundo*) appartenant à deux courants politiques différents. Comme dans le cas précédent, l'immigration est associée à l'absence de légalité, à un phénomène en croissance mais d'une certaine manière hors de contrôle, représentant une menace à laquelle on doit faire face.

Griebel et Vollmann[21] adoptent la même approche comparatiste entre quotidiens, *Taz* et *Die Welt*, cette fois au sein du panorama allemand. Ils découvrent ainsi comment divers degrés d'ouverture et de flexibilité envers le migrant s'associent à la variation de la perception des sentiments de dangerosité ou, à l'inverse, d'utilité du migrant. Mentionnons un cas de prosodie positive présenté par Salahshour[22] qui se concentre sur la communication quotidienne de la presse en Nouvelle-Zélande sur une période couvrant 2007 et 2008. Tout en adoptant le cadre théorique des métaphores cognitives proposés par Lakoff et Johnson,[23] son

---

[21] Griebel, T., Vollmann, E., « We can('t) do this. A corpus-assisted critical discourse analysis of migration in Germany », *Journal of Language and Politics*, 18 (5), pp. 671-697, 2019, DOI : 10.1075/jlp. 19006. gri.

[22] Salahshour, N., « Liquid metaphors as positive evaluations : A corpus-assisted discourse analysis of the representation of migrants in a daily New Zealand newspaper », *Discourse*, Context and Media 13, pp. 73-81, 2016. DOI : 10.1016/j.dcm.2016.07.002.

[23] Lakoff, G., Johnson, M., *Metaphors We Live By*, University of Chicago Press, Chicago, 1980.

travail se focalise sur les métaphores liquides qui représenteraient les migrants, faisant émerger un aspect économique positif de la migration.

2. 1. 3. Le corpus MIFI : prétraitement et composition

Le corpus, à la base de notre recherche, a été construit à partir du site du Ministère de l'Immigration, de la Francisation et de l'Intégration – MIFI[24],en téléchargeant automatiquement le matériel qui a ensuite été soumis à un prétraitement manuel. Le corpus est monolingue-français, bien que le site dispose également d'une version anglaise qui pourra être utilisée pour des études futures. Le site web, promulgué par le gouvernement en place, fait référence à la période de son activité, soit 5 ans : le corpus issu est donc considéré comme synchronique. Nous considérons le cadre communicatif comme étant formé par l'énonciateur (le MIFI) et le co-énonciateur (à la fois le migrant et le citoyen Québecois).

En observant l'architecture du site web, le contenu apparaît divisé en trois macro-sections, ou volets, nommées : « I-Immigrer au Québec », « II-Favoriser l'Intégration », « III-Connaitre le Ministère », ce qui semble renvoyer les participants à la scène d'énonciation du discours : si le troisième volet représente le Ministère, qui est l'énonciateur, les deux premiers volets coïncident avec un sous-groupe du

[24] https ://www.immigration-quebec.gouv.qc.ca/fr/accueil.html [dernière consultation 10/06/2021].

co-énonciateur, le migrant, mais qui est saisi à deux moments différents : avant son arrivée (I-Immigrer au Québec) et après son arrivée (II-Favoriser l'intégration). Au co-énonciateur est destiné un discours différent selon l'endroit où il se trouve, encore chez lui ou déjà dans la terre d'accueil. Grâce aux perspectives nouvelles que cette hypothèse ouvre, nous avons décidé en phase de prétraitement d'enrichir le corpus avec des métadonnées qui ont permis de garder la partition en trois volets pendant l'analyse.

Le corpus a été nettoyé : tout d'abord ont été supprimées les références à l'utilisation des liens extra-textuels, des pièces jointes (poids du document en « kilobites » ou « megabites ») et des informations pour l'utilisateur (par exemple « haut de page »). De même, les contenus de liens hypertextuels externes au MIFI (par exemple ceux qui renvoient aux actualités) n'ont pas été inclus car, étant produits par un énonciateur différent, ils n'ont pas été considérés comme intéressants pour notre thématique de recherche. De même, le contenu des annexes n'a pas été pris en compte, car il s'agit entièrement de formulaires techniques.

Le analyses ont été conduites grâce au logiciel TXM[25]

[25] Heiden S., Magué J-P., Pincemin B., « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », in Bolasco, S., Chiari, I., Giuliano, L. (dir.), « Proceedings of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010 », v. 2, pp. 1021-1032, Edizioni Universitarie di Lettere Economia Diritto, Roma, 2010. http ://textometrie.ens-lyon.fr/spip.php?article60&lang=fr.

qui inclut l'annotateur TreeTagger ;[26] il a été possible d'annoter automatiquement chaque terme avec sa partie du discours. Les annotations ont ensuite été corrigées manuellement. Les dimensions du corpus sont détaillées dans le tableau 1 qui suit :

Occurrences	262405
Lemmas	5626

Tableau 1. Composition du corpus par occurrences et lemmes

Un index réduit de richesse lexicale, de 0,021, nous renseigne sur un style répétitif et dépourvu des relations de synonymie.

2.1.4. La représentation de celui qui migre et les limites de la scène d'énonciation

Le terme *migrant* possède un noyau fixe, *-migr*, qui correspond au concept de déplacement, et grâce à un processus de préfixation ou suffixation peut devenir *immigrant* ou émigrant, formes diverses qui portent en elles le point de vue du locuteur ou de l'énonciateur. Si le terme *migrant* de ce point de vue s'affiche comme neutre et indique celui

[26] Schmid H., Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994 ; Schmid, H., Improvements in Part-*of-Speech Tagging with an Application to German, Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995.

qui se déplace, le prefixe *-im* qui donne *immigrant*, renvoie à une scène d'énonciation ou l'énonciateur tient à souligner sa position depuis l'intérieur, d'où il voit arriver l'Autre. De manière opposée le mot émigré avec le préfixe *-é* indique que l'énonciateur veut mettre en évidence la position de la communauté de départ, qui voit un des siens s'éloigner.

Suivant cette réflexion, nous nous sommes attardés sur le choix lexical à opérer dans notre cas d'étude : notre première démarche dans la recherche de la définition linguistique que le MIFI fait de la catégorie du migrant se fait grâce à une CQL (Corpus Query Language) [word = « .*migr.* »%c] qui questionne n'importe quel terme avec la racine **migr.* sans distinction entre majuscule et minuscule. Les résultats affichés dans le tableau 3 attestent 26 formes pour 1483 occurrences, qui ont ensuite été regroupées par lemmes et par parties du discours. Ce travail a été conduit de manière semi-automatique : l'annotation par partie du discours a été faite automatiquement par le biais de TreeTagger intégré dans TXM et ensuite les annotations ont été revues manuellement. Parmi les résultats affichés dans le tableau qui suit (tableau 2) nous observons plusieurs formes : *immigration*, *immigrant*, *immigrer*, *immigré*, *migration*, *migrant*, *migratoire*. Il y a un écart considérable dans la fréquence des trois premières formes et aussi une différente distribution des parties du discours, comme c'est le cas pour *immigrant* qui peut être adjectif et nom. L'analyse détaillée des différents usages des formes conjuguées que nous allons conduire par la suite nous fournira d'autre pistes de réflexion.

Formes	POS	Occurrences
immigration	NOM	951
immigrant	ADJ	307
migration	NOM	13
immigrant	NOM	93
immigrer	VER	62
immigré	ADJ	18
migrant		8
migratoire		1
Autres usages (sigles, entreprises, associations,etc…)		14
		1483 occurrences

Tableau 2. Requête CQL [word = « .*migr.* »%c] – Distribution regroupée par parties du discours et lemmes.

À une première lecture, nous observons comment, parmi la totalité des occurrences, le pourcentage des termes porteurs d'un point de vue neutre (*migr.**) est de 1,5% à côté des termes porteurs d'une perspective de l'intérieur (*immigr.**) du 98,5 %. Le discours est donc polarisé dans la mesure où l'énonciateur fait prévaloir sa position, qui accueille et donne à son discours une orientation.

Analysons en détail les rares formes neutres affichées dans le tableau qui suit (tableau 3) : *migration(s), migrant, migratoire.*

	I volet	II volet	III volet	TOTAL
migratoire	0	1	0	1
migrations	0	0	2	2
migration	2	5	4	11
migrants	0	4	4	8

Tableau 3. Distribution sur les trois volets du corpus des termes : *migration(s), migrant, migratoire.*

Les concordances extraites sont affichées dans le tableau qui suit (tableau 4) :

…la réalité des personnes immigrantes et de leur *parcours migratoire*.
…*migrations* interprovinciales au Quebec
Tendances récentes en *migrations* internationales
…intégration des *migrants*
…journée internationale des *migrants*

Tableau 4. Concordances des termes : *migration(s), migrant, migratoire*

Le terme *migration* au singulier, qui a le taux d'occurrences le plus élevé, n'est pas affiché dans le tableau car il se réfère de fait à un clonage informatique (« migration ») des contenus d'une plateforme à l'autre. La migration ici est non seulement associée à un aspect purement géographique mais aussi à un concept qui est fondamental dans l'analyse : celui de l'intégration.

L'analyse approfondie des formes marquées par une perspective de l'intérieur : *immigration*, *immigrant*, *immigrer*, *immigré* permet d'observer que les termes font référence à un concept (*immigration*), à des individus (*immigrant*, *immigré*) et à une action (*immigrer*), tout en marquant des différences lorsque la même forme indique deux parties différentes du discours, *immigrant* en tant que nom et *immigrant* en tant qu'adjectif. Le plus haut taux d'occurrences est de la forme *immigration* pour un total de 951. Grâce à la lecture des concordances nous savons que 357 occurrences sont écrites avec une majuscule et indiquent l'acronyme du ministère MIFI – Ministère de l'Immigration, de la Francisation et Intégration. Des 594 occurrences restantes nous avons extrait les concordances que nous avons regroupées par structures syntaxiques :

1) NOM + *de* (complément de spécificité)
   demande *d'immigration*, ***démarches*** *d'immigration*, ***matière*** *d'immigration*, ***procédures*** *d'immigration*, ***processus*** *d'immigration*, ***programme*** *d'immigration*, ***catégories*** *d'immigration*, ***statut*** *d'immigration*, ***projet*** *d'immigration* ;
2) NOM + *en* (complément d'argument)
   Conseiller *en immigration*, ***consultant*** *en immigration*, ***Représentant*** *en immigration* ;
3) NOM + *sur* (complément d'argument)
4) ***Loi*** *sur l'immigration*, ***Règlement*** *sur l'immigration*, ***statistiques*** *sur l'immigration* ;
5) *NOM + de* (complément de spécificité)
   Portrait *de l'immigration* ;
6) *NOM + a*
   Candidats *à l'immigration.*

Les groupes sémantiques permettent d'affirmer que : l'aspect juridique est très fort (*loi*, *statut*, *règlement*, *procédures*) et s'accompagne d'un aspect « paperasse » (*demande*, *démarches*) et d'une dimension de planification (*programme*, *projet*, *conseiller*, *consultant*, *candidats*). L'immigration est une *matière* et elle est aussi mesurable (*statistiques*).

La deuxième forme qui a le plus haut taux d'occurrences est *immigrant* (voire tableau 3), pour un total de 400 occurrences qui peuvent être nom (307) et adjectif (93) (tableau 5 et tableau 6).

lemme_*immigrant*_ ADJ	Forme conjuguée	Occ.
	immigrantes	242
	immigrante	58
	immigrants	4
	Immigrants	3

Tableau 5. Détail des formes conjuguées du lemme *immigrant (masc./sing.)* en tant qu'adjectif

lemme_*immigrant*_ *NOM*	Forme conjuguée	Occ.
	immigrants	73
	immigrant	17
	immigrante	2
	immigrantes	1

Tableau 6. Détail des formes conjuguées du lemme *immigrant (masc./sing.)* en tant que nom

Le terme *immigrantes* conjugué au féminin pluriel en tant qu'adjectif a une présence bien plus importante que les autres termes : une lecture des concordances nous permet de constater que, dans la totalité des cas, le nom qui y est associé est *personnes,* ce qui donne vie au syntagme *personnes immigrantes* que l'énonciateur choisit pour définir la catégorie des migrants. Il s'agit d'un choix de style qui nous suggère l'hypothèse d'un renforcement volontaire de l'aspect humain du phénomène migratoire. L'hypothèse est alors que le choix du pluriel indique la préférence à une représentation en tant que groupe au détriment de l'aspect individuel. En effet, lorsqu'on regarde le terme au féminin singulier, *immigrante*, 58 occurrences sont encore liées à *personne immigrante*, mais pour 6 occurrences les concordances montrent deux noms qui indiquent une pluralité : *population* et *famille*.

Concordances	Occurrences
famille immigrante	2
population immigrante	4
personne immigrante	47
clientèle immigrante	1
main d'œuvre immigrante	4

Tableau 7. Concordances du terme *immigrante* (fém./sing.) en tant qu'adjectif

Toujours à l'intérieur des concordances d'*immigrante* en tant qu'adjectif au féminin singulier, un hapax attire notre attention, en l'occurrence *clientèle immigrante* : « Les ser-

vices offerts à la *clientèle immigrante* visent à accélérer le processus d'intégration de ces personnes et à favoriser leur pleine participation, en français, à la société québécoise. » Cette occurrence apparait dans le troisième volet – celui consacré à la représentation de l'énonciateur. L'immigrant devient un client de la société québécoise qui lui offre des services pour son intégration. Il nous semble de pouvoir interpréter de la même façon les concordances de la *main-d'œuvre immigrante* (tableau 8).

Concordances (Volet III)
Voici quelques conseils pratiques pour optimiser l'intégration à votre entreprise de la *main-d'œuvre immigrante* […]
La ministre Diane De Courcy encourage les acteurs régionaux à promouvoir la région pour attirer la *main-d'œuvre immigrante* sur la Côte-Nord
Organiser un déjeuner-causerie sur l'importance pour les régions du Québec de faire appel à de la *main-d'œuvre immigrante*.
le MIFI souhaite offrir un accompagnement personnalisé en région […] pour leur permettre de se développer grâce au talent de la *main-d'œuvre immigrante* et des minorités ethnoculturelles.

Tableau 8. Concordances de l'expression *main d'œuvre immigrante*

Grâce aux concordances, il est possible de remarquer que le destinataire de ces portions de texte est toujours le Québécois et jamais le migrant : comme si la main-d'œuvre

immigrante était un phénomène que seuls les non immigrants peuvent comprendre. Un autre détail à souligner est que les occurrences de cette forme particulière se trouvent uniquement dans le troisième volet, celui consacré au ministère (tableau 9).

	Freq. Tot	I volet	II volet	III volet
immigrante_ ADJ	58	40	14	4

Tableau 9. Distribution des occurrences du mot *immigrante* (fém./sing) en tant qu'adjectif sur les trois volets du corpus

La dernière forme, *migrant* au masculin singulier, toujours pour la partie du discours d'adjectif, n'est pas très représentée. Parmi les 7 occurrences, nous relevons en trois occurrences l'emploi de la majuscule qui indique l'entreprise « Investissement Québec — Immigrants Investisseurs inc. » qui délivre des certificats. Les quatre occurrences restantes, bien qu'en nombre réduit, nous informent sur les catégories dans lesquelles les migrants sont classés, comme on le voit dans le tableau ci-dessous : élèves, étudiants, *professionnels*, *travailleurs.*

Demander à des jeunes élèves *immigrants* de partager ce qu'ils aiment le plus de leur culture d'origine…
compréhension écrite et en production écrite des étudiants *immigrants*…
faciliter la reconnaissance des compétences des *professionnels immigrants*…
compétences des *travailleurs immigrants* qualifiés…

Tableau 10. Concordances du terme *immigrants* (masc./plur.) en tant qu'adjectif.

L'analyse des formes qui caractérisent la partie du discours du nom (voir tableau 6) relève une totalité de 73 occurrences pour le mot *immigrants* au pluriel . Nous avons d'abord calculé les co-occurrences de cette forme (tableau 11).

Co-occurrent	Fréq.	Co-Fréq.	Index	Dist. moyenne
Présence	23	16	29	5,6
admis	76	17	20	0,4
intégration	419	26	16	3,2
linguistique	34	11	15	1,5
Programme	318	17	10	5,7
caractéristiques	7	5	9	1,0
ressources	21	6	8	0,2
portraits	21	6	6	5,3
établis	26	6	7	1,0

Tableau 11. Cooccurrences du terme *immigrant* (masc./sing.) en tant que nom.

Les co-occurrents qui apparaissent appartiennent au champ sémantique de l'évaluation où les compétences linguistiques semblent jouer un rôle fondamental dans l'intégration. Comme ce concept a été souvent associé aux migrants dans ce corpus, nous suivons cette piste et grâce à

l'analyse des concordances (tableau 12), nous sommes en mesure de confirmer cette hypothèse : la compétence linguistique devient explicite et se concrétise dans des cours de durée et de niveaux divers, et donne vie à un concept fondamental : l'intégration linguistique.

d'intégration linguistique pour les *immigrants*
cours de français adaptés aux besoins variés des *immigrants*
aide financière pour *l'intégration linguistique* pour les *immigrants*
les *cours de français* à la disposition des *immigrants*
Les services du Ministère proposent aux *immigrants* des *cours de durée et de niveaux divers*
Les *enfants d'immigrants*, quelle que soit leur langue maternelle, doivent normalement fréquenter un établissement de la commission *scolaire francophone*

Tableau 12. Concordances du terme *immigrant* (masc./sing.) en tant que nom.

Ce panorama initial nous a montré les limites du champ où l'énonciateur et le destinataire vont communiquer. La prochaine étape consistera à retracer les actes subis ou commis par l'immigrant.

Nous sommes en mesure de déterminer une CQL qui renvoie toutes les formes qui se réfèrent au migrant en tant qu'être humain et non en tant que concept : du groupe des termes qui en résultent, nous avons calculé les cooccurrences, avec la mesure statistique du Log- Likelihood

(Dunning, 1993) et une fenêtre de +/- 8 mots des deux côtés. Nous avons ensuite trié les résultats par parties du discours ; nous affichons les verbes dans le tableau qui suit (tableau 13).

En observant cette liste, se dégage une prosodie sémantique liée à une épreuve à surmonter, à un barème à atteindre, à un accès non totalement libre et sans contrepartie. Des verbes comme *réussir*, *admettre*, et des formes passives comme *acquis* et *recensés*, tiennent à un processus de sélection.

Occurrent	P. du discours	Fréq.	CoFréq.	Indice
réussie	VER :pper	16	16	24
perdre	VER :infi	6	6	9
admis	VER :pper	76	17	9
admises	VER :pper	14	8	8
liées	VER :pper	60	14	8
souligne	VER :pres	17	8	8
recensées	VER :pper	9	6	7
habitent	VER :pres	9	6	7
arrivées	VER :pper	18	7	6
habite	VER :pres	4	4	6
acquis	VER :pper	10	5	5

Tableau 13 : requête CQL [word = « immigr.* » & word != « immigration »%c & word != « immigrer »%c]. Résultats : Co-occurrents filtrés par verbes

Afin de valider notre hypothèse nous avons extrait les concordances dont nous affichons les résultats les plus fréquents dans le tableau qui suit (tableau 14) :

une **intégration *réussie*** des *personnes immigrantes*
immigrants ***admis* au Québec**
personnes immigrantes, ***admises* au Québec**
Il est de la responsabilité de la *personne immigrante* de connaître les **conditions *liées*** à son statut et de s'assurer de les respecter
résoudre/cerner les **difficultés** liées à la reconnaissance des compétences des *personnes immigrantes*
***souligne* l'apport exceptionnel** d'une *personne immigrante* au développement culturel et artistique du Québec sur la scène nationale ou internationale
les femmes immigrées recensées au Québec ***minorités visibles recensées au Québec***
Pour être sélectionné comme *immigrant* travailleur, vous devez avoir ***acquis*** une scolarité et des compétences professionnelles
les *personnes immigrantes* ayant acquis leurs compétences à l'étranger peuvent se heurter, pour obtenir un emploi dans une profession ou un métier réglementé au Québec

Tableau 14 : Concordances des co-occurrents de la CQL [word = « immigr.* » & word != « immigration »%c & word != « immigrer »%c].

Ce travail permet d'affirmer que l' immigration s'accomplit par le biais d'une intégration linguistique mesurée

par un niveau de compétences bien déterminé. Le migrant est accueilli, mais en contrepartie, il a des obligations à respecter pour ne pas perdre son statut d'immigrant et les conditions qui y sont liées.

2.1.4.1. Le « nous » communauté d'accueil et le « vous » communauté d'origine du migrant

L'analyse de la représentation que l'énonciateur fait de lui-même et du co-énonciateur dans une perspective de distribution discursive mérite aussi une grande attention. Grâce à la CQL = [mot = « nous »%c | mot = « vous » %c] nous avons pu calculer les cooccurrences des pronoms sujets à la première et à la deuxième personne du pluriel, « nous » et « vous ». Pour la visualisation, nous nous sommes servis du calcul des progressions disponibles dans TXM qui permet de visualiser la présence en termes de densité dans les trois volets à travers lesquels nous avons au préalable partitionné le corpus (tableau 3). La répartition est plutôt marquée selon les sections : le « vous » est très présent dans la première section « Immigrer au Québec », alors que dans la deuxième « Favoriser l'intégration », il subit une chute au profit du terme « nous » qui obtient sa fréquence la plus élevée dans la troisième section « Connaitre le Ministère ». Cette distribution nous suggère l'hypothèse d'un usage exclusif des deux pronoms sujet où il n'y a pas coprésence.

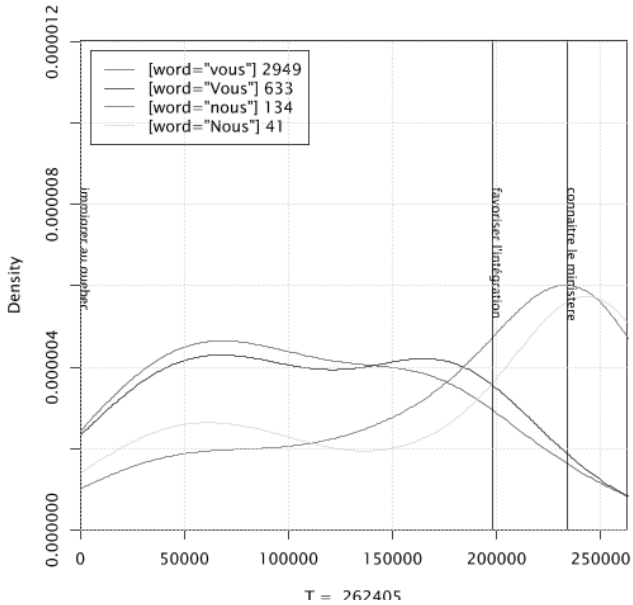

Image 1 : distribution « nous » et « vous » dans le corpus

Après ces considérations générales, procédons à l'analyse des cooccurrences des deux pronoms afin de retracer des prosodies discursives. Grâce à l'annotation par parties du discours, les résultats sont d'abord filtrés en retenant uniquement les verbes, qui sont ensuite classés suivant leur modalité et leur temps : le « nous » dans le tableau 15 et le « vous » dans le tableau 16. Les lignes ont été reportées partiellement en considérant que l'indice atteint son seuil de discrimination aussi selon le volume des occurrences totales : donc pour le terme « nous » qui apparait dans 175 occurrences, la limite de l'indice a été établie sur 5, alors

que pour le terme « vous » qui parait en 3582 occurrences la valeur se positionne sur 10.

Occurrent	POS	Fréq.	CoFréq.	Indice
invitons	VER :pres	17	17	33
traitions	VER :impf	10	10	19
sommes	VER :pres	24	12	17
communiquer	VER :infi	48	14	15
reçues	VER :pper	28	10	12
proposons	VER :pres	6	6	11
engageons	VER :pres	5	5	9
communiquant	VER :ppre	8	5	8
remplissant	VER :ppre	4	4	7
rapporter	VER :infi	4	4	7
rappelons	VER :pres	4	4	7
signaler	VER :infi	6	4	6
utiliser	VER :infi	59	8	6
consulter	VER :infi	64	8	6
utiliserons	VER :futu	3	3	5
soumettant	VER :ppre	3	3	5
recommandons	VER :pres	3	3	5
pouvons	VER :pres	3	3	5
mettons	VER :pres	3	3	5
jugeons	VER :pres	3	3	5
encourageons	VER :pres	3	3	5
approuvons	VER :pres	3	3	5

Tableau 15 : requête CQL [word= »nous »%c] Totale 175 occurrences – cooccurrences filtres par verbes

Le « nous » présente une prosodie sémantique positive qui tend à établir une relation et des règles qui servent à la maintenir. L'énonciateur se présente comme disponible et accueillant - *invitons*, *traitions* – tout en gardant un pouvoir juridique présent dans des termes comme *approuvons*, *jugeons*. Parallèlement, la représentation discursive montre un côté réconfortant visible à travers les termes *encourageons* et *recommandons*. Les verbes expositifs (Roulet, 1978) tels que *communiquer*, *signaler*, *rapporter*, sont très variés et font référence au rôle institutionnel du co-énonciateur. Le tableau 16 ci-dessous consacré au terme *vous* montre des résultats très différents.

Cooccurrences	Fréq.	CoFréq.	Indice
devez_VER :pres	429	419	274
avez_VER :pres	354	335	205
pouvez_VER :pres	214	203	125
êtes_VER :pres	170	168	113
devrez_VER :futu	144	133	78
souhaitez_VER :pres	96	91	56.3563
désirez_VER :pres	67	66	44.4382
pourrez_VER :futu	60	60	41.9708
faire_VER :infi	431	192	30.0919
transmettre_VER :infi	90	67	27.8579
recevrez_VER :futu	42	41	27.1484
aurez_VER :futu	49	45	26.5273
voulez_VER :pres	37	37	25.8792

serez_VER :futu	38	37	24.3933
aider_VER :infi	45	38	19.5795
parrainez_VER :pres	41	35	18.3895
renseigner_VER :infi	31	29	17.8018
établir_VER :infi	75	50	17.6156
remplir_VER :infi	107	62	17.1256
informer_VER :infi	60	42	16.105
trouverez_VER :futu	23	23	16.086
communiquera_VER :futu	22	22	15.3865
faites_VER :pres	37	30	14.6223
préparer_VER :infi	42	32	14.1476
adresser_VER :infi	30	26	14.1175
retourner_VER :infi	27	24	13.5966
pourriez_VER :cond	32	26	12.7834
effectuer_VER :infi	66	40	12.2012
invitons_VER :pres	17	17	11.8893
pourrait_VER :cond	43	30	11.6299
sera_VER :futu	160	71	11.5922
parrainer_VER :infi	53	34	11.4811
disposez_VER :pres	19	18	11.4011
aviser_VER :infi	21	19	11.1487
embaucher_VER :infi	55	34	10.818
réussir_VER :infi	58	35	10.6869

Tableau 16 : requête CQL [word= »vous »%c] Totale 3582 occurrences – cooccurrences filtres par verbes

Les deux formes verbales qui apparaissent très fréquemment sont *pouvoir* (349 occurrences) et *devoir* (573 occurrences) et suggèrent une représentation bipolaire où les actions du migrant oscillent. Cette hypothèse est renforcée par la présence des temps du futur *sera*, *trouverez*, *aurez*, *recevrez* qui ont une cohérence si on les relie à *souhaitez* et *désirez* : la prosodie discursive est celle d'une ouverture future de possibilités concrètes. Toutefois, cette disponibilité est soumise à des conditions que nous tirons de l'autre pôle – celui du devoir – et qui se concrétisent dans une série d'éléments, de conditions et d'attributs visibles à travers les actions *faites*, *préparer*, *effectuer*, *disposez*, *adresser*, établir, *remplir*. Le *vous* est donc un sujet actif qui *fait* et qui *pourrait avoir* : la dimension est prédictive, normative, dépourvue de tonalités émotionnelles, mais avec une forte charge prescriptive.

2.1.5. Conclusions

Si l'étude du discours de presse autour des migrants est fondamentale en raison de son pouvoir d'influence sur l'opinion publique, l'analyse d'autres typologies de discours public est tout aussi intéressante dans cette perspective car ces typologies contribuent à former l'identité du migrant. La communication ministérielle, par le biais du site web, se situe entre la description de la réalité et l'énumération de normes, d'instruction et de procédures à suivre, tout en esquissant le profil culturel du migrant, et

aussi le rapport entre le migrant (qui est le co-énonciateur) et la communauté d'accueil (qui est l'énonciateur).

Le lexique qui caractérise le discours public du Ministère n'est pas riche et est composé principalement de noms, suivis de verbes et d'adjectifs, éléments qui indiquent un style répétitif et essentiel.

La stratégie discursive mise en place par l'énonciateur est de souligner sa position prédominante dans le dialogue entre celui qui arrive et celui qui accueille. La scène d'énonciation est ainsi polarisée non parce que l'énonciateur est celui qui accueille mais parce qu'il a la nécessité de le souligner par les choix linguistiques des termes immigr.*.

Le lemme neutre *migration* et ses formes conjuguées ont une valeur géographique (*migrations* interprovinciales, *migrations* internationales, parcours *migratoire*) et événementielle (journée internationale des migrants). Au contraire, la forme *immigration* comporte une dimension conceptuelle beaucoup plus marquée : elle définit les contours du dialogue qui semble être un pacte entre la communauté d'accueil et les individus cherchant à intégrer cette communauté. Des structures syntactiques nous montrent des champs sémantiques définis : l'aspect normatif du processus d'immigration est mis en évidence, de même que la nécessité d'une réglementation administrative. L'immigration est par ailleurs également liée à une dimension dynamique (*processus*, *projet*, *programme*, *procédures*, *démarches*). Il y là aussi quelque chose de l'ordre du mesurable (*statistique*) et qui devient un champ dont on peut être experts et conseillers.

Les formes linguistiques apparaissant de manière prépondérante pour identifier celui qui migre sont au nombre de deux : *personnes immigrantes* et *immigrants*.

Si la forme *personnes immigrantes* vise à renforcer la dimension humaine du migrant, en même temps la fréquence des formes conjuguées au pluriel indique la préférence pour la représentation d'une collectivité et non d'un individu, même quand la forme linguistique est au singulier et se réfère à un nom commun (*famille immigrante*, *population immigrante*).

Le renforcement de ce concept indique un aplatissement de la diversité d'origine du migrant qui est défini ici uniquement par l'action d'arriver et non par sa communauté d'origine. La seule manière de différencier les immigrants est de le faire selon leur catégorie sociale : élèves, étudiants, professionnels, travailleurs.

Si cette tendance révèle un aspect positif, en supprimant des occasions de discrimination basées sur la provenance, de l'autre côté elle renforce la position de l'énonciateur qui semble être toujours le plus puissant dans cet accord qu'est le pacte migratoire.

La deuxième forme, *immigrant/s*, est associée au concept d'intégration linguistique qui se concrétise par des cours offerts, mais surtout par une participation active du migrant. Le migrant est décrit comme une ressource pour l'énonciateur qui le motive dans son parcours de reconstruction identitaire et d'acquisition d'identité. En effet, le migrant qui arrive avec une identité vide, ouverte, disposée aux changements doit acquérir sont statut à travers

une constante évaluation de ses compétences et de sa pleine participation à la vie de la communauté d'accueil. Le migrant est donc représenté comme un être actif et dynamique, dans la mesure où il remplit ses obligations, c'est-à-dire celles qui lui garantissent l'intégration dans la société québécoise.

La prosodie discursive est positive dans le sens où elle est liée à une épreuve à surmonter, à un barème à atteindre, à un accueil conditionné. Des verbes comme *réussir*, *admettre* et des formes verbales passives comme *acquis* et *recensés* tiennent à un processus de sélection. Il faut reconnaître également que l'immigrant est identifié à une collectivité, rarement à un individu avec des traits spécifiques, sauf à travers les fonctions qu'il occupe dans la communauté québécoise comme étudiant ou travailleur. Le trait universel qui le définit est le mouvement – de sa communauté d'origine vers sa communauté d'accueil. Ses activités ne sont jamais statiques, il est pris dans un processus de démarches, de demandes et d'activités ayant pour but l'intégration et dans lequel la compétence langagière – l'apprentissage de la langue française – est fondamentale. À la lumière de cette conception de l'immigration, l'immigrant se transforme en client auquel l'énonciateur offre ses services.

La narration est prise en charge par un énonciateur qui produit une bipartition très nette des pronoms le long des volets : le *vous* est très présent dans la première section « Immigrer au Québec », alors que dans la deuxième « Favoriser l'intégration » il subit une chute au profit du terme

nous qui atteint sa fréquence la plus élevée dans la troisième section « Connaitre le Ministère ». Les pronoms pluriels *nous* et *vous* ont un usage exclusif, leur coprésence est presque nulle : du moment que le terme choisi pour définir le co-énonciateur est *immigrant*, nous sommes invités à en déduire que lorsque le *nous* est employé, cela se réfère à la communauté d'accueil.

L'énonciateur se présente comme disponible, accueillant (*invitons*, *traitions*), et même réconfortant (visible par les termes *encourageons* et *recommandons*), tout en gardant un pouvoir normatif qui transparaît dans des mots comme *approuvons*, *jugeons*. Le *vous*, qui renvoie donc à la communauté d'arrivée, présentée en bloc et sans distinction, semble osciller entre deux pôles d'action, le *pouvoir* et *devoir*, et la forte présence d'une dimension future et conditionnelle associée au parcours du migrant, module la définition de cette communauté dans une perspective de projection, de parcours sans cesse redéfini, de catégorie fluide.

L'architecture du site web reflète une volonté de diviser la représentation de l'identité du co-énonciateur en deux moments temporellement distincts : le migrant est dans une condition *X* avant de partir et puis dans une condition *Y* une fois arrivé. Toutefois, le troisième passage manque, celui de l'intégration, bien que le mot « intégration » soit présent dans le sigle qui indique l'énonciateur. Dans ce cas d'étude il n'y a aucune trace de termes faisant référence à une nouvelle identité, ou à un nouveau statut, après avoir accompli le processus d'intégration : le migrant reste celui qui ne cesse de bouger, d'être en mouvement. La langue

garde la trace de cette asymétrie qui devient identitaire : d'abord purement linguistique, elle se fait ensuite conceptuelle.

2. 2. LE DISCOURS LITTERAIRE

2.2.1. Le Digital Literary Criticism

Nous avons déjà relevé que c'est à l'intérieur d'un des premiers livres sur la traduction automatique que paraissait le chapitre « The Computer in Literary Studies »[1]. Deux raisons peuvent expliquer cela : d'une part, le fait que l'une des premières applications des Humanités Numériques aux textes est effectivement celle de l'analyse textuelle qui constitue aussi une des bases de la critique littéraire ; d'autre part de manière plus générale- sans doute moins consciente- le fait que la critique littéraire représente un défi ultime pour les machines en raison des tâches d'interprétations qui lui sont propres. Tout ce qui relève de l'interprétation personnelle, de la sensibilité du critique littéraire et de sa capacité à établir des connexions entre le texte et le monde, connexions tissant les fondements de ses analyses, a toujours été lié à des aspects problématiques pour la gestion numérique et pour le *Digital Literary Criticism*. L'application de calculs quantitatifs à des techniques des humanités traditionnelles existe et a eu un certain succès. C'est le cas, par exemple le cas de la stylométrie[2] qui fournit une aide dans l'attribution de l'autorialité ou

---

1 Booth, A. D., *Machine Translation of Languages*, *op. cit*.

2 Tannery, P. « La stylométrie ses origines et son présent » *Revue Philosophique de La France et de l'Étranger*, 47, Presses Universitaires de France, Paris, pp. 159-69, 1899.

pour la reconnaissance d'un certain style. La stylométrie se base sur l'idée que chaque texte – ou ensemble de textes – présente un ensemble de termes typiques, surtout dans des proportions spécifiques qui peuvent aider à en reconnaitre la paternité. Et encore : il est possible de tracer des similarités dans les œuvres de plusieurs auteurs ; L'Image 1 en témoigne en affichant une similarité stylistique entre plusieurs auteurs du panorama français contemporain. Pour une classification de cette sorte, l'analyse quantitative est une ressource puissante qui ne laisse pas l'ombre d'un doute. Néanmoins, l'application de l'informatique à la critique littéraire – quand celle-ci touche à l'interprétation – rencontre des réticences.

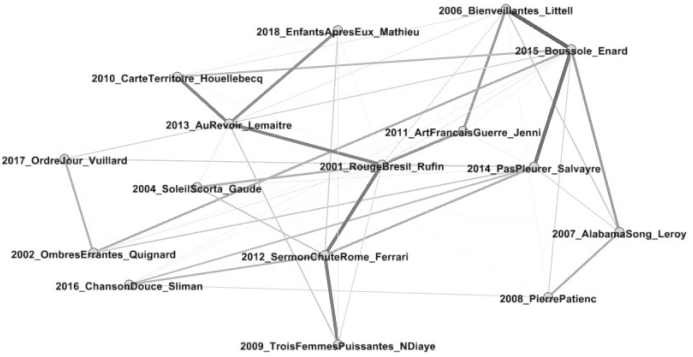

Image 1. Similarité stylistique calculée avec « Stylo » réalisé en langage R.[3]

[3] Eder, M., Rybicki J., Kestemont M.,« Stylometry with R : A Package for Computational Text Analysis », *The R Journal*, 8(1), 2016.

Les positions défavorables à l'application de l'informatique à la critique littéraire se basent en fait sur l'impossibilité pour une machine de posséder des capacités interprétatives, de simuler la sensibilité typique de l'être humain : une impossibilité, donc, non seulement à mettre ensemble toutes les informations de contexte liées à une pratique interprétative, mais aussi à ajouter un trait personnel, intuitif en quelque sorte, et dont le résultat est une interprétation critique littéraire. À ce sujet, David Hoover affirme que « computer-assisted textual analysis has a long, rich story, despite the fact that, as has often been noted, it has not been widely adopted in contemporary literary studies »[4]. Le débat est en cours et illustre des positions variées. Ainsi, certains voient avec enthousiasme dans cette évolution un virage vers les sciences dures. C'est le cas de Matthew Jockers qui affirme, dans le chapitre intitulé « Revolution »[5], que : « Now, slowly and surely, the same elements that have had such an impact on the sciences are revolutionizing the way that research in the humanities get done »[6], étant convaincu que la méthodologie littéraire n'est pas "in essence"[7] différente de la méthodologie scientifique. Jockers a développé une méthode en mesure de détecter, à l'intérieur d'un roman la

---

4 Hoover, D.L., « Textual Analysis », in Siemens R., Price K. (dir.), *Literary Studies in the Digital Age : An Evolving Anthology*, Modern Language Association, New York, 2013.

5 Jockers, M.L., *Macroanalysis : Digital Methods and Literary History*, University of Illinois Presses, 2013.

6 *Ibidem*, p. 10.

7 *Ibidem*, p. 13.

courbe émotionnelle, grâce à l'usage d'un lexique annoté selon la polarisation émotive des mots. Cette application a été développée en suivant l'intuition selon laquelle la forme de l'intrigue d'une œuvre de fiction peut être déduite aussi en analysant les courbes des sentiments (Image 2). Or, cette application s'inspire à une intuition qui vient de la littérature : en fait Jockers aurait suivi une affirmation de Kurt Vonnegut : « The fundamental idea is that stories have shapes which can be draw on graph paper, and that the shape of a given society's stories is at least as interesting as the shape of its pots or spearheads »[8]. Si donc l'intuition dérive de la critique littéraire la réalisation se fait avec des méthodes informatiques qui travaillent sur des données en grande quantité.

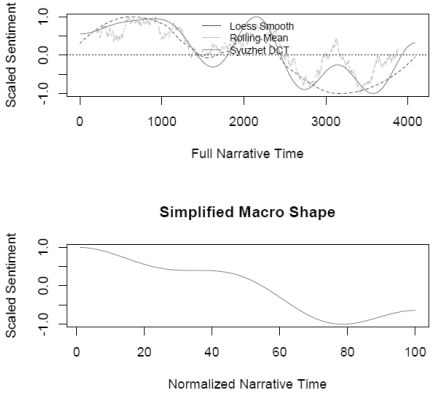

Image 2. Courbe émotionnelle du roman *La carte et le territoire* de Michel Houellebecq

[8] Vonnegut, K., A men without a Country, Seven Stories Press, New York, 2005.

D'autres affirment que certaines questions ne peuvent être traitées avec les mêmes méthodes dans les sciences humaines et dans les sciences naturelles, comme la physique ou la biologie. C'est le cas de Stephen Ramsay, qui, dans *Reading Machines*[9] nous assure que, si certains problèmes dans les sciences humaines, comme l'identification des auteurs, peuvent véritablement s'appuyer avec bonheur sur les méthodes développées par les sciences naturelles, pour la plupart des entreprises de critique littéraire – comme la caractérisation de la subjectivité de Virginia Wolf dans son roman *The Waves*, par exemple – il n'est pas possible d'identifier de manière claire, automatique, un ensemble de faits qui soient analysables. Encore Jerome McGann, dans son livre *Radiant Textuality*, paru en 2001, affirme :

> Digital technology used by humanities scholars has focused almost exclusively on methods of sorting, accessing and disseminating large bodies of materials, and on certains specialized problems in computational stylistics and linguistics. In this respect the work rarely engages those questions about interpretation and self-aware reflection that are the central concerns for most humanities scholars and educators […] the general field of humanities education and scholarship will not take the use of digital technology seriously until one demonstrates how its tools improve the ways we explore and explain aesthetic works – until, that is, the expand our interpretational procedures.[10]

---

9 Ramsay, S., *Reading Machines : Toward an Algorithmic Criticism*, University of Illinois Presses, 2011.

10 McGann, J., *Radiant Textuality : Literature after the World Wide Web*,

Entre ces deux extrêmes, de nombreux chercheurs fournissent des panoramas convaincants de ce que permet la numérisation, en traitant de l'évolution actuelle des Humanités Numériques et en particulier des études littéraires[11], en offrant des points de vue variés et enrichissants sur ces sujets. Au cours des dernières années, un déclic a permis l'avancement de ce débat, ou au moins sa transposition sur un niveau inédit : il s'agit du paradigme du *distant reading* théorisé par Franco Moretti.

2.2.2. Le changement du paradigme : le distant reading

Pour que les nouvelles méthodes de recherche soient pleinement utilisées, il faut un changement de paradigme dans la structure mentale de la communauté scientifique chargée d'observer et d'interpréter les données. Thomas Khun[12] parle d'un changement de paradigme nécessaire dans la pensée chaque fois qu'une révolution scientifique se produit. Galileo et Newton, parmi d'autres exemples cités par l'auteur, ont pu exister grâce à une révolution de la pensée et ont contribué à leur tour au changement de perspective dans l'interprétation des données. Dans la na-

Palgrave, 2001.

[11] Schreibman, S., Siemens, R., Unsworth, J. (dir.) *A Companion to Digital Humanities*, Malden, MA/Oxford/Carlton, Blackwell Publishing, Victoria, pp. 254-270, 2004 ; Siemens R., Price K. (dir.), *Literary Studies in the Digital Age : An Evolving Anthology*, Modern Language Association, New York, 2013.

[12] Khun, T. S., *Structure of Scientific Revolution*, *op. cit.*

ture cyclique des révolutions scientifiques que préconise Khun, il existe une rupture avec le paradigme précédent du fait que les données à la disposition de la communauté scientifique ne correspondent plus à la théorie prédominante qui les explique et à partir de laquelle la communauté scientifique élabore, avec créativité et par le biais de nouvelles technologies, une nouvelle théorie de référence. Une analogie est suggérée avec l'introduction des méthodes et des outils TAL dans la critique littéraire : les nouvelles méthodes et données résultant de l'introduction de nouvelles techniques et instruments de recherche doivent être lues et interprétées à la condition d'une modification du paradigme de la pensée. Cette réflexion introduit la « révolution » soulevée par Franco Moretti avec son paradigme du *distant reading* paru pour la première fois en 2000[13] ensuite élaborée et diffusée dans sa riche activité scientifique. [14] Le *distant reading* se veut complémentaire – et non alternatif – à l'école prédominante du *close reading* qui représente la tradition de l'approche à la critique littéraire. L'expression *close reading*, apparue pour la première fois dans le texte *Critical Practicism*[15] et ensuite reprise et formalisée par l'école du New Criticism, a comme condition une lec-

[13] Moretti, F., « Conjectures on world literature », *New left review* 1, London. URL = https ://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature, 2000. [dernier accès : 22/08/2021].

[14] Moretti, F., *Graphs, Maps, Trees : Abstract Models for a Literary History*, Verso Books, London, 2007.

[15] Richard, I.A., *Practical Criticism*, K. Paul, Trench, Trubner Company, London, 1929.

ture rapprochée, à la recherche de formes et d'expressions évocatrices, identifiées par la sensibilité du critique qui, dans ce cadre devient le pivot fondamental, porteur de connaissances et capable d'opérer des connexions sémantiques. Au contraire, la proposition que Franco Moretti avance avec le *distant reading* s'inspire de l'hypothèse selon laquelle il existe dans la littérature des mouvements, des structures, des caractéristiques et des formes qui ne sont visibles qu'à partir d'une certaine distance du texte, et qui ne sont visibles que si on analyse les textes en les regardant de loin. La distance est une condition de connaissance :

> where distance, let me repeat it, is a condition of knowledge : it allows you to focus on units that are much smaller or much larger than the text : devices, themes, tropes, – or genres and system. And if, between the very small and the very large, the text itself disappears, well, it is one if those cases when one can say, Less is more.[16]

Moretti, à travers un mouvement d'abstraction et de réduction, analyse la littérature à l'aide de méthodes importées des sciences naturelles : les graphes dérivés de l'histoire quantitative, les cartes prises de la géographie et enfin les arbres s'inspirant de la théorie de l'évolution. Ces graphes donnent la mesure des mouvements littéraires à grande échelle : le roman anglais du XIXème siècle en tant que miroir des changements de la société est déduit de la courbe marquée par son succès éditorial ; ou encore le

[16] Moretti F. « Conjectures on world literature », *op. cit*.

changement de la valeur sociale du tissu urbain est visible grâce à la reproduction de l'espace de vie des personnages d'une œuvre. De plus, l'utilisation de méthodes importées des sciences exactes donne la mesure d'un changement profond de perspective. Si, au niveau méthodologique, ce changement de perspective autorise le mélange de deux modèles traditionnellement distants, il implique, au niveau du contenu, la création d'un cadre pour interpréter les nouvelles données. Le résultat du traitement des données conduit à des formes et à des motifs qui ne sont pas indépendants de la méthode appliquée : seule la distance permet l'identification de certaines caractéristiques qui ne sont pas en contraste avec les hypothèses du *close reading*.

Le succès de ce paradigme a suscité une ample littérature et a enrichi le panorama du *distant reading* : Janinski parle de la « construction of abstract models » [17] ou Jockers de « macroanalytic approach »[18] ou encore Drucker parle de « the idea of processing content in or information about a large number of textual items without engaging in the reading of the actual text ».[19] La grande quantité de données disponibles aujourd'hui est l'environnement naturel

---

[17] Jasinski, J., *Sourcebook on Rhetoric : Key Concepts in Contemporary Rhetorical Studies*, SAGE Publications, London, 2001.

[18] Jockers, M.L., « On Distant Reading and Macroanalysis », 2011 https ://www.matthewjockers.net/2011/07/01/on-distant-reading-and-macroanalysis/ [dernier accès : 12/10/2021].

[19] Drucker, J., *Distant Reading and Cultural Analytics. Intro to Digital Humanities*, 2013, http ://dh101.humanities.ucla.edu/?page_id=62 [dernier accès 12/10/2021].

du *distant reading* et des les méthodes TAL qui se prêtent parfaitement à être utilisées dans ce cadre théorique.

Un aspect important du *distant reading* est la visualisation des données ou des résultats : en effet, les méthodes choisies (cartes, graphes ou réseaux) ont une fonction profondément explicative qui permet d'identifier les *patterns*. Selon Edward Tufte chaque visualisation de données doit posséder les caractéristiques suivantes :

1) show the data 2) induce to think about the substance 3) avoid distortion 4) present many data in a small space 5) make large datasets coherent 6) encourage comparisons 7) have several levels of details 8) serve a clear purpose 9) be closely integrated with the statistical and descriptive part.[20]

Parmi les présupposés à la base du *distant reading*, il y a aussi le fait que la littérature ait des connexions profondes avec certains aspects de la société. Moretti montre comment par exemple une représentation du changement de l'espace social des villes est représenté dans la fiction, non parce qu' elles sont décrites ouvertement, mais grâce à la reconstruction des mouvements et des vies des personnages. Notre intérêt s'est focalisé sur la manière dont le rapport entre les cultures est mis en scène au sein du discours littéraire. Quelles sont les structures syntactiques qui composent les discours ? Quels sont les mots choisis et

[20] Tufte, E., *The Visual Display of Quantitative Information*, Hoepli, 2000 (1983).

quels imaginaires véhiculent-ils pour définir et décrire une culture ?

2.2.3. Présentation du corpus : Le prix Goncourt

La littérature a toujours eu un rapport profond avec la société en termes de représentativité : entre le début du XIXème siècle et le XXème siècle – la grande époque des éditeurs – la littérature était parmi les formes privilégiées de divertissement culturel sous une forme écrite : elle était porteuse des modes de la société, des courantes de pensée ou encore des débats actuels. Aujourd'hui avec la révolution numérique le monde a changé : le divertissement culturel passe aussi par d'autres formes de narration et la littérature a dû partager avec d'autres son rôle prédominant, sans toutefois perdre sa particularité. À la suite de ces réflexions, nous avons sélectionné notre corpus à partir de romans ayant remporté le Prix littéraire Goncourt. La représentativité d'un corpus ainsi collecté est garantie par l'importance que le Prix Goncourt revêt à l'intérieur de la société française : en effet, ce prix littéraire ne se limite pas évaluer la qualité littéraire du texte, mais il prend en compte également les goûts du public, les événements nationaux et internationaux qui touchent la société, les questions qui animent les débats publics. De plus, le Prix Goncourt peut être considéré intéressant à deux autres titres : le règlement prévoit que le prix ne peut être attribué qu'une seule fois dans la vie d'un écrivain et que celui doit être francophone. Ces deux caractéristiques

garantissent une proportion équilibrée et donc une ample variété du panorama littéraire contemporain. Nous nous sommes concertés sur une période qui va de 2000 à 2018. Ce plan établi, nous avons collecté et élaboré les textes et nous avons construit un corpus qui possède les dimensions suivantes (table 1) :

Nombre de mots	2187140
Lemmas	34186

Table 1. Taille et composition du Corpus Goncourt

Le corpus a été tokenisé et prétraité : toutes les métadonnées présentes dans les éditions comme les informations éditoriales ou les références à d'autres ouvrages de l'auteur ont été supprimées. Ensuite le corpus a été étiqueté automatiquement par parties du discours avec l'outil Treetagger[21]

Etiquettes part of speech	Occurrences
NOM	364858
ADJ	106425
VER :impf	86224
VER :pres	77967
VER :pper	61355
VER :infi	57586
NAM	52905
VER :impf	86224

[21] Schmid, H., *op. cit*.

VER :pres	77967
VER :pper	61355
VER :simp	34751
VER :ppre	9127
VER :cond	8123
VER :futu	4085
VER :subp	2506
VER :ppre	9127
VER :cond	8123
VER :subi	1250
VER :impe	26

Table 2. Etiquetage par parties du discours du Corpus Goncourt

2.2.4. Analyses et résultats

2.2.4.1. La distribution géographique de la fiction et les structures syntactiques

Parmi les techniques TAL, le premier outil que nous avons utilisé pour mener notre recherche est le NER, acronyme de *Named Entity Recongnition* qui nous permet d'extraire automatiquement du texte les entités géographiques et les entités nommées (êtres humains et, dans certains cas, noms d'institutions publiques). Nous avons procédé à cette extraction[22] que nous avons pu visualiser avec le logi-

[22] L'extraction a été faite par le biais d'un script Python avec la librairie Spacy.

ciel Recogito.[23] La carte ci-dessous (Image 3) montre cette extraction en distinguant au niveau visuel les occurrences relatives au pays ou aux villes et endroits spécifiques. Nous observons comment la distribution entre les pays et les villes est inégale. L'Europe est citée de manière ponctuelle, spécifique : les villes et les endroits sont détaillés. L'Amérique, la Russie et l'Afrique ne bénéficient pas du même type de description. Le discours littéraire montre une prédilection pour l'Europe, en particulier pour la France. Les points correspondent principalement aux villes, parfois à des zones géographiques et cela montre au moins une connaissance du territoire approfondie et une volonté de construire une narration qui fournit ce genre de détail.

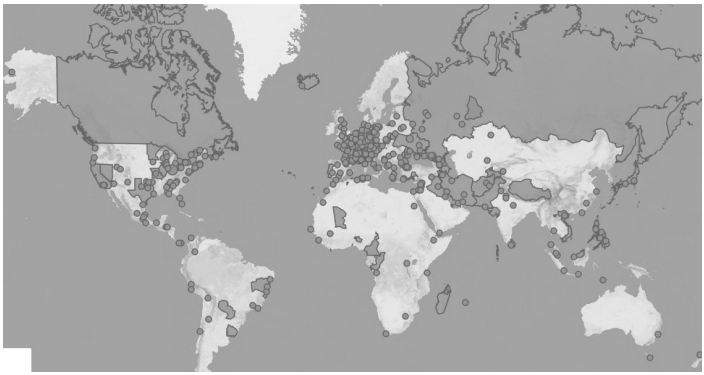

Image 3 – Représentation sur carte des entités géographiques du Corpus Goncourt réalisée avec Recogito.

Ce premier survol nous donne aussi la possibilité de

[23] https ://recogito.pelagios.org/help/about.

nous concentrer sur une liste de pays paraissant un nombre élevé de fois comme nous pouvons l'observer dans le tableau ci-dessous (table 3).

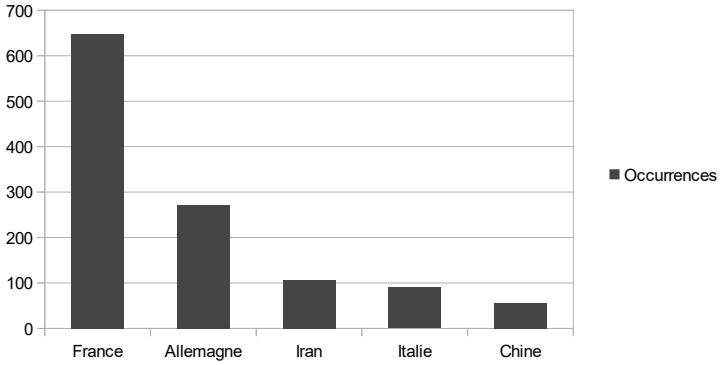

Table 3 – Présences des pays dans le corpus Goncourt.

Notre manière de procéder est guidée par l'hypothèse selon laquelle autour de la nomination des pays il peut y avoir des informations latentes, que nous voulons enquêter au niveau syntactique. Être le sujet ou l'objet d'une action, ou être associé à des adjectifs, sont autant d'éléments qui peuvent signifier qu'une entité géographique est associée à un imaginaire. Ce type de réflexion est autorisée, bien évidemment, à l'intérieur d'une grande quantité de données ne tenant pas compte de la particularité d'un style : plus on s'éloigne, plus nous sommes en mesure de voir des patterns linguistiques. Pour vérifier cette hypothèse, l'analyse syntactique des textes par le biais des Dépendances

Universelles[24] se révèle très utile. Comme nous l'avons déjà envisagé dans la première partie, les UD (Dépendances Universelles) s'inspirent de la grammaire valencielle de Lucien Tesnière qui permet d'attribuer à chaque terme une relation avec un autre terme en raison de sa capacité d'attraction syntactique. L'analyse automatique permet de créer une structure arborescente – chaque mot a, de cette manière une relation directe avec un autre. L'élaboration automatique s'est faite grâce à un script réalisé avec Python, ensuite nous avons traité le corpus et nous avons obtenu une version annotée que nous avons analysée semi-automatiquement. Pour trier les résultats et procéder à une lecture approfondie, nous nous sommes servis du logiciel de textométrie TXM et du dictionnaire Dictionnaire Electronique des Synonymes (DES)[25] réalisé par le CRISCO (Centre de recherche interlangues sur la signification en contexte)[26]. Nous présentons plus loin les résultats obtenus pour les cas les plus intéressants.

La première observation que nous pouvons formuler est que la présence du terme « France » possède un écart considérable par rapport aux autres pays. Cela est motivé probablement par une naturelle inclination à parler d'un pays qu'on connait bien (un pays où on habite, probablement) de la part des énonciateurs du discours littéraire.

---

[24] de Marneffe M.C, Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., Manning C., « Universal Stanford Dependencies : A cross-linguistic typology », LREC 2014.

[25] https ://crisco2.unicaen.fr/des/.

[26] http ://crisco.unicaen.fr/.

Nous nous sommes concentrés d'abord sur le cas de la France en tant que sujet ou objet, en analysant les verbes et ce afin d'identifier des regroupements sémantiques motivés par des thématiques. Une telle démarche suppose qu'on s'attarde sur l'idée de dégager un sens dans une approche globale, c'est-à-dire que l'analyse vise à trouver un sens à l'ensemble. Le tableau qui suit (tableau 4) montre l'ensemble des mots ayant un rapport direct avec le terme « France » ; en particulier ces mots ont été classés, grâce à l'étiquetage des Dépendances Universelles, en deux groupes : les verbes ayant la France comme sujet et les verbes ayant la France comme complément d'objet. Nous précisons ici que la fréquence de chaque terme n'est pas affichée car cette fréquence n'est pas pertinente dans notre analyse puisque nous nous intéressons aux relations et non pas à la distribution.

Verbes qui ont « France » comme sujet	Verbes qui ont « France » comme objet
'attendre'	'aller'
'bouillonner'	'appeler'
'céder'	'concerner'
'commencer'	'constituer'
'connaitre'	'construire'
'contenir'	'couper'
'croire'	'crier'
'disparaitre'	'demander'
'dire'	'enflammer'
'dresser'	'faire'

'engager'	'garder'
'ériger'	'incarner'
'exaspérer'	'mécontenter'
'faire'	'mentionner'
'frapper'	'organiser'
'offrir'	'parcourir'
'participer'	'peupler'
'produire'	'pousser'
'reconstruire'	'priver'
'redevenir'	'quitter
'replier'	'représenter'
'résister'	'ravager'
'savoir'	'rédiger'
'sauver'	'réformer'
'être'	'regagner'
'sortir'	'rencontrait'
'tenir'	'revoir'
'valoir'	être
'vivre'	'sillonner'
	'surnomma'
	'traitera'
	'traverser'
	'viser'
	'vivre'
	'voir'

Tableau 4. Verbes qui ont « France » comme objet et sujet.

La première observation que nous sommes en mesure de faire est qu'il existe un groupe majoritaire de verbes orienté vers une dimension factuelle, visant une idée de construction, une forme d'engagement, active ou proactive en quelque sorte reliée à une représentativité. La liste des termes comme : « commencer », « dresser », « engager », « ériger », « participer », « produire », « reconstruire » montrent un trait factif, que l'on retrouve aussi dans le cas de France en tant qu'objet avec les termes : « constituer », « construire », « organiser », mais dans des proportions moindres. En particulier, au sein de cette idée de construction nous trouvons une tendance à la résistance ou à un renouvellement : « reconstruire », « redevenir » « résister », ou « reformer », « regagner ». Un autre groupe sémantique peut être relevé dans la sphère sémantique du *dire* rendu visible par les termes : « appeler », « crier », « demander », « dire » ; cette présence est prépondérante dans le cas de France en tant qu'objet. Mais encore : nous observons une sphère sémantique du sacrifice et de la privation : « priver », « quitter », « ravager », « frapper ». Ceci associé à un rôle résistant de la France pourrait rendre l'idée de force. Davantage : une sphère sémantique liée à la géographie est également prise en considération : « traverser », « sillonner », « replier », « aller ».

Quant à l'Allemagne, notre intérêt se penche sur les noms qui y sont reliés en tant que modificateurs[27] fournis-

[27] La définition UD pour la catégorie nmod est la suivante : « The nmod relation is used for nominal dependents of another noun or noun phrase

sant une autre typologie d'information : il ne s'agit pas là du rôle de l'Allemagne dans la représentation de la narration, mais d'attributs. En analysant le tableau qui suit (tableau 5) nous remarquons comment une sphère sémantique est liée au combat : « alliés », « arme », « cargaisons », « évacuation », « réarmement », « détenus », « victoire », « guerre ». L'idée selon laquelle l'Allemagne est liée aux deux dernières guerres mondiales contribue probablement à l'enracinement d'un tel imaginaire. Une autre sphère que nous avons pu relever est celle caractérisée par un trait d'humanité : « asociaux », « gens », « mère », « personne », « vieillard », sphère qui se relie à l'expression des sentiments ou d'actions chez des humains : « confiance », « discussion ». Nous observons également la présence d'une sphère qui tient à l'avenir, à l'incertitude : « avenir », « destin », « certaine », « destination ».

Noms qui sont des modificateurs de « Allemagne »
'alliés'
'ans'
'arme'
'asociaux'
'avenir'
'cargaisons'
'certaine'

and functionally corresponds to an attribute, or genitive complement » source : https ://universaldependencies.org/u/dep/nmod.html.

‘clef’
‘collection’
‘comédie’
‘conditions’
‘confiance’
‘côtés’
‘danger’
‘défilé’
‘destin’
‘destination’
‘détenus’
‘discussion’
‘effort’
‘études’
‘évacuation’
‘foins’
‘fois’
‘gens’
‘guerre’
‘haine’
‘Heine’
‘heure’
‘libération’
‘littérature’
‘médical’

'mère'
'milieux'
'partis'
'partition'
'pays'
'Personne'
'photos'
'politique'
'poste'
'produit'
'réarmement'
'relations'
'retour'
'route'
'rues'
'sauveur'
'tort'
'vacances'
'vermine'
'victoire'
'vieillard'
'voyage'
'vraie'

Tableau 5. Noms qui sont des modificateurs de « Allemagne »

L'Iran – qui est le premier pays en termes de présence après Allemagne et France – montre un tableau verbal typique, comme nous voyons de l'image ci-dessus (fig. 2).

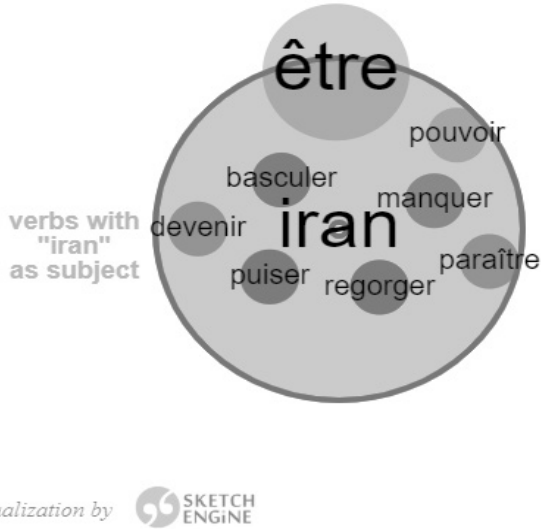

Image 4. Verbes qui ont « Iran » comme sujet.

L'Iran est le sujet d'actions qui semblent être opposées : « puiser », « regorger » ou « paraitre » et « devenir » qui se complète avec « basculer ». La sphère sémantique est caractérisée par une idée de mouvements entre des extrêmes ; on affiche les modificateurs de la catégorie des noms pour compléter cette hypothèse (tableau 6). Il en sort un niveau émotionnel très chargé : « charmes », « dou-

ceur », « mélancolie » sont des termes qui, pris ensemble, donnent un imaginaire poétique. On ajoute d'autres termes comme : « éther », « exil », « odeur » qui renforcent cet imaginaire. Le résultat nous semble une sphère sémantique très orientée par un aspect émotif, intérieur et moins par une dimension concrète et réelle comme pour l'Allemagne et la France. Bien que la présence du mot « réalité » semble contredire ce que nous venons d'affirmer, nous remarquons que ce mot n'est pas accompagné par d'autres qui peuvent compléter une sphère sémantique. Un autre groupe sémantique identifiable est lié à la tradition : « 'patronyme' », « peuple », « tradition », « union » sont des termes qui renvoient à un ensemble partagé de normes et de connaissance, à une culture commune.

NOUN – nmod de « Iran »
'avril'
'avenir'
'cadeau'
'charmes'
'communistes'
'douceur'
'éther'
'exil'
'grandeur'
'guerre'
'liberté'

‘magnifique’
‘mélancolie’
‘milieu’
‘odeur’
‘Ouest’
‘patronyme’
‘peuple’
‘poste’
‘profonde’
‘réalité’
‘recherche’
‘retour’
‘retrouve’
‘roses’
‘séjour’
‘sommet’
‘tradition’
‘union’

Tableau 6. Noms modificateurs de « Iran ».

Poursuivons avec l’exemple d’ « Italie » qui montre un champ sémantique spatiale : « arrivée », « errance », « carte », « campagnes », « plaines », « nord », « quarts », « centre » et aussi lié au temps : « années », « jours », « fois », « souvenir ». Soulignons les termes « faiblesses » et « convalescence », « urgence » qui suggèrent un état d’es-

soufflement. Les deux termes « conversation » et « amoureux » peuvent faire penser à un imaginaire latent qui renvoie à une dimension de relations humaines forte souvent attribuée à l'Italie.

NOUN – nmod de « Italie »
'ambassadeur'
'amoureux'
'années'
'armes'
'arrivée'
'campagnes'
'carte'
'centre'
'convalescence'
'conversations'
'errances'
'extraits'
'faiblesses'
'fois'
'guerres'
'jour'
'nord'
'partie'
'parution'
'père'

‘plaines’
‘quarts’
‘revers’
‘souvenir’
‘teintés’
‘urgence’

Tableau 7. Noms modificateurs de « Italie ».

Les noms associés à la Chine montrent des traits stéréotypiques assez marqués. Il ne s’agit pas d’une orientation négative, plutôt neutre, toutefois elle est composée par des termes qui sont souvent associés à l’Orient : « opium », « porcelaine », « encre », « vases ». Contrairement à l’Iran la Chine semble être représentée par ce qu’elle produit et non par des sentiments ou d’émotions.

NOUN – nmod de « Chine »
‘carte’
‘départ’
‘empereurs’
‘encres’
‘Est’
‘fuite’
‘goût’
‘lettrés’
‘mer’

'mouchette'
'numéro'
'opium'
'porcelaine'
'proximité'
'vases'
'vie'

Tableau 8. Noms modificateurs de « Chine ».

2.2.4.2. Le lexique des mots étrangères

Le lexique joue un rôle fondamental dans le discours interculturel. Le fait d'employer des mots provenant d'une langue autre a une fonction culturelle. Ainsi comme le cas du lexique de l'interculturalité de l'UNESCO (voir 1.3.2. Le dialogue interculturel : l'importance de la parole) où l'on introduit des mots de la culture africaine ou japonaise pour faire passer un concept qui est lié à l'autre peut se passer aussi en littérature, dans le discours littéraire le mot étranger n'as jamais uniquement un sens dénotatif, mais il s'enrichi avec une dimension connotative. Nous avons pu vérifier cette hypothèse grâce au calcul des mots-clés du Corpus en comparant le lexique du corpus Goncourt avec le lexique d'un autre corpus de la même typologie – le Corpus FranText[28] dédié à la littérature française entre le XI et

[28] Développée par l'ATILF (Analyse et Traitement Informatique de la

le XX siècle. Notre intérêt se dirige vers le comportement des mots étrangers visibles grâce à une analyse systématique des contextes immédiats de parution – c'est-à-dire les concordances. Nous avons remarqué que les mots étrangers sont souvent utilisés –les personnages mis à part – aussi pour des objets alimentaires (boissons et nourritures) et pour les vêtements. Autre à ces éléments parfois le « mot de l'autre » est utilisé au-delà d'une exigence dénotative, mais avec une valeur connotative. Les mots allemands sont employés pour indiquer des rôles militaires dans une hiérarchie bien définie : bien que les termes en question soient traduisibles dans d'autres langues, le fait de les utiliser de cette manière signifierait que le lexique aide à bâtir l'imaginaire d'une entité culturelle : dans ce cas le binôme guerre/Allemagne serait renforcé.

2.2.5.Conclusions

Le discours interculturel dans la littérature est considéré comme la manière – et les techniques – pour représenter les autres cultures ou pays. En sachant que le locuteur du discours littéraire est la France (en raison de la spécificité du Corpus Goncourt qui a été collecté) il est possible définir la représentation que elle fait de soi-même et des autres, qui sont les pays qui suscitent le majeur intérêt et comment ils sont racontés.

Une interprétation en perspective *distant reading* ne

Langue Française) https ://www.frantext.fr/.

donne pas des détails, et il serait possible de l'élargir à l'infini avec l'ajout d'autres données par exemple historiques, démographiques ou de croissance économique.

Nos conclusions sont que d'abord en littérature est visible l'axe franco – allemand qui est caractérisé depuis vingt ans l'Europe. Ensuite nous pouvons affirmer qu'il y a une correspondance entre la manière de parler et la provenance de l'énonciateur du discours littéraires : une connaissance majeure des endroits pousse à les nommer de manière plus ponctuelle. Ensuite des possibles regroupements sémantiques informent sur le rôle qui tiennent les pays mais aussi sur leurs attributs. La France construit, est active, l'Allemagne porte mémoire de son histoire et de sa dimension humaniste. L'Iran est représenté comme poétique et de l'Italie sont mises en valeurs ses caractéristiques spatiales et esthétiques. La présence des mots étrangères n'a pas une caractère purement dénotatif mais se charge d'une valeur connotative qui contribue à bâtir un certain imaginaire.

Bibliographie

Aa. Vv. ALPAC, « Languages and Machines : Computers in Translation and Linguistics », A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC : National Academy of Sciences, National Research Council, 1966 ; http ://www.nap. edu/books/ARC000005/html ;

Aa.Vv. (IBM), *Introduction to Computers in the Humanities*, White Plains, New York : IBM, 1971 ;

Aa.Vv. (IBM), *Literary Data Processing*, White Plains, New York : IBM, 1971 ;

Aa.Vv. (IBM), *Computers in Anthropology and Archeology*, White Plains, New York : IBM, 1971 ;

Aa.Vv., *Compétences interculturelles, Cadre conceptuel et opérationnel*, Organisation des Nations Unies pour l'éducation, la science et la culture, Paris, 2013 ;

Abdessadek, M., « Identité et migration : le modèle des orientations identitaires », *L'Autre* 3 (13) : pp. 306-317, 2012 ;

Al Fajri, M. S., « Hegemonic and minority discourse around immigrants : a corpus-based critical discourse analysis », *Indonesian Journal of Applied Linguistics* 7(2) : 381-390, 2017 ;

Alcaraz-Mármol, G., Soto Almela J., « The semantic prosody of the words inmigración and inmigrante in the Spanish written media : A corpus-based study of two national newspapers », *Revista Signos Estudios de Linguistica*, 49(91) pp. 145-167, 2016. DOI : 10.4067/S0718-09342016000200001 ;

Alvarado, R., « The Digital Humanities Situation », *The Transducer*, 2011 [dernier accès http ://transducer.ontoligent.com/?p=717] ;

Anctil D., Tremblay O., « Les collocations : des combinaisons de mots privilégiées », dans *Correspondance* Vol.21, n° 3, CCDMD, 2016 ;

Austin, J. L., *How to do things with words*, Clarendon Press, Oxford, 1962 ;

Baker, P., Gabrielatos, C., Khosravinik, M., McEnery, T., Wodak, R., « A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press », *Discourse and Society,* 19(3) : pp. 273-306, 2008 ;

Baker, P., Gabrielatos, C., McEnery, T., *Discourse analysis and media attitudes : the representation of Islam in the British press*, Cambridge University Press, Cambridge, 2013 ;

Baker, P., *Using Corpora in Discourse Analysis*. Continuum, New-York / London, 2006 ;

Barbera, M., *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione*. ASAR, Milano, 2013 ;

Berry, D.M. (dir.), *Understanding Digital Humanities*, Palgrave Macmillan, London, 2012 ;

Bertels A., Spleelman D., « La contribution des cooccurrences de deuxième ordre à l'analyse sémantique », *Corpus*, 11, 2012, DOI : https ://doi.org/10.4000/corpus.2184 ;

Beynon, M., Russ, S., McCarty, W. « Human Computing? Modelling with Meaning », *Literary and Linguistic Computing*, 21(2), pp. 145-147, 2006 ;

Bolasco S., *L'analisi automatica dei testi. Fare ricerca con il text mining*, Carocci, Roma, 2013 ;

Booth, A.D. (dir.), *Machine Translation*, North-Holland Publishing Company, Amsterdam, pp. 173-94, 1967 ;

Booth, A.D., Booth, K.H.V., « The beginnings of MT », in Hutchins, W.J., (dir.), *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 253–61, 2000 ;

Booth, A.D., Booth, K.H.V., *Automatic Digital Calculators*, Butterworth, London, 1953 ;

Bowles, Edmund A. « Towards a Computer Curriculum for the Humanities », *Computers and the Humanities*, 6(1), Springer, 1971, pp. 35-38 ;

Burnard, L., « On the Hermeneutic Implications of Text Encoding », in Fiormonte, D., Usher, J. (dir.), « New Media and the Humanities : Research and Applications », *Humanities Computing Unit*, Oxford, pp. 31-38, 2001 ;

Busa, R., « The Annals of Humanities Computing : The Index Thomisticus », *Computers and the Humanities*, 14, pp. 83-90, 1980 ;

Busa, R., « Complete Index Verborum of Works of St. Thomas », *Speculum : A Journal of Medieval Studies*, XXV/1 , pp. 424-425, 1950 ;

Busa, R., « Foreword : Perspectives on the Digital Humanities », in Schreibman, S., Siemens, R., Unsworth J., (dir.), *A Companion to Digital Humanities*, Blackwell, Victoria, 2004 ;

Busa, R., *La terminologia Tomistica dell'Interiorità : Saggi di metodo per un'interpretazione della metafisica della presenza*, Fratelli Bocca, Milano, 1949 ;

Busa, R., *S. Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Fratelli Bocca, Milano, 1951 ;

Charaudeau, P., *Le discours politique. Les masques du pouvoir,* Vuibert, Paris, 2005 ;

Chaumartin F. R., Lemberger, P., *Le traitement automatique des langues : comprendre les textes grâce à l'intelligence artificielle*, Dunod, Paris, 2021 ;

Cheng, W., « Corpus-Based Linguistic Approaches to Critical Discourse Analysis », in Chapelle, C.A. (dir.), *The encyclopedia of applied linguistics*, Wiley-Blackwell Publishing, Oxford, pp. 1-8, 2013 ;

Chiari, I., *Informatica e lingue naturali. Teorie e applicazioni computazionali per la ricerca sulle lingue*, Aracne, Roma, 2004 ;

Chiari, I., *Introduzione alla linguistica computazionale*, Laterza, Bari, 2007 ;

Chomsky, N., *Syntactic Structures*, Mouton & Co, The Hague, 1957 ;

Condamines, A., « Linguistique de corpus et terminologie », *Langages, La terminologie : nature et enjeux*, 157, p. 36-47, 2005 ;

Dacos, M., Mounier, P., *Humanités Numériques : État Des Lieux et Positionnement de La Recherche Française Dans Le Contexte International,* Research Report Institut français, 2015. https ://hal.archives-ouvertes.fr/hal-01228945 [dernier accès 18/09/2021] ;

de Marneffe M.C, Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., Manning C., « Universal Stanford Dependencies : A cross-linguistic typology », LREC 2014 ;

Demorgon, J., « L'interculturel entre réception et invention. Contextes, médias, concepts », *Questions de communication*, 4, pp. 43-70, 2003 ;

Drucker, J. : Distant Reading and Cultural Analytics. Intro to Digital Humanities, 2013, http ://dh101.humanities.ucla.edu/?page_id=62 [dernier accès 12/10/2021] ;

Ducrot, O., *Dire et ne pas dire. Principes de sémantique linguistique*, Hermann, Paris, 1972 ;

Ducrot, O., Todorov, T., *Dictionnaire encyclopédique des sciences du langage*, Seuil, Paris, 1972 ;

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics* 19(1), pp. 61-74, 1993 ;

Durand, M., *Histoire du Québec*, Imago, Paris, 2003 ;

Eco, U., *Lector in fabula*, Bompiani, Milano, 1979 ;

Eder, M., Rybicki J., Kestemont M.,« Stylometry with R : A Package for Computational Text Analysis », *The R Journal*, 8(1), 2016 ;

Fairclough, I., Fairclough, N. *Political Discourse Analysis : A Methods for Advanced Students*, Routledge, London, 2012 ;

Fairclough, N. *Critical discourse analysis : the critical study of language*, Longman, London/ New-York, 1995 ;

Fairclough, N., *Language and Power,* Longman, London, 2001 ;

Fargues, P., « Migration and Identity : The Paradox of Reciprocal Influences », *Esprit*, 1 : pp. 6-16, 2010. DOI : 10.3917/espri.1001.0006 ;

Firth J.R., *Papers in Linguistics*, Longman, London, 1957 ;

Fistetti, F., *Théorie du multiculturalisme. Un parcours entre philosophie et sciences sociales*, La Découverte, Paris, 2009 ;

Freddi, M., *Linguistica dei corpora*, Carocci, Roma, 2014 ;

Gabrielatos, C., « Keyness analysis : Nature, metrics and techniques », in Taylor C., Marchi, A. (dir.), *Corpus approaches to discourse : A critical review.* Routledge, Oxford, 2018 ;

Ganascia J.-G.,« The Logic of the Big Data Turn in Digital Literary Studies », *Frontiers in Digital Humanities*, 2 (7), 2015 ;

Gold, M.K. (dir.) *Debates in the Digital Humanities*, Minneapolis : University of Minnesota Press, 2012 ;

Griebel, T., Vollmann, E., « We can('t) do this. A corpus-assisted critical discourse analysis of migration in Germany », *Journal of Language and Politics*, 18 (5), pp. 671-697, 2019, DOI : 10.1075/jlp. 19006.gri ;

Griece, P., *Logic and conversation* , Cole, P., Morgan, J. L. (dir.), *Syntax and Semantics, III : Speech Acts*, Academic Press, New York, 1975 ;

Heiden S., Magué J-P., Pincemin B., « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », in Bolasco, S., Chiari, I., Giuliano, L. (dir.), « Proceedings of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010 », v. 2, pp. 1021-1032, Edizioni Universitarie di Lettere Economia Diritto, Roma, 2010. http :// textometrie.ens-lyon.fr/spip. php?article60&lang=fr ;

Hockey, S., « The History of Humanities Computing », in Schreibman, S., Siemens, R., Unsworth, J. (dir.), *A Companion to Digital Humanities*, Blackwell, Oxford, 2004 ;

Hoover, D.L., « Textual Analysis », in Siemens R., Price K. (dir.), *Literary Studies in the Digital Age : An Evolving Anthology*, Modern Language Association, New York, 2013 ;

Hunston, S., *Corpora in Applied Linguistics*, Cambridge University Press, Cambridge, 2002 ;

Hutchins W.J. (dir.), *Early Years in Machine Translation.*

Memoirs and Biographies of Pioneers, John Benjamins Publishing Company, Amsterdam/ Philadelphia, 2000 ;

Hutchins, W.J., (dir.), *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2000 ;

Jasinski, J., *Sourcebook on Rhetoric : Key Concepts in Contemporary Rhetorical Studies*, SAGE Publications, London, 2001 ;

Jockers, M.L., *Macroanalysis : Digital Methods and Literary History*, Presses of Illinois University, 2013 ;

Jones, S.E., *Roberto Busa, and the emergence of Humanities Computing : The Priest and the Punched Cards*, Routeledge, London, 2016 ;

Khun, T. S., *Structure of Scientific Revolution*, University of Chicago Press, Chicago, 1962 ;

Kim, K. H., « Examining US news media discourses about North Korea : A corpus-based critical discourse analysis », *Discourse and Society*, 25 (2), pp. 221-244, 2014. DOI : 10.1177/0957926513516043 ;

Kirschenbaum, M.G., « Digital Humanities as/is a Tactical Term », in Gold, M.K., (dir.), *Debates in the Digital Humanities*, University of Minnesota Press, Minneapolis MN, 2012, http ://dhdebates.gc.cuny.edu/debates/ ;

Koller, V., Mautner, G., « Computer applications in critical discourse analysis », in Coffin, C., Hewings, A. O'Halloran, K., (dir.), *Applying English Grammar*, Arnold, London, 2004 ;

Kurdi Z., *Traitement automatique des langues et linguistique informatique*, ISTE éditions, Paris, 2017 ;

Labelle, M., « Multiculturalisme, interculturalisme, antiracisme : le traitement de l'altérité », *Revue Européenne des Migrations Internationales*, 31 (2), pp. 31-54, 2015 ;

Lacrampe-Camus, I., « Reconfiguration des ancrages et construction des origines dans un contexte de double migration. Jeunes d'origine équatorienne entre l'Espagne et Londres », *Cahiers des Amériques latines*, 91, pp. 153-170, 2019 ;

Ladmiral, J.-R., Lipiansky, E.-M., *La Communication interculturelle*, Les Belles Lettres, Paris, 2015 ;

Lakoff, G., Johnson, M., *Metaphors We Live By*, University of Chicago Press, Chicago, 1980 ;

Le Deuff, O., *Les humanités digitales : historique et développements*, ISTE Editions, London, 2018 ;

Lebart L., Pincemin B., Rioux M., *Analyse des données textuelles*, Presses de l'Université du Québec, Montréal, 2019 ;

Lenci, A., « Distributional semantics in linguistic and cognitive research », *Italian Journal of Linguistics*, 20, 2008 ;

Lenci, A., Montemagni, S., Pirrelli, V., *Testo e Computer. Elementi di linguistica computazionale*, Carocci, Roma, 2016 ;

Léon, J., « De la TA à la linguistique computationnelle et au TAL », *Histoire de l'automatisation des sciences du langage,* ENS Éditions, Lyon, 2015 ;

Levi Strauss C., « L'apport des sciences sociales à l'hu-

manisation de la civilisation technique », *Courrier de l'UNESCO*, 2008 (Archives inédits UNESCO 8 août 1956). [https ://fr.unesco.org/courier/2008-5/aportacion-ciencias-sociales-humanizacion-civilizacion-tecnica, [dernier accès 1/1/2021] ;

Levison, M. « The Mechanical Analysis of Language », *NPL*, pp. 562 -574, 1962 ;

Levison, M., « The Computer in Literary Studies », Booth, A.D. (dir.), *Machine Translation*, North-Holland Publishing Company, Amsterdam, pp. 173–94, 1967 ;

Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012 ;

Locke, W.N., Booth, A.D. (dir.) *Machine Translation of Languages : Fourteen Essays*, MIT, Cambridge, 1955 ;

Lombardo Vallauri, E., *La lingua disonesta*, Il Mulino, Bologna, 2019 ;

Louw, B., « Irony in the text or insincerity of the writer? The diagnostic potential of semantic prosodies », Baker, M., Francis, G., Tognini-Bonelli, T., (dir.), *Text and Technology : In Honour of John Sinclair*, John Benjamins, Amsterdam, 1993, pp. 30-50 ;

Lovelace, A.A., « Notes by the Translator », Morrison, P., Morrison, E. (dir), *Charles Babbage and his Calculating Engines. Selected Writings by Charles Babbage and Others*, Dover Publications, New York, pp. 245-295, 1961. [Prémiere parution : Taylor, R. (dir.), *Scientific Memoirs, Selections from The Transactions of Foreign Academies and Learned Societies and from Foreign Journals*, 1843] ;

Lunenfeld, P., Burdick A., Drucker, J., Presner, T., Schnapp, J., *Digital_Humanities*, The MIT Press, Cambridge, 2012 ;

Maingueneau, D., « Le tour ethnolinguistique de l'analyse du discours », *Langages. Ethnolinguistique de l'écrit*, 26 (105), pp. 114-125, 1992 ;

Maingueneau, D., Charaudeau, P., *Dictionnaire d'analyse du discours,* Seuil, Paris, 2002 *;*

Maingueneau, D., *Les termes clés de l'analyse du discours*, éditions du Seuil, Paris, 2009 ;

Marche, S., « Literature Is Not Data : Against Digital Humanities », *Los Angeles Review of Books*, 2012 ;

May, P., *Philosophies du multiculturalisme*, Les Presses de Sciences Po, Paris, 2016 ;

Mayaffre D., « *Analyse du discours politique et Logométrie : point de vue pratique et théorique* », *Langage et Société*, Maison des Sciences de L'homme, Paris, pp. 91-121, 2005 ;

Mayaffre D., Viprey J.-M., (dir.), *La cooccurrence, du fait statistique au fait textuel*, Revue *CORPUS,* 11, 2012 ;

Mayaffre, D., « L'entrelacement lexical des textes. Cooccurrences et lexicométrie », *Journées de Linguistique de Corpus*, Lorient, pp. 91-102, 2007 ;

McCarty, W. « A Telescope for the mind? » in M.K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis MN : University of Minnesota Press, 2012 ;

McCarty, W. *Humanities Computing*, Palgrave, London, 2005 ;

McCarty, W., « Humanities Computing as Interdiscipline », séminaire dans le cycle 'Is Humanities Computing an Academic Discipline?' IATH, University of Virginia, 5 Nov. 1999, http ://www. iath.virginia.edu/hcs/mccarty.html [dernier accès le 27/09/2021] ;

McCarty, W., « Tree, Turf, Centre, Archipelago – or Wild Acre? Metaphors and Stories for Humanities Computing », *Literary and Linguistic Computing*, 21 (1), pp. 1-13, 2006 ;

McCarty, W., Kirschenbaum, M., « Institutional Models for Humanities Computing », *Literary and Linguistic Computing,* 18 (4), pp. 465-89, 2003 ;

McGann, J., *Radiant Textuality : Literature after the World Wide Web*, Palgrave, 2001 ;

Menabrea, L.F., « Sketch of the Analytical Engine Invented by Charles Babbage. Translated by Ada Augusta, Countess of Lovelace », in Morrison, E. (dir.), *Charles Babbage and his Calculating Engines. Selected Writings by Charles Babbage and Others*, Dover Publications, New York, pp. 225-245, 1961 ; [Premiere publication : *Bibliothèque Universelle de Genève*, 82 (Oct. 1842) ; trad. Lovelace, A.A. in Taylor, R., (dir.), *Scientific Memoirs, Selections from The Transactions of Foreign Academies and Learned Societies and from Foreign Journals*, 1843 ;

Meunier, J.G., « Humanités numériques ou computationnelles : Enjeux herméneutiques », *Sens Public*, 2014, http ://www.sens-public.org/articles/1121/ [dernier accès 26/9/2021] ;

Moretti, F., « Conjectures on world literature », *New left review* 1, London. URL = https ://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature, 2000. [dernier accès : 22/08/2021] ;

Moretti, F., *Graphs, Maps, Trees : Abstract Models for a Literary History*, Verso Books, London, 2007 ;

Mounier, P., « Manifeste des *Digital Humanities* », *Journal des anthropologues* [En ligne], pp. 122-123, 2010, mis en ligne 01/12/2012, URL : http ://journals.openedition.org/jda/3652 ; DOI : https ://doi.org/10.4000/jda.3652 ;

Mounier, P., *Les humanités numériques. Une histoire critique*, Éditions de la Maison des sciences de l'homme, Paris, 2018 ;

Ollivier-Yaniv, C., « La communication gouvernementale, un ordre en négociation », *Revue française d'administration publique*, 3(171), pp. 669 – 680, 2019 ;

Orlandi, T. « Per un curriculum europeo di informatica umanistica », Fiormonte, D. (dir.), *Informatica umanistica dalla ricerca all'insegnamento*, Bulzoni Editore, Roma, pp. 19-25, 2003 ;

Partington, A. « Utterly content in each other's company : Semantic prosody and semantic preference », *International Journal of Corpus Linguistics,* 9(1), pp. 131-56, 2004 ;

Pincemin, B., « Concordances et concordanciers : de l'art du bon KWAC », *XVIIe colloque d'Albi Lagages et signification – Corpus en Lettres et Sciences sociales : des do-*

cuments numériques à l'interprétation, pp. 33-42, Albi, Jul. 2006 ;

Pincemin, B., « Sémantique interprétative et textométrie– Version abrégée », *Corpus*, 10, 2011 ;

Ramsay S., *Reading Machines : Toward an Algorithmic Criticism*, University of Illinois Presses, 2011 ;

Rastier, F. (dir.), *Sémiotique narrative et textuelle*, Larousse, Paris, 1973 ;

Rastier, F., *Faire sens. De la cognition à la culture*, Classiques Garnier, Paris, 2018 ;

Rastier, F., *La mesure et le grain. Sémantique de corpus*, Paris, Champion, 2011 ;

Rastier, F., Sémantique interprétative, Presses universitaires de France, Paris, (1987), 2009 ;

Richard, I.A., *Practical Criticism*, K. Paul, Trench, Trubner Company, London, 1929 ;

Salahshour, N., « Liquid metaphors as positive evaluations : A corpus-assisted discourse analysis of the representation of migrants in a daily New Zealand newspaper », *Discourse*, Context and Media 13, pp. 73-81, 2016. DOI : 10.1016/j.dcm.2016.07.002 ;

Sbisà, M., « Presupposizioni e contesti », *La svolta contestuale*, C. Penco, (dir.) Milano-New York 2002, pp. 221-39 ;

Sbisà, M., *Detto non detto. Le forme della comunicazione implicita,* Laterza, Bari, 2010 ;

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Confer-*

ence on New Methods in Language Processing, Manchester, 1994 ;

Schmid, H., « Improvements in Part-of-Speech Tagging with an Application to German », *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995 ;

Schreibman, S., Siemens, R., Unsworth, J. (dir.) *A Companion to Digital Humanities*, Malden, MA/Oxford/Carlton, Blackwell Publishing, Victoria, pp. 254-270, 2004 ;

Searle, J., *Speech Atcs. An essay in the philosophy of language*, Cambridge University Press, Cambridge, 1969 ;

Sinclair J. M., *Reading concordances*, Pearson, Londres, 2003 ;

Sinclair, J. M. « The search for units of meaning », *Textus,* 9(1), pp. 75-106, 1996 ;

Sinclair, J. M., *Corpus, concordance, collocation*. Oxford University Press, Oxford, 1991 ;

Sparck Jones, K., « R.H. Richens. Translation in the nude', in W.J. Hutchins (ed.), *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins, Amsterdam/Philadelphia, 2000 ;

Stubbs, M., *Text and corpus analysis :Computer-assisted studies of language and culture*. Wiley-Blackwell, London, 1996 ;

Stubbs, M., *Words and phrases : Corpus studies of lexical semantics*. Blackwell, London, 2001 ;

Svensson, P. « The Landscape of Digital Humanities », *Digital Humanities Quarterly,* 4 (1), 2010 ;

Svensson, P., « From Optical Fiber To Conceptual Cyber-

infrastructure », *Digital Humanities Quarterly,* 5 (1), 2011 ;

Svensson, P., « Envisioning the Digital Humanities », *Digital Humanities Quarterly*, 6 (1), 2012 ;

Svensson, P., « Humanities Computing as Digital Humanities », *Digital Humanities Quarterly*, 3 (3), 2009 ;

Tannery, P. « La stylométrie ses origines et son présent » *Revue Philosophique de La France et de l'Étranger*, 47, Presses Universitaires de France, Paris, pp. 159-69, 1899 ;

Tauveron M., « De la cooccurrence généralisée à la variation du sens lexical », *Corpus,* 11, 2012 ;

Terras, M., Vanhoutte, E., Nyhan, J., *Defining Digital Humanities : A Reader*, Routledge, London/New York, 2013 ;

Thaller, M. (2006), « Waiting for the Next Wave : Humanities Computing in 2006 », *CliP*, King's College London, London, 2006 ;

Thaller, M., « Bridging the Gap ; Splitting the Bridge? Studying Humanities Computer Science in Cologne », *CLiP*, Duisburg, 2001 ;

Tomasin, L., *L'impronta digitale : cultura umanistica e tecnologia*, Carocci, Roma, 2017 ;

Tufte, E., *The Visual Display of Quantitative Information*, Hoepli, 2000 (1983) ;

Van Dijk, T.A. (dir.), *Discourse studies : A multidisciplinary introduction*, SAGE Publications, London, 2011. DOI : 10.4135/9781446289068 ;

van Dijk, T.A., *Communicating Racism : Ethnic Prejudice in Thought and Talk*, Sage, London, 1987 ;

van Dijk, T.A., *Handbook of Discourse Analysis : Discourse analysis in society*, Academic Press, London, 1985 ;

van Dijk, T.A., *Racism and the Press*, Routledge, London, 1991 ;

Victorri B., Fuchs C., *La polysémie : construction dynamique du sens*, Hermès, Paris, 1996 ;

Vonnegut, K., *A men without a Country*, Seven Stories Press, New York, 2005 ;

Warwick, C., Terras, M., Nyhan, J. (dir.), *Digital Humanities in Practice*, Facet Publishing/UCL, London, 2012 ;

Weaver, W. *Scene of Change. A Lifetime in American Science*, Scribner, New York, 1970 ;

Weaver, W., « Translation », in Locke, W.N., Booth, A.D. (eds) *Machine Translation of Languages. Fourteen Essays*, Cambridge, MA : The MIT Press, pp. 15-23, 1965 [Original publication 1949] ;

Yngve, V.H., « Early research at M.I.T. In search of adequate theory », in Hutchins, W.J. (dir.) *Early Years in Machine Translation. Memoirs and Biographies of Pioneers*, John Benjamins, Amsterdam/Philadelphia, 2000 ;

Finito di stampare
nel mese di ottobre 2021
da Gesp – Città di Castello (PG)

I LIBRI DI
EMIL
www.ilibridiemil.it