UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE SUR LES REPRÉSENTATIONS DES BIAIS EN TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN COMMUNICATION

PAR

CAROL-ANN LÉVESQUE

FÉVRIER 2024

REMERCIEMENTS

Je tiens à remercier sincèrement l'ensemble des personnes dont le soutien et les encouragements constants ont rendu possible la réalisation de ce mémoire.

Je tiens à offrir un remerciement tout spécial à Michelle Stewart qui a été une directrice extraordinaire. Merci de m'avoir épaulé durant ces dernières années et d'avoir embarqué dans cette aventure avec moi. Ta grande curiosité, ton esprit critique, ton expertise et ta patience ont été des atouts inestimables dans la réalisation de ce projet.

Merci à mes ami(e)s, ma famille et mes collègues dont les encouragements n'ont pas manqué. Plus particulièrement, merci à Claudie, Stéphane et Sara pour votre amitié et vos esprits vifs.

Merci à Tiago, pour ta bienveillance, ta présence et ton support inconditionnel. Tu as été une grande source de bonheur tout au long de ce projet.

Et surtout, merci infiniment à mes parents Marthe et André, sans qui rien de tout cela n'aurait été possible.

DÉDICACE

À Hannah, qui sait inspirer les esprits à penser les grands changements de ce monde.

AVANT-PROPOS

« Le langage est la maison de l'être. »
Martin Heideigger, 1976

Le présent mémoire est le résultat d'une pensée grandissante sur un monde en pleine révolution technologique, voire technique. S'il s'agissait au début d'analyser les équipes éthiques des géants du Web, le déploiement à grande échelle des technologies en traitement automatique du langage naturel a eu l'effet de complètement réorienter notre recherche. Alors que nous étions en analyse de l'équipe éthique en place chez Google en 2019, la publication de la recherche *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big ?* par des membres de son équipe a tout changé. C'est suite à cette dernière qu'il y a eu des renvois massifs chez la firme américaine. Pourtant la publication se voulait complètement alignée à la responsabilité de toute équipe éthique Web : remettre en question le déploiement d'une technologie dont les impacts, tant sociaux qu'environnementaux, ne semblaient pas avoir été pris en considération.

Nous nous sommes donc intéressés aux risques (environnementaux et biais) liés au développement massif des technologies en TALN. En voulant analyser les représentations de ces derniers dans la recherche scientifique, nous avons réalisé l'exhaustivité du projet que nous étions en train de dessiner. Par souci de faisabilité, nous avons choisi de restreindre notre recherche aux biais présents dans les modèles de langage. Plus particulièrement, puisque la vitesse de développement de ces technologies est en constante accélération, nous nous sommes intéressés à la façon dont la recherche traite de ce sujet complexe, évolutif et fondamentalement humain. Avec ce mémoire, nous souhaitons proposer une analyse qualitative qui privilégie un raisonnement normatif et considère que les biais en TALN sont bien plus qu'une anomalie technique du système, mais bien une réelle source de discrimination et de préjudices pour certaines communautés.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

IA	Intelligence artificielle
LLMs	Large Language Models
ML	Modèle de langage
NLP	Natural Language Processing
TALN	Traitement automatique du langage naturel

RÉSUMÉ

Le développement massif des technologies en traitement automatique du langage naturel (TALN), ainsi que l'obsession grandissante de la taille des modèles de langage utilisés, s'accompagne d'enjeux sociaux et environnementaux trop souvent occultés. À travers une méta-synthèse qualitative, notre mémoire présente une analyse du développement technologique en TALN et explore comment différentes études traitent des biais dans ce domaine en plein essor. Les résultats de notre analyse ont été mis en relation, puis approfondis avec les concepts encadrant notre recherche, tels que le langage opérationnel, les biais des représentations de l'IA et l'éthique relationnelle. Nous avons constaté que les études sur les biais au sein de notre corpus privilégient largement des méthodologies quantitatives pour traiter de ce sujet social, souvent au détriment d'une analyse approfondie du concept de biais lui-même et de sa relation au langage. La recherche en NLP devrait davantage se nourrir des approches interdisciplinaires et inclusives, afin de mieux comprendre et approfondir les aspects socioculturels et les préjudices liés aux biais.

Mots clés : Traitement automatique du langage naturel, biais, modèle de langage, intelligence artificielle

ABSTRACT

The massive development of technologies in Natural Language Processing (NLP), along with the growing obsession with the size of language models used, comes with social and environmental challenges that are often overlooked. Through a qualitative meta-synthesis, our master's thesis presents an in-depth analysis of technological development in NLP and explores how different studies address biases in this rapidly evolving field. The results obtained from our analysis were compared and further examined in relation to the concepts framing our research, such as operational language, biases in AI representations, and relational ethics. We have observed that studies on biases within our corpus predominantly favor quantitative methodologies when addressing this social topic, often at the expense of in-depth analysis of the bias concept itself and its relationship to language. Research in NLP should further embrace interdisciplinary and inclusive approaches to better comprehend the sociocultural aspects and biases-related prejudices.

Keywords : Natural Language Processing, bias, language models, artificial intelligence

INTRODUCTION

Dans son livre *Critical Theory of Technology : An Overview*, le philosophe Andrew Feenberg (2005) explique que l'organisation de nos sociétés autour de la technique implique à la fois une nouvelle forme de pouvoir et une gestion technique des êtres humains. Cette décontextualisation et transformation de l'objet de l'expérience en données est un exercice de pouvoir dans lequel l'humain devient à la fois opérateur et objet (Feenberg, 2005). Ce changement de nature de la technique et l'accélération constante de son développement s'est accompagné d'une multitude de définitions. Dans une lecture du 18 juillet 1962, traduite de l'allemand à l'anglais *Traditional Language and Technological Language*, le philosophe Martin Heidegger s'est intéressé à repenser les conceptions communes des technologies modernes. Il explique qu'à travers l'âge industriel moderne, il a été possible de dénoter deux révolutions technologiques. La première, représentée par la transition des technologies à moteur. La seconde, par la montée de l'automatisation et de la cybernétique (Heidegger, 1998). Mais, aussi largement décrites, Heidegger exprime que nous percevons difficilement le caractère distinctif des technologies modernes hérité de ces révolutions. Dans cette volonté de les définir plus précisément, il a démontré que malgré certaines conceptions des technologies modernes qui en apparence peuvent sembler justes, une difficulté persiste à vouloir révéler leur singularité (*Ibid*.). Il a poursuivi sa réflexion ainsi :

> Therefore, we will consider now the function and character of modern science within modern technology in the attempt to bring what is peculiar to modern technology […] a characterization of modern technology as an applied science is inadequate. Instead of this, one now talks of a "mutual support" in the relationship between science and technology. […] Now, what is that thing in which modern natural science and technology agree and is thus the same ? What is peculiar to both ? […] This is determined, more or less deliberately, by the leading question : How must nature be projected in advance as a region of objects so that natural processes are made calculable in advance ? (*Ibid.*, p.136)

Pour Heidegger, cette interrogation amène à penser la question du caractère de la réalité de la nature, que seul ce qui est vrai peut être mesuré, et seul ce qui peut se mesurer peut être. Avec la technique moderne, il s'agit de rendre disponible l'énergie et de réquisitionner la nature en tout temps. C'est pour lui, précisément par le fait qu'elle défit l'avenir que la technique moderne trouve sa singularité. (*Ibid.*).

S'il est pertinent d'aborder l'essence de la technique moderne, nous soutenons l'importance de questionner l'apport du siècle dernier dans le développement de ce qu'Heidegger avait pointé comme étant la seconde révolution technologique de l'âge industriel moderne : la montée de l'automatisation et la cybernétique.

La professeure en sociologie Orit Halpern (2014) explique dans son livre *Beautiful Data* que c'est suite à la Seconde Guerre mondiale, que le mathématicien Norbert Wiener a inventé la cybernétique. Son idée était claire : transformer l'humain en transformant son rapport à la machine. Il considérait cette transformation comme un passage nécessaire vers un âge nouveau, celui de l'organisation de nos sociétés qui, jusqu'à la moitié du 20e siècle, avaient été investies dans l'exploration. Halpern exprime que ce que Wiener revendiquait à ce moment n'était pas sans complexité, il imaginait une nouvelle forme de visualisation de l'être (Halpern, 2014). Elle ajoute :

> Wiener indicated a desire to see an older archival order, adjoined to modern interests in taxonomy and ontology, rendered obsolete by another mode of thought invested in prediction, self-referentiality, and communication. […] Wiener dreamed of a world where there is no "unknown" left to discover, only an accumulation of records that must be recombined, analyzed, and processed. Wiener argued that in observing too closely and documenting too "meticulously", one is unable to deduce patterns, to produce in his words a "flow of ideas". (*Ibid., p.12*)

Pour Wiener, il s'agissait d'entraîner la machine à communiquer et à anticiper, dans le but de contrôler et prédire les actions futures. (*Ibid.*) En d'autres mots, il imaginait un monde où l'information serait conceptualisée comme une entité extérieure, distincte du corps où elle est originellement intégrée (Hayles, 1999). Ce désir d'automatisation de nos sociétés s'est accompagné de l'idée selon laquelle certains des éléments humains des plus fondamentaux, tels que le langage, pouvaient être réduits à des méthodes informatiques (Golumbia, 2009). La Seconde Guerre mondiale, avec des événements comme le décryptage de Enigma par Turing, aura certainement nourri l'analogie entre le langage et le code (*Ibid.*). Puis, de cette longue histoire qui a lié des idéaux technologiques au calcul et à la gouvernance, est né l'intérêt majeur de la linguistique computationnelle.

> The idea that the human brain might just be a computer […] entailed a real investigation of what exactly was meant by the "human brain" and more pointedly "the human mind" in the first place. [...] Perhaps the easiest and most obvious way to show that a computer was functioning like a human brain would be to get the computer to produce its results the same way a human being (apparently) does : by producing language. (*Ibid.,* p.83)

Depuis ses débuts, le développement de la linguistique computationnelle a démontré un désir incessant de voir les ordinateurs utiliser le langage de manière humaine. En réponse à cette nécessité de traiter les langues naturelles de manière automatisée, le domaine du traitement automatique du langage naturel (TALN), plus communément appelé en anglais Natural Language Processing (NLP), a émergé dans les années d'après-guerre (*Ibid*.). Les chercheurs en linguistique computationnelle ont ainsi entrepris de développer des

méthodes, des algorithmes et des modèles capables de permettre aux ordinateurs de générer et manipuler le langage humain.

Au cours des dernières années, une évolution extrêmement rapide de l'IA a ouvert de nouveaux horizons dans le domaine du TALN. En constante quête d'amélioration et d'intégrations d'avancées majeures à leurs produits, des acteurs importants tels que OpenAI, Microsfot, Google, Facebook et Amazon ont tous investi massivement dans le développement des grands modèles de langage, connus sous le nom de Large-language models (LLMs) en anglais.

> This is one of the so-called neural networks. Large-language models are trained by pouring into them billions of words of everyday text, gathered from sources ranging from books to tweets and everything in between. The LLMs draw on all this material to predict words and sentences in certain sequences (Hern et Milmo, 2023).

Entraînés sur des quantités de données colossales provenant de partout sur le Web, les grands ML servent aux assistants virtuels (chatbot), à la traduction automatique, la catégorisation d'informations, l'extraction d'informations, la synthétisation de documents, la reconnaissance vocale, etc. Depuis novembre 2022, il est maintenant possible de faire l'expérience d'un LLMs sous la forme de ChatGPT (Generative Pretrained Transformer), un chatbot de dernière génération lancé par OpenAI.

Les technologies de langage nous entourent de façon quotidienne depuis plusieurs années, mais il semble que le monde en ait réellement pris conscience avec l'arrivée de ce chatbot génératif. Les conversations sur les promesses et surtout les dangers de ces outils linguistiques se sont faites de plus en plus fréquentes, en passant des experts en technologie aux représentants du gouvernement (Bender, 2023). Emily M. Bender, professeure en linguistique computationnelle à l'université de Washington, explique cette situation par le fait que ces systèmes de transcription automatique sont formés sur d'immenses volumes de données, les exposant ainsi à d'importants biais impossibles à filtrer (*Ibid*.). Ces biais, liés au genre, à la race, à l'orientation sexuelle, à la classe sociale et à la culture, peuvent être exacerbés dans les résultats générés par ces systèmes et affecter négativement, voire discriminer, certains groupes de personnes (Bender *et al.*, 2021). Pour reprendre les mots de Bender, c'est ce qui revient à produire des systèmes d'oppression (Bender, 2023).

Les biais en TALN représentent un enjeu important de ce développement technologique et doivent être considérés comme un élément fondamental pour la recherche. C'est pourquoi nous proposons, dans ce mémoire, d'analyser les représentations des biais dans différentes études en traitement automatique du langage naturel. À travers l'analyse d'un corpus d'articles sélectionnés relatifs aux biais en TALN, notre but est de comprendre la façon dont les auteurs s'engagent avec ce sujet complexe, en passant des approches

utilisées pour examiner les biais dans les modèles de langage aux objectifs visés de ces recherches et plus encore. Nous abordons également les conséquences potentielles des biais sur nos sociétés et nos modes de communication, en plus de permettre de mieux comprendre les points forts et les points faibles de la recherche sur les biais en TALN. Finalement, nous pensons que ce travail peut aider à adopter un regard critique sur ces technologies et à favoriser une utilisation responsable et éthique de celles-ci.

CHAPITRE 1
PROBLÉMATIQUE

Dans ce chapitre, nous présentons une analyse historique de l'évolution de la linguistique computationnelle jusqu'à l'émergence des technologies actuelles en TALN. Il est important de comprendre comment s'est développé ce domaine, et nous proposons de revisiter les grands noms qui l'ont composé et qui ont offert les bases possibles à son développement. Ces bases sont porteuses des idéologies qui ont contribué à orienter son développement. Nous allons par la suite approfondir la relation entre le domaine du TALN et les grands modèles de langage qui nous mènera à une recension des écrits permettant de comprendre le contexte dans lequel s'inscrit notre recherche. Suivant cette exploration, nous formulons notre question de recherche, qui constitue le fil conducteur de notre étude. Enfin, nous abordons la pertinence communicationnelle de notre recherche, en plus de définir les objectifs de recherche que nous souhaitons atteindre.

1.1 L'évolution de la linguistique computationnelle

La linguistique computationnelle, développée au milieu du 20e siècle, s'est présentée comme l'un des défis et des intérêts majeurs des scientifiques de l'informatique. C'est dans ces mêmes années que Noam Chomsky a complètement transformé la discipline académique de la linguistique, en plus de fournir involontairement ce qui allait constituer les bases nécessaires au développement de la linguistique computationnelle (Golumbia, 2009). Dans sa publication de 1957 *Syntactic Structures*, le linguiste a formulé sa théorie de la grammaire générative, selon laquelle la capacité humaine à utiliser et comprendre la langue découle d'un système mental qui génère les structures syntaxiques. (*Ibid.*) Il a soutenu l'idée selon laquelle le cerveau est similaire à l'un des derniers développements technologiques de cette époque, soit l'ordinateur. Sa perspective combinait et renforçait une théorie nécessaire pour la LC selon laquelle le langage pouvait être considéré comme une forme de calcul. (*Ibid.*)

Dans un de ses articles[1] datant des années 50, Chomsky a présenté un modèle pouvant être utilisé pour décrire les langues naturelles de manière formelle. En considérant les langues humaines au même niveau que les langues formelles[2], ce travail a contribué à soutenir et à renforcer les avancées en matière de linguistique computationnelle (*Ibid.*). Chomsky, de son côté, s'est pourtant dissocié de l'idée selon laquelle

[1] L'article *On Certain Formal Properties of Grammars* a été une contribution essentielle à la théorie du langage formel.

[2] Le langage formel est un moyen de communication précis qui utilise des symboles spécifiques, des règles strictes et des structures claires pour exprimer des informations. En d'autres termes, dans un langage formel, seules les règles de construction et d'organisation des symboles ont de l'importance, et le sens des mots ou des symboles n'est pas pris en compte lors de leur manipulation (Encyclopædia Universalis, 2023).

les machines seraient aptes à utiliser le langage humain et a précisé que son travail ne soutenait pas cette idée (*Ibid.*). Il s'est dit surpris du nombre de travaux en grammaire générative qui ont alimenté l'intérêt grandissant de certains domaines informatiques. Il a soutenu que les avancées technologiques pouvaient avoir des impacts nuisibles en orientant la recherche vers des problèmes créés par ces mêmes technologies, et a aussi rejeté la réduction du langage humain au contrôle computationnel. Pour le professeur associé à l'université américaine *Virginia Commonwealth,* David Golumbia, il est étonnant que sa position sur ce sujet ait été aussi largement ignorée par les professionnels. (*Ibid.)* Il explique que Chomsky a vu, tout au long de sa vie intellectuelle, les autres tirer ce qu'ils voulaient de son travail. Golumbia précise « […] it seems clear that what happens in this case is not just computer scientists but an entire community of technologically-minded intellectuals seeing in Chomsky's work precisely the potential to do what Chomsky disclaims » (Golumbia, 2009, p.38). Il poursuit en expliquant que les pionniers de l'informatique, tels que Turing[3], Shannon et Warren Weaver, n'avaient pratiquement aucune formation en linguistique ou en études linguistiques. Le travail de Chomsky leur a simplement fourni les bases nécessaires pour soutenir l'idée selon laquelle certains aspects fondamentaux de l'humain, tel que le langage, pouvaient se réduire au domaine de l'informatique (*Ibid.*).

Golumbia explique que pour ces pionniers, la meilleure façon de prouver que l'ordinateur fonctionne comme un cerveau humain a été de faire en sorte que l'ordinateur produise du langage. C'est en quoi le mathématicien et ingénieur Warren Weaver a joué un rôle important dans le développement de la linguistique computationnelle en tant que discipline. Il a publié un rapport en 1949 intitulé *Translation* qui a posé les bases de la traduction automatique (*Ibid.*). Il a proposé l'idée que la traduction automatique pourrait être considérée comme un problème mathématique complexe qui nécessite une combinaison d'approches statistiques et formelles pour le résoudre.

> Weaver's memorandum starts from a cultural observation : "a multiplicity of languages impedes cultural interchange between the peoples of earth, and is a serious deterrent to international understanding". […] The view becomes characteristic of computationalist views of language : that human society as a whole is burdened by linguistic diversity, and that political harmony requires the removal of linguistic difference. […] Weaver suggests that a "book written in Chinese is simply a book written in English which was coded into the "Chinese Code". (*Ibid,* p.88-90)

[3] La machine de Turing a été développée durant la Seconde Guerre mondiale dans le but de décoder Enigma, un système de chiffrement utilisé par les forces armées allemandes pour transmettre des informations sensibles. Cette machine était entraînée pour effectuer des opérations algorithmiques, elle n'a jamais traduit ou interprété le langage. Elle a plutôt développé des schémas statistique et mathématique pour révéler la transmission linguistique prévue. Elle a tout de même établi l'analogie entre décoder et parler en plus de nourrir l'idée selon laquelle le langage humain devait être un code (Golumbia, 2009, p.89)

Mais les propositions de Weaver n'ont pas été appuyées de tous. Dans une de leur correspondance publiée par Weaver lui-même, Norbert Wiener a soulevé plusieurs problèmes associés à la traduction automatique (*Ibid.*). Selon Wiener, le langage humain est souvent ambigu et peut avoir plusieurs interprétations différentes, même lorsqu'il semble être univoque. Éventuellement, l'idée de Weaver selon laquelle la traduction est une opération similaire au décodage a aussi été rejetée par des ingénieurs en informatique (*Ibid.*). Évidemment, peu de linguistes auraient consenti à l'idée selon laquelle le langage est une sorte de code.

> […] Nevertheless, even today, prominent subset of computationalists (although rarely ones directly involved in computer language processing) continue to insist that formal languages, programming languages, and ciphers are all the same kinds of things as human languages, despite the manifest differences in form, use, and meaning of these kinds of systems. Surely the mere fact that a particular word – language – is conventionally applied to these objects is not, in and of itself, justification for lumping the objects together in a metaphysical sense ; yet at some level, the intuition that language is code-like underwrites not just MT but its successor (and strangely more ambitious) programs of CL and NLP. (*Ibid.,* p.93*)*

Avec le temps, les domaines de la linguistique computationnelle et du traitement automatique du langage naturel se sont diversifiés et ont évolué dans de nombreuses directions différentes. Néanmoins, leur objectif fondamental reste inchangé : faire en sorte que les ordinateurs utilisent le langage de la même façon que les humains (*Ibid.*). Pour David Golumbia, la linguistique computationnelle a simplifié le langage en le réduisant à des modèles mathématiques et algorithmiques. Il suggère que le langage a des implications politiques et culturelles plus larges et que l'adoption de diverses méthodes informatiques pour le reproduire peut renforcer des structures de pouvoir déjà en place et perpétuer des inégalités (*Ibid*.).

Au cours de la dernière décennie, nous avons pu voir un avènement marqué des grands modèles de langage. Initialement, les grands modèles de langage utilisaient des approches de modèles statistiques basées sur des grammaires formelles. Une grammaire formelle est un ensemble de règles qui décrivent la structure d'une langue, en plus de définir les constructions grammaticales valides et les relations entre les mots et les phrases (Aznar *et al.*, 2020). Cependant, avec le développement marqué de l'apprentissage profond (deep learning), les modèles de langage ont commencé à être entraînés sur une quantité colossale de données en utilisant des réseaux de neurones[4] pour modéliser les relations entre les mots et les phrases.

[4] L'utilisation de cette métaphore (réseaux de neurones) pour parler d'un modèle d'apprentissage automatique est révélatrice de l'objectif énoncé par Golumbia, soit de faire en sorte que les ordinateurs utilisent le langage de la même façon que les humains (Golumbia, 2009).

La professeure en linguistique informatique de l'Université de Washington Emily Bender explique au sujet des grands modèles de langage formés sur des réseaux de neurones artificiels que :

> The Transformer allows you to say for each given word, I'm going to talk about it not as a string but as a thing in vector space that represents what other words it co-occurs with. We get more and more elaborate representations of words that represent more and more information about what other words they've co-occurred with. (Bender et Mitchell, 2021)

Par contre, il est important de préciser que les LLMs n'ont a aucune réelle compréhension du langage (Bender *et al.*, 2021). Les réseaux de neurones possèdent une structure prédictive plus élaborée et, en raison de la grande quantité de données sur lesquelles ils sont entraînés, peuvent être justes un bon nombre de fois tout en n'ayant absolument rien compris (Bender et Mitchell, 2021). Il est fallacieux de parler de ce type de technologie en termes de compréhension, comparable à l'entendement humain, alors qu'il s'agit d'entraînement, d'identification, de modélisation, de prédiction et d'extraction d'information à partir de données.

1.1.1 Technologie et langage

L'utilisation du langage dans un contexte informatique comporte des risques bien documentés. Plus on décrit les grands modèles de langage comme étant capables de comprendre le langage et d'en capturer le sens, plus on incite les utilisateurs à voir en eux une sorte d'intelligence, en plus de contribuer à leur anthropomorphisation. Alors que réellement, c'est une reconnaissance de modèles à grandes échelles (*Ibid*.). Un système qui a été entraîné uniquement à reconnaître la forme d'un mot ou d'une phrase n'a en principe aucun moyen d'apprendre son sens (Bender et Koller, 2020).

> Languages are symbolic systems, and symbols are pairings of form and meaning (or, per de Saussure, signifier and signified). But GPT-3 in its training was only provided with the form part of this equation and so never had any hope of learning the meaning part. (Bender, 2022)

Autrement dit, les modèles de langage sont capables de percevoir l'aspect matériel du signe (l'image, la suite de lettres), mais ne comprennent pas réellement le signifié, l'aspect conceptuel du signe, son sens. (Bender *et al.*, 2021). En revanche, ils apprennent une certaine réflexion du sens dans la forme linguistique (Bender et Koller, 2020).

Dans son ouvrage la Condition de l'homme moderne, paru en 1958, la philosophe allemande Hannah Arendt a abordé les dangers potentiels de la domination technique sur la vie humaine (Arendt, 1998). Elle a considéré l'automatisation comme le désir de libérer l'homme du travail, en plus d'avoir comme

conséquence de réduire son langage à un ensemble de symboles mathématiques. Selon elle, le travail et le langage sont des concepts fondamentaux à notre capacité d'agir politiquement en tant qu'individu (Halpern, 2014). Le langage, au cœur de la vie politique, est un outil essentiel à la communication et un moyen pour les individus de formuler leur propre compréhension du monde (Arendt, 1998).

> For the sciences today have been forced to adopt a "language" of mathematical symbols which, though it was originally meant only as an abbreviation for spoken statements, now contains statements that in no way can be translated back into speech. The reason why it may be wise to distrust the political judgment of scientists qua scientists is not primarily their lack of "character"—that they did not refuse to develop atomic weapons—or their naïveté—that they did not understand that once these weapons were developed they would be the last to be consulted about their use—but precisely the fact that they move in a world where speech has lost its power. (*Ibid*., p.37)

Selon Arendt, tout ce que les humains font, expérimentent et connaissent ne prend sens que s'il est possible d'en parler. Nous ne pouvons faire l'expérience du sens que si nous sommes en mesure d'en parler et de donner du sens à nos actions (*Ibid*.). Le désir d'automatisation et d'opérationnalisation d'actions fondamentalement humaines affaiblit le lien primordial entre le langage et la faculté de jugement (Andrejevic, 2020).

Puisque les ordinateurs nous invitent à voir le langage en leur terme, il se voit déterminé par les capacités techniques de la machine (Golumbia, 2009). Le philosophe Martin Heidegger a considéré que le langage technologique est ce qu'il y a de plus menaçant à ce qui est propre au langage : le « dire ». « Dire » va bien au-delà de simplement énoncer des phrases ; c'est une connexion profonde entre le langage, la pensée et notre perception du monde. C'est le moyen par lequel l'humanité peut véritablement s'engager avec le monde, en exprimant et en comprenant des significations qui existent, parfois même au-delà de la réalité matérielle des choses (Heidegger, 1998). Lorsque nous utilisons le langage pour faire référence à quelque chose, nous utilisons des mots pour représenter un objet, un concept ou une idée qui n'est pas physiquement présente au moment de la communication. Pour le philosophe Georg Wilhelm Friedrich Hegel, il est impossible de saisir ou d'exprimer pleinement le référent d'un mot par le langage, car celui-ci repose sur des symboles et des concepts abstraits, alors que le référent lui-même est une expérience concrète et sensorielle (Andrejevic, 2020). En d'autres termes, nous ne pouvons jamais "dire" ou communiquer pleinement l'objet auquel nous nous référons, car l'objet est toujours plus complexe et multiforme que les mots que nous utilisons pour le décrire. À cette différence, le langage de l'automatisation reste opérationnel dans la mesure où les machines et les ordinateurs sont capables d'exécuter des tâches en utilisant des instructions et des règles établies dans le langage de programmation. C'est un langage qui ne saisit pas la complexité et la richesse des objets et des idées auxquelles il se rapporte, qui ne peut se référer à des concepts abstraits.

Malgré ces distinctions qui peuvent sembler évidentes, le langage généré par des ordinateurs reste porteur de sens et de pouvoir. C'est pourquoi il nous est important de considérer les relations de pouvoir que le langage entretient. Dans une étude réalisée sur les biais en TALN, les auteurs ont expliqué comment les différences et les variations de style dans divers langages peuvent être reliées aux différences de genre, de classe sociale, d'origine ethnique, de facteurs socioculturels et bien plus encore (Blodgett *et al.*, 2020). Les catégorisations effectuées dans différents systèmes en TALN envers certains groupes de personnes peuvent renforcer les stéréotypes et contribuer à perpétuer les inégalités sociales (*Ibid*.).

> […] many groups have sought to bring about social changes through changes in language, disrupting patterns of oppression and marginalization via so-called "gender-fair" language (Sczesny et al., 2016; Menegatti and Rubini, 2017), language that is more inclusive to people with disabilities (ADA, 2018), and language that is less dehumanizing (e.g., abandoning the use of the term "illegal" in everyday discourse on immigration in the U.S. (Rosa, 2019)). The fact that group labels are so contested is evidence of how deeply intertwined language and social hierarchies are. Taking "gender-fair" language as an example, the hope is that reducing asymmetries in language about women and men will reduce asymmetries in their social standing. Meanwhile, struggles over language use often arise from dominant social groups' desire to "control both material and symbolic resources"—i.e., "the right to decide what words will mean and to control those meanings". (*Ibid.)*

La relation entre le langage et les hiérarchies sociales est un exemple précis mettant en évidence le potentiel de préjudices possibles avec les grands modèles de langage. Le langage évolue constamment et est utilisé par certains groupes ou communautés de manière stratégique pour déstabiliser ou encore, à des fins de réappropriation du discours dominant. Les mouvements sociaux produisent de nouvelles normes et de nouvelles façons de communiquer, ce qui représente un défi considérable pour les LLMs (Bender *et al.*, 2021).

1.1.2 Le TALN et les grands modèles de langage

Jusqu'à présent, nous avons évoqué à quelques reprises le concept de traitement automatique du langage naturel (NLP), ainsi que celui de grand modèle de langage (LLMs). Nous proposons dans cette partie d'y revenir plus en profondeur pour permettre une meilleure compréhension de ce qui leur est constitutif et de la relation entre les deux.

Le TALN, dont les fondements relient la linguistique, l'informatique et les mathématiques, est un domaine de l'IA qui vise le développement d'outils capable de modéliser, traiter et générer du langage. Un modèle de langage (Language Model) est une technologie de base en TALN, c'est un outil statistique qui sert à déterminer la probabilité qu'une chaîne de mots surgisse dans une phrase (Castello et Lajeunesse, 2019). Depuis les dernières années, l'une des principales tendances dans le domaine du TALN est l'augmentation

massive de la taille[5] de ces modèles de langage, qu'on appelle ainsi les grands modèles de langage (Bender *et al.*, 2021). Les LLMs sont des réseaux de neurones qui utilisent des techniques d'apprentissages profonds (deep learning). Ces réseaux, modélisant des neurones artificiels, restent mystérieux à plusieurs égards. Ils sont en mesure de découvrir les structures d'un problème donné, pouvant être relié au langage, au son, aux images, etc. Ils ont donc une capacité de généralisation impressionnante, mais la nature des régularités découvertes reste inconnue pour les concepteurs (Mallat, 2019). Leur structure prédictive très élaborée permet de concevoir les mots en termes de leur intégration à une phrase au lieu de simplement prédire un mot en fonction du précédent. Or, comme mentionné par la numéricienne Aurélie Jean, cette technique est susceptible de développer des éléments biaisés contenus en elle-même et de ce fait, d'amplifier les conséquences des biais qui ne sont pas pour autant explicites (Jean, 2019).

> La complexité des réseaux neuronaux augmente encore dans le cas de l'apprentissage profond, le fameux deep learning, avec des invariants si solidement installés dans le réseau que toute rééducation de l'algorithme devient quasiment impossible. (*Ibid*., p.131)

En 2020, Timnit Gebru, qui était alors à la co-direction de l'équipe éthique de Google[6], a publié avec la professeure de l'université de Washington Emily Bender l'article *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big ?* Cette publication avait pour but de questionner si une réflexion importante avait été mise en place concernant l'utilisation des grands modèles de langage à grande échelle et leurs risques potentiels. Parmi ces risques et pour ne nommer que ceux-ci, elles ont mis en lumière les coûts environnementaux et financiers, les biais encodés, ainsi que les diverses possibilités de préjudice associées aux grands modèles de langage (Bender *et al.*, 2021).

Les autrices ont abordé différents types de biais pouvant être intégrés aux grands modèles de langage, soit les biais linguistiques, culturels, cognitifs et présents dans les données d'entraînement. Elles ont démontré que ces biais peuvent entraîner des préjudices d'allocation et de représentation (*Ibid*.). Il y a préjudice d'allocation lorsqu'un système octroie ou refuse à certains groupes une opportunité ou une ressource. Ce sont des biais immédiats, facilement quantifiables et de nature transactionnelle. Ce genre de préjudice peut

[5] L'augmentation de la taille des modèles de langage est mesurée par la taille des données d'apprentissage et le nombre de paramètres intégrés à un modèle (Bender *et al*., 2021).

[6] Timnit Gebru a été renvoyée de chez Google suite à la publication de cet article. Le directeur du département d'intelligence artificielle de Google, Jeff Dean, a affirmé que cet article n'était pas à la hauteur de leur standard de publication et a précisé dans un courriel envoyé aux employés de Google : "It ignored too much relevant research — for example, it talked about the environmental impact of large models, but disregarded subsequent research showing much greater efficiencies. Similarly, it raised concerns about bias in language models, but didn't take into account recent research to mitigate these issues." (Newton, 2020) Le renvoi de Gebru, extrêmement respectée dans son domaine, a suscité l'indignation de plusieurs et soulevé de nombreuses questions. Une lettre du Congrès américain signée par neuf de ses membres a été envoyée à Google demandant une explication de la situation entourant son congédiement (Hao, 2020). Des centaines de collègues ont signé une lettre ouverte demandant son retour et un engagement de Google à respecter l'intégrité de la recherche (Singh, 2020).

être en apparence qu'une simple différence de traitement et aller jusqu'au déni complet de certains services. (Crawford, 2017) En ce qui concerne les préjudices de représentation, ils renforcent la subordination de certains groupes en raison de leur identité. En d'autres mots, les préjudices de représentation surgissent lorsque le système représente certains groupes sociaux de manière défavorable, les dénigre ou encore, ne reconnaît tout simplement pas leur existence. Ce sont des biais qui peuvent être constatés à long terme puisqu'ils sont plus difficiles à formaliser en raison de leur caractère diffus et culturel. (*Ibid*.)

Avec le lancement public de chatGPT en novembre 2022 par OpenAI et de Bard[7] en mars 2023 par Google, un bon nombre de dangers contre lesquels les autrices mettaient en garde sont désormais bien réels. Il est maintenant possible de « jouer » avec des LLMs sous la forme d'agents conversationnels virtuels (chatbot). Ils sont utilisés dans le but de générer des textes, répondre à une variété de questions, composer des essais et bien plus encore. Depuis son déploiement à grande échelle, ChatGPT a créé un réel buzz, atteignant 100 millions d'utilisateurs après deux mois de mise en ligne (Skopeliti et Milmo, 2023).

De son côté, Bard a spécifiquement été lancé dans l'intention de concurrencer le populaire modèle linguistique ChatGPT d'OpenAI. Le PDG de Google, Sundar Pichai, a décrit Bard comme un service d'IA conversationnelle capable de simplifier les choses difficiles et de fournir des réponses de « qualité » (Vincent, 2023). Pourtant, dans une démonstration de Bard en février 2023, le chatbot a donné une réponse erronée concernant le télescope spatial James Webb de la Nasa (Milmo, 2023). Lorsque cette erreur a été signalée, Google a affirmé la nécessité de tester rigoureusement ces modèles avant qu'il soit disponible au grand public. Ce qui peut sembler ironique considérant leur empressement à lancer Bard publiquement moins d'un mois plus tard en mars 2023, avec comme intention principale de compétitionner ChatGPT.

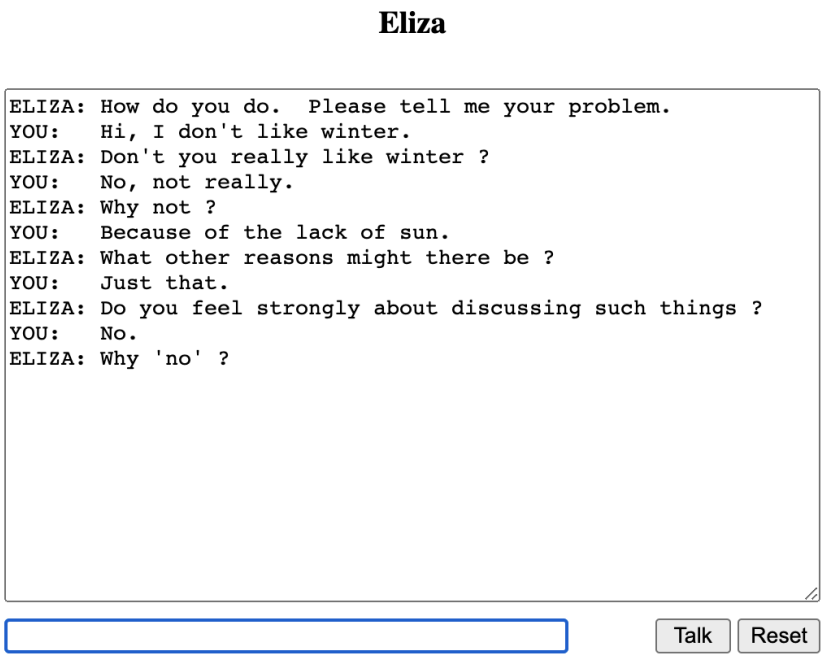
Or, il serait faux de croire qu'une fois rendus publics, ces agents conversationnels seraient sans erreurs. Par exemple, dans des questions posées par l'ingénieur Dan Atkinson sur le régime alimentaire en Angleterre au 11^{e} siècle, le chatbot de OpenAI a précisé qu'il s'agissait d'une alimentation de pommes de terre et d'autres légumes. Cependant, les pommes de terre n'existaient pas sur le territoire européen avant le 16^{e} siècle (Skopeliti et Milmo, 2023). Aktinson se dit inquiet de la confiance aveugle allouée à ces technologies et notre capacité à y voir une forme d'intelligence très rapidement (*Ibid*.).

Si ces événements technologiques nous semblent tout nouveaux, certains moments dans l'histoire de la linguistique computationnelle se sont rapprochés de notre actualité. Le professeur émérite au MIT, Joseph

[7] En date du mois d'avril 2023, Bard est seulement accessible aux États-Unis et en Angleterre.

Weizenbaum, avait dévoilé le premier chatbot[8] au monde en 1966 : Eliza (Naughton, 2023). Il l'avait créé dans le but de démontrer que les communications entre les humains et les ordinateurs ne pouvaient qu'être superficielles. Eliza était programmée selon un script médical, représentant un psychothérapeute fictif, pour prendre en compte ce qui a été écrit et produire une réponse (*Ibid.*).

Figure 1.1 – Exemple d'échange avec l'agent conversationnel[9]



Weizenbaum a voulu démontrer que même si les machines semblent en apparence pouvoir « communiquer », ce n'est qu'une illusion (*Ibid.*). Même si nous sommes maintenant des décennies plus tard et que les gains en TALN sont considérables au niveau des quantités de données entraînées et de la rapidité d'élaboration des réponses, ChatGPT reste, tout comme Eliza, une illusion de langage. Les relations humaines, les possibilités d'entendement et de créativité ne sont pas l'essence de la machine. Avec les nouvelles

[8] Le terme chatbot n'existait pas en 1966 et est apparu dans les années 1990 (Naughton, 2023).

[9] Le chatbot Eliza est toujours accessible, on peut y accéder sur https://masswerk.at/elizabot/. Cet échange est tiré d'une conversation que nous avons nous-mêmes produite (2023).

technologies déployées en TALN, qu'il s'agisse d'enjeux environnementaux importants[10], de biais encodés ou de travailleurs exploités filtrant des contenus haineux[11], les préjudices ne font que s'accélérer. L'intégrité de la recherche sur les biais en TALN et ses risques associés est plus que jamais décisive. Les études en TALN peuvent contribuer à identifier les types de biais, à développer des outils pour atténuer ces derniers, à informer la population sur ce genre de technologies et bien plus encore. C'est pourquoi nous considérons important de contribuer à adopter un regard critique dans l'analyse des représentations des biais dans différentes études en TALN.

1.2 Recension des écrits

Notre recension des écrits a été effectuée en ligne via Zeta Alpha, un moteur de recherche consacré spécifiquement aux recherches réalisées en IA. Il recense des sources importantes en linguistique computationnelle telles que arXiv, ACL, ACL Anthology et ACM. Ces sources significatives sont constituées de recherches scientifiques en TALN, de publications présentées lors de conférences importantes dans le domaine et bien plus encore. Puisqu'une bonne majorité de la recherche en TALN est réalisée en anglais, nous avons choisi de conduire notre revue de littérature dans cette langue en appliquant différents mots-clés tels que : « NLP bias », « NLP fairness » et « NLP toxicity ». Ce qui nous a permis d'identifier trois tendances considérables de la recherche en TALN.

La première tendance porte sur l'identification des biais et des risques associés dans les systèmes en TALN (Bansal, 2022 ; Bender *et al.*, 2021 ; Blodgett *et al*., 2020.) Ces recherches offrent une représentation complète des différents types de biais, en plus de conceptualiser de manière précise ce terme. Ce sont des études qui démontrent comment ces risques associés affectent certaines populations et communautés, en plus de démontrer de manière plus concrète comment ces systèmes peuvent conduire à des résultats discriminatoires ou injustes. En d'autres mots, les auteurs vont jusqu'à aborder les origines des biais et leur apparition dans les systèmes. Ces biais, touchant la race, le genre, la religion, l'invalidité et bien plus encore, peuvent être liés à différents facteurs, tel que les données utilisées pour entraîner les modèles, les choix de conception, etc. Bender *et al*. (2021) ont démontré, au travers de situations concrètes, de quelle manière il était possible de retrouver ces préjudices dans les systèmes en TALN et quels en étaient leurs impacts. Elles

[10] Dans la recherche majeure de Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP* publiée en 2019, les auteurs ont évalué les coûts de formation et de développement des grands modèles de langage en termes d'émissions estimées de CO_2. Ils ont estimé que l'entraînement d'un seul modèle de base BERT (modèle de langage en TALN développé par Google) nécessitait autant d'énergie qu'un vol transaméricain. Ces modèles sont coûteux à entraîner et à développer sur le plan environnemental en raison de l'empreinte carbone nécessaire pour les former et les alimenter (Strubell *et al.*, 2019).

[11] Selon une enquête dirigée par TIME, OpenAI a fait appel à des travailleurs kenyans pour rendre ChatGPT moins toxique en les payant moins de 2$ de l'heure. OpenAI, célébrant les gains d'efficacité de ses technologies, ne mentionne jamais l'importance et l'iniquité de cette main-d'œuvre dans le développement de ses produits (Perrigo, 2023).

ont donné l'exemple d'un Palestinien dont le message Facebook qui disait « bon matin » en arabe a été traduit à tort pour "hurt them" en anglais. Il a été arrêté par la police pour une erreur majeure du système de traduction automatique (Bender *et al*., 2021). Ce sont des articles qui invitent les lecteurs à prendre du recul, en plus des questionner les coûts associés. Qu'il s'agisse de coût environnemental, financier et humain, il est important de considérer les risques et les impacts du déploiement à grande échelle des technologies en TALN.

La seconde tendance s'intéresse à l'intersectionnalité des biais en TALN (Hessenthaler *et al*., 2022 ; Cheng *et al*., 2022 ; Lalor *et al.,* 2022). En tenant compte de la manière dont les différents biais peuvent se combiner et interagir, les chercheurs visent à mieux comprendre comment ces biais peuvent affecter différents groupes de personnes de manière disproportionnée. Ces publications soulèvent le manque de compréhension des relations entre les différentes formes de biais. La recherche *Bridging Fairness and Environmental Sustainability in Natural Language Processing* de Hessenthaler *et al.* (2022) a démontré le manque de considération de la relation entre les biais et les enjeux environnementaux en TALN. Un focus exclusif sur les biais et l'équité peut occulter la consommation énergétique des modèles en TALN. À l'inverse, une focalisation sur les impacts environnementaux peut obstruer la place que jouent les méthodes pour réduire les biais dans la consommation énergétique des modèles. Il est important de considérer l'intersectionnalité des risques pour tenir compte des multiples dimensions qui peuvent s'influencer mutuellement. De leur côté, Cheng *et al.*, (2022) ont montré que les biais sont souvent corrélés entre eux et que les approches indépendantes de réduction des biais ne peuvent pas, pour ces raisons, être suffisantes. L'atténuation intersectionnelle des biais est plus souhaitable que le débiaisage individuel.

La troisième tendance, davantage technique, concerne les recherches qui visent l'évaluation, la mesure et l'atténuation des biais en TALN (Antoniak et Mimno, 2021 ; Guo *et al.*, 2022). Ces dernières, souvent davantage axées sur l'entraînement et la configuration des ML, comportent de nombreux calculs mathématiques d'algèbre linéaire, de statistique et bien plus encore. Les chercheurs travaillant à la création de méthodes pour réduire les biais en TALN, tentent de développer ou suggérer des moyens pour identifier les biais dans les modèles et les données d'entraînement. Ils développent également des métriques pour quantifier les biais dans les modèles et tentent de créent des bases de données équitables[12] pour atténuer les développements de biais. De cette tendance, nous identifions plusieurs publications (Dev *et al*., 2021 ; van der Wal *et al.*, 2022) qui expriment les difficultés pour la recherche de mesurer avec précision les biais en

[12] Une base de données plus équitable en NLP est construite de manière à minimiser les biais et inégalités qui peuvent être présents dans les données. Par exemple, Antoniak et Minmo (2021) ont émis des recommandations pour la sélection des ensembles de données initiaux en NLP visant à comprendre d'où proviennent les données, à mettre l'accent sur la compréhension des risques potentiels associés à la source, à effectuer des évaluations manuelles, etc. (Antoniak et Minmo, 2021)

préjudices en TALN, en démontrant l'incapacité des chercheurs à définir ces concepts et les mesures utilisées pour les identifier. Ces mesures ne sont pas explicitées, du moins, il n'est pas possible de savoir ce qu'elles évaluent et dans quelle proportion elles sont sujettes à des erreurs. Ces publications visent donc à proposer un cadre pratique servant de lignes directrices pour bien définir et comprendre comment mesurer les biais en préjudices en TALN.

Nous avons constaté que la recherche sur les biais en TALN est complexe et nécessite une approche méthodologique adaptée à chaque contexte d'étude. Cela peut conduire à des résultats présentés de différentes manières en fonction des objectifs de la recherche et de son public cible. Selon les diverses disciplines s'intéressant au sujet, nous remarquons que la contextualisation des biais est abordée à travers les tendances mentionnées précédemment qui englobent différentes perspectives, à savoir sociales, éthiques, intersectionnelles et techniques. La perspective sociale et éthique, que nous considérons comme liées, contextualise les biais en mettant l'accent sur la façon dont les biais peuvent affecter de manière disproportionnée certaines populations et communautés, ainsi que sur la manière dont ces systèmes peuvent engendrer des résultats discriminatoires ou injustes. La perspective intersectionnelle contextualise les biais en reconnaissant que les individus et les groupes sont affectés de manière différente en raison de l'intéraction complexe entres les différents biais, qui peuvent ultimement se combiner. Enfin, la perspective technique contextualise les biais en les considérant comme des problèmes à résoudre à travers des méthodes techniques et des outils d'évaluation.

Il va sans dire que c'est un domaine d'étude en pleine effervescence dans lequel il est possible de souligner certaines difficultés que peuvent rencontrer les chercheurs. Ces derniers peuvent être influencés par des biais ou des hypothèses personnels, ce qui peut conduire à utiliser des méthodes qui ne sont pas alignées à leurs objectifs de recherche. Par exemple, en supposant que les biais se retrouvent dans les données, un chercheur pourrait se concentrer sur leur évaluation et ne pas considérer l'apport possible des algorithmes dans son objet d'étude. La recherche de Blodgett et al. (2020) a démontré, en analysant 146 publications sur les biais en TALN, que 32% de celles-ci ne sont pas motivées par aucune raison normative, mais bien par la performance du système. Même lorsque les recherches exprimaient des motivations précises, elles étaient souvent imprécises à développer en quoi les comportements du système décrits comme « biaisés » sont préjudiciables, de quelle manière et envers quel groupe. Puis, les auteurs ont démontré que les articles sur les systèmes en TALN conceptualisent les biais de différentes manières, ce qui conduit à des résultats extrêmement variés, et ce, même lorsqu'il s'agit d'une même problématique (*Ibid*.). De plus, certains auteurs sont confrontés à des pressions externes ou corporatives et c'est ce que nous avons pu constater avec des publications telles que *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big ?*

(Bender et *al*., 2020*).* Google, qui investit massivement dans les développements des technologies en TALN, avait demandé le retrait de cette publication pour finalement renvoyer ses employées qui ont décidé de respecter l'intégrité de la recherche en allant de l'avant avec sa diffusion (Hao, 2020).

1.3 Question de recherche

La littérature sur les biais en TALN est abondante et en constante évolution. En raison d'un développement technologique continu, c'est un domaine de recherche qui a pris une importance majeure au courant des dernières années. Nous pouvons constater le dynamisme de ce champ de recherche et sa pertinence actuelle en considérant tous les enjeux qui l'accompagnent, ainsi que la variété des chercheurs provenant de différents domaines (linguistique, informatique, sociologie, etc.) qui s'y intéressent. Nous soutenons l'importance communicationnelle de la juste compréhension des objets techniques et de leur déploiement à grande échelle. Il est nécessaire de préserver notre capacité à saisir les représentations, les enjeux et les débats qui façonnent le domaine du TALN.

En considérant plusieurs lacunes pointées par des auteurs qui ont étudié le sujet, notre recension des écrits nous a amenés à remettre en question la façon dont différentes études conçoivent et présentent les biais en TALN. C'est donc compte tenu de ce qui précède que plusieurs réflexions sur le sujet ont mené à établir précisément la question de recherche suivante : De quelle façon les différentes recherches en traitement automatique du langage naturel étudient-elles les biais de ce domaine de l'intelligence artificielle ? Cette question de recherche principale reste passablement vaste puisqu'elle sera raffinée et encadrée par des sous-questions spécifiques, soit : Quelles similitudes et différences peut-on observer dans la manière dont les chercheurs étudient les biais dans notre corpus ? Quels constats est-il possible de tirer des différentes approches utilisées ? Quelles limites peut-on observer dans la façon de s'engager avec les biais ? Nous considérons que ces sous-questions offrent un angle d'analyse et un cadrage permettant de communiquer de manière compréhensible la façon dont les études sont construites.

1.4 Pertinence communicationnelle

Ce mémoire, relevant de la recherche en communication, s'inscrit dans une période de réflexion importante sur les impacts du développement des technologies en IA. Il contribuera à une meilleure compréhension du sujet, en plus d'encourager des pratiques de conception plus équitables et promouvoir une utilisation responsable. Les modèles en TALN sont de plus en plus utilisés dans divers champs d'application qui impactent directement notre façon de communiquer. Il suffit de voir l'engouement concernant la sortie de ChatGPT pour constater l'ampleur de l'arrivée de ces technologies dans nos sociétés. C'est pourquoi il est important, d'un point de vue social et communicationnel, de considérer les potentiels risques d'utilisation

de ce genre de technologies, ainsi que la recherche qui les constitue. Les systèmes en TALN reforment notre manière d'interagir, d'écrire, de chercher l'information, de s'interroger, et bien plus encore. Nous soutenons qu'il est de notre intérêt de s'intéresser et de tenter de comprendre, non pas comment cette technique fonctionne, mais comment son déploiement à grande échelle nous affecte. Il est nécessaire de valoriser la compréhension des enjeux reliés aux technologies qui sont déployées à l'heure actuelle.

1.5 Objectifs de recherche

Avec ce projet de mémoire, nous souhaitons présenter un portrait approfondi, détaillé et nuancé de la façon dont se développe la recherche sur les biais en TALN, en mettant de l'avant des éléments caractéristiques et significatifs des représentations de ces derniers. Si les préjudices qu'il est possible de retrouver dans les systèmes en TALN sont le reflet de ce que l'on peut retrouver dans nos sociétés, et ce, de manière beaucoup plus insidieuse et difficile à voir, nous sommes en droit de questionner la vitesse de déploiement, voire compétitive, de ces technologies. L'accueil enthousiaste et la fétichisation de cette technologie par le marché, les médias et certains consommateurs suscitent des interrogations quant aux implications sociales et éthiques qui en découlent. Notre objectif est donc de recueillir des données pertinentes à notre objet d'étude nous permettant de proposer une analyse qualitative compréhensive et approfondie.

CHAPITRE 2
CADRE THÉORIQUE

Notre mémoire est structuré par plusieurs concepts se reliant les uns aux autres. Le cadre théorique que nous allons présenter permettra de saisir ces relations, tout en situant notre travail dans un contexte plus large qui englobe les théories et les structures qui le constituent. Dans le prochain chapitre, nous poserons les bases théoriques de notre recherche qui nous permettront de guider notre analyse. Nous commencerons par l'automatisation, l'opérationnalisme et le langage opérationnel des processus de communication qui incorporent des idéologies présentent dans le développement des technologies en TALN. Le concept de biais algorithmique suivra ensuite et sera présenté sous différents angles permettant de lui rendre sa complexité et ses diverses compositions. Puis, nous terminerons par aborder l'éthique technologique qui est un concept d'une importance indéniable pour l'angle de recherche que nous avons choisi.

2.1 L'automatisation

Le développement technologique massif est indissociable du concept d'automatisation, qui propose de surpasser les limites humaines avec un développement massif de divers types de technologies (Andrejevic, 2020). L'automatisation consiste à créer, mettre en place et utiliser des systèmes, des machines, des logiciels ou d'autres technologies qui peuvent effectuer des tâches de manière autonome, sans nécessiter une intervention directe de la part des êtres humains (*Ibid*.). Mark Andrejevic, professeur d'études sur les médias à l'université de Monash, explique dans son ouvrage *Automated Media* que les formes d'automatisation dont il est question à l'heure actuelle ne sont pas simplement mécaniques, mais bien informationnelles. (*Ibid.)* L'IA nous propose d'automatiser le travail mental en remplaçant le rôle de l'humain dans la communication et le traitement de l'information en alimentant continuellement la promesse d'une plus grande rapidité, d'une plus grande efficacité et d'une plus grande puissance (*Ibid.)*. Entraînés sur des données, ces systèmes se déploient, précise Chomsky, comme des machines de « plagiat hi-tech » qui s'approprient sans autorisation le travail créatif de millions de personnes (Naughton, 2023).

L'historique de l'automatisation s'appuie sur un long parcours de déqualification de l'humain dans ses fonctions les plus naturelles (Andrejevic, 2020). Andrejevic explique :

> Automated media offload sociality onto digital systems just as industrialized mass production allowed for the offloading of physical labor onto machines, assembly lines, and, eventually, robots. Just as physical de-skilling paved the way for mechanized automation, social de-skilling anticipates new forms of data-driven automation. This has been the refrain of critics, ranging from the philosopher Hubert Dreyfus (2013) to cultural critic Nicholas Carr (2010) and MIT

> professor Sherry Turkle (2010). The automation of physical labor reserved for humans the role of planner and controller: the overseers in charge of the mechanized production line. Under changed societal conditions, automation also promised the possibility of liberation from labor as drudgery. By contrast, the automation of communicative processes envisions a surpassing of the pace and scale of human thought and interaction, which is why the technological imaginary tends toward post-humanism. (*Ibid,* p.7)

L'auteur explique que si les systèmes automatisés peuvent dépasser nos capacités physiques et mentales, éviter l'obsolescence implique à ce jour une incorporation de la machine dans nos vies. Le paradoxe ici, pointé par Andrejevic, est qu'une telle fusion de l'humain avec la machine nous est présentée comme si le sujet pouvait être préservé, alors que ce sont ses propres capacités que l'on tente de remplacer (*Ibid.)*. Les systèmes en TALN incorporent ce qu'il qualifie de logique de l'automatisation : une collecte automatisée de données permettant le traitement automatisé de ces dernières et menant à des réponses automatisées (*Ibid.)*.

L'automatisation considère l'IA, en tant que grand « autre » rempli de données, qui serait capable de donner un sens au monde qui échappe à l'humain (*Ibid*).

> The technological fantasy of automated information processing is that for the first time in human history, instead of simply conceding the impossibility of absorbing and making sense of all available information or relegating the process to an inaccessible metaphysical position, humans anticipate the possibility of building such a position for themselves – and putting it to work in the form of an enslaved mechanical god (*Ibid*., p.2)

Ce traitement automatisé de l'information permettrait aux humains de jouer un rôle plus actif dans la gestion de l'interprétation de vastes quantités d'informations grâce à la technologie. En devenant l'opérateur et l'objet même de ces technologies, l'automatisation positionne l'humain dans une relation technique au monde (Feenberg, 2005). Cette relation technique au monde tend à éviter les réflexions subjectives et à définir l'intelligence humaine en termes d'automatisation. La fantaisie technologique dont parle Andrejevic représente la promesse que si nous avons assez de données, nous pouvons tout automatiser. Puis, en utilisant des systèmes automatisés pour donner du sens à ces données, il n'y a plus de réelle compréhension, seulement une suite logique d'opérations (Andrejevic, 2020).

2.1.1 L'opérationnalisme

Andrejevic, en référence au théoricien de la communication Harold Innis, décrit l'opérationnalisme comme étant l'un des biais de l'automatisation (Andrejevic, 2020).

> "Operationalism" refers to the displacement of narrative accounts and explanations by automated responses. Automated systems do not seek to understand but to act: they are not representational but operational (*Ibid*., p.18).

L'opérationnalisme est une approche qui vise à simplifier et à rationaliser les processus automatisés en termes d'opérations concrètes. Les concepts fondamentaux qui permettent de comprendre et d'expliquer la signification et la nature des choses sont écartés au profit d'aspects observables et quantifiables (*Ibid*.). Par exemple, plutôt que de se baser sur des critères théoriques, grammaticaux et linguistiques, l'opérationnalisme en TALN mesure la qualité d'un système en se concentrant uniquement sur ses performances dans des tâches spécifiques. Les ML entraînés sur de grands ensembles de données peuvent écrire des textes, les traduire et bien plus encore, sans pour autant qu'il y ait une compréhension dans les choix qui sont effectués. Pour Andrejevic, il n'est pas surprenant que la transparence soit un critère si important à l'heure actuelle puisque nous tentons de trouver une façon de se représenter ce qui ne peut l'être (*Ibid*.). Nous essayons de comprendre l'impossible, c'est-à-dire, les interactions complexes entre des milliards de données, ainsi que les résultats qu'elles produisent (*Ibid*.).

> Even if the databases were to be thrown open by force, the decision-making process has retreated into the neural nets and emergent processes of data mining and machine learning. The trajectory of automated operationalism does not just raise important issues of accountability; it also collapses the space of representation and thus of politics. […] In practical terms, the goal is to develop automated systems that make decisions that govern life, liberty, and opportunity. In this respect, automation embraces the logic of immediation (the invisibility or disappearance of the medium) that parallels the promise of virtual reality. […] operationalism offloads the labor of civic life onto automated systems – it envisions the perfection of social life through its obliteration. Always-on connectivity has culminated in the pathology of communication overload, which, in turn, provides data for the machine-learning systems that promise to solve the problems they have created by automating our sociality for us (*Ibid.,* p.109).

En d'autres mots, l'opérationnalisme encourage l'automatisation des processus sociaux, en utilisant des méthodes quantitatives et des technologies pour réguler les interactions humaines. Ce qui réduit la participation active des individus dans la vie sociale, tout en confiant cette responsabilité à des systèmes automatisés qui en éliminent les aspects les plus subjectifs. La complexité de nos processus de communication est réduite et transformée en opérations mesurables de symboles mathématiques. Andrejevic soutient que l'effondrement de la représentation dans l'opérationalisme rend extrêmement difficile de comprendre les décisions qui sont prises par des systèmes automatisés. En d'autres mots, en automatisant le social, il devient excessivement compliqué de discerner les raisons des résultats obtenus et nous concédons à soustraire le jugement humain de toute prise de décision (*Ibid*.).

2.1.1.1 Langage opérationnel

En suivant la logique de l'automatisation, le « langage » des machines se veut opérationnel. C'est-à-dire qu'en excluant la possibilité de penser l'écart entre les mots et les choses, le langage opérationnel ne laisse pas la place à l'interprétation et supprime l'espace nécessaire pour penser le sens. Cette aspiration à vouloir rendre le langage dépourvu de complexité, de sens cachés et d'ambiguïtés est fondamentale à l'opérationnalisme (Andrejevic, 2020).

Le langage opérationnel repose sur l'idée même que le langage ne serait pas une caractéristique spécifique à l'humain et donc, partageable avec la machine (Heidegger, 1998). Évidemment, une telle affirmation n'est possible que dans la mesure où l'on considère que ce qui est singulier au langage peut se réduire à une simple transmission de signaux (*Ibid*.). La machine peut disposer d'un code, mais il sera toujours distinctif du langage qui donne l'espace nécessaire pour penser ce qui est inexprimable (*Ibid*.). Le langage opérationnel est marqué par l'impossibilité de sa complétude. En revanche, ce que promet le langage opérationnel, c'est la possibilité de modéliser le social. C'est l'idée d'un accès immédiat au réel lui-même dans un langage inaccessible : celui du code. Les données agissent comme des signaux quantifiables dépourvus de toute interprétation et signification (Rouvroy, 2018).

Or, Andrejevic précise que la notion même de « langage » est erronée, voire trompeuse dans ce contexte. Les machines ne sont pas des sujets linguistiques et c'est le code qu'elles produisent qui est opérationnel. « […] – it does not refer to an absent referent but collapses the signifier into the signified: this is the logic of code and of the operational image/symbol. » (Andrejevic, 2020, p.128). Il s'agit donc de créer du sens à partir de ce qui ne passe même plus par le langage et qui relève seulement de la transcription. Il ne s'agit pas non plus d'interprétation, mais bien de repérage et modélisation statistique des données (Rouvroy, 2018). Cette attribution de « langage » pour décrire des systèmes automatisés contribue à ce que plusieurs qualifient de *IA hype*[13]. En gagnant de la visibilité et en s'appliquant dans divers contextes, il est important d'ajuster notre manière de référer aux systèmes NLP pour ne pas tromper le public envers leur capacité. Pour bien comprendre ce point, il faut distinguer l'utilisation du langage chez l'humain. Quand l'humain utilise le langage, il le fait avec une intention communicationnelle.

> Communicative intents are about something that is *outside of language* . When we say *Open the window!* Or *When was Malala Yousafzai born ?*, the communicative intent is grounded in the real world the speaker and listener inhabit together. Communicative intents can also be

[13] *AI hype* est utilisé pour décrire cet enthousiasme excessif envers les capacités de l'IA (Bender et Koller, 2020).

> about abstract worlds, e.g. bank accounts, computer file systems, or a purely hypothetical world in the speaker's mind (Bender et Koller, 2020).

Ce monde réel qui permet au langage d'exister s'appuie sur le contexte, y compris des indices sociaux, culturels et émotionnels. L'intention communicationnelle implique les mots utilisés, mais aussi le ton, le langage corporel, le contexte. Le code, quant à lui, est dépourvu de cette richesse contextuelle et d'ambiguïté. Alors que l'ambiguïté est une composante du langage humain et non une source d'erreur (*Ibid*.).

2.2 Biais algorithmique

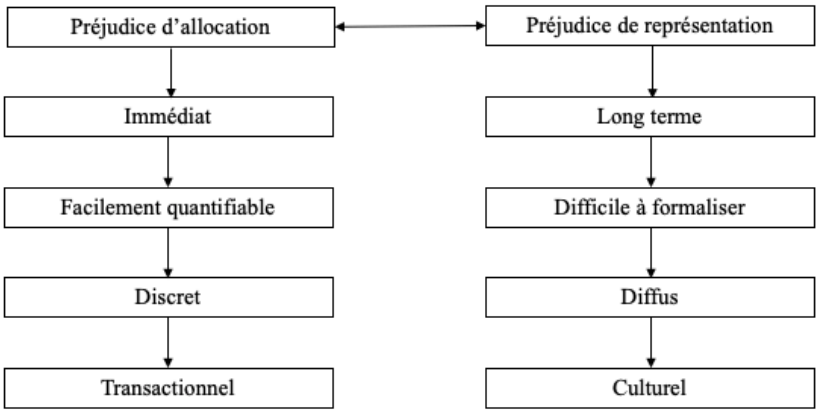
> « Data will always bear the marks of its history.
> That is human history held in those data sets. »
> Kate Crawford, 2017

En passant du concept d'opérationnalisme d'Andrejevic à celui de biais algorithmique de Crawford, nous élargissons notre cadre théorique pour englober les nuances complexes de ces derniers. Alors que l'opérationnalisme met en lumière la manière dont les systèmes algorithmiques sont déployés et utilisés au quotidien, en se concentrant notamment sur les implications en matière de surveillance, de contrôle sociale et de perturbation des significations symboliques du langage humain, le concept de biais algorithmique examine de près les injustices intégrées dans ces systèmes. En enchainant ces concepts, nous pouvons mieux comprendre comment les biais algorithmiques peuvent être exacerbés lors de la mise en œuvre opérationnelle des systèmes en TALN. Or, selon Kate Crawford, chercheuse australienne spécialiste de l'IA, le terme « biais » reste trop étroit pour les types de répercussions et problèmes dont il est question. Les formes de discrimination ne sont pas qu'intégrées dans les systèmes, elles font aussi parties de leur construction et manière de voir le monde (Corbyn, 2021). Certes, dans le but de conceptualiser le terme de biais, il est important de préciser ce que nous entendons par ce dernier. Nous considérons que le concept de biais est utilisé pour décrire un large éventail de comportements nuisibles du système qui, pour une multitude de raisons possibles, créent différents types de préjudices, envers différents groupes de personnes et communautés (Blodgett et *al*., 2020).

Nous ajoutons à cette conceptualisation l'importance de prendre en considération les types de préjudices possibles créés par les biais. Pour les catégoriser, nous utiliserons les termes de Crawford (2017) qui en distingue deux types, soit les préjudices d'allocation (allocational harm) et les préjudices de représentation

(representational harm) dont nous avons parlé un peu plus tôt dans ce mémoire[14] (Crawford, 2017). La figure 2.1 présente les caractéristiques associées à ces deux catégories.

Figure 2.1 – Préjudice d'allocation et de représentation selon Crawford (2017)[15]



En raison du caractère diffus des préjudices de représentation qui sont plus difficiles à formaliser, Crawford a développé une catégorisation de ces derniers en cinq termes : la dénigration, les stéréotypes, la reconnaissance, la sous-représentation et l'ex-nomiation (Crawford, 2017).

- Dénigration

La dénigration survient lorsque les systèmes utilisent des étiquettes culturellement offensantes ou inappropriées pour identifier des personnes. Un des exemples bien connus de dénigration fut le cas de Google en 2015 dont l'algorithme de reconnaissance d'image a identifié deux personnes noires comme étant des gorilles (Hern, 2018).

- Stéréotype

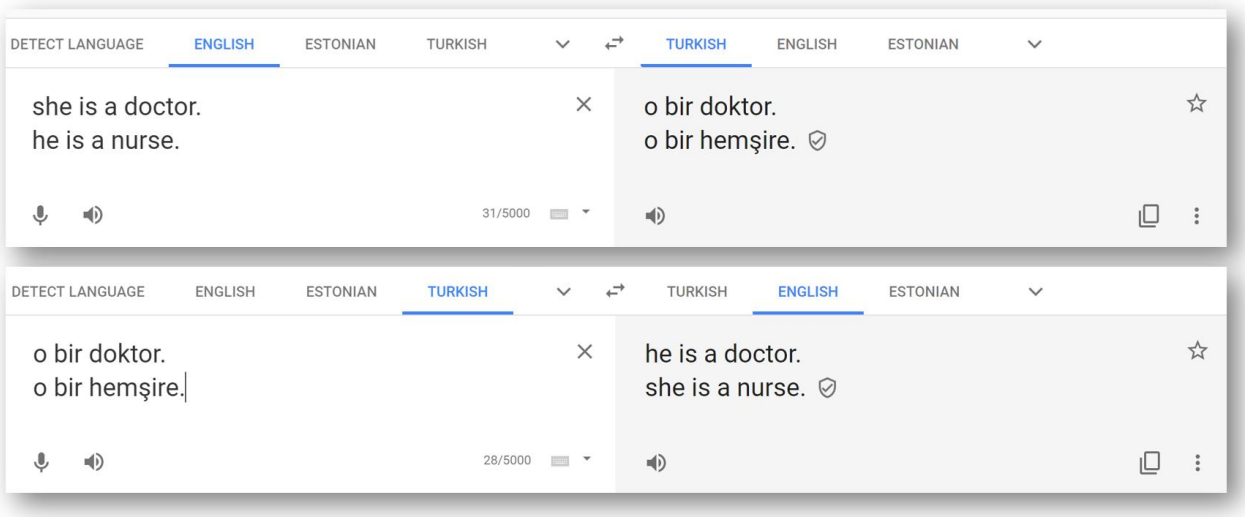
Les stéréotypes en IA surviennent lorsque des associations préjudiciables sont faites en raison du genre, de la culture, de la religion, des professions, etc. Une recherche de 2016 sur les stéréotypes de genres a

[14] La page 11 du présent mémoire présente les définitions des préjudices d'allocation et de représentation.

[15] La figure 2.1 est basée et traduite de la conférence *Neural Information Processing Systems* (NIPS) de Crawford en 2017.

démontré que Google Translate inversait les pronoms et les professions en traduisant d'un langage neutre comme le turc (Crawford, 2017).

Figure 2.2 – Exemple de mauvaise traduction réalisée par Google Translate en 2016[16]



- Reconnaissance

Un préjudice de reconnaissance se produit lorsque le système ne reconnaît tout simplement pas une personne, invisibilisant ainsi son humanité et sa dignité. C'est un indicateur fort à savoir si le système fonctionne pour tous ou s'il a été pensé pour un groupe de personnes privilégié (Kumaraswamy, 2017). Joy Buolamwini, fondatrice de Algorithmic Justice League, en a fait l'expérience en 2017 lorsqu'un algorithme de reconnaissance faciale n'a pas été capable de reconnaître son visage et a mieux performé lorsqu'elle s'est mise un masque blanc (Tucker, 2017).

- Sous-représentation

La sous-représentation survient lorsqu'un groupe ou une communauté est complètement non représenté par l'algorithme. En guise d'exemple, Crawford parle de la recherche d'image sur les PDG dans un moteur de recherche qui n'a donné qu'une seule femme et qui était principalement constituée d'hommes blancs (Crawford, 2017).

[16] La figure 2.2 est tirée de la conférence *Neural Information Processing Systems* (NIPS) de Crawford en 2017.

- Ex-nomination

> L'ex-nomination est un terme forgé par le philosophe français Roland Barthes. Crawford explique comment certaines catégories humaines dominantes, telles que la blancheur, la masculinité et l'hétérosexualité, ont été traitées comme des normes humaines centrales. Ce processus de masquage a donc permis de neutraliser certaines idéologies et normes, de rendre leurs processus de construction invisibles, de les situer au-dessus de tout besoin de catégorisation, en plus de devenir tout simplement constitutifs de la culture occidentale (Crawford, 2009).

Évidemment, ces termes utilisés par catégoriser les préjudices de représentation, peuvent parfois sembler lointain pour l'évolution rapide des technologies en IA. Certains pourraient argumenter que ces exemples sont désuets, puisque les grandes entreprises technologiques ont rapidement voulu réparer ces erreurs lorsqu'elles étaient soulevées. Or, il est important de ne pas banaliser ou invisibiliser les préjudices passés ayant impacté diverses communautés sous prétexte qu'ils sont maintenant résorbés.

Dans une série d'articles intitulés *Artificial Intelligence 101*, visant à dresser un portrait général de la réglementation actuelle en IA, l'autrice Mar Estrach explique que les biais dans les systèmes d'IA sont souvent le reflet de problèmes socialement ancrés dans nos sociétés (Estrach, 2022). Elle précise que la grande majorité des concepteurs qui travaillent sur ces systèmes ont un profil similaire et sont, pour un bon nombre d'entre eux, issus des mêmes sphères économiques et sociales (*Ibid*.). Lorsque les différents systèmes d'IA sont mis en œuvre, ils se déploient dans des contextes sociaux souvent très différents de leur entraînement initial. La manière de concevoir ce qui est équitable peut être extrêmement différente dans divers groupes et communautés. Nous avons tendance à croire que l'objectivité potentielle de la machine, en d'autres mots la neutralité technique, est quelque chose de tangible et qu'il serait possible de créer des systèmes sans biais dans une société composée de ces derniers (*Ibid*.).

Safiya Noble, professeure d'études de genre et d'études afro-américaines à l'Université de Californie, a expliqué dans son livre *Algorithms of Opression : How Search Engines Reinforce Racism* (2018) qu'il est trompeur de croire en l'objectivité de la machine. Les systèmes automatisés sont le reflet, de manière intentionnelle ou non, de la discrimination systémique et ont pour effet de renforcir les préjudices envers certains groupes (Noble, 2018). Selon Noble, les systèmes automatisés vont devenir l'un des enjeux les plus majeurs au niveau des droits humains au cours du siècle prochain. Elle exprime :

> I strongly believe that, because machine-learning algorithms and projects are using data that is already biased, incomplete, flawed, and [we are] teaching machines how to make decisions

> based on that information. We know [that's] going to lead to a variety of disparate outcomes. Let me just add that AI will be harder and harder to intervene upon because it will become less clear what data has been used to inform the making of that AI, or the making of those systems. There are many different kinds of data sets, for example, that are not standardized, that are coalescing to make decisions (*Ibid*.).

En affirmant qu'il deviendra de plus en plus difficile de savoir quelles données ont été utilisées pour bâtir certains modèles en IA, Noble touche un point particulièrement présent dans les LLMs. Par exemple, le modèle GPT-3 a été entraîné à l'aide de bases de données textuelles provenant de l'internet constituant 570 Go de données, ce qui équivaut à environ 300 milliards de mots (Hughes, 2023). Les données d'entraînement de ces modèles sont la propriété exclusive de OpenAI qui a choisi de ne pas les rendre publiques pour des raisons de confidentialité et de sécurité (Gebru *et al*., 2023). La taille des données sur lesquelles un système est entraîné ne garantit pas la diversité (Bender et *al*., 2021) et rend, précise Noble, les interventions difficiles de par son exhaustivité et sa complexité. Évidemment, un entraînement basé sur de grandes quantités de texte sur le Web n'est pas représentatif de toute l'humanité. Cette automatisation, basée sur des quantités de données massives, risque de perpétuer les points de vue dominants, en plus d'accroître les déséquilibres de pouvoir (*Ibid.)*.

2.2.1 Biais épistémologiques des représentations de l'IA

Il existe plusieurs biais épistémologiques sous-jacents aux représentations de l'intelligence artificielle qui sont importants à considérer pour permettre une compréhension plus complète et représentative de la réalité. Pour les besoins de ce travail, nous en avons soulevé deux que nous considérons comme jouant un rôle fondamental dans la manière de façonner la connaissance, la manière de penser et de raisonner en IA, soit la neutralité technique et le solutionnisme technologique.

2.2.1.1 La neutralité technique

La neutralité technique, que nous avons déjà mentionnée précédemment, est une approche constitutive des représentations de l'IA qui invite à penser la technique comme un outil impartial à la disposition des humains. Elle est aussi l'essence fondamentale de la poursuite de systèmes en IA qui seraient dépourvus de tous biais. Cette idée de la « bonne intention » de l'usage de l'outil qui en déterminerait le caractère bon ou mauvais est complètement inadéquate, voire même dangereuse pour permettre une réflexion valide sur le sujet.

Vincent Bontems, philosophe des sciences et des techniques, explique que pour Gilbert Simondon[17] les objets techniques :

> […] sont des dispositifs qui possèdent une certaine autonomie et par conséquent qui ne vont pas se mettre sous la coupe d'une régulation aussi simple que de prolonger nos intentions. Les techniques ont un fonctionnement, elles vont diffuser des actions, elles vont diffuser des valeurs, elles vont communiquer. C'est trop facile et c'est un mensonge destiné à rassurer les gens que de présenter ça comme si c'était toujours du niveau des outils. (Bontems, 2017)

Le sociologue français Jacques Ellul va lui aussi remettre en question cette représentation de la neutralité technique. Pour lui, croire que c'est l'usage que l'on fait de la technique qui détermine si c'est bon ou mauvais consiste à méconnaître la technique (Ellul, 2012). Dans son livre *Le bluff technologique* paru en 1988, il explique que la technique n'est pas une histoire d'intention et qu'elle comprend en elle-même des conséquences négatives ou positives. Même s'il est vrai que l'usage peut orienter les effets de la technique, elle contient indéniablement des potentialités qui, indépendantes aux fins poursuivies, seront inévitablement exploitées (*Ibid*.). Aujourd'hui, les fins poursuivies par la technique ne sont pas clairement définies ou alors, tout simplement mal formulées. Elle se développe dans une logique d'auto-accroissement qui échappe trop souvent à la maîtrise des humains.

Alors que les technologies en TALN développent des modèles capables de générer des textes et de « communiquer », la présentation de ces systèmes comme étant des outils à notre disposition empêchent de penser la façon dont ils sont déployés. Il ne suffit pas de se concentrer sur l'entraînement des données pour rendre justice à la complexité des biais et leur encrage insidieux dans les systèmes. En masquant les inégalités réelles reproduites par la technique qui bien trop souvent s'avèrent inévitables, la neutralité technique est une approche qui ne permet pas de penser en profondeur les phénomènes et préjudices liés au développement technologique. De plus, c'est aussi cette représentation mensongère de l'IA qui entretient l'hypothèse selon laquelle les systèmes automatisés pourraient un jour se débarrasser de tous les biais et atteindre une neutralité technique insaisissable (Andrejevic, 2020). L'idée même de l'obtention de la neutralité technique est non seulement erronée, mais aussi fondatrice de la promesse d'objectiver le monde (Rouvroy, 2018). Ce qui nous est promis est enraciné à l'idéologie technique : la neutralisation de l'état de fait en présentant le monde sous forme de données (*Ibid*.).

> Data and data sets are not objective; they are creations of human design. We give numbers to their voice, draw inferences from them, and define their meaning through our interpretations.

[17] Gilbert Simondon est un philosophe de la technique du 20e siècle. Son œuvre majeure *Du mode d'existence des objets techniques* (1958) porte sur les objets techniques et leur relation aux individus.

> […] in a society that it is already biased an AI unbiased system can hardly be created, and since we can do better than what is done as of today, we should do better. […] maybe the greatest challenge is changing the cultural attitude of the science community. […] there is a tendency to believe objectiveness is always present in investigations when, often, the result itself is not objective. (Estrach, 2022)

2.2.1.2 Le solutionnisme technologique

Le solutionnisme technologique est un courant de pensée qui soutient l'idée que la technologie peut apporter des solutions pour résoudre les grands problèmes du monde (Laugée, 2015). Nous avons choisi de le cadrer pour notre analyse, puisque nous considérons ce concept comme étant fondamentalement liés aux ambitions propagées par l'idéologie technique de l'IA. C'est le chercheur d'origine biélorusse Evgeny Morozov qui a inventé le concept en 2014 dans son ouvrage *To Save Everything, Click Here : The Folly of Technological Solutionism* pour décrire la tendance à considérer la technologie comme la solution à tous les problèmes (Morozov, 2014).

Les grands ML, formés sur des réseaux de neurones, ont connu des gains de performance et de précision considérables dans plusieurs opérations en traitement automatique du langage. Cependant, ces gains se sont aussi accompagnés d'une plus grande consommation d'énergie et d'un plus grand besoin de ressources pour produire ces modèles. La recherche majeure de Strubell et al., *Energy and Policy Considerations for Deep Learning in N*LP (2019), a démontré que les coûts de formation et de développement des grands modèles de langage[18], en termes d'émissions estimées de CO_2, nécessitaient autant d'énergie qu'un vol transaméricain. Les biais en TALN, ainsi que les impacts environnementaux, sont des concepts intersectionnellement liés (Hessenthaler *et al.*, 2022). Plus les systèmes en TALN possèdent des biais, plus ils nécessitent de *fine-tuning* (mise au point) pour les réentraîner et les ajuster. Autrement dit, plus ils nécessitent une grande consommation d'énergie et de ressources (*Ibid*.).

Le journaliste français Guillaume Pitron s'est intéressé à l'industrie digitale et a mené un travail colossal pour comprendre les coûts environnementaux liés au numérique. Il a exprimé que cette industrie, qui nous semble immatérielle, est pourtant poussée par une forte consommation de ressources[19] (Pitron, 2022). Le solutionnisme technologique se retrouve dans cette idée selon laquelle les nouvelles technologies seraient

[18] Leur étude était basée sur le grand modèle de langage BERT développé par Google.

[19] Dans son livre *L'enfer numérique : voyage au bout d'un like* (2021), Guillaume Pitron expose comment les fondements matériels des technologies numériques, tels que les centres de stockage de données, les câbles sous-marins, ainsi que les métaux rares utilisés, contribuent au coût écologique majeur cette industrie.

une force intrinsèquement liée à la résolution des problèmes auxquels elles ont elles-mêmes contribué, telle que la pollution. Pitron explique que :

> Il faut en effet constater que pas un jour ne passe sans qu'on nous annonce l'avènement d'une nouvelle technologie censée permettre une gestion plus efficiente de l'information, un stockage plus efficace, etc. Ce serait une sorte de cercle vertueux, la bonne technologie venant corriger les effets néfastes de la technologie dont on s'est rendu compte après-coup des défauts. Il s'agit effectivement d'un véritable solutionnisme technologique. Il y a un problème de concordance des temps dans le débat qui oppose les critiques et ceux qui soutiennent ces technologies. Le critique regarde ce qu'il se passe aujourd'hui quand le technologue regarde ce qu'il se passe demain. Quand on fait le constat d'un problème à un instant T, en 2021, souvent la réponse du technologue consiste à dire : « Pourquoi vous vous intéressez à cette question-là puisque le futur lui a déjà donné tort, puisque les technologies futures vont régler le problème ? » (*Ibid*.).

La question temporelle de son affirmation est extrêmement pertinente et essentielle à la compréhension du concept. Pitron explique qu'en agissant ainsi, on ne s'intéresse plus au présent, mais bien à ce dit futur qui serait possiblement en mesure de régler de nombreux enjeux, et ce, en reposant sur une idée de progrès inexistant pour le moment. Ce procédé rhétorique est constitutif du solutionnisme technologique et créer une réalité alternative dangereuse face à cette incapacité de parler du présent. Cette logique est particulièrement vraie en TALN, où il est constamment question d'une nouvelle technologique mieux entraînée, pouvant corriger les biais de la précédente, permettant un entraînement plus efficace, permettant de réduire la consommation d'énergie, etc. Cette confiance excessive dans le développement technologique occulte certains impacts importants, tels que les répercussions sociales et environnementales, en plus d'empêcher la possibilité de les penser dans le moment présent. Pour le chercheur Evgeny Morozov, cet état d'esprit conduit souvent à des solutions superficielles et incomplètes puisqu'elles ne tiennent pas compte de la complexité des problèmes du monde réel. Ce déterminisme technologique ignore le rôle de l'action, des valeurs et des institutions humaines dans le construction de l'avenir (Morozov, 2014). Il est donc important de ne pas ignorer les conséquences du présent par les potentialités de futur.

2.3 L'éthique technologique

Le changement de nature de la technique et l'accélération constante de son développement ne s'accompagnent d'aucune éthique disponible pour penser ces transformations du vivant. Le concept d'éthique, appliqué au domaine technologique, n'est pas un ensemble d'énoncés inaltérables et il nous faut en élargir sa vision (Pommier, 2014). Le professeur agrégé et docteur en philosophie de l'université Paris, Éric Pommier, est un spécialiste du philosophe allemand Hans Jonas, qui s'est longuement intéressé à la question de l'éthique technologique. Pommier rapporte que pour Jonas :

> Ce qu'il y aurait de nouveau dans les nouvelles technologies, c'est que cette fois-ci la technique a des effets à l'échelle de la terre, globale et à l'échelle des générations futures. Elle s'étend de manière démesurée dans le temps et dans l'espace. Elle aurait une dimension frénétique, irrésistible. On ne peut pas s'arrêter d'améliorer nos techniques (*Ibid*.).

Cette amélioration constante, perçue comme étant toujours mieux que la précédente, n'est pas toujours synonyme d'avancement pour nos sociétés. C'est en quoi Jonas va exprimer qu'il nous faut penser une éthique de la responsabilité qui prenne en compte les effets de nos actions, mêmes celles qui sont faites avec de bonnes intentions (Jonas, 1992). Il y a des perspectives éthiques au passé, au présent et au futur. Certaines techniques ont eu un caractère destructif volontaire, d'autres, ont eu un caractère destructif involontaire ou secondaire tel que la pollution (Bontems, 2017). L'enjeu éthique du futur serait de concevoir que la perspective de croissance infinie ne peut plus constituer une vision acceptable. L'éthique technologique se doit de prendre en compte les limites de la planète et de se concentrer sur des objectifs de bien-être collectif plutôt que sur une croissance économique perpétuelle. Si le progrès ne peut être atteint qu'en ralentissant le rythme effréné de l'innovation technologique, il faut prendre le temps de mieux comprendre les besoins des individus et des communautés (*Ibid*.). Reconsidérer notre rapport à l'innovation et à la technologie implique de mieux répondre aux besoins réels de la société, en plus d'en améliorer la qualité de vie.

2.3.1 Principe responsabilité

Dans son ouvrage *Le principe responsabilité* (1979), Hans Jonas propose une reformulation de l'éthique pour la civilisation technologique autour de l'idée de responsabilité. Il exprime que nous n'avons pas seulement une responsabilité sur nos actions passées, mais aussi sur celles à venir. Ce qu'il qualifie d'obligation à l'égard de la postérité (Jonas, 1992).

> […] c'est précisément à ce qui n'est pas encore que l'éthique cherchée a affaire et son principe de responsabilité doit être indépendant aussi bien de toute idée d'un droit que de celui d'une réciprocité. […] Or c'est d'une obligation de ce type qu'il s'agit avec l'obligation à l'égard de l'humanité future, car elle veut dire en premier lieu que nous avons l'obligation de l'existence de l'humanité future - indépendamment même de la question de savoir si notre propre postérité en fait partie - et en second lieu il s'agit également de l'obligation de son être tel (*Ibid.*, p.66).

Il est important de garder en tête que nos actions et décisions actuelles en matière de technologie engendrent des conséquences qui ne pourront être résorbées. De toute évidence, il n'y a certes pas de définition exacte de l'éthique appliquée au numérique. Cependant, on constate le besoin d'encadrer et de réaffirmer certaines valeurs, principes qui devraient guider le développement des nouvelles technologies. Il ne s'agit pas nécessairement de remplacer l'éthique telle qu'elle est connue dans sa conception morale ou même de bonne volonté, mais de concevoir une éthique qui deviendrait opératoire dans le domaine technique (Pommier,

2014). Si l'éthique du numérique est un concept qui vient cadrer cette recherche, c'est qu'il ne s'agit pas seulement de parler des objets techniques, mais bien de considérer que c'est un mode d'être. L'idée selon laquelle tout peut être mesuré, calculé, pour être manipulé et utilisé est une tendance de fond qui modifie considérablement la nature de la technique et façonne notre manière de penser l'éthique (*Ibid.*). En ouvrant de nouvelles possibilités aux humains jusque-là inexplorées, il est important de valoriser la réflexion concernant les valeurs que nous souhaitons inclure dans la technique. En s'étendant de manière effrénée, il faut développer un devoir de responsabilité face à cette utopie de la transformation constante et considérer cette responsabilité à l'égard de l'humanité comme un fondement de l'éthique.

2.3.2 L'éthique relationnelle

L'apprentissage automatique, maintenant intégré dans plusieurs sphères de nos vies, transforme des problèmes politiques et sociaux en résolution mathématique. Alors que les systèmes algorithmiques utilisés comportent de nombreux défis, ils s'accompagnent aussi d'importants travaux sur l'équité, la représentation, les données inclusives, et bien plus encore. Pour la chercheuse en sciences cognitives Abeba Birhane (2021), bien que ces travaux constituent une partie du remède, la voie de l'équité se doit d'examiner la situation dans son ensemble. Qu'il s'agisse de croyances qui ne sont pas remises en question sur les ensembles de données, d'injustices actuelles, d'asymétries de pouvoir, elle considère qu'il y a une vision dominante dans le milieu de l'IA qui considère que tout problème peut se résoudre en utilisant des données et des algorithmes (Birhane, 2021).

Birhane explique que la plupart des solutions proposées à des problèmes donnés en AI sont de nature technologique et ne considèrent pas de manière adéquate et inclusive les communautés impactées de manière disproportionnée. L'éthique est souvent inscrite comme une entité allant au-delà des solutions technologiques et une forme de rationalité dominante est devenue le standard de notre vision du monde (*Ibid*.).

> Dichotomous thinking—such as subject versus object, emotion versus reason—persists within this tradition. Ethical and moral values and questions are often treated as clearly separable (and separate) from ''scientific work'' and as something with which the scientist need not contaminate their ''objective'' work. In its desire for absolute rationality, Western thought wishes to cleave thought from emotion, cultural influence, and ethical dimensions. Abstract and intellectual thinking are regarded as the most trustworthy forms of understanding, and rationality is fetishized. Data science, and the wider discipline of computer science, have implicitly or explicitly inherited this worldview. These fields, by and large, operate with rationalist assumptions in the background. The view of the data scientist/ engineer is mistaken as ''the view from nowhere''—the ''neutral'' view. Misconceptions such as a universal, relatively static, and objective knowledge that can emerge from data are persistent. (*Ibid*., p.3).

Selon Birhane, la science des données est une incarnation du rationalisme dans sa façon de préserver une pensée binaire dépourvue de toutes ambiguïtés. Le raisonnement abstrait et sans contexte est privilégié par rapport à l'expérience immergée dans de réelles relations, rendant ainsi plus susceptible la reproduction de résultats discriminatoires et nuisibles (*Ibid*.). C'est pourquoi Birhane propose un changement fondamental d'une éthique qu'elle considère à l'heure actuelle comme étant rationnelle à une éthique relationnelle (*Ibid*.). Cette approche se concentre sur l'importance des relations et de nos interactions dans le développement des savoirs humains. Ces derniers ne sont pas seulement basés sur une logique rationnelle, mais bien sur des éléments vivants et interconnectés. Autrement dit, notre expérience et notre compréhension du monde sont intrinsèquement liées à différents facteurs culturels, historiques, politiques, de genres, et bien plus encore, qui ne peuvent être détachées de leur contexte d'origine (*Ibid*.).

L'éthique relationnelle invite à penser comment nos actions et décisions doivent être conséquentes et alignées aux réels besoins des autres. Considérez le bien-être et les nécessités des autres sont des éléments fondamentaux de notre vie morale. En AI, c'est une éthique qui s'applique à la base de tout nouveau processus de développement et d'ajustement technologique. Pour toute solution à un problème donné, le point de départ devrait être les personnes directement impactées. Il faut centrer les solutions sur ceux qui sont disproportionnellement affectés, ce qui signifie de questionner ce que les systèmes font, de rendre compte de leurs conséquences en plus de penser leur raison d'être. Il faut s'intéresser à la relation qui se développe entre les machines et les humains, en plus de rendre compte des besoins des personnes les plus vulnérables dans la conception des systèmes d'IA. Il est important de considérer comment les humains s'engagent dans le monde de manière significative et, trop souvent, imprédictible (*Ibid*.).

2.4 Le travail critique : une composante essentielle en IA

Les défis en AI sont souvent formulés en termes de problème/solution. L'une des conséquences directes de cette manière de concevoir le monde est de ne pas considérer que certains problèmes ne se conçoivent tout simplement pas de cette manière. Le travail critique en AI peut évidemment servir à pointer des biais, des injustices et des préjudices pouvant avoir des impacts majeurs pour des groupes de personnes et diverses communautés, mais il peut aussi et surtout servir de levier pour examiner, questionner et réfléchir. Il faut rendre compte de l'importance de ne pas tout concevoir en termes de « solution » et redonner l'importance aux phases de diagnostic, de questionnements et de conversation (Birhane, 2021).

En 2020, suite à la publication *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big ?* qui, nous le rappelons, avait pour but de questionner si des réflexions importantes avaient été faites concernant l'utilisation de LLMs et son déploiement à grande échelle, le directeur du département d'IA de

Google Jeff Dean avait écrit « Highlighting risks without pointing out methods for researchers and developers to understand and mitigate those risks misses the mark on helping with these problems. » (Dean, 2020).

Or, le travail critique est une composante fondamentale pour une IA équitable. La journaliste d'investigation américaine et spécialiste des études algorithmiques, Julia Angwin, soutient l'importance de bien diagnostiquer un problème pour le résoudre :

> If you think about the turn of the century and industrialization, we had, I don't know, 30 years of child labor, unlimited work hours, terrible working conditions, and it took a lot of journalist muckraking and advocacy to diagnose the problem and have some understanding of what it was, and then the activism to get laws changed. I feel like we're in a second industrialization of data information. I think some call it the second machine age. We're in the phase of just waking up from the heady excitement and euphoria of having access to technology at our fingertips at all times. That's really been our last 20 years, that euphoria, and now we're like, "Whoa, looks like there're some downsides." I see my role as trying to make as clear as possible what the downsides are, and diagnosing them really accurately so that they can be solvable (Angwin et Paglen, 2019).

Les périodes de diagnostic ne s'accompagnent pas toujours de solutions aux problèmes qu'elles soulèvent. L'idée selon laquelle on ne peut pas mettre en évidence les problèmes sans proposer de solution est une conception extrêmement dommageable. Cette propension à considérer les choses en termes de problème/solution repose sur un binarisme et un sens de causation simpliste selon lesquels l'injustice, les biais et l'équité sont des concepts statiques qu'il nous serait possible de régler au moyen de technologies « neutres » (Birhane, 2021). Pourtant, rien n'est plus en changement et en constante évolution que ces concepts, qui ne sont pas amnésiques du contexte dans lesquels ils existent et de leur réalité temporelle, sociale et culturelle. Il est important de consacrer des efforts pour identifier, analyser et comprendre avec précision les risques et les dommages liés à l'IA. Ces démarches critiques sont essentielles et analysent manière complexe et profonde des sujets qui ne se pensent pas toujours en termes de leur résolution, mais bien de leur fondement.

CHAPITRE 3
MÉTHODOLOGIE

Dans ce chapitre, nous allons présenter la méthodologie de recherche que nous avons élaborée pour faciliter les discussions en relation avec les concepts clés de notre cadre théorique. Comme nous l'avons vu au chapitre précédent, les concepts clés d'automatisation, de biais algorithmiques et de l'éthique technologique fournissent la base théorique qui servira de référence pour analyser les résultats. Puisque nous souhaitions ancrer notre cadre dans le monde réel, il allait de soi de privilégier une méthode nous permettant d'analyser les résultats tout en préservant leur contexte unique de création.

Ainsi, nous commencerons par présenter notre choix méthodologique dans le but d'établir la pertinence de cette approche en rapport à nos objectifs de recherche. Nous poursuivrons par la présentation des caractéristiques méthodologiques pour assurer la transparence et la reproductibilité de notre étude. Par exemple, il sera question des sources de données utilisées, des critères d'inclusion et d'exclusion du corpus, de la fiabilité, etc. Par la suite, nous discuterons de la récolte des articles et présenterons notre choix de corpus final. Finalement, nous ferons l'explication de notre grille d'analyse, en plus de justifier en quoi les catégories d'analyse choisies sont soutenues par notre cadre théorique

3.1 Choix de la méthodologie et justification

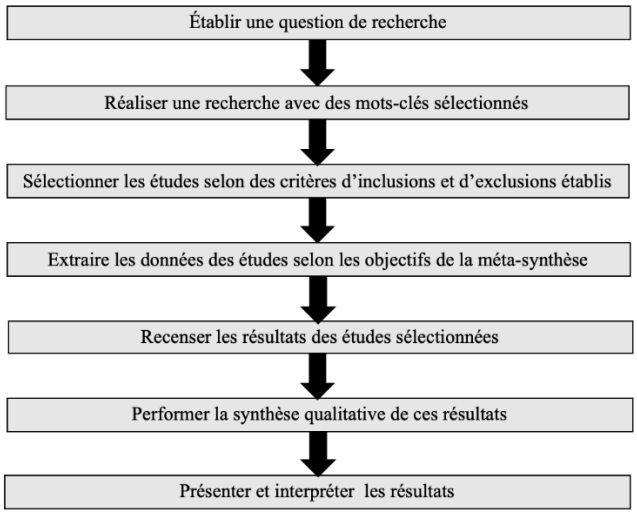
En considérant la nature des technologies de langage, plusieurs recherches sur le sujet se sont arrêtées sur des méthodologies quantitatives. Dans ce mémoire, puisque nous proposons une analyse de la façon dont différentes recherches en TALN étudient les biais, nous souhaitons nous concentrer sur la compréhension approfondie de ce phénomène et de ses diverses composantes. C'est pourquoi nous considérons la méthodologie qualitative comme étant la plus alignée avec nos objectifs de recherche. Plus précisément, nous utiliserons la méthodologie qualitative de la méta-synthèse.

La méta-synthèse est une méthode de recherche qui consiste à mettre en commun des éléments fondamentaux de plusieurs études de recherche sur un sujet donné afin de produire une synthèse globale. En combinant divers concepts, elle permet de créer une vision plus complète du sujet de recherche (Beaucher et Jutras, 2007). Elle peut inclure des études qualitatives et quantitatives et permet une analyse plus approfondie des tendances, des relations de réciprocité et d'oppositions entre les résultats et des données atypiques (Finfgeld, 2003). Autrement dit, la métasynthèse rend possible l'interprétation des résultats de différentes catégories d'études sélectionnées mises en relation. Elle permet d'aborder plus largement les

tendances analytiques ; elle n'a pas pour but d'accumuler des données ou de faire émerger de nouvelles théories (Beaucher et Jutras, 2007).

Sur le plan méthodologique, c'est une analyse qui nécessite un enchaînement structuré. En nous basant sur les études de Beaucher et Jutras (2007), Chowdhury et Turin (2019) et de Finfgeld (2003), nous respecterons les étapes suggérées pour s'engager dans les processus.

Figure 3.1 - Étapes de la méta-synthèse qualitative selon Chowdhury et Turin (2019)[20]



D'abord, comme dans toute recherche, il faut poser notre question de recherche. Cette question, que nous avons déjà présentée dans le premier chapitre de ce mémoire, permet d'établir les caractéristiques méthodologiques importantes guidant notre recherche et notre sélection d'articles. Cette sélection doit être effectuée selon des critères d'inclusion ou d'exclusion qui permettent de garder une certaine généralité tout en assurant de restreindre les éléments impertinents selon le sujet étudié. Il convient d'identifier et de justifier en quoi ces études vont ensemble et font partie du même phénomène. Beaucher et Jutras (2007) soutiennent que :

> […] ces critères ne servent pas nécessairement à trouver un corpus d'études semblables, mais bien des études qui sont similaires par leur objet, tout en pouvant être différentes dans leurs

[20] Figure créée à partir de de l'étude de Chowdhury et Turin (2019) *Synthesizing Quantitative and Qualitative Studies in Systematic Reviews: The Basics of Meta-analysis and Meta-synthesis.*

> approches et dans leurs résultats. En effet, les différences peuvent s'avérer intéressantes à explorer et la métasynthèse permet cet exercice (*Ibid*., p.67).

Puis, il y a la phase de lecture des articles sélectionnés, suivie d'une phase de classification selon une grille d'analyse précise. Ce qui nous mène à l'analyse des données collectées. Durant cette étape, il est important de faire valoir certaines considérations de la recherche, telle que les résultats provenant de méthodologies différentes et de justifier ces choix. Ensuite, en suivant la grille d'analyse préétablie, il est possible de catégoriser chaque étude avant d'en synthétiser les résultats, ce qui permet d'établir des relations entre les données. Cette analyse des résultats dans la méthasynthèse s'opère à deux niveaux. En premier, il faut conserver une représentation fidèle de chacune des études. En second, il nous faut trouver les éléments similaires et distinctifs de ces dernières. Les résultats permettent l'émergence d'un nouveau construit qui représente l'ensemble du corpus à l'étude (*Ibid*.). Finfgeld (2003) soutient l'importance d'identifier et d'analyser les relations de réciprocité et d'oppositions entre les résultats de différentes catégories d'analyse, sans oublier d'accorder une attention particulière aux données atypiques afin de valider leur fiabilité. Elle maintient que :

> […] it is these types of analyses and discussions that are useful for preserving and emphasizing the uniqueness of individual studies while building a comprehensive whole around reciprocal relationships and parallel lines of argument (*Ibid*., p.901).

De plus, lors de la phase d'analyse, il est essentiel de faire usage de preuves pour supporter les résultats et la crédibilité de la recherche. L'utilisation de citations des sources originales pour appuyer certaines affirmations, en plus de la contextualisation des données, permet d'expliquer et de justifier des variations pouvant survenir entre les études du corpus. Enfin, la métasynthèse est une méthode créative nécessitant un souci de la transparence. Il est important de clarifier le processus et les choix effectués, c'est pourquoi nous allons détailler au point 3.2 les caractéristiques méthodologiques guidant notre mémoire.

3.2 Caractéristiques méthodologiques

Les caractéristiques méthodologiques représentent les méthodes et les procédures utilisées pour collecter nos informations et surtout, justifier notre choix de corpus. Nous considérons que ces précisions sont essentielles pour garantir la qualité et la transparence de cette recherche.

3.2.1 Source des données

Pour mener cette étude, il était fondamental d'user d'une ressource pertinente aux recherches en TALN. C'est pourquoi nous avons choisi d'utiliser la plateforme de l'Association for Computational Linguistics

(ACL) Anthology. Cette archive numérique héberge des recherches académiques spécialisées dans le domaine du TALN et en linguistique computationnelle. Donnant accès à des milliers d'articles, c'est une précieuse ressource qui s'avère extrêmement alignée et pertinente pour notre objet de recherche (ACL Anthology, 2023). ACL Anthology :

1. Donne un accès gratuit aux recherches qui sont hébergées, ce qui garantit une accessibilité du travail académique ;
2. Actualise constamment les recherches présentes sur sa plateforme, ce qui assure le partage des dernières avancées dans la matière ;
3. Couvre une multitude de sujets importants en NLP dont la traduction automatique, l'entraînement de modèle, la reconnaissance vocale et bien plus encore ;
4. Contiens une diversité d'articles allant de conférences aux articles scientifiques, en plus d'avoir un processus de révision scientifique rigoureux ;
5. Permets d'accéder rapidement aux articles les plus pertinents avec des options de recherches avancées et une interface facile d'utilisation (*Ibid*.)

ACL Anthology est donc représentative du domaine de recherche en TALN en raison de sa large couverture des publications en NLP (articles, conférences, revues, etc.), de la variété des sujets qui y sont hébergés et de la rigueur académique qui est mise de l'avant (*Ibid*.).

Dans la variété des sujets hébergés en TALN, il est possible de retrouver ceux qui touchent au ;

- Natural Language Generation ;
- Natural Language Understanding ;
- Discourse and Conversation Analysis ;
- Machine Translation ;
- Automatic Speech Recognition ;
- Text and Corpus Analysis ;
- Machine Learning and NLP ;

Ces différents sujets englobent les diverses étapes du traitement automatique du langage naturel qui sont susceptibles d'être affectées par les biais présents dans les données et les modèles utilisés, allant de la « compréhension » à la génération, ainsi que la traduction et l'analyse. C'est en quoi nous considérons qu'ACL Anthology offre une bonne représentativité du sujet et de ses champs de recherche (*Ibid*.).

3.2.2 Critères d'inclusion et d'exclusion du corpus

Beaucher et Jutras (2007) précisent l'importance d'avoir des critères d'inclusions bien définis dans la sélection du corpus à l'étude. Puisque c'est une étape cruciale de notre recherche, ces critères permettent d'encadrer les sélections effectuées, d'avoir une stratégie claire en plus de minimiser la possibilité de faire des choix biaisés.

3.2.2.1 Dates de publication

En raison du développement massif des technologies en TALN lors des dernières années, il est possible de constater l'abondance des recherches dans ce domaine.

L'idée dominante du *state of the art* conditionne de nouvelles avancées technologiques rapides, considérées meilleures et produisant des résultats améliorés. Par conséquent, les chercheurs de ce domaine doivent constamment s'adapter à un rythme effréné. Certaines recherches peuvent avoir une portée plus large et rester d'actualité, tandis que d'autres se voient rapidement dépassées. Les études en TALN peuvent devenir obsolètes rapidement menant ainsi à une abondance de recherche sur le sujet. C'est donc en raison de son évolution rapide que nous avons choisi de retenir seules les études produites entre 2020 et 2023. Plus particulièrement, nous avons choisi de cadrer à partir l'année 2020 puisque c'est au courant de cette dernière que OpenAI a rendu public GPT-3. GPT-3 marque le début des très grands modèles de langage avec un entraînement basé sur plus de 175 milliards de données (Vu, 2022). C'est un jalon important dans le domaine du TALN puisque l'arrivée de ces modèles plus volumineux et performants a ouvert la porte à de nouvelles possibilités, tout en soulevant de nombreux enjeux et questions éthiques.

3.2.2.2 Fiabilité

Il peut sembler évident d'avoir un corpus composé de sources fiables. Par contre, tout dépendamment d'un objet de recherche et des choix méthodologiques qui peuvent être faits, nous estimons important de le préciser dans nos critères d'inclusions. Nous considérons une source fiable toute recherche provenant spécifiquement de revues scientifiques fiables, de conférences reconnues, ayant un comité de lecture ou une évaluation par les pairs. Les articles soumis à ACL Anthology sont pratiquement tous évalués par des experts du domaine qui assurent la qualité de la recherche et de la méthodologie utilisée, ainsi que la pertinence des conclusions. Ce processus garantit que seuls des articles rigoureux sont publiés.

De plus, l'équipe de ACL Anthology, qui assure la qualité académique des articles accessibles sur leur plateforme, peut demander des corrections. De façon générale, on retrouve cinq types de corrections demandées :

1. Une correction de métadonnée, ce qui correspond à la modification d'une information de la recherche telle que le titre, les auteurs, etc. ;
2. Une clarification des erreurs commises dans l'ouvrage original sous forme de courtes notes ou déclarations ;
3. Une révision ou un remplacement du travail scientifique original ne mettant pas en évidence les erreurs de la version précédente ;
4. Une rétraction de l'article si des erreurs graves sont découvertes ;
5. Une suppression complète survient en cas de plagiat, de problèmes éthiques et juridiques (ACL Anthology, 2021).

La plateforme utilisée pour conduire notre recherche assure la rigueur académique des contenus qui sont hébergés. Nous aurons tout de même un regard critique lors de la sélection de notre corpus pour assurer la validité des articles sélectionnés.

3.2.2.3 Pertinence

Dans le but de sélectionner les articles de notre corpus, nous ferons usage de l'option de recherche « *relevance »* disponible sur ACL Anthology. Nous avons choisi d'utiliser la pertinence comme indicateur de qualité pour assurer la cohérence entre notre objet de recherche et les résultats d'articles obtenus.

Cette option de recherche sur le site de ACL Antholoy est basée sur l'algorithme de Google qui établit un ordre des meilleurs résultats selon la demande de recherche. Cette classification se base sur différents facteurs tels que la qualité d'une recherche, la correspondance au mot-clé utilisé, le titre, la popularité, etc. (*Ibid*.) Il est intéressant de faire usage de cet indicateur pour classer et encadrer notre sélection.

3.2.2.4 Nombre d'articles sélectionnés

Ce mémoire consiste à réaliser une méta-synthèse des représentations des biais dans différentes études en TALN, il convient donc de statuer sur un nombre d'articles à sélectionner. Puisque nous avons opté pour une méthodologie qualitative, il est important d'avoir un nombre d'articles plutôt restreint pour nous permettre de nous concentrer sur une analyse approfondie et détaillée. Nous souhaitons examiner rigoureusement les articles sélectionnés dans le but d'analyser les similitudes, les nuances, les données

atypiques et plus encore. En nous inspirant de l'étude de Hessenthaler *et al.* (2022) que nous considérons méthodologiquement bien effectuée, nous ferons une première sélection de 20 articles. Ensuite, nous appliquerons nos critères d'inclusion et d'exclusion pour obtenir une sélection finale alignée aux caractéristiques méthodologiques que nous avons mises de l'avant.

3.2.2.5 Degré de technicité

Dans le but de conserver le volet compréhensif de ce mémoire, toutes les recherches relevant d'un degré de technicité élevé seront exclues de notre corpus. Puisque ces dernières sont davantage axées sur la technique elle-même, l'entraînement des LLMs ou leurs configurations, elles comportent de nombreux calculs mathématiques d'algèbre linéaire, de statistique et bien plus encore. Pour des raisons d'entendement et de conformité de la discipline dans laquelle s'inscrit cette recherche, tous les articles dont les résultats et discussions seront axés sur les performances mathématiques des modèles (Hassan *et al*., 2021) ou encore, sur les techniques utilisées en machine learning pour atténuer les biais - bias subspace subtraction, data augmentation, adversarial training, transfer learning - seront également écartés. Il va sans dire que de ne pas avoir recours à un « traducteur » de contenu en TALN est un choix méthodologique que nous devons justifier. L'objectif de notre corpus est de nous permettre d'avoir une vue d'ensemble et l'inclusion d'études très techniques pourrait déséquilibrer notre analyse et rendre la synthèse plus difficile. De plus, nous considérons que les arguments critiques de certaines recherches discutent déjà des articles plus techniques et des manques qui leur sont associés.

Ainsi, pour garantir que notre analyse englobe l'ensemble du champ pertinent sans risquer de passer à côté d'éléments centraux ou de rendre compte de manière trop partielle de la discipline, nous nous concentrons sur une sélection d'articles qui offrent une perspective équilibrée et accessible aux lecteurs. Cette approche nous permettra de fournir une vue d'ensemble complète tout en maintenant la cohérence et l'accessibilité de nos résultats, conformément aux objectifs de notre recherche.

3.2.2.6 Type d'article

Puisque nous accorderons une grande importance à la singularité de chaque étude, toute publication rassemblant une multitude d'articles ou présentant le compte-rendu d'un événement ou d'une conférence sera exclue de notre corpus.

3.2.2.7 Méthodologie des articles

Nous tenons à préciser qu'aucune sélection d'articles ne sera effectuée sur la base de la méthodologie employée par les auteurs. Nous considérons le choix méthodologique très intéressant, puisqu'il peut démontrer la manière dont les différents auteurs abordent et choisissent de s'intéresser aux biais dans leurs études. Autrement dit, la méthodologie ne constitue pas un critère d'inclusion ou d'exclusion pour la sélection des articles de notre corpus, mais contribuera à notre analyse dans la façon d'étudier les biais en TALN.

3.3 Récolte des articles

Pour la récolte de nos articles, effectuée en avril 2023, nous avons interrogé le moteur de recherche ACL Anthology en inscrivant le mot-clé « NLP bias » et en activant l'option de recherche en anglais de « *relevance »*. Évidemment, le choix d'effectuer notre requête en anglais était de mise puisqu'il s'agit d'une plateforme internationale et d'un champ de recherche particulièrement développé dans cette langue. Contrairement à notre recension des écrits et par choix méthodologique, nous n'avons pas voulu utiliser plus d'un mot-clé pour lancer notre recherche. Il fait partie de notre étude de considérer les types de biais qui seront intégrés dans ces articles et pour ne pas influencer la recherche dans une certaine direction, nous avons tenu à ne pas être trop spécifiques dans notre requête. Nous avons par la suite examiné les 20 premiers résultats obtenus. En appliquant nos critères d'inclusion et d'exclusion, nous avons retiré :

- Cinq articles qui ne correspondaient pas aux dates de publication retenues (Sun *et al.*, 2019 ; Bender et Friedman, 2018 ; Costa-jussà *et al.*, 2019 ; Rudinger *et al.*, 2017 ; Webster *et al.*, 2019) ;

- Trois articles présentant un compte rendu de conférences en TALN, ne respectant pas le type d'articles requis (Costa-jussà *et al.*, 2020 ; Costa-jussa *et al.*, 2021 ; Hardmeier *et al.*, 2022) ;

- Trois articles relevant d'un degré de technicité trop élevé ont été exclus (Hassan *et al.*, 2021 ; Ahn et Oh, 2021 ; Li *et al.*, 2022) ;

- Un article en double apparu en double dans notre sélection a été retiré (Shah *et al.*, 2020).

3.4 Corpus sélectionné

Cette approche nous a permis d'obtenir une sélection finale de huit publications pertinentes (Hutchinson *et al.*, 2020 ; Schick *et al.*, 2021 ; Matthews *et al.*, 2021 ; Lalor *et al.*, 2022 ; Dev *et al.*, 2022 ; Ruder *et al.*,

2022 ; Shah *et al.*, 2020 ; Blodgett *et al.*, 2020) qui composeront notre corpus officiel[21]. Pour justifier que ce nombre d'articles est suffisant pour notre recherche, nous nous sommes basés sur l'étude de Sandelowski et Barroso (2007) *Handbook for Synthesising Qualitative Research*. Les autrices ont abordé les préoccupations dans la littérature des échantillons jugés « trop petits » dans la recherche qualitative qui ne seraient ni fiables ni généralisables. Ce à quoi elles clarifient la différence entre la généralisation nomothéique, l'objectif de la recherche quantitative, et la généralisation idiographique. La généralisation nomothétique consiste à produire des résultats généraux et à identifier des modèles, des régularités, ou encore des relations. Tandis que la généralisation idiographique se concentre sur la compréhension et l'explication de cas individuels uniques et des contextes spécifiques qui leur sont associés. En se concentrant sur un petit échantillon, il est possible d'approfondir les nuances de chaque étude et d'obtenir une vision plus détaillée que ne le permettrait une approche plus large ou un corpus plus large. Un petit échantillon permet de préserver la pertinence contextuelle, c'est-à-dire que les résultats restent applicables au contexte dans lequel ils ont été produits (Sandelowski et Barroso, 2007). Or, en considérant ce qui vient d'être énoncé, la diversité des études disponibles, ainsi que nos critères d'inclusion et d'exclusion établis, c'est un nombre d'articles suffisant pour permettre une méta-synthèse qualitative rigoureuse. Nous disposons d'un nombre de données adéquates, en plus d'éviter la surcharge d'information qui risquerait de compromettre la qualité, la clarté et la profondeur de l'analyse souhaitée.

Dès lors identifié, nous avons par la suite lu chacune de ces publications avec pour objectif de les annoter et organiser manuellement selon différentes catégories d'analyse.

3.5 Grille d'analyse

La méthode de Chowdhury et Turin (2019) recommande d'extraire les données et de les organiser de manière cohérente sous des catégories à l'étude. Nous avons opté pour une grille d'analyse nous permettant d'organiser les informations pertinentes extraites des études et faciliter la comparaison et la synthèse des résultats. Notre grille comporte plusieurs catégories et a été conçue de manière à inclure différentes dimensions et aspects du corpus à analyser.

Nous y avons inscrit les catégories centrales de notre question de recherche, pour ensuite nous permettre observer si certains éléments sont présents, de quelle manière, en plus d'accorder une grande importante aux citations des sources originales. Notre cadre théorique, basé sur les concepts d'automatisation, de biais algorithmique et d'éthique technologique, soutiendra l'analyse des diverses catégories de notre grille. Le

[21] L'annexe A de ce présent mémoire présente la bibliographie intégrale des 20 premiers résultats obtenus lors de la sélection de notre corpus.

point 3.5.1 fera l'explication de ces diverses catégories sélectionnées, en plus de démontrer comment nos concepts informeront notre analyse.

Tableau 3.1 – Grille d'analyse

Grille d'analyse du corpus								
Titre :								
Catégories à l'étude	Biais	Relation entre biais et langage	Type de préjudice	Objectif	Méthodologie	Résultats	Limitation	
Classifications possibles	Conceptualisés, non conceptualisés	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, vague, non spécifiée	
Attributions								
Citations des sources originales								

3.5.1 Explication des catégories d'analyse

Dans notre grille d'analyse, nous avons inclus plusieurs catégories pour examiner et évaluer de manière approfondie la façon dont les différentes recherchent étudient les biais en TALN. Ces catégories fournissent un cadre structuré pour examiner les différentes dimensions et présentations de chaque étude, en plus de faciliter la compréhension de leurs résultats. Pour chaque catégorie à l'étude, nous avons utilisé une classification spécifique permettant de catégoriser les études incluses dans notre méta-analyse. Ces classifications jouent un rôle essentiel dans l'organisation des études en fonction de la qualité et du type d'information fournie.

Pour chacune des neuf catégories à l'étude, nous présenterons dans cet ordre : l'objectif de la catégorie, les classifications possibles pour cette catégorie, ainsi qu'une explication concernant l'utilisation de ces classifications.

1. Biais

Objectif de la catégorie : Cette catégorie vise à vérifier si une conceptualisation formelle de ce que les auteurs entendent par biais a été effectuée. Le concept de biais peut être interprété de différentes manières en fonction du contexte et des objectifs de l'étude. Une conceptualisation de ce terme permet de clarifier la recherche, de l'encadrer, en plus d'avoir une méthodologie alignée aux biais à l'étude. Notre cadre conceptuel, qui aborde les biais tant aux termes des comportements nuisibles du système que des biais épistémologiques des représentations de l'IA, nous permettra d'analyser les conceptualisations offertes et de discuter des manques qui peuvent sous-tendre la vision de certaines représentations techniques des auteurs.

Classifications possibles : Conceptualisé, non conceptualisé

Utilisation de la classification : Nous utiliserons la classification « conceptualisé » pour toute étude ayant offert une définition claire et précise du terme de biais et l'utilisation qui en est faite dans leur contexte d'étude. En revanche, nous utiliserons la classification « non conceptualisé » lorsque le terme n'est pas défini ou compris dans un cadre conceptuel.

2. Relation entre biais et langage

Objectif de la catégorie : Cette catégorie examine si les auteurs expriment un lien explicite entre les biais en NLP et le langage. Puisque le langage reflète les biais sociaux et culturels, les biais peuvent se manifester à travers des stéréotypes de genre, des discriminations raciales, des préjugés socio-économiques et bien plus encore (Blodgett *et al*., 2020). Étudier les biais dans le langage permet de comprendre comment ces biais se propagent et peuvent être perpétués par les systèmes. Le concept de langage opérationnel nous servira de base pour analyser les aspirations des auteurs à vouloir rendre le langage (code) dépourvu de toute complexité et ambiguïté, en plus de vérifier si le langage est traité comme une simple transmission de signaux ou s'il s'agit d'une caractéristique spécifique de l'humain.

Classifications possibles : Explicite, vague, non spécifiée

Utilisation de la classification : Nous utiliserons la classification « explicite » lorsqu'une relation entre les biais étudiés et le langage est présentée. En d'autres mots, si les biais étudiés en NLP se réfèrent à la relation de ces systèmes avec le langage. La classification « vague » sera utilisée lorsque la relation entre ces deux instances s'avère moins précise. Elle peut manquer de détails spécifiques ou être exprimée de manière générale, ce qui rend son interprétation moins précise ou sujette à différentes interprétations. De plus, la classification « non spécifié » sera employée lorsqu'il n'y a pas de lien ou de corrélation apparente, lorsque les informations ne présentent pas de relations identifiables.

3. Type de préjudice

Objectif de la catégorie : Cette catégorie observe si les auteurs abordent les types de préjudices (représentation ou allocation) causés par les biais qu'ils analysent. Puisque le concept de biais algorithmique fait partie de la fondation de notre projet, il est important pour notre analyse de considérer les types de préjudices créés par ces biais. En détaillant les préjudices spécifiques (stéréotypes, dénigration, sous-représentation, etc.), il est possible d'encadrer l'étude et de bien comprendre les effets des biais étudiés. Ce qui inclut l'identification des groupes de population qui sont touchés de manière disproportionnée, les inégalités perpétuées, les discriminations renforcées, etc.

Classifications possibles : Représentation, allocation, non spécifié

Utilisation de la classification : Pour cette classification, nous utiliserons la taxonomie utilisée par Crawford (2017) que nous avons présentée dans le chapitre 2 de ce présent mémoire. Nous utiliserons « représentation » lorsque les recherches se pencheront sur les préjudices renforçant la subordination de certains groupes en raison de leur identité. Nous rappelons que ces derniers surgissent lorsque le système représente certains groupes sociaux de manière défavorable, les dénigre ou encore, ne reconnaît tout simplement pas leur existence. Nous emploierons « allocation » lorsqu'ils étudieront les préjudices permettant d'octroyer ou de refuser à certains groupes une opportunité ou une ressource. Ces types de préjudices sont immédiats et pouvant aller jusqu'au déni total de certains services. Puis, la classification « non spécifié » sera inscrite lorsque les comportements décrits comme nuisibles du système ne détailleront pas les préjudices possibles.

4. Objectif

Objectif de la catégorie : Cette catégorie vise à analyser si les auteurs abordent explicitement les objectifs de leur étude. Les objectifs sont révélateurs de la façon de percevoir les biais, ainsi que de l'orientation de la recherche. Différents objectifs auront certainement différentes méthodologies pour parvenir aux résultats souhaités. Surtout, nous souhaitons analyser les objectifs en rapport à l'éthique et la logique d'automatisation, dont les concepts cadrent notre recherche. Il est intéressant d'examiner si les auteurs visent à simplifier et améliorer des processus automatisés en termes d'opérations concrètes en utilisant un raisonnement abstrait, ou encore, s'ils privilégient une finalité en rapport à l'expérience humaine réelle.

Classifications possibles : Explicite, vague, non spécifié

Utilisation de la classification : Nous utiliserons la classification « explicite » lorsque les objectifs de recherche seront clairement énoncés. La classification « vague » sera utilisée lorsque les auteurs restent peu précis dans la formulation des objectifs, offrent des intentions floues et manquant de détails spécifiques. Puis, la classification « non spécifié » sera employée lorsqu'il n'y a aucun objectif identifiable et présenté par les auteurs.

5. Méthodologie

Objectif de la catégorie : Cette catégorie permet de recenser la méthodologie utilisée par les auteurs. Certains sont dans une analyse compréhensive et d'autres, davantage dans la mesure et l'évaluation. Cette catégorie est intéressante puisqu'elle indique une manière différente de traiter un sujet et permet de comprendre comment les études ont été menées et comment les résultats ont été obtenus. Le concept d'opérationnalisme

sera constitutif de notre analyse, en aidant à examiner si les approches employées visent à simplifier les problématiques en se concentrant sur des opérations concrètes et des aspects quantifiables, ou si elles cherchent à comprendre et à expliquer les concepts fondamentaux qui les sous-tendent.

Classifications possibles : Qualitative, quantitative, mixte

Utilisation de la classification : La classification « qualitative » sera attribuée aux études présentant une méthodologie basée sur des caractéristiques qualitatives, des attributs non mesurables. Elles se concentrent sur la description, la compréhension, etc. La classification « quantitative » sera utilisée pour les études utilisant des méthodes d'analyse mesurables et quantifiables, reposant sur l'utilisation de méthodes statistiques et d'analyses numériques. Puis, la classification « mixte » sera inscrite pour les études démontrant à la fois des caractéristiques qualitatives et quantitatives.

6. Résultats

Objectif de la catégorie : Cette catégorie observe la contribution des résultats obtenus par les études de notre corpus. En classifiant les résultats des différentes recherches, il nous est possible d'aller au-delà des résultats isolés et d'avoir une vue d'ensemble plus complète sur les différentes contributions au champ de recherche. Les concepts d'éthique technologique et de solutionnisme technologique seront mobilisés pour permettre une discussion plus approfondie en s'intéressant à la relation de l'humain à la machine dans les résultats produits.

Classifications possibles : Compréhension des biais, performance du système

Utilisation de la classification : La classification « compréhension des biais » sera utilisée lorsque les résultats permettent une meilleure compréhension du concept de biais, des facteurs contribuant à son développement, ou encore, proposent un cadre unifiant pour la recherche sur les biais en TALN. La classification « amélioration des performances » sera attribuée aux résultats se concentrant sur des techniques visant l'amélioration, l'atténuation, la détection et l'évaluation des biais dans les modèles NLP.

7. Limitation

Objectif de la catégorie : Cette catégorie regarde la transparence scientifique des auteurs. L'explicitation des limitations d'une étude permet d'encadrer la portée des résultats, ainsi que leur interprétation. De plus,

elles peuvent montrer des zones inexplorées et sont révélatrices de l'état de la recherche actuelle. Nous mobiliserons le travail critique, cadré comme composante essentielle du travail en IA. Les limitations d'études peuvent servir à examiner, questionner, pointer des lacunes, etc. Cette catégorie permet de ne pas tout concevoir en termes de « solution » et de donner de l'espace nécessaire pour penser les manques et la façon de considérer certains problèmes.

Classifications possibles : Explicite, non spécifiée

Utilisation de la classification : Une classification « explicite » sera indiquée lorsque les auteurs auront présenté de manière évidente et claire les limitations de leur recherche. Il sera intéressant d'observer chez les études ayant explicité des limitations si elles portent sur la méthodologie, les données, la généralisation, etc. Puis, une classification « non spécifiée » sera inscrite lorsqu'aucune limitation ne sera précisée.

3.5.2 Analyse des articles

Pour mener l'analyse de nos articles, nous avons commencé par la phase de lecture. Cette première lecture nous a permis de nous familiariser avec le contenu, les biais, les objectifs, la méthodologie et bien plus encore. Après cette première lecture, nous avons effectué une relecture approfondie en utilisant notre grille d'analyse préalablement. Notre grille d'analyse a été conçue pour examiner plusieurs aspects clés des articles. En utilisant cette dernière, nous avons pu systématiquement catégoriser chaque élément à l'étude, en notant les informations pertinentes et en préservant les citations directes des sources originales pour une référence ultérieure. Cette phase de classification manuelle nous a permis de regrouper les articles en fonction de leurs caractéristiques communes, facilitant ainsi la comparaison et la mise en évidence des tendances générales. Nous avons pu identifier les similarités entre les études, en plus de repérer les manques potentiels dans la manière d'étudier les biais en TALN. Il est important de souligner que notre approche d'analyse est qualitative, ce qui signifie que nous avons cherché à préserver la singularité de chaque étude tout en identifiant les relations générales entre elles. Il est possible de consulter, dans l'annexe B, la grille d'analyse complétée pour chacune des études de notre corpus. Dans le prochain chapitre, la présentation de nos résultats mettra en évidence les similitudes, les différentes, les manques, ainsi que les conclusions importantes. Cette méta-synthèse nous permettra de formuler des conclusions globales sur l'état actuel de certaines recherches des plus pertinentes sur les biais en TALN.

CHAPITRE 4
RÉSULTATS

Dans ce chapitre, nous allons présenter les résultats de l'étude en commençant par mettre en évidence la fiabilité des résultats et les limitations méthodologiques associées. Les résultats généraux seront ensuite exposés et suivis de notre méta-synthèse. Nous avons choisi de structurer cette dernière en suivant les catégories à l'étude de notre grille d'analyse. Cette section représente le fruit de notre collecte de données et nous servira de base pour l'analyse et la discussion du chapitre suivant.

4.1 Fiabilité des résultats et limitations méthodologiques

La présente méta-synthèse vise à consolider la façon dont huit études indépendantes étudient les biais dans le domaine du NLP. Les recherches incluses ont été sélectionnées en considérant plusieurs critères d'inclusions et d'exclusions que nous avons pris soin d'expliciter. Nous avons par la suite utilisé une grille d'analyse avec diverses catégories (biais, type de préjudices, objectif, méthodologie, etc.) et classifications possibles. Malgré un grand souci de transparence et les efforts pour assurer une approche robuste, certaines limites et biais doivent être pris en considération.

Dans une recherche qualitative, plusieurs biais peuvent s'introduire. Nous devons considérer, sur un sujet d'étude polarisant et d'actualité, que des biais d'interprétations peuvent s'introduire. Ces derniers surviennent lorsque le chercheur influence involontairement les résultats en fonction de ses propres croyances lors de l'analyse des données qualitatives. Dans le but de réduire ce risque, nous avons eu une grande préoccupation de transparence de notre démarche tout au long de notre recherche et nous avons mis en place une grille d'analyse suivant une classification préétablie et explicitée. De plus, il est important de considérer que cette recherche vise à présenter une méta-synthèse dans la manière d'étudier les biais chez huit études sélectionnées et que les résultats obtenus sont le fruit de cette mise en commun spécifique. Or, une généralisation de ces résultats hors de son contexte de recherche serait difficilement applicable, et ce, même si certaines recommandations ou constats généraux peuvent être apportés.

4.2 Résultats généraux

Le tableau 4.1 présente les résultats généraux du nombre d'articles associés aux diverses classifications de chaque catégorie à l'étude. La présentation de ces résultats sous forme de tableau sert à faciliter la comparaison entre les différentes études incluses, permettant ainsi de mettre en évidence rapidement les similitudes et les différences observées dans les résultats.

Tableau 4.1 – Résultats généraux

Catégories et classifications	*Nombre d'articles*
BIAIS	
Conceptualisé	3
Non conceptualisé	5
RELATION ENTRE BIAIS ET LANGAGE	
Explicite	3
Vague	5
Non spécifiée	0
TYPE DE PRÉJUDICES	
Représentation	0
Allocation	2
Représentation et allocation	4
Non spécifiés	2
OBJECTIF	
Explicite	3
Vague	5
Non spécifié	0
MÉTHODOLOGIE	
Quantitative	3
Qualitative	3
Mixte	2
RÉSULTATS	
Compréhension des biais	4
Performance du système	4
LIMITATION	
Explicite	5
Non spécifiée	3

4.3 Méta-synthèse

L'analyse de notre corpus visait à examiner la façon dont différentes recherches en TALN étudient les biais. Nous avons choisi de mettre en commun des éléments fondamentaux, tels que la conceptualisation des biais, les types de préjudices, les objectifs, les méthodologies utilisées et plus encore, pour ainsi les analyser au travers de chaque étude selon une grille préétablie. Notre but était de mettre en évidence les similitudes, les différences, les limites et plus encore, dans la conception et la conduite de ces études par différents auteurs.

4.3.1 Principaux constats

En regardant le tableau 4.1, il est possible d'effectuer plusieurs observations préliminaires que nous allons examiner plus en détail. Dans le but d'offrir une analyse plus approfondie et structurée, nous avons choisi de présenter ces principaux constats de notre méta-synthèse par catégorie à l'étude dans l'ordre qui suit :

- Biais
- Biais et langage
- Type de préjudices
- Objectif
- Méthodologie
- Résultats
- Limitation

4.3.2 Biais

La majorité des études de notre corpus ne s'engagent pas de manière explicite sur ce qui constitue un biais. Dans l'ensemble, plus de la moitié des études de notre corpus n'ont pas conceptualisé le terme de biais. Nous avons constaté que, bien que ces recherches abordaient différents types de biais et leurs risques potentiels associés, elles n'offraient pas de définitions claires et concises de ces biais.

Tableau 4.2 – Citations des sources originales pour les biais non conceptualisés

Source	***Citation***	***Classification***
Hutchinson et al., 2020)	*In addition, since text classifiers are trained on large datasets,* ***the biases they exhibit may be indicative of societal perceptions of persons with disabilities. Depending on how such models are deployed, this could potentially result in reduced autonomy, reduced freedom*** *[...] bias with respect to different disability groups has been relatively under-explored.*	*Non conceptualisé*
Lalor et al., 2022	*Accordingly, in this study we perform a broad benchmark analysis of intersectional bias encompassing the following key characteristics: [...] : gender, race, age, education, and income.* ***[...] Intersectional biases arising as a result of interacting demographics have been studied in the broader machine learning literature, either from a theoretical perspective.***	*Non conceptualisé*
Matthews et al., 2021	*Ideally,* ***profession words would not reflect a strong gender bias. However, in practice, they often do. According to such a metric, doctor might be male biased or nurse female biased based on how these words are used in the corpora from which the word embedding model was produced.*** *Thus, this gender bias metric of profession words as calculated from the Word2Vec model can be used as a measure of the gender bias learned from corpora of natural language.*	*Non conceptualisé*
Schick et al., 2021	*With model sizes continually increasing (Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020; Fedus et al., 2021), ever-larger pretraining datasets are necessary both to prevent overfitting and to provide access to as much world knowledge as possible. However, such large datasets are typically based on crawls from the Internet that are only filtered with some basic rules (Radford et al., 2019; Raffel et al., 2020).* ***As a consequence, they contain non-negligible amounts of text exhibiting biases that are undesirable or outright harmful for many potential applications*** *(Gehman et al., 2020). Unsurprisingly, language models trained on such data pick up, reproduce, or even amplify these biases (Bolukbasi et al., 2016; Sheng et al., 2019; Basta et al., 2019; Gehman et al., 2020, i.a.).*	*Non conceptualisé*
Ruder et al., 2022	*We will refer to this prototype as NLP's SQUARE ONE—and to the bias that follows from it, as the SQUARE ONE BIAS. We* ***argue this bias manifests in a particular way: Since research is a creative endeavor, and researchers aim to push the research horizon, most research papers in NLP go beyond this prototype, but only along a single dimension at a time****. Such dimensions might include multilinguality, efficiency, fairness, and interpretability, among others. The effect of the SQUARE ONE BIAS is to baseline novel research contributions, rewarding work that differs from the prototype in a concise, one-dimensional way.*	*Non conceptualisé*

Bien que ces études reconnaissent l'existence des biais et leurs conséquences négatives potentielles, elles laissent inexplicité la définition de biais et ce qu'ils entendent par ce terme. L'étude de Hutchinson *et al*. (2020) met en évidence les biais pouvant s'insérer dans les classificateurs de textes et la manière dont ils peuvent refléter les perceptions sociétales des personnes handicapées. Les auteurs soulèvent les inquiétudes quant à l'impact potentiel des modèles biaisés sur la réduction de l'autonomie et liberté des personnes affectées, mais manquent à définir ce concept et les types de biais dont il est question dans l'étude. Lalor *et*

al. (2022) étudient les biais intersectionnels en soulignant qu'ils résultent de l'interaction de plusieurs dimensions démographiques. Les auteurs soutiennent qu'ils s'appuient sur la littérature existante concernant les biais intersectionnels, mais n'explicitent pas le concept et s'en remettent à la déduction des lecteurs pour cadrer leur recherche. Matthews *et al*. (2021) reconnaissent que les mots associés à diverses professions peuvent refléter un biais de genre féminin ou masculin. Par contre, ils n'approfondissent pas explicitement ce qu'est un biais de genre. Schick *et al*. (2021) parlent des biais indésirables et nuisibles pouvant s'insérer dans les grands ensembles de données. Les auteurs vont jusqu'à mentionner que les systèmes NLP entraînés sur ces genres de données sont susceptibles de reproduire et amplifier ces biais, tout en laissant indéterminés les biais dont il est question et n'apportant aucune définition précise. De leur côté, Ruder *et al.* (2022) mettent en évidence les biais découlant d'un cadre d'analyse de recherche en NLP, mais ne fournissent pas de conceptualisation de ce terme. Ce manque explicite de conceptualisation et de définition des biais en TALN ne permet pas de bien encadrer la recherche, les préjudices associés, ainsi que les méthodes employées pour les analyser.

En effet, l'absence de définition approfondie de la notion de biais constitue une réduction des résultats pour plusieurs raisons. Premièrement, sans une définition claire et explicite du terme de biais, il devient difficile de comprendre précisément ce que les chercheurs considèrent comme des biais dans leurs études. Cela peut entraîner une confusion quant aux critères utilisés pour identifier et évaluer les biais, ce qui compromet la validité des conclusions tirées de ces études. Deuxièmement, l'absence de définitions claires peut conduire à une interprétation subjective des résultats par les chercheurs et les lecteurs, ce qui pourrait influencer la manière dont les résultats sont présentés et compris. Enfin, cela limite également la possibilité de comparer et de synthétiser les résultats entre différentes études, car l'absence de normes communes pour définir les biais rend difficile la comparaison des méthodes et des résultats. Nous pourrions aussi dire que ce manque de conceptualisation renforce la représentation selon laquelle les systèmes automatisés pourraient se débarrasser des biais qui ne seraient pas des éléments humains fondamentaux, mais bien des problèmes pouvant se résoudre pour atteindre une neutralité technique insaisissable (Andrejevic, 2020). Une réflexion plus approfondie sur la nature des biais et leurs impacts permettrait de sortir de l'approche de la neutralité technique qui ne permet pas de penser en profondeur les phénomènes de ce développement technologique.

Les études offrant une conceptualisation des biais au sein de notre corpus incluent un raisonnement normatif et technique. En effet, deux des études ayant conceptualisé les biais s'inspirent d'une même étude et donc, du même raisonnement.

Tableau 4.3 – Citations des sources originales pour les biais conceptualisés (normatif)

Source	***Citation***	***Classification***
Blodgett et al., 2020	*Indeed, the term "bias" (or "gender bias" or "racial bias")* ***is used to describe a wide range of system behaviors, even though they may be harmful in different ways, to different groups, or for different reasons.***	*Conceptualisé*
Dev et al., 2022	*Bias in language models is commonly defined as* ***"skew that produces a type of harm"*** *(Crawford, 2017)* ***towards different social groups, though it is a complex notion that is often not well-defined in existing literature*** *(Blodgett et al., 2020; Delobelle et al., 2022; Talat et al., 2022).*	*Conceptualisé*

L'étude de Dev. *et al*. (2022) s'inspire de celle de Blodgett *et al.* (2020) pour définir les biais. Il est possible de constater que ces définitions impliquent toutes deux de penser les biais en termes de préjudices pour certains groupes sociaux. Cette perspective suggère un aspect normatif en reconnaissant que les biais impactent, pour différentes raisons, négativement certaines populations. Tandis que l'étude de Shah et al. (2020) offre une conceptualisation technique des biais en offrant quatre origines possibles au sein des systèmes NLP.

Tableau 4.4 – Citations des sources originales pour les biais conceptualisés (technique)

Source	***Citation***	***Classification***
Shah et al., 2020	*In essence,* ***biases are priors that inform our decisions*** *(a dialogue system designed for elders might work differently than one for teenagers). Still, undetected and unaddressed,* ***biases can lead to negative consequences: There are aggregate effects for demographic groups, which combine to produce predictive bias.*** *[...] We identify four points within the standard supervised NLP pipeline where bias may originate: (1) the training labels (label bias), (2) the samples used as observations — for training or testing (selection bias), (3) the representation of data (semantic bias), or (4) due to the fit method itself (overamplification).*	*Conceptualisé*

Cette perspective technique se concentre sur les aspects opérationnels des biais dans les modèles de langage et fournit un cadre pour l'identification et l'analyse des biais dans les systèmes de traitement du langage naturel. Évidemment, il est possible de souligner qu'en identifiant les origines des biais de manière purement technique, ce cadrage ne considère pas de manière inclusive les réelles relations et contextes menant à ces derniers.

4.3.3 Relation entre biais et langage

La majorité des études de notre corpus présente une relation vague et souvent de nature technique entre les biais et le langage. Les études présentant une relation explicite entre les biais et le langage ont intégré une littérature pertinente permettant d'approfondir les normes et hiérarchies sociales pouvant façonner la langue.

Tableau 4.5 – Citations des sources originales offrant une relation explicite entre biais et langage

Source	***Citation***	***Classification***
Dev et al., 2022	***Dehumanizing language uses techniques such as moral disgust, denial of agency, or likening members of a target group to non-human entities (Markowitz and Slovic, 2020) to reinforce normative identities—often as indication of a biological hierarchy of 'species' within humankind**. […] **A single instance of language may represent/cause multiple forms of harm** (e.g., some Stereotyping harms may also be Dehumanization harms). Does the measure provide a method for measuring multiple harms separately as well as in aggregate (e.g., are subsets of the underlying data tagged along multiple axes)? What language and culture is the bias and measure most relevant to? […] Language data for bias measures is sourced primarily in two ways: by extracting from existing textual data or by generating from specific templates. While the first has the advantage of being more similar to "real samples" that models see, the latter has the advantage of testing for specific artifacts by construct.*	*Relation explicite entre biais et langage*
Blodgett et al., 2020	*Turning frst to (R1), **we argue that work analyzing "bias" in NLP systems will paint a much fuller picture if it engages with the relevant literature outside of NLP that explores the relationships between language and social hierarchies.** Many disciplines, including sociolinguistics, linguistic anthropology, sociology, and social psychology, study how language takes on **social meaning and the role that language plays in maintaining social hierarchies.** […] As a result, many groups have sought to bring about social changes through changes in language, disrupting patterns of oppression and marginalization via so-called "gender-fair" language (Sczesny et al., 2016; Menegatti and Rubini, 2017), language that is more inclusive to people with disabilities (ADA, 2018), and language that is less dehumanizing (e.g., abandoning the use of the term "illegal" in everyday discourse on immigration in the U.S. (Rosa, 2019).*	*Relation explicite entre biais et langage*

Dev *et al*. (2022) et Blodgett *et al*. (2020) traitent de la relation entre le langage et les hiérarchies sociales, ainsi que des préjudices potentiels associés au langage. Ils reconnaissent l'impact du langage sur le renforcement des identités normatives et le maintien des hiérarchies sociales. Leurs auteurs soulignent également l'importance de comprendre le langage dans un contexte plus large que celui du NLP. Ils soulignent la pertinence de disciplines telles que la sociolinguistique, l'anthropologie linguistique, la sociologie et la psychologie sociale pour examiner comment le langage façonne les significations sociales et la dynamique du pouvoir. Les textes reconnaissent que le langage n'est pas seulement un outil de communication neutre, mais qu'il a des implications sociales et culturelles. Au niveau culturel, le langage

est lié à l'expérience collective de la condition humaine et au niveau social, il reflète les relations et les appartenances des humains (Sherzer, 2012). L'une des contributions majeures des auteurs réside dans la reconnaissance que le langage va bien au-delà d'un simple moyen de communication. En effet, il agit comme un agent actif dans la consolidation des identités normatives et dans la reproduction des structures sociales préexistantes (Guy et Labov, 1997). Une approche instrumentale qui considère le langage de manière purement opérationnelle ne permet pas de s'intéresser aux mécanismes qui sous-tendent les inégalités sociales et culturelles. En effet, le langage peut refléter et perpétuer des stéréotypes de genre, être biaisé sur des bases raciales, refléter des stéréotypes sociaux profondément enracinés, etc (*Ibid*.). Il est donc crucial de concevoir le langage dans toute sa complexité, sans effacer des analyses son contexte de création, ses ambiguïtés et son évolution.

Tableau 4.6 – Citations des sources originales offrant une relation explicite (technique) entre biais et langage

Source	***Citation***	***Classification***
Hutchinson et al., 2020	*Neural text embedding models (Mikolov et al., 2013) are critical first steps in today's NLP pipelines. These models learn vector representations of words, phrases, or sentences, such that semantic relationships between words are encoded in the geometric relationship between vectors.* ***Text embedding models capture some of the complexities and nuances of human language. However, these models may also encode undesirable correlations in the data that reflect harmful social biases****. […] It is important to* ***recognize that social norms around language are contextual and differ across groups.*** *[…] models for detecting abuse can be used to nudge writers to rethink comments which might be interpreted as toxic (Jurgens et al., 2019). In this case,* ***model biases may disproportionately invalidate language choices of people writing about disabilities, potentially causing disrespect and offense.***	*Relation explicite entre biais et langage*

Bien que Hutchinson *et al*. (2020) présentent une relation un peu plus technique, ils soulignent également que les normes sociales entourant le langage sont contextuelles et varient d'un groupe à l'autre. Ce qui peut être considéré comme un langage acceptable ou respectueux dans une communauté ou une culture peut être offensant ou irrespectueux dans une autre. Par conséquent, ils prennent en considération que lorsque des modèles sont utilisés pour détecter les abus ou évaluer la qualité du langage, ils doivent être conçus et évalués avec soin pour éviter d'invalider de manière disproportionnée les choix linguistiques de certains groupes. Les auteurs considèrent l'impact potentiel que ces biais peuvent avoir sur les communautés marginalisées ou sous-représentées. Si les modèles de détection des abus sont biaisés, ils risquent d'aggraver cette marginalisation en censurant de manière disproportionnée le langage de ces groupes. De plus, si ces

mêmes modèles ne tiennent pas compte de la diversité des normes linguistiques, ils peuvent considérer comme abusif ou inapproprié le langage utilisé par certains groupes, même si ce langage est valide dans son propre contexte culturel.

Les études présentant une relation vague entre les biais et le langage offrent une représentation technique et traite davantage le langage comme une simple transmission de signaux quantifiables (Rouvroy, 2018) au lieu de s'engager pas dans une analyse plus approfondie.

Tableau 4.7 – Citations des sources originales offrant une relation vague entre biais et langage

Source	***Citation***	***Classification***
Shah et al., 2020	*This problem arises from* ***the tendency of statistical models to pick up on non-generalizable signals during the training process****. In the case of domains, these non-generalizations are words, phrases, or senses that occur in one text type, but not another. However, this kind of variation is not just restricted to text domains: it is a fundamental property of human-generated language: we talk differently than our parents or people from a different part of our country, etc. (Pennebaker and Stone, 2003; Eisenstein et al., 2010; Kern et al., 2016).* ***In other words, language reflects the diverse demographics, backgrounds, and personalities of the people who use it.***	*Relation vague entre biais et langage*
Matthews et al., 2020	***Gender bias in NLP has been well studied in English, but has been less studied in other languages.*** *In this paper, a team including speakers of 9 languages - Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof - reports and analyzes measurements of gender bias in the Wikipedia corpora for these 9 languages. We develop extensions to profession-level and corpus-level gender bias metric calculations originally designed for English and apply them to 8 other languages, including languages that have grammatically gendered nouns including different feminine, masculine, and neuter profession words.*	*Relation vague entre biais et langage*
Lalor et al., 2022	*Other recent studies* ***have empirically shown that the biases inherent in language models for gender and race intersections might exceed those observed for gender and race alone*** *(Tan and Celis, 2019), and that only debiasing along a single dimension can be problematic.*	*Relation vague entre biais et langage*
Schick et al., 2021	*When trained on large,* ***unfiltered crawls from the Internet, language models pick up and reproduce all kinds of undesirable biases that can be found in the data****: They often generate racist, sexist, violent, or otherwise toxic language.*	*Relation vague entre biais et langage*
Ruder et al., 2022	*Overall,* ***almost 70% of papers evaluate only on English, clearly highlighting a lack of language diversity in NLP*** *(Bender, 2011; Joshi et al., 2020). […] While studies have demonstrated the ability of word embeddings to capture linguistic information in English,* ***it remains unclear whether they capture the information needed for processing morphologically rich languages*** *(Tsarfaty et al., 2020). […] Another bias in our models relates to word order. In order for n-gram models to capture interword dependencies, words need to appear in the n-gram window. This will occur more frequently in languages with relatively fixed word order compared to languages with relatively free word order (Bender, 2011).*	*Relation vague entre biais et langage*

Bien que l'étude de Shah *et al*. (2020) reconnaissent que le langage reflète la diversité démographique, les différents contextes et personnalités de ceux qui l'utilisent, elle se concentre principalement sur l'aspect technique du problème et sur la tendance des modèles statistiques à détecter ces variations. Les auteurs abordent les défis rencontrés au cours du processus d'apprentissage, lorsque les modèles statistiques ont tendance à capturer et à apprendre des signaux non généralisables, plutôt que de discuter des aspects normatifs ou des implications sociales de la diversité linguistique. Matthews *et al*. (2020) abordent les biais de genre en mettant l'accent sur les mesures d'analyse et de calculs métriques. Malgré un sujet porteur d'implications sociales, les auteurs examinent seulement une relation à la langue basée sur l'application cohérente de technique de mesure, plus précisément, au développement d'extensions de calculs métriques conçus à l'origine pour la langue anglaise. Lalor *et al*. (2022) discutent de résultats d'études qui suggèrent que les biais intersectionnels de genre et de race peuvent être plus importants en termes techniques que les biais liés uniquement au genre ou à la race. L'accent est mis sur les défis et les considérations pour débiaiser plusieurs dimensions dans les modèles linguistiques, et non sur la relation du langage et des hiérarchies sociales pour les groupes situés à l'intersection de ces axes d'oppression. Shick *et al*. (2021) parlent de la relation technique entre les modèles linguistiques et les biais présents dans les données d'apprentissage. Lorsque les modèles linguistiques sont entraînés sur de grands ensembles de données non filtrées provenant de l'internet, ils apprennent à partir des modèles et des biais inhérents qui par conséquent peuvent reproduire et générer un contenu qui reflète les biais présents dans les données d'apprentissage. Les auteurs n'engagent pas avec la littérature et abordent cette relation de manière très technique en se concentrant sur la façon dont les modèles de langage « apprennent ». Puis, Ruder *et al*. (2022) mentionnent le manque de diversité linguistique dans le domaine du TALN et se concentrent sur l'incapacité de certains modèles à capturer efficacement les informations linguistiques nécessaires pour les langues qui ont des structures de mots plus complexes. Ils restent dans une représentation très technique et n'entrent pas dans la littérature linguistique pertinente pour exprimer cette diversité linguistique et les singularités qui la composent. Ils auraient pu, par exemple, discuter davantage de cette capture d'informations linguistiques en NLP qui se distingue du processus d'apprentissage de langage chez l'humain qui s'appuie sur l'intersubjectivité et l'attention partagée (Bender et Koller, 2020). Autrement dit, l'humain n'apprend pas la signification dans la forme des mots et il aurait été intéressant d'élaborer en quoi consiste cette dite « capture d'information » pour un système NLP.

4.3.4 Type de préjudices

De notre corpus, seulement deux études ne mentionnent pas les types de préjudices découlant des biais étudiés.

Tableau 4.8 – Citations des sources originales pour les types de préjudices non spécifiés

Source	***Citation***	***Classification***
Shick et al., 2021	*As a consequence, they contain non-negligible amounts of text exhibiting biases that are* ***undesirable or outright harmful for many potential applications.***	*Non spécifiés*
Ruder et al., 2022	*We argue that the SQUARE ONE BIAS* ***has several negative effects****, most of which amount to the study of one of the above dimensions being biased by ignoring the others. Specifically, by focusing only on exploring the edges of the manifold, we are not able to identify the non-linear interactions between different research dimensions. […] We discuss the impact of this prototype on our research community, and the bias it introduces.* ***We then discuss the negative effects of this bias.***	*Non spécifiés*

En effet, il est possible de constater que Shick *et al*. (2021) restent très imprécis, et ce, tout au long de leur étude, sur les effets négatifs et possibles préjudices associés aux biais qu'ils analysent. Ils abordent les effets indésirables et nuisibles des biais dans une multitude d'applications, sans pour autant définir et expliciter ces dernières. De son côté, l'étude de Ruder *et al*. (2022) suggère que l'existence d'un modèle prototypique de recherche en NLP introduit automatiquement des biais dans la manière de conduire ces recherches. Les auteurs expliquent la tendance de la recherche actuelle à se concentrer sur l'avancement d'une dimension spécifique de l'expérience prototypique, plutôt que d'explorer un large éventail d'aspects simultanément. Ces biais étudiés en faveur d'avancées unidimensionnelles sont ce que les auteurs appellent le *SQUARE ONE BIAS*. Ils soutiennent que le *SQUARE ONE BIAS* a plusieurs effets négatifs, dont la plupart reviennent à biaiser l'étude en se concentrant sur une dimension et ignorant les autres dimensions pouvant interagir avec cette dernière. Évidemment, puisque cette étude explore les biais de la recherche, la taxonomie de (Crawford, 2017) ne permet pas de catégoriser les préjudices découlant de ces derniers. Par contre, nous pouvons tout de même constater que d'aborder les effets négatifs en termes de biais, et non de préjudices possibles, revient à ignorer les implications potentielles de ce type de biais dans la façon de conduire la recherche en NLP sur les communautés visées, les prises de décision, l'élaboration de politiques, et bien plus encore.

Toutes les autres études incluses dans notre corpus abordent des types de préjudices pouvant être catégorisés sous la taxonomie de Crawford (2017). Matthews *et al*. (2020) et LaLor *et al*. (2022) se concentrent spécifiquement sur les préjudices d'allocation en donnant des exemples concrets et permettant de comprendre la nature transactionnelle de ces derniers.

Tableau 4.9 – Citations des sources originales pour les préjudices d'allocation

Source	***Citation***	***Classification***
Matthews et al., 2020	*They demonstrated that word embedding software trained on a corpus of Google news could associate men with the profession computer programmer and women with the profession homemaker. Systems based on such models, trained even with "representative text" like Google news,* ***could lead to biased hiring practices if used to, for example, parse resumes and suggest matches for a computer programming job.***	*Allocation*
Lalor et al. 2022	*[...] allocational harms - these "arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups." Allocational harm is often aligned with downstream tasks/interventions guided by the NLP model. [...]* ***Biases in predictions for healthcare-related variables*** *(Psychometrics)****, or personality type variables*** *(MBTI, FIPI)* ***can affect an individual's health care plan, personalized interventions, job prospects, etc. Biased predictions for drug rating sentiment can affect which drugs a future user chooses to take****. [...] Collectively, the results underscore the allocational harm implications of NLP models on several downstream tasks - ones that even well-designed and well-intentioned debiasing strategies cannot overcome. This can be problematic in the era of personalized marketing and precision health, with NLP-based persona-generation playing a bigger role. For tasks like numeracy and literacy,* ***this can affect how a patient is treated by a medical staff during a hospital visit (i.e., a false positive high literacy prediction for a person who has trouble understanding his or her medical record). For the personality indicators, inconsistent predictions may lead to biased decisions in the workplace (e.g., a manager looking to form a team of extroverts).*** *[...] We build on the emergent literature on intersectional biases by assessing datasets encompassing up to five demographic dimensions, in conjunction with state-of- the-art word embeddings and debiasing methods, on downstream tasks where biased predictions can lead to allocational harm (§3.1).*	*Allocation*

Matthews *et al.* (2020) étudient les associations biaisées faites par des logiciels d'intégration de mots (word embedding) entraînés sur des textes d'actualités provenant de Google. Ces associations biaisées entre professions et genres peuvent mener les systèmes NLP à favoriser injustement certains candidats et ainsi, réduire les opportunités de travail de certains groupes. Puis, Lalor *et al.* (2022) se basent sur la littérature concernant les préjudices d'allocation pour aborder les conséquences des biais qu'ils étudient. Ils vont même jusqu'à parler d'évidences empiriques dans la façon dont les modèles NLP peuvent avoir des conséquences concrètes sur les opportunités et les ressources allouées à certains groupes dans le domaine de la santé et du marketing personnalisé.

De leur côté, Hutchinson *et al.* (2020) et Shah *et al*. (2020) discutent des deux types de préjudices, soit d'allocation et de représentation. Il est possible de constater la présence dominante des biais d'allocation

qui se retrouvent dans la majorité de notre corpus, et ce, même dans les études qui abordent aussi les préjudices de représentation.

Tableau 4.10 – Citations des sources originales pour les préjudices d'allocation et de représentation

Source	***Citation***	***Classification***
Hutchinson et al., 2020	*If models inappropriately condition on mentions of disability, this could impact people writing, reading, or seeking information about a disability. Depending on how such models are deployed,* ***this could potentially result in reduced autonomy, reduced freedom of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals****.[...] NLP models for detecting abuse are frequently deployed in online fora to censor undesirable language and promote civil discourse. Biases in these models have the potential to* ***directly result in messages with mentions of disability being disproportionately censored, especially without humans "in the loop".*** *Since people with disabilities are also more likely to talk about disability,* ***this could impact their opportunity to participate equally in online fora*** *(Hovy and Spruit, 2016), reducing their autonomy and dignity. Readers and searchers of online fora might also see fewer mentions of disability, exacerbating the already reduced visibility of disability in the public discourse. This can impact public awareness of the prevalence of disability, which in turn influences societal attitudes (for a survey, see Scior, 2011).*	*Représentation et allocation*
Shah et al., 2020	*Predictive models in NLP are sensitive to a variety of (often unintended) biases throughout the development process. As a result, fitted models do not generalize well, incurring performance and reliability losses on unseen data. They also have socially undesirable effects* ***by systematically underserving or mispredicting certain user groups.*** *[...] Still, undetected and unaddressed, biases can lead to negative consequences:* ***There are aggregate effects for demographic groups****, which combine to produce predictive bias.*	*Représentation et allocation*

Bien que la recherche de Hutchinson *et al.* (2020) comporte davantage de préjudices de représentation en abordant la perpétuation des stéréotypes et des inégalités des personnes atteintes d'un handicap, elle aborde aussi les préjudices d'allocation face à l'octroi de ressources. Par exemple, les auteurs parlent des modèles NLP qui, utilisés pour détecter les abus dans les forums, peuvent avoir une incidence sur l'opportunité de certains groupes en situation de handicap de participer et discuter librement dans ce type de forum. Puis, Shah *et al.* (2020) discutent des biais dans les données et les algorithmes pouvant perpétuer des représentations erronées de certains groupes démographiques et donc, un traitement inégal. Ces préjudices de représentation mènent à certains préjudices d'allocation en négligeant systématiquement ces groupes au niveau des ressources pouvant leur être octroyées. Il est intéressant de soulever que Shah *et al.* (2020), qui présentaient une relation entre les biais et le langage vague, restent évasifs dans les effets concrets de ces préjudices. Par exemple, lorsque les auteurs affirment « […] biases can lead to negative consequences: There are aggregate effects for demographic groups […] » (Shah *et al.,* 2020), ils reconnaissent l'existence

d'effets négatifs résultant de ces biais, mais ils n'approfondissent pas spécifiquement ces conséquences. Il aurait été intéressant de développer la portée de ces préjudices d'allocation. Néanmoins, il est crucial de noter que cette recherche, ainsi que celle de Hutchinson *et al*. (2020) mettent en lumière l'importance de la prise de conscience de ces préjudices, à la fois en termes de représentation et d'allocation.

Enfin, deux études du corpus abordent les préjudices d'allocation et de représentation en termes de taxonomie permettant de décrire les biais étudiés.

Tableau 4.11 – Citations des sources originales des préjudices d'allocation et de représentation (taxonomie)

Source	***Citation***	***Classification***
Dev et al., 2022	*The relevant* ***harms can be subdivided into representational or allocational harms, depending on whether there is a generalization of harmful representations of groups or if there is a tangible, disparate distribution of resources between groups, respectively*** *(Crawford, 2017).*	*Représentation et allocation*
Blodgett et al., 2020	*We used a previously developed* ***taxonomy of harms for this categorization, which differentiates between so-called allocational and representational harms*** *(Barocas et al., 2017; Crawford, 2017). Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups; representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether. […] Adapting and extending this taxonomy, we categorized the 146 papers' motivations and techniques into the following categories: Allocational harms. Representational harms: Stereotyping that propagates negative generalizations about particular social groups. Differences in system performance for different social groups, language that misrepresents the distribution of different social groups in the population, or language that is denigrating to particular social groups.*	*Représentation et allocation*

Les études de Dev *et al*. (2022) et de Blodgett *et al.* (2020) offrent tous deux une guidance dans la manière de conduire la recherche sur les biais NLP. De par la nature de leurs recherches, il est sans surprise que leur approche s'accompagne d'une description des types de préjudices à considérer dans l'étude des biais en NLP. En plus d'avoir toutes deux conceptualisé le concept de biais de manière explicite, ces recherches mettent en évidence l'importance de reconnaître et d'analyser les biais de manière approfondie en considérant les préjudices qui leur sont associés. Leurs recherches jouent un rôle essentiel dans la sensibilisation à savoir comment l'automatisation peut affecter nos ressources et nos opportunités, ou encore, refléter et amplifier des préjugés déjà existants dans nos sociétés. Elles contribuent à rendre cette taxonomie

plus largement disponible en encourageant son utilisation lors de l'analyse des biais et en démontrant la pertinence de considérer la totalité des implications du phénomène étudié.

4.3.5 Objectif

La grande partie des études de notre corpus restent vagues quant à l'explicitation de leur objectif de recherche. En effet, il est possible de constater que certains objectifs restent très techniques dans les performances du système, sans raisonnement éthique ou encore, très généraux.

Tableau 4.12 – Citations des sources originales offrant des objectifs vagues

Source	***Citation***	***Classification***
Lalor et al., 2022	*There is **a need for a more systematic analysis of how current state-of-the-art language models and mitigation strategies perform** with regards to intersectional bias in down-stream tasks.*	*Vague*
Shick et al., 2021	*Building training datasets with more care and deliberation, an alternative solution discussed by Bender et al. (2021), is important, especially for improving linguistic and cultural diversity in online and other forms of communication. However, for large language models that are available for common global languages, it is desirable to also have other mechanisms to address bias because dataset curation and documentation is extremely resource intensive, given the amount of data required. It can also necessitate building different training sets and, accordingly, training different models for each desired behavior, which can result in high environmental impact (Strubell et al., 2019). In this paper**, we therefore propose an approach that, instead of trusting that a model will implicitly learn desired behaviors from the training data, makes explicit how we expect it to behave at test time: If the model is told which biases are undesired—and it is able to discern their presence—it should be able to avoid them** even if they are present in some of the texts it has been trained on.*	*Vague*
Matthews et al., 2020	*While our **goal in this study was to identify a defining set and profession set that could more easily be used across many languages and for which the T-test results indicated no statistically significant difference in results over the English Wikipedia corpus**, it would be interesting to repeat this analysis with additional variations in the defining set and profession set.*	*Vague*
Hutchinson et al., 2020	*This paper focuses on the representation of persons with disabilities through the lens of technology. Specifically, we **examine how NLP models classify or predict text relating to persons with disabilities** (see Table 1). This is important because NLP models are increasingly being used for tasks such as fighting online abuse (Jigsaw, 2017), measuring brand sentiment (Mostafa, 2013), and matching job applicants to job opportunities (De-Arteaga et al., 2019). In addition, since text classifiers are trained on large datasets, the biases they exhibit may be indicative of societal perceptions of persons with disabilities. […]**Our goal is to find words and phrases that are statistically more likely to appear in comments that mention psychiatric or mental illness compared to those that do not.***	*Vague*

Ruder et al., 2022	*We provide* ***historical and recent examples of how the square one bias has led researchers to draw false conclusions or make unwise choices, point to promising yet unexplored directions on the research manifold, and make practical recommendations*** *to enable more multi-dimensional research.*	*Vague*

Par exemple, Lalor *et al*. (2022) expliquent que certaines études récentes ont démontré qu'il peut être problématique de débiaiser les modèles NLP en fonction d'une seule dimension (ex. : race, genre), puisque les biais intersectionnels intégrés dans ces modèles seraient plus complexes et nombreux. Leur recherche vise à répondre à la nécessité d'une analyse systémique complète des performances et de l'efficacité des modèles NLP les plus récents dans les stratégies d'atténuation des biais intersectionnels dans des applications pratiques. L'objectif reste très technique et empirique, tout en laissant inexplicité pourquoi il est problématique de débiaiser les modèles en fonction d'une seule dimension et comment le système NLP contribue à renforcer certaines injustices. Shick *et al.* (2021), face à la nécessité de disposer d'alternatives d'apprentissage moins énergivore, ont comme objectif de proposer une approche pour rendre explicites les comportements souhaitables au modèle NLP pour qu'il puisse être en mesure d'éviter les biais indésirables. En anthropomorphisant ces modèles et en se concentrant sur les performances du système, les auteurs restent imprécis sur les objectifs à savoir s'il s'agit de réduire l'impact environnemental ou encore, de réduire les biais. Puisque les biais ne sont pas définis, ni la façon dont un système plus performant pourrait bénéficier d'une consommation d'énergie moindre, l'objectif reste vague. Matthews *et al.* (2020) combine plusieurs idées, telles que l'identification d'un ensemble de définitions et de professions pouvant être plus facilement utilisé par de nombreuses langues (autres que l'anglais) et dont le T-test n'indique pas de différence significative en rapport au corpus de Wikipédia en anglais. Le T-test est un test statistique couramment utilisé pour comparer les moyennes de deux échantillons afin d'évaluer si la différence observée est susceptible d'être due au hasard ou si elle est statistiquement significative (Prabhakaran, 2020). L'objectif n'explicite pas comment les résultats du T-test, utilisé pour évaluer les différences entre ces ensembles, s'inscrivent dans un cadre théorique plus large, et permet de discuter les implications potentielles des résultats attendus. Hutchinson *et al*. (2020) cherchent à trouver les mots et les phrases qui sont statistiquement plus susceptibles d'apparaître dans les commentaires abordant la maladie mentale. Ils n'énoncent pas clairement les objectifs ou les résultats spécifiques qu'ils cherchent à atteindre. Ils se concentrent sur la représentation des personnes en situation de handicap à travers le prisme de la technologie, mais ils restent vague sur la façon dont les systèmes NLP contribuent à définir les perceptions sociales de ces dernières. Finalement, bien que Ruder *et al.* (2022) cherchent à identifier des directions de recherche inexplorées et établir des recommandations pratiques, leur objectif reste vague sur ce qu'ils ont l'intention d'accomplir dans ces domaines et ce qu'ils visent à atteindre.

Ces objectifs que nous considérons « vagues » restent dans une logique d'automatisation de l'atténuation des biais et de résultats statistiques plus significatifs. C'est-à-dire, une logique de collecte automatisée des données et de leur traitement menant à des réponses automatisées améliorées (Andrejevic, 2020). Les objectifs restent au niveau des performances et les auteurs n'engagent pas de raisonnement éthique qui permettrait de penser la relation de ces améliorations aux personnes directement impactées. Autrement dit, les objectifs sont davantage centrés sur les systèmes et ne rendent pas compte en quoi ces améliorations répondent aux besoins des personnes les plus vulnérables (Birhane, 2021).

Seulement trois études de notre corpus ont, de manière explicite, abordé les objectifs de leur recherche.

Tableau 4.13 – Citations des sources originales explicitant les objectifs

Source	***Citation***	***Classification***
Dev et al., 2022	*This **paper is motivated by two main goals**. The first goal is to **define a practical framework for harms that is both theoretically-motivated and empirically useful for describing bias measures**. […] The second goal **is to define a collection of documentation questions around bias measures that helps others capture measure limitations and align operationalizations of "biases" to harms**. […] To achieve these goals, we organize a practical framework of harms, a tagged collection of 43 existing bias measures and the associated harms, a set of documentation questions, and a collection of case studies.*	*Explicite*
Blodgett et al., 2020	*We provide a list of all 146 papers in the appendix. […] Once identifed, we then read each of the 146 papers **with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating "bias."** […] We then **describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems.** We argue that such work should examine the relationships between language and social hierarchies; we call on researchers and practitioners conducting such work to articulate their conceptualizations of "bias" in order to enable conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why; and we recommend deeper engagements between technologists and communities affected by NLP systems. We also provide several concrete research questions that are implied by each of our recommendations.*	*Explicite*
Shah et al., 2020	*[…] much work has focused on bias effects and symptoms rather than their origins. While it is essential to address the effects of bias, it can leave the fundamental origin unchanged (Gonen and Goldberg, 2019), requiring researchers to rediscover the issue over and over. **The "bias" discussed in one paper may, therefore, be quite different than that in another. A shared definition and framework of predictive bias can unify these efforts, provide a common terminology, help to identify underlying causes, and allow coordination of countermeasures** (Sun et al., 2019). However, such a general framework had yet to be proposed within the NLP community. **To address these problems, we suggest a joint conceptual framework**, depicted in Figure 1, outlining and relating the different origins of bias. […]We hope this paper will help researchers spot, compare, and address bias in all its various forms.*	*Explicite*

Les auteurs de ces études énoncent nettement les objectifs de leur recherche, donnent une orientation claire sur la façon de les atteindre, en plus d'être guidés par des préoccupations normatives. Ce qui signifie que leurs objectifs sont établis en fonction de ce qui est souhaitable pour mener, analyser et améliorer la recherche d'un sujet extrêmement complexe. Nous pouvons constater, puisque ces études ont pour objectif de proposer un cadre conceptuel commun pour la recherche, qu'elles visent à instaurer une base solide et unifiée permettant d'orienter les chercheurs vers une compréhension commune et cohérente de ce domaine d'étude.

4.3.6 Méthodologie

Nous pouvons constater que les études utilisant des méthodes quantitatives travaillent sur un cadre purement technique des biais contenus dans les modèles et offrent une considération limitée du contexte entourant la formation des données utilisées dans leur recherche. Elles utilisent une approche visant à simplifier le sujet étudié et à rationaliser les processus automatisés en termes d'opérations concrètes (Andrejevic, 2020). Bien que ces recherches soient essentielles dans l'amélioration des performances et offrent des représentations importantes de la façon dont les systèmes NLP peuvent être nuisibles, les approches opérationnelles de leur méthodologie peuvent négliger les questions et analyses plus complexes et nuancées soulevées par la littérature sociale.

Par exemple, Hutchinson *et al.* (2020) ont pratiqué une méthode d'analyse des biais sur un ensemble de 1000 phrases extraites d'un sous-corpus de la plateforme Reddit créé par Voigt *et al.* (2018). Ils se sont concentrés sur des mesures de co-occurrence de 56 expressions linguistiques utilisées pour parler de personnes avec divers types d'handicap, afin d'examiner la fréquence à laquelle ces mots apparaissent. Cette relation statistique ne prend pas explicitement en compte les facteurs contextuels entourant l'utilisation du langage relatif aux handicaps.

Tableau 4.14 – Citations des sources originales méthodologie quantitative

Source	***Citation***	***Classification***
Hutchinson et al., 2020	***For each category, we show the average score diff for recommended phrases vs. non-recommended phrases along with the associated error bars.*** *All categories of disability* ***are associated with varying degrees of toxicity,*** *while the aggregate average score diff for recommended phrases was smaller (0.007) than that for non-recommended phrases (0.057).*	*Quantitative*

Or, en l'absence du contexte socioculturel dans lequel ces associations s'inscrivent, il peut s'avérer difficile de faire la distinction entre un véritable biais et un usage descriptif du langage, voire une réappropriation de certains termes par ces groupes marginalisés. Par contre, il est important de préciser que les auteurs ont admis avoir travaillé cette problématique de manière technique et reconnaissent qu'une utilisation statistique de la définition de l'équité devrait s'accompagner de justifications normatives et sociales.

Matthews *et al.* (2020), de leur côté, ont utilisé une méthode quantitative pour mesurer les biais de genre de diverses professions présents dans les corpus Wikipédia de neuf langues différentes. Ils se sont appuyés sur un ensemble de paires de mots sexués créé en anglais par Bolukbasi *et al.* (2016) pour définir les relations de genre. Cet ensemble comprend, par exemple, les mots « she/he », « daughter/son », etc. Les auteurs ont ensuite adapté et traduit cet ensemble dans huit langues, dont des langues qui ont des noms de professions grammaticalement genrés, avec des mots féminins, masculins et neutres. Puis, ils ont entraîné le modèle algorithmique Word2Vec, largement utilisé en TALN pour apprendre les représentations de mots, avec ces ensembles afin d'analyser les biais de genre présents dans les corpus Wikipédia de ces langues. Leur objectif était de créer des ensembles de professions dans différentes langues, pouvant être utilisés en apprentissage automatique pour atténuer les biais de genre, démontrant des résultats statistiquement similaires à ceux obtenus en anglais.

Tableau 4.15 – Citations des sources originales méthodologie quantitative

Source	***Citation***	***Classification***
Matthews et al., 2020	*In some languages like Chinese, Arabic, and Wolof, there are different words for younger and older sister or brother. We also considered and discarded many other profession words such as bartender, policeman, celebrity, and electrician.* ***For example, we discarded bartender because it is not a legal profession in some countries.***	*Quantitative*

Bien que les auteurs aient retiré certaines professions illégales dans certains pays de leurs ensembles, ils n'ont pas ajouté de métiers spécifiques à ces pays ou encore, considéré qu'ils pouvaient exister de manière clandestine. En utilisant l'anglais comme point de référence et de comparaison pour mener cette étude, les ensembles ne couvrent pas toutes les nuances et variations culturelles spécifiques à chaque langue. Les normes de genre et les stéréotypes peuvent différer considérablement d'une culture à l'autre, ce qui rend difficile la comparaison directe des résultats entre les langues.

Lalor *et al.* (2022) ont effectué une évaluation comparative de plusieurs modèles NLP en fonction de dix tâches de classification dans le but d'évaluer leur performance en matière de biais intersectionnels à travers cinq dimensions démographiques (le sexe, la race, l'âge, l'éducation et le revenu). Les tâches observées impliquaient l'analyse de contenu généré par les utilisateurs à partir de plateformes telles que Twitter, Reddit, etc. Les auteurs ont établi diverses combinaisons de dimensions graphiques. Ils ont par la suite déterminé des groupes spécifiques pouvant faire l'objet de préjugés dans ces dimensions, ainsi que des groupes privilégiés, susceptibles de faire l'objet de moins ou peu de préjugés. Puis, ils ont comparé les résultats entre ces groupes en utilisant des mesures d'équité pour quantifier et évaluer toute disparité.

Tableau 4.16 – Citations des sources originales méthodologie quantitative

Source	***Citation***	***Classification***
Lalor et al., 2022	*There are several definitions of fairness in the literature (Mehrabi et al., 2021), each with corresponding methods of assessment. In this work* ***we rely on two prior metrics from the literature, and also present a new metric, adjusted disparate impact, to account for base rates in the dataset.***	*Quantitative*

Suivant ces mesures, la méthode se concentre sur des combinaisons de dimensions démographiques, mais ne peut tenir pleinement compte de la complexité de l'intersectionnalité. L'intersectionnalité reconnaît que les expériences des individus en matière de préjugés sont façonnées par les interactions entre les multiples dimensions de leur identité. En examinant les dimensions démographiques séparément ou dans des combinaisons prédéfinies, la méthode peut négliger les expériences distinctes des individus ayant des identités intersectionnelles complexes. De plus, cette méthode peut ne pas tenir pleinement compte des facteurs contextuels qui contribuent aux préjugés. Les préjugés sont influencés par des dynamiques sociales, culturelles et historiques, ainsi que par des déséquilibres de pouvoir et des structures systémiques (Delouvée, 2018). Ces facteurs contextuels sont essentiels pour comprendre les expériences singulières de préjugés auxquelles les individus sont confrontés.

En comparaison, il est possible d'observer chez les trois études de notre corpus ayant utilisé des méthodologies qualitatives qu'elles incluent d'importantes considérations quant aux facteurs contextuels qui contribuent aux biais. Leur méthode offre un cadre davantage normatif pour analyser les biais et leur déploiement au sein des systèmes NLP. Ce sont des études qui visent l'amélioration dans la façon de conduire la recherche sur les biais en NLP et qui, de ce processus intrinsèquement normatif, s'engagent moins avec le volet technique de ce sujet et s'ancrent davantage dans la littérature.

Tableau 4.17 – Citations des sources originales méthodologie qualitative

Source	***Citation***	***Classification***
Dev et al., 2022	***Our practical framework of harms builds upon existing taxonomies of representational harms*** *(e.g. Blodgett (2021) and establishes specific heuristics (Appendix A)* ***to disentangle the characteristics of five non-mutually exclusive categories****. While these harms have previously been taxonomized, we ground the definitions of harms into documentation questions and heuristics to help practitioners align NLP bias measures with specific harms.*	*Qualitative*
Ruder et al., 2022	***We annotate the 461 papers that were presented orally at ACL 2021****, a representative cross-section of the 779 papers accepted to the main conference. […] Overall, almost 70% of papers evaluate only on English, clearly highlighting a lack of language diversity in NLP (Bender, 2011; Joshi et al., 2020). […]* ***We discuss the impact of this prototype on our research community, and the bias it introduces. We then discuss the negative effects of this bias.*** *We also list work that has taken steps to overcome the bias.*	*Qualitative*
Blodgett et al., 2020	***Our survey includes all papers known to us analyzing "bias" in NLP systems—146 papers in total****. We omitted papers about speech, restricting our survey to papers about written text only. […] Once identifed, we then read each of the 146 papers with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating "bias."* ***We used a previously developed taxonomy of harms for this categorization, which differentiates between so-called allocational and representational harms*** *(Barocas et al., 2017; Crawford, 2017).*	*Qualitative*

Par exemple, Dev *et al.* (2022) ont organisé un cadre pratique des biais et préjudices en NLP, en plus d'un ensemble de questions pouvant aider et guider les chercheurs dans le développement des mesures de biais. Leur méthode est soutenue par des concepts provenant de la psychologie sociale et de la linguistique et vise un alignement plus profond du contexte culturel menant aux biais et préjudices associés. Leur approche ne repose pas sur l'utilisation de techniques pratiques, mais bien théoriques. De leur côté, Blodgett *et al.* (2020) ont réalisé une analyse critique de 146 études en catégorisant les motivations et techniques utilisées pour mesurer ou atténuer les biais. En suivant une taxonomie existante et en l'adaptant à leurs objectifs, leurs auteurs ont utilisé six catégories pour classifier leur corpus. Ils ont choisi de présenter sous forme de données statistiques les catégories d'analyse de leur corpus pour ensuite en offrir une analyse approfondie. Leur méthode accorde une grande importance à la relation entre les biais, les préjudices et le langage. Elle permet aux auteurs d'analyser ces études techniques avec un raisonnement éthique et social explicite. Puis, Ruder *et al*. (2022) ont annoté 461 études selon les dimensions suivantes : multilinguisme, équité et biais, efficacité et interprétabilité. Tout comme Blodgett *et al*. (2020), ils ont choisi de présenter leurs résultats sous forme de pourcentage en présentant le nombre d'articles appartenant aux diverses dimensions à l'étude. Les

auteurs ont par la suite offert une analyse historique incluant des exemples récents appuyés de littérature scientifique explicitant comment ces données impactent et biaisent la recherche en NLP.

Puis, deux études de notre corpus ont utilisé des méthodologies mixtes. Il est possible de constater, malgré un usage combiné de méthodologies, que ces recherches penchent toujours majoritairement vers une approche quantitative technique avec une dimension qualitative moindre. Ce qui, indéniablement, façonne la manière d'analyser les biais.

Par exemple, Schick *et al*. (2021) ont évalué de manière quantitative la capacité des modèles de langage à détecter et corriger les caractéristiques indésirables des résultats qu'ils produisent. Le tout, en proposant un algorithme de décodage réduisant la probabilité d'un modèle de produire du texte biaisé. Suite à ces évaluations, les auteurs ont présenté une analyse qu'ils ont qualifiée de qualitative. Cette analyse consiste à examiner manuellement la qualité du texte généré afin d'évaluer l'efficacité du modèle GPT2-XL à produire des réponses appropriées.

Tableau 4.18 – Citations des sources originales méthodologie mixte

Source	***Citation***	***Classification***
Schick et al., 2021	***Qualitative Analysis*** *Table 4 shows five selected prompts from the challenging subset of RealToxicityPrompts as well as continuations generated by GPT2-XL with regular decoding and with self-debiasing using λ = 10; all texts are generated with greedy decoding and a beam size of 3. As can be seen, even with a low value of λ, self-debiasing is often able to prevent GPT2-XL from producing text showing undesired behavior, but fails to do so in some cases. Table 4 also illustrates the problem of imperfect classifications by Perspective API: the self-debiased output for the second prompt is wrongly classified as being a threat, and that for the fourth prompt as being toxic and sexually explicit.*	*Mixte*

Il est possible de constater que cette section de leur étude offre davantage une présentation descriptive des résultats de l'expérience qu'une analyse qualitative. L'absence de cadre analytique ne permet pas d'interpréter les données ou encore, d'approfondir conceptuellement les explications des résultats en rapport aux biais et conserve la prédominance technique de la recherche.

Enfin, Shah *et al.* 2020 ont développé un cadre conceptuel sur les origines des biais en se basant sur la littérature pertinente en TALN et certains travaux en sciences sociales. Bien que les auteurs cherchent à présenter ces phénomènes sociaux complexes tout en s'ancrant dans la littérature, ils ont aussi choisi

d'analyser les origines des biais dans les systèmes de manière technique en fournissant des définitions quantitatives des biais prédictifs.

Tableau 4.19 – Citations des sources originales méthodologie mixte

Source	***Citation***	***Classification***
Shah et al., 2020	*We base our framework on an extensive survey of the relevant NLP literature, informed by selected works in social science and adjacent fields. […] Our primary contributions include: (1) a conceptual* ***framework for identifying and quantifying predictive bias and its origins within a standard NLP pipeline****, (2) a survey of biases identified in NLP models, and (3) a survey of* ***methods for countering bias in NLP*** *organized within our conceptual framework.*	*Mixte*

Selon les auteurs, ces définitions techniques permettent d'identifier les biais au sein des systèmes NLP, en plus de développer des mesures de prévention alignées à leurs origines. Par contre, ce choix méthodologique de traiter des problèmes sociaux de manière technique ne permet pas d'interroger les implications normatives des origines des biais dans ces systèmes et limite la profondeur de l'analyse.

4.3.7 Résultats

Les résultats présentés par les études de notre corpus ont été classifiés en deux catégories, soit « performance du système » ou « compréhension des biais ». Nous pouvons constater un nombre égal de recherches provenant de chacune de ces catégories. En effet, notre corpus offre quatre études dont les résultats sont axés sur l'évaluation, la quantification et l'atténuation des biais dans les modèles de langage, ainsi que quatre études dont les résultats permettent une meilleure compréhension du concept de biais.

Dans les résultats généraux en lien avec la performance du système, on retrouve l'étude de Hutchinson *et al.* (2020) qui a permis de démontrer la présence de biais envers les personnes en situation de handicap dans trois modèles NLP disponibles. Ensuite, l'étude de Shick *et al.* (2021) qui, basée sur le fait que les modèles NLP ont la capacité de diagnostiquer jusqu'à un certain degré les biais qu'ils produisent, a produit un algorithme permettant d'entraîner les modèles à mieux apprendre les comportements indésirables du système. Puis, l'étude de Lalor *et al.* (2022) a prouvé que certaines méthodes pour débiaiser les modèles NLP performent pauvrement lorsqu'il s'agit de prendre en considération des biais intersectionnels composés de plusieurs dimensions telles que le genre, l'âge, la race, etc. Enfin, l'étude de Matthews *et al.* (2020) qui, à l'aide d'une méthodologie existante adaptée, a quantifié les biais de genre dans différentes langues.

Tableau 4.20 – Citations des sources originales pour les résultats axés sur la performance du système

Source	***Citation***	***Classification***
Hutchinson et al., 2020	*We have presented evidence that these concerns extend to biases around disability**, by demonstrating bias in three readily available NLP models that are increasingly being deployed in a wide variety of applications.** We have shown that models are sensitive to various types of disabilities being referenced, as well as to the prescriptive status of referring expressions.*	*Performance du système*
Schick et al., 2021	*[...] we first demonstrate a surprising finding: **Pretrained language models recognize, to a considerable degree, their undesirable biases and the toxicity of the content they produce. We refer to this capability as self-diagnosis.** Based on this finding, we **then propose a decoding algorithm that, given only a textual description of the undesired behavior, reduces the probability of a language model producing problematic text.***	*Performance du système*
Lalor et al., 2022	*We also look at known debiasing methods for these models and show that **while the debiased versions maintain predictive performance (as expected), they do not help with mitigating biases.** While most models are relatively fair when looking at a single demographic characteristic, **accounting for intersectional groups leads to less fair models and wider ranges of bias because of the combinatorial considerations of the intersectional groups**.*	*Performance du système*
Matthews et al., 2020	*We have extended an influential method for computing gender bias from Bolukbasi et al., a technique that had only been applied in English. We made key modifications that allowed us to extend the methodology to 8 additional languages, including languages with grammatically gendered nouns. **With this, we quantified how gender bias varies across the Wikipedia corpora of 9 languages** and discuss future work that could benefit immensely from a computational linguistics perspective.*	*Performance du système*

Il est intéressant de noter que ces résultats issus de la performance du système proviennent de quatre études de notre corpus dont le concept de biais n'était pas formellement conceptualisé. Ce manque de conceptualisation peut entraîner l'utilisation de méthodes inappropriées pour détecter et évaluer les biais qui sont sous-entendus par la recherche. Sans une définition claire du biais dans son contexte de recherche et de ce qu'il englobe, les résultats de l'étude peuvent refléter de manière plus ou moins fidèle les véritables problèmes/solutions présentés. Nous pouvons aussi pointés que les résultats de ces études s'ancrent dans la critique implicite du solutionnisme technologique. L'idée selon laquelle les technologies peuvent résoudre tous les problèmes tend à encourager une confiance excessive dans les solutions technologiques pour des problèmes socialement enracinés qui ne s'accompagnent d'aucune définition claire du concept de biais lui-même (Morozov, 2014). Néanmoins, ces résultats, pris dans leur contexte, permettent une application pratique et une amélioration continue de la recherche.

Puis, les études dont les résultats généraux sont en lien avec la compréhension des biais ont, pour la majorité d'entre elles, conceptualisé le concept de biais. Ces études s'adressent davantage aux chercheurs en NLP et offrent des résultats pouvant guider de manière éthique et améliorer la recherche dans ce domaine.

Tableau 4.21 – Citations des sources originales pour les résultats axés sur la compréhension des biais

Source	***Citation***	***Classification***
Blodgett et al., 2020	*By surveying 146 papers analyzing "bias" in NLP systems, we found that (a)* ***their motivations are often vague, inconsistent, and lacking in normative reasoning****; and (b)* ***their proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP.*** *To help researchers and practitioners avoid these pitfalls, we proposed* ***three recommendations that should guide work analyzing "bias"*** *in NLP systems, and, for each, provided several concrete research questions.*	*Compréhension des biais*
Shah et al., 2020	*We present a comprehensive overview of the recent literature on predictive bias in NLP. Based on this survey,* ***we develop a unifying conceptual framework to describe bias sources and their effects (rather than just their effects).*** *This framework allows us to group and compare works on countermeasures.*	*Compréhension des biais*
Dev et al., 2022	*We organize a framework to define and* ***distinguish between different types of harms****—presented through heuristics and documentation questions—****to guide more intentional development of bias measures.*** *Our proposed documentation template also facilitates combining, comparing, and utilizing different bias measures, and continuously revisiting them to update limitations and comparative understanding with other measures.*	*Compréhension des biais*
Ruder et al., 2022	*We highlighted the associated SQUARE ONE BIAS,* ***which encourages research to go beyond the prototype in a single dimension. We discussed the problems resulting from this bias****, by studying the area statistics of a recent NLP conference as well as by discussing historic and recent examples.*	*Compréhension des biais*

Par exemple, Blogdett *et al.* (2020) ont trouvé, dans un corpus de 146 études sur les biais en NLP, que les motivations étaient souvent floues et que les techniques quantitatives proposées pour mesurer les biais n'étaient pas alignées aux objectifs de la recherche. Ce à quoi les auteurs ont émis trois recommandations pour guider les recherches analysant les biais en NLP consistant à s'engager davantage dans la littérature à l'extérieur du domaine du NLP, à mieux décrire en quoi les systèmes décrits comme biaisés sont dangereux et en examinant plus en profondeur la relation entre les langages et les communautés affectées par l'usage de ces systèmes. De leur côté, Shah *et al.* (2020) ont créé un cadre conceptuel visant à identifier les origines des biais au sein des systèmes NLP, ainsi que leurs effets. Leur cadre offre une meilleure compréhension

des mécanismes techniques sous-jacents qui conduisent à l'apparition de biais dans les applications NLP. En analysant les sources potentielles de biais, telles que les données d'entraînement, ils cherchent à fournir des outils pour atténuer ces effets indésirables. Dev *et al.* (2022) ont aussi présenté un cadre conceptuel, mais cette fois-ci, permettant de mieux aligner les mesures des biais à l'étude aux types de préjudices qui leur sont associés. Puis, Ruder *et al.* (2022) ont présenté les biais introduits par le prototype de recherche qui étudie une seule dimension d'un problème complexe et invitent les chercheurs à aller vers une recherche multidimensionnelle en NLP. Les résultats visant une compréhension plus profonde de la complexité des biais sont essentiels pour encourager la responsabilisation des chercheurs. Ces derniers jouent un rôle clé dans la conception et le développement des systèmes d'IA. En comprenant les biais et en fournissant des cadres d'analyse plus adaptés, ces recherches peuvent améliorer la fiabilité et la capacité de généralisation des résultats découlant des performances des systèmes NLP.

Il est intéressant de constater que les chercheurs présentant des résultats en rapport aux performances du système ont majoritairement une relation technique dans la façon de traiter les biais. Par exemple, Matthews *et al.* (2020) affirme que « […] what we can measure, we can more easily begin to track and improve […]. » Tandis que les chercheurs dont les résultats portent sur la compréhension des biais offrent une perspective normative plus approfondie sur la complexité du concept de biais et de la relation qui les unit aux humains (Birhane, 2021). Or, il est possible de constater un manque d'équilibre entre ces différentes approches. Il serait pertinent que les chercheurs en NLP se nourrissent davantage de ces deux types d'analyse et, au lieu de les voir comme deux champs de recherche différents, de les considérer comme complémentaire l'un à l'autre. En intégrant à la fois les approches techniques qui permettent de mesurer et d'améliorer les performances du système et les perspectives normatives qui approfondissent la compréhension des biais, les chercheurs pourraient produire des résultats plus robustes et éthiquement responsables dans le domaine du TALN.

4.3.8 Limitation

Plus de la moitié des études de notre corpus ont explicité les limitations de leur recherche. Il est possible de constater deux facteurs communs dans les limitations présentés de ces études sur les biais en NLP : la capacité de généralisation et la méthodologie. Nous considérons les limitations présentées comme un levier essentiel pour mettre en évidence des manques qui ne s'accompagnent pas toujours d'une solution, mais sont révélateurs de l'état actuel de la recherche en TALN.

Tableau 4.22 – Citations des sources originales pour les limitations de recherche explicites (capacité de généralisation)

Source	***Citation***	***Classification***
Hutchinson et al., 2020	*One limitation of this paper is its* ***restriction to the English language*** *and* ***US sociolinguistic norms****.*	*Explicite*
Matthews et al., 2020	*One important limitation to* ***note is that for many languages, if a word is expressed with a multi-word phrase (e.g. astronomer in Arabic), the word count reported by Word2Vec for this phrase will be zero.*** *For each language, there is a tokenizer that identifies the words or phrases to be tracked. In many cases, the tokenizer identifies words as being separated by a space. The Chinese tokenizer however attempts to recognize when multiple characters that are separated with spaces should be tracked as a multi-character word or concept.*	*Explicite*
Dev et al., 2022	*We acknowledge that our framework of harms has been* ***created from a US-centric perspective and has been influenced by the Social Dominance Theory*** *(Sidanius and Pratto, 2001), which can be limiting from a global perspective and does not include cultural harms.*	*Explicite*

Les études de Hutchinson *et al.* (2020), Matthews *et al.* (2020) et Dev *et al.* (2022) ont toutes présenté des limitations sur la capacité de généralisation de leur recherche. Il est intéressant de constater que l'ensemble de leurs limitations pointent vers la spécificité culturelle, c'est-à-dire qu'elles sont toutes liées à des contraintes culturelles ou linguistiques spécifiques qui peuvent restreindre la généralisation et l'applicabilité de l'étude à d'autres cadres respectifs. Hutchinson *et al*. (2020) se limitent à la langue anglaise et aux normes sociolinguistiques américaines. Cela signifie que les résultats, les conclusions et les implications de leur recherche peuvent ne pas s'appliquer directement à d'autres langues ou contextes socioculturels en dehors des États-Unis. Matthews *et al.* (2020) présentent une limite concernant Word2Vec, un modèle de langage développé par Google et entraîné à reconnaître les relations sémantiques entre les mots. Les auteurs soulignent que les performances de l'outil peuvent varier d'une langue à l'autre en raison de l'incapacité du modèle à reconnaître des caractéristiques linguistiques spécifiques présentes dans des langues telles que l'arabe et le chinois. Puis, Dev *et al.* (2022) affirme que leur cadre conceptuel a été créé d'un point de vue centré sur les États-Unis, ce qui implique que le cadre pourrait ne pas englober tous les préjudices et toutes les perspectives culturelles, et que son applicabilité pourrait être limitée lorsqu'il s'agit d'analyser les problèmes dans les différents contextes culturels du monde.

Tableau 4.23 – Citations des sources originales pour les limitations de recherche explicites (méthodologie)

Source	***Citation***	***Classification***
Lalor et al., 2022	*In this work, our scope is debiasing embeddings, not debiasing classifiers. While there is much work in the area of debiasing classifiers,* ***here we restrict our focus to the debiasing of embeddings.***	*Explicite*
Schick et al., 2021	*One major limitation of our evaluation* ***is that it relies to a large extent on attribute scores assigned by Perspective API; this means not only that we cannot thoroughly test the effectiveness of our method for many relevant biases that are not measured by the API,*** *but also that our labels are error-prone. For example, Perspective API may fail to detect more subtle forms of bias and be overreliant on lexical cues (Gehman et al., 2020).*	*Explicite*

Puis, les études de Lalor *et al.* (2022) et Schick *et al.* (2021) ont de leur côté explicité des limitations méthodologiques. Dans le cas de Lalor *et al*. (2022), la restriction exprimée consiste à limiter l'objectif de l'étude à débiaiser une seule composante du modèle, soit les « embeddings », et à exclure les classificateurs.[22] Ce qui signifie que les chercheurs travaillent spécifiquement sur la réduction des biais liés aux représentations numériques des mots ou des phrases (embeddings), mais qu'ils n'abordent pas les biais qui peuvent survenir lors de l'utilisation d'algorithmes pour catégoriser ou étiqueter les données (classificateurs). Or, dans le contexte de leur étude qui s'intéresse aux biais intersectionnels, cette restriction pourrait exclure certains aspects importants de la détection et réduction des biais qui peuvent aussi survenir au moment de la classification des données. En ce qui concerne Schick *et al.* (2021), les auteurs ont exprimé une limitation liée à Perspective API, un outil d'apprentissage automatique qu'ils ont utilisé pour identifier la capacité des modèles de langage a diagnostiquer les contenus toxiques et biais indésirables. Cependant, puisque Perspective API ne mesure pas certains biais importants, tels que les biais de genre, l'évaluation de la méthode de débiaisement peut être incomplète et ne pas refléter la complexité des biais présents dans un ML. Conscients de ces limitations et pour offrir une analyse plus complète et inclusive, les auteurs ont additionnellement utilisé un ensemble de données conçu pour mesurer le degré de présence de neuf différents types de biais sociaux dans les modèles de langage. Il est important que les chercheurs prennent

[22] Les « embeddings » sont des représentations numériques de mots ou de phrases qui tentent de capturer le sens et le contexte des mots. Les classificateurs sont des algorithmes utilisés pour catégoriser ou classer des données dans différents groupes ou catégories sur la base de certaines caractéristiques ou attributs (Bujokas, 2020).

en compte ces limitations lors de l'interprétation des résultats pour envisager d'autres méthodes d'évaluation complémentaires et mieux comprendre l'efficacité réelle de la méthode de débiaisement proposée.

Enfin , Blogett *et al.* (2020), Shah *et al.* (2020) et Ruder *et al.* (2022) n'ont pas fait mention des limites de leur recherche. Ces études, dont l'objectif principal visait à présenter des recommandations et cadres d'analyse pour guider les chercheurs en NLP, ont mis l'accent sur la présentation de ces résultats plutôt que sur les limitations spécifiques des études incluses dans leur recherche. Nous pourrions caractériser ce manque commun de « focus sur les recommandations » : c'est-à-dire que les auteurs de ces études considèrent que le but premier de leur travail est de fournir des recommandations pratiques et peuvent supposer que les lecteurs de leurs travaux comprendront que des limites existent inévitablement, mais qu'elles sont inhérentes et implicitement comprises dans les résultats présentés.

CHAPITRE 5
ANALYSE ET DISCUSSION

Dans ce présent mémoire, la question de recherche que nous avions établie un peu plus tôt était : De quelle façon les différentes recherches en traitement automatique du langage naturel étudient-elles les biais de ce domaine de l'intelligence artificielle ? À cette dernière, nous avions ajouté les trois sous-questions suivantes : Quelles similitudes et différences peut-on observer dans la manière dont les chercheurs étudient les biais dans notre corpus ? Quels constats est-il possible de tirer des différentes approches utilisées ? Quelles limites peut-on observer dans la façon de s'engager avec les biais ? Ces questions visaient à observer la façon dont différentes recherches étudiaient les biais en NLP. Dans le chapitre précédent, nous avons présenté les résultats de notre grille d'analyse par catégorie à l'étude. Nous proposons maintenant d'examiner ces résultats dans une analyse complète et alignée à nos questions de recherche. Dans le but d'enrichir notre analyse, plusieurs concepts de notre cadre seront mobilisés dans ce chapitre pour ancrer et soutenir notre discussion.

5.1 Analyse globale

En examinant les résultats que nous avons présentés précédemment, nous avons établi une analyse globale de la façon dont les différentes recherches incluses dans notre corpus ont étudié les biais. Pour commencer, il est possible de constater que le manque de conceptualisation des biais au sein de notre corpus a impacté la façon de mener certaines recherches. En effet, les objectifs de recherche, les méthodologies utilisées, ainsi que les résultats présentés ont perdu en cohérence lorsque les biais étudiés n'étaient pas bien cadrés et développés par les auteurs. Ce manque de clarté dans la définition des biais peut conduire à des évaluations erronées des modèles en NLP et peut entraîner des généralisations inappropriées des résultats de recherche. Blodgett *et al.* (2020) et Dev *et al.* (2022) ont aussi exprimé que ce manque de conceptualisation et de contextualisation des biais repose souvent sur des présomptions non explicitées de la part des auteurs. Ce qui conduit fréquemment au développement de mesures de biais qui s'avèrent inadéquates, voire impossibles à évaluer quant à leur capacité à réellement saisir les distinctions significatives en matière de biais.

Bien que la majorité des études de notre corpus ont fait mention des types de préjudices pouvant découler des biais intégrés dans les modèles étudiés, nous avons constaté une faible relation explicitée entre le concept de langage et de biais. En effet, la plupart des études de notre corpus sont restées vagues et très techniques quant à la façon dont le langage qui se déploie dans ces systèmes renforcit les préjudices mentionnés. Parmi

les études qui ont abordé ce sujet, certaines ont souligné que le langage utilisé pour entraîner les modèles peut refléter les biais présents dans les données d'apprentissage. Par exemple, si les données d'entraînement contiennent des préjugés envers certaines catégories de personnes, le modèle apprendra à reproduire ces biais dans ses prédictions. Toutefois, ces études (ex. : Shick *et al.,* 2021) n'ont souvent pas approfondi comment ces mécanismes de renforcement des préjugés opèrent concrètement. Elles n'ont pas suffisamment examiné les liens entre les caractéristiques linguistiques spécifiques et les biais qui en résultent. Seules les études ayant conceptualisé les biais (ex. : Blodgett *et al.,* 2020 ; Dev *et al,* 2022) se sont engagées dans la relation qui les unit au langage. Ceci pouvant s'expliquer par la façon beaucoup plus normative dont les auteurs ont traité ce sujet. Néanmoins, ce constat souligne une vision du langage très opérationnelle, offrant la possibilité de modéliser le social et de le rendre statique (Rouvroy, 2018). Autrement dit, il existe donc toujours un biais opérationnel dans les études sur les biais en TALN, puisque le fondement même de ce domaine repose sur une conception opérationnelle du langage, où le langage est traité principalement comme un outil fonctionnel pour réaliser des tâches spécifiques plutôt que comme un système complexe influencé par des facteurs socioculturels et historiques. Ce manque de relation entre le langage et les biais soulève la nécessité d'une compréhension plus approfondie des mécanismes linguistiques qui sous-tendent les biais et les préjudices dans les modèles d'apprentissage automatique.

De plus, en considérant que les études utilisant une méthodologie mixte au sein de notre corpus avaient une dimension quantitative prédominante, il est possible d'observer la forte utilisation du quantitatif pour mener ce type d'étude. L'utilisation de ces méthodes quantitatives pour réguler les interactions humaines est une forme d'opérationnalisme qui soutient l'automatisation de processus sociaux (Andrejevic, 2020). Il est permis de déduire que la recherche en TALN, de par la nature des systèmes étudiés, tend à traiter des problèmes sociaux (biais) de manière technique. Ce même constat résonne dans les études qualitatives (Blodgett *et al.,* 2020 et Dev *et al.,* 2022) de notre corpus qui visent à offrir un cadre et des recommandations de recherche aux études quantitatives et qui soutiennent que les chercheurs gagneraient à s'engager plus en profondeur dans les dimensions sociales et humaines des systèmes étudiés. Par exemple, Blodgett *et al.* (2020) ont démontré que la grande majorité des études utilisant des méthodes quantitatives n'explicitent pas de littérature pertinente à l'extérieur du domaine du NLP pour fonder la pertinence des mesures utilisées pour identifier, calculer et atténuer les biais. Ce qui a pour effet qu'un bon nombre d'études utilisent des techniques qui ne sont pas alignées à leur objet de recherche. Dans le but d'éviter ce manque d'alignement, les auteurs proposent trois recommandations pour la recherche traitant des biais en TALN ; qu'elle soit fondée sur une littérature pertinente et externe au NLP ; qu'elle fournisse des explications détaillées sur la façon dont les comportements du système décrits comme biaisés sont nuisibles et envers quel groupe ; qu'elle examine l'utilisation de ces technologies en s'engagent dans de réelles expériences vécues par les

personnes et communautés affectées par ces systèmes (Blodgett *et al.,* 2020). S'engager dans de réelles expériences implique de questionner ce que les systèmes font, de prendre en compte leurs conséquences et d'interroger leurs raisons d'être. En d'autres mots, penser la relation de l'humain à la machine signifie de prendre comme point de départ à un problème donné les personnes directement impactées, plutôt que de se précipiter vers des solutions techniques (Birhane, 2021).

En ce qui concerne les résultats découlant des diverses méthodes de notre corpus, ils ont pu être catégorisés en termes de « performance du système » et de « compréhension des biais ». Nous remarquons que les résultats axés sur les performances du système sont associés à des recherches n'ayant pas conceptualisé les biais étudiés. Ce qui a pour effet de ne pas bien cadrer l'applicabilité des résultats obtenus et la validité de la technique utilisée pour étudier ces biais. L'accent excessif sur la performance technologique sans avoir intégré une réflexion approfondie sur les biais peut conduire à considérer la technologie comme une solution à tout (Morozov, 2014). Du côté des résultats visant la compréhension des biais, nous constatons une maîtrise du concept de biais beaucoup plus en profondeur, mais une faible interaction avec les résultats issus des performances du système.

Enfin, plus de la moitié des études de notre corpus ont présenté les limites de leur recherche. Il est intéressant d'observer que ces limitations, touchant à la capacité de généralisation des résultats et à la méthodologie, sont pour la plupart en lien avec une recherche centrée dans la langue anglaise, selon des normes sociales et linguistiques américaines. Cette tendance soulève des préoccupations quant à la représentativité des résultats obtenus dans notre corpus et à leur applicabilité à des contextes linguistiques et culturels différents. En se limitant à la langue anglaise et aux normes américaines, les études comprises dans notre étude ont majoritairement négligé les spécificités des autres langues et cultures, entraînant un manque important de diversité dans la façon de conduire la recherche en NLP. L'autrice Mar Estrach que nous avons abordée dans notre cadre a soutenu ce point en dressant le portrait actuel de la réglementation en IA. Elle a démontré que les concepteurs qui travaillent sur ces systèmes ont un profil similaire et sont souvent issus de sphères sociales et culturelles similaires (Estrach, 2022). Pourtant, les systèmes d'IA sont mis en œuvre dans des contextes sociaux très différents de leur entraînement initial, ce qui souligne la nécessité de prendre en compte la diversité linguistique et culturelle pour assurer une utilisation plus représentative et adaptée de ces technologies.

5.1.1 Similitudes et différences

Dans les similitudes observées dans la façon de conduire la recherche sur les biais en NLP, nous avons pu constater une vision dominante du langage opérationnel. En effet, la majorité des liens entre les concepts de

biais et de langage établis dans notre corpus étaient vagues, techniques et axés sur la performance du système. Les concepts fondamentaux permettant d'expliquer les biais dans le langage étaient écartés et remplacés par des aspects observables (Andrejevic, 2020). Par exemple, Shah *et al*. (2020) ont abordé la tendance des modèles statistiques à prendre en compte des signaux non généralisables (mots, phrases, sens, etc.) durant leur entraînement. Il est possible de voir dans cet exemple que les concepts fondamentaux sur ce que constituent les biais reproduits par les ML sont complètement simplifiés et rationalisés en termes d'opération (*Ibid*.). Cette considération de la capacité des modèles à capturer et reproduire le langage amène la recherche à mesurer la qualité d'un système en se basant sur les performances dans des tâches spécifiques, plutôt que de véritables critères linguistiques et sociaux.

Nous pouvons également observer, dans notre corpus, une prédominance à mener la recherche en considérant que le problème de biais peut être résolu par le développement et l'utilisation de nouvelles technologies. Par exemple, Schick *et al*. (2021) ont souligné l'importance de bâtir des ensembles de données plus rigoureusement de manière à améliorer la diversité culturelle et linguistique, mais ont soutenu la nécessité d'avoir d'autres mécanismes pour adresser les biais des grands modèles de langage puisqu'ils exigent trop de données. Cette confiance dans le développement de nouvelle méthode plus adaptée occulte la possibilité de penser les répercussions sociales et environnementales de l'utilisation même de ces modèles (Pitron, 2021). Le solutionnisme technologique est un biais épistémologique fort dans les représentations de l'IA et de son développement. Il est intéressant de voir dans les recherches de notre corpus qu'il ne s'agit jamais de remettre en question l'utilisation même de ces technologies, mais bien de guider la recherche à aller plus en profondeur dans le concept de biais, à évaluer et mesurer les systèmes NLP. La confiance résolue placée dans le développement de nouvelles méthodologies et techniques toujours mieux adaptées pour traiter le sujet, révèle la propension importante à négliger l'exploration des retombées sociales et environnementales découlant de l'usage de ces modèles.

L'une des différences majeures dans la façon de conduire la recherche sur les biais en NLP dans notre corpus réside dans la façon de s'engager avec les biais. En effet, nous avons noté que les études ayant conceptualisé et cadré les biais étudiés se sont intéressées aux relations entre nos interactions et le développement de savoirs humains (Birhane, 2021). Par exemple, Blodgett *et al*. (2020) ont fortement explicité l'importance d'aborder la relation entre le langage et les hiérarchies sociales dans les recherches en TALN, incitant les chercheurs à s'intéresser aux expériences humaines des communautés affectées par le déploiement de ces technologies. En revanche, les études s'étant peu engagées avec la littérature sur les biais ont pour la plupart proposé des solutions techniques à des problèmes sociaux sans considérer de manière adéquate les communautés impactées par ces systèmes. Schick *et al*. (2021) ont proposé un algorithme pour entraîner les

ML à reconnaître les comportements indésirables reproduits par les systèmes. Or, ce raisonnement sans contexte est privé d'expériences réelles et rend susceptible la reproduction de résultats nuisibles envers les communautés affectées (Birhane, 2021).

5.1.2 Les différentes approches

Dans les différentes approches utilisées par notre corpus, nous avons constaté une certaine dominance pour les études quantitatives. Il est intéressant d'observer qu'un sujet empreint de composantes normatives est, en grande majorité, traité de manière extrêmement technique. Les types de recherches effectuées avec des méthodes qualitatives ou quantitatives offrent des résultats très différents qui nous permettent d'affirmer que la pensée dichotomique persiste dans ce domaine. Soit, il s'agit d'étudier ce sujet avec un raisonnement très normatif, plongé dans la compréhension et la complexité du sujet (Dev *et al.,* 2022 ; Blodgett *et al.,* 2020 ; Ruder *et al.,* 2022) ; soit il s'agit de l'analyser de manière technique dans son rapport au système (Matthews *et al.,* 2020 ; Lalor *et al.,* 2022 ; Hutchinson *et al.,* 2020). Les recherches considérant les normes, questions éthiques et morales sont souvent séparées du « travail scientifique rationnel » et n'appartiennent pas aux mêmes approches (Birhane, 2021). En intégrant davantage de méthodes qualitatives, ou encore en intégrant les résultats des études qualitatives visant une meilleure compréhension des biais, les études quantitatives pourraient gagner en pertinence et en contextualisation. Elles pourraient ainsi mieux saisir les implications éthiques, les préjugés ou les discriminations potentielles liées aux systèmes TALN, en plus de développer des stratégies alignées à leurs objectifs pour les atténuer. La complémentarité entre les approches qualitative et quantitative permettrait de mieux appréhender la complexité des problématiques sociales liées au TALN et de réduire la perspective neutre détachée de toute implication morale de la recherche. Le rationalisme et l'idée de résultats statiques pouvant améliorer et potentiellement rendre ces modèles sans biais ne peuvent guider ces recherches sur des sujets fondamentalement humains et en constante évolution.

Avec cette dominance du quantitatif au sein de notre corpus, il est aussi possible de noter que les recherches techniques, telles que celle de Lalor *et al.* (2022), ont tendance à analyser les biais en termes de problèmes pour lesquels une solution se doit d'exister (*Ibid.).* En utilisant des méthodes statistiques pour traiter ces problèmes sociaux, les auteurs misent sur le fait que des concepts en constante évolution et liés à une réalité humaine bien réelle pourraient être complètement détachés de leur contexte de création d'origine. Sans compréhension approfondie de leurs fondements et contextes, il s'agit de traiter le problème en surface. Cependant, cela ne signifie pas que les démarches quantitatives sont inadéquates ou à disqualifier. Au contraire, elles fournissent des outils indispensables pour quantifier et mesurer les biais dans le systèmes en NLP, offrant ainsi une base solide pour l'identification et la correction de ces biais. C'est en quoi, nous notons que la recherche en NLP devrait davantage se nourrir de différentes approches et méthodologies

complémentaires. D'une part, une approche interdisciplinaire qui s'engage avec la littérature issue de domaines variés, tels que la sociologie et la linguistique, pourrait apporter des perspectives et des connaissances essentielles pour mieux comprendre les enjeux socioculturels sous-jacents aux problèmes traités. D'autre part, une approche inclusive qui privilégierait l'intégration des communautés affectées par les systèmes NLP dans la recherche permettrait de mieux comprendre comment les préjudices causés par les biais se déploient.

5.1.3 Les limites face aux biais

En observant les résultats de notre grille, nous avons soulevé plusieurs limites importantes quant à la façon d'étudier les biais. Évidemment, nous avons déjà évoqué précédemment le manque de définition des biais étudiés et un faible engagement de la relation qui unit les biais au langage qui ont pour effet de laisser à la déduction du lecteur l'encadrement de la recherche, en plus d'affaiblir la portée des résultats. À ceci, nous ajoutons un manque de relation entre les types de préjudices mentionnés par les auteurs et les communautés qui sont réellement impactées par le déploiement de ces modèles de langage. Par exemple, l'étude de Hutchinson *et al.* (2022) qui s'intéresse aux biais présents dans certains modèles NLP visant les personnes en situation de handicap, affirme que les choix linguistiques biaisés des modèles de langage pourraient être offensifs et irrespectueux. Le point de départ pour régler ce problème ne devrait pas être au niveau de l'ajustement technologique, mais bien au niveau des personnes directement visées. Les solutions aux problèmes techniques devraient être centrées sur les humains impactés et rendre compte de la relation entre ces systèmes et les communautés affectées (Birhane, 2021). C'est ainsi que l'éthique relationnelle permet de prendre en considération les véritables besoins des autres dans le développement technologique. En intégrant une perspective éthique, la recherche bénéficierait d'être guidée par une compréhension approfondie des expériences et des besoins des personnes affectées par ces développements. Il est donc primordial d'élargir la portée des évaluations des biais au-delà des métriques traditionnelles en adoptant des approches plus nuancées qui explorent une variété de contextes d'utilisation et qui tiennent compte des conséquences à long terme. Pour s'y faire, l'accent doit être mis sur la transparence[23] et la traçabilité tout au long du cycle de vie des modèles de langage. Documenter et rendre public les décisions prises de la collecte des données à la conception des algorithmes permettrait non seulement de mieux comprendre l'origine des biais, mais aussi de responsabiliser les parties prenantes face aux conséquences réelles de ces derniers.

[23] La transparence en NLP vise l'architecture des ML, ainsi que les données utilisées pour l'entraînement des modèles. OpenAI, par exemple, n'a pas rendu publics les ensembles de données utilisées pour entraîner les modèles qui soutiennent ChatGPT (Schaul *et al.*, 2023).

De plus, les limitations offertes par diverses études de notre corpus permettent d'observer une recherche extrêmement centrée sur la langue anglaise, ainsi que les normes sociales, culturelles et linguistiques américaines. Nous en avons fait mention précédemment, mais nous tenions à expliciter cette limite que nous considérons extrêmement révélatrice face à l'inclusion sociale, voire mondiale, de cette technologie. Par exemple, Dev *et al.* (2022) affirme avoir créé un cadre d'analyse centré sur les perspectives américaines. Hutchinson *et al.* (2020) explique avoir restreint son étude à la langue anglaise et ses normes sociolinguistiques. Puis, Matthews *et al.* (2020) exprime que l'utilisation de Word2Vec lors de leur recherche a été limitée par certaines langues (chinois et arabe) dont l'outil ne reconnaissait pas certains caractères spéciaux. Évidemment, l'utilisation d'un outil déployé et entraîné en anglais, pour analyser des biais de genre dans diverses langues, ne permet pas de saisir la complexité et la singularité linguistique de ces langues. Analyser les biais de genre dans différentes langues en utilisant un outil qui élimine le sens, les ambiguïtés, les nuances et les éléments constitutifs au langage, ne permet pas de penser la diversité et la richesse culturelle inhérente à chacune de ces langues. Ce qui soulève des préoccupations quant à la généralisation des résultats obtenus et à leur application dans des contextes variés.

Ce que l'omniprésence des études centrées sur la langue anglaise et la culture américaine témoigne, c'est une inégalité inquiétante dans la représentation et la considération des autres cultures et langues en NLP. Cette tendance réduit la portée des recherches sur les biais en NLP et crée des obstacles pour les personnes et les groupes linguistiquement marginalisés. Pour une véritable inclusion sociale de la technologie, il est crucial de promouvoir une recherche plus équilibrée et équitable qui prend en compte la diversité linguistique et culturelle mondiale. Il est essentiel que les chercheurs en TALN prennent conscience de cette limitation et s'efforcent d'inclure un échantillon plus varié de langues et de contextes culturels dans leurs recherches. Cela permettra d'améliorer la pertinence et la portée des résultats, en fournissant une meilleure compréhension de l'implication des systèmes étudiés dans un contexte global et diversifié. En élargissant les perspectives linguistiques et culturelles, la recherche en TALN pourrait ainsi contribuer de manière plus significative aux enjeux sociaux et humains liés aux technologies du langage naturel.

5.2 Discussion et perspective future

Enfin, nous souhaitons soutenir comme perspective future que la vision de croissance infinie du développement des grands ML ne peut constituer une vision acceptable. Ce développement technologique massif s'accompagne d'un coût environnemental important, trop souvent occulté, et nous soutenons que l'éthique technologique doit prendre en compte les limites de notre planète et de l'entendement humain (Bontems, 2017). L'équité, au travers de l'étude des biais, et l'impact environnemental sont des axes de recherche importants en NLP. Pourtant, il est intéressant de soulever que dans notre corpus seulement deux

études ont fait une brève mention de la question environnementale (Ruder *et al*., 2022 et Schick *et al.,* 2021). Or, comme présenté par l'étude de (Hessenthaler *et al.*, 2022), ce manque d'interaction entre ces deux axes de recherche est problématique. Une recherche simplement focalisée sur les biais peut ne pas prendre en considération les impacts environnementaux dans les solutions proposées et à l'inverse, une recherche focalisée sur la réduction de la consommation énergétique des modèles peut ne pas prendre en considération les entraînements nécessaires à l'atténuation des biais. L'empreinte carbone des larges réseaux de neurones dépend de certaines configurations de l'algorithme, du type de matériel informatique utilisé pour l'entraîner et de la production d'électricité nécessaire lors de cet entraînement (Patterson *et al.*, 2021). Par contre, un entraînement plus efficace, qui hypothétiquement pourrait avoir une consommation énergétique moindre, ne garantit pas une considération sur les performances du système en termes de biais.

S'il est vrai que nous avons une responsabilité prospective au développement technologique, l'environnement est un concept central pour réitérer l'importance de prendre en compte les effets de nos actions dans leur totalité (Bérubé, 2007). Les impacts néfastes des changements climatiques impactent davantage les communautés marginalisées (Bender *et al.*, 2021). Dans cette perspective, intégrer le concept de décroissance à notre réflexion pourrait offrir une voie alternative pour réduire l'empreinte écologique de l'intelligence artificielle. Alors que les grands ML sont présentement prédominants dans la langue anglaise, tout comme nous avons pu le constater dans notre corpus, et qu'ils nécessitent une consommation d'énergie importante, il est juste de dire que ces technologies ne sont pas destinées aux populations qui seront le plus affectées par les changements environnementaux. Encourager une transition vers une économie de la décroissance implique de reconsidérer notre modèle de développement basé sur une croissance exponentielle, en privilégiant des solutions plus économes et respectueuses de l'environnement. C'est en quoi nous soutenons l'importance pour la recherche future à prendre en compte l'intersectionnalité importante des éléments fondamentaux qui se doivent de composer davantage la recherche en TALN.

CONCLUSION

« Les mots justes trouvés au bon moment sont de l'action. »
Arendt, 1958

Dans ce mémoire, nous souhaitions analyser la façon dont différentes recherches en TALN étudient les biais de ce domaine. Au travers une méta-synthèse qualitative, nous avons analysé la façon dont huit études se sont engagées avec le sujet, ce qui nous a permis de démontrer certaines similitudes, différences et limites de la recherche actuelle. Nous avons débuté notre recherche en s'intéressant historiquement au développement de la linguistique computationnelle. En passant aux effets de la seconde guerre mondiale sur ce sujet, au développement de la cybernétique, à la naissance de l'obsession d'un monde transcrit sous forme de données, nous considérons avoir poser les bases du développement technologique de ce domaine.

Puis, nous avons développé le cadre conceptuel soutenant notre recherche. Le concept d'automatisation, incluant l'opérationalisme et le langage opérationnel, nous a permis de cadrer la nature des transformations actuelles qui ne sont pas seulement d'ordre mécanique, mais bien informationnel (Andrejevic, 2020). Cette proposition d'automatisation du travail mental, le dépossédant de toutes composantes sociales, s'accompagne de nombreuses conséquences qu'il nous a été possible de constater au travers des biais étudiés par notre corpus. Ensuite, le concept de biais algorithmique nous a offert la possibilité de développer de manière cohérente et théorique notre sujet d'étude. En ajoutant les biais épistémologiques des représentations de l'IA, la neutralité technique et le solutionnisme technologique, nous avons cadré que les biais ne sont pas toujours technologiques avec des répercussions sociales, mais bien inclus dans la conception même de l'IA. En terminant avec le concept d'éthique technologique, nous avons abordé le principe responsabilité, l'éthique relationnelle et l'importance du travail critique, qui permettent de penser la voie de l'équité dans son ensemble. Cette fondation solide incluant divers penseurs et leurs théories, nous a permis de donner l'angle souhaité à notre recherche, en plus de poser les bases fondamentales de notre pensée. Nous avons intégré ces concepts à nouveau dans notre analyse, servant de points de référence importants pour guider et appuyer certains constats.

Pour mener notre étude, nous avons déterminé les catégories que nous souhaitions analyser dans chacune des recherches de notre corpus. Pour chaque catégorie à l'étude, une classification a été développée et intégrée à notre grille d'analyse. Notre grille a permis de structurer et de donner une orientation claire à notre recherche. Elle a été constituée de manière à soutenir notre cadre théorique et d'offrir une comparaison plus facile entre notre corpus, nous permettant de regrouper les données similaires et distinguer plus facilement les différences. Il nous était important, considérant la méthodologie choisie, de conserver la

singularité de chacune des études de notre corpus. C'est pour cette raison que des tableaux incluant des citations des sources originales ont été intégrés tout au long des résultats, de manière à supporter les choix de classification effectués. Puis, c'est lors de notre analyse et discussion que nous avons été vers des résultats et conclusions plus générales et moins axées sur les recherches individuelles.

En analysant la façon de conduire la recherche sur les biais en TALN au sein de notre corpus, nous avons constaté que certaines études ont souffert d'un manque de clarté dans la définition des biais, ce qui a affaibli la portée de leurs objectifs, méthodologies et résultats. Autrement dit, le manque de cadrage d'un sujet aussi vaste et complexe impacte la qualité des méthodes choisies pour évaluer ou mesurer les systèmes NLP. De plus, un manque d'explication claire de la relation entre le langage et les biais dans les systèmes étudiés n'a pas permis d'approfondir la complexité humaine et sociale des systèmes étudiés. Les recherches qui se sont principalement concentrées sur des approches quantitatives ont souvent négligé des dimensions importantes dans les méthodes d'analyse choisies et les données utilisées. Nous avons aussi observé le manque d'interaction entre les résultats des études de notre corpus, catégorisés en termes de performances du système et de compréhension des biais. Les résultats découlant des études visant la compréhension des biais devraient davantage résonner dans les études visant les performances du système, permettant d'ajouter un raisonnement éthique examinant les principes, valeurs et normes qui devraient guider les actions et les décisions prises. Enfin, la plupart des études ont reconnu les limites de leur recherche, notamment en termes de généralisation des résultats et de la méthodologie utilisée. Une préoccupation majeure concerne la représentativité des résultats, car la plupart des recherches sont centrées sur la langue anglaise et les normes américaines, négligeant ainsi la diversité linguistique et culturelle de systèmes pourtant déployés mondialement.

Ce que ces résultats nous ont permis de soutenir, c'est qu'il est essentiel pour la recherche en TALN de mieux cadrer les biais étudiés. Une meilleure définition et identification permettra d'opter pour des techniques alignées aux objectifs de la recherche. Plusieurs recherches qualitatives, telles que celle de Blodgett *et al.* (2020) et Dev *et al*. (2022) offrent des cadres d'analyse extrêmement pertinents permettant de guider tout chercheur en NLP qui souhaite s'engager avec le sujet. De plus, nous soutenons qu'une meilleure combinaison des recherches techniques et normatives pourrait permettre d'approfondir la relation complexe des biais au langage et d'analyser en profondeur les interactions de ces systèmes avec les humains qui les utilisent.

5.3 Contribution de la recherche

Dès le début de cette recherche, nous avons positionné que ce mémoire s'inscrit dans une période de réflexion et d'analyse importante du développement de certaines technologies en IA, et ce, particulièrement dans le domaine du TALN. Le déploiement public de technologie telle que ChatGPT au courant de la dernière année est venu bouleverser plusieurs sphères de nos vies, qu'elles soient professionnelles, académiques ou personnelles. En 2020, Timnit Gebru a été renvoyée de chez Google suite à la publication de l'article *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big ?* Cet article fut l'un des premiers à aborder publiquement les dangers associés au NLP et à mettre en lumière un sujet qui, jusque-là, était très méconnu du public. Le directeur du département d'intelligence artificielle de Google, Jeff Dean, avait affirmé que cet article ignorait plusieurs recherches importantes et qu'il soulevait des préoccupations concernant les biais présents dans les ML sans prendre en considération les recherches récentes visant à atténuer ces problèmes (Newton, 2020).

Ce que nous pouvons constater aujourd'hui, c'est la complexité de ce sujet et qu'il ne s'agit pas de dire que des méthodes d'atténuation des biais existent, mais bien de considérer la relation de ces biais aux communautés affectées. Puisque nous avons noté une faible conceptualisation des biais, nuisant au développement de technique d'analyse alignée au sujet de recherche, nous pourrions même questionner la portée et l'applicabilité réelles des recherches visant à atténuer des problèmes sociaux dont la complexité humaine n'est souvent pas prise en considération. Il y a une réalité humaine dans ces systèmes, dans ceux qui la conçoivent, dans les travailleurs exploités qui ont entraîné ces systèmes (Hao, 2023), dans les données utilisées, etc. Cette étude critique est une contribution aux analyses des biais en NLP et permet de penser le problème en termes de fondement, en plus de soutenir que le concept de biais n'est pas statique et technologiquement saisissable, mais bien en constante évolution. La recherche sur les biais en TALN se doit de considérer davantage les personnes impactées par ces systèmes, en plus d'accorder une attention particulière à la contribution des activités technologiques aux changements climatiques, en particulier sur les populations qui sont davantage touchées par ces derniers en raison de leur position géographique.

Enfin, nous aimerions conclure ce mémoire en soutenant qu'un modèle de langage est trop grand lorsque l'ensemble des données utilisées à sa création est trop important pour être documenté (Gebru *et al.*, 2023). Pour des raisons éthiques et environnementales, nous ne devrions pas nous concentrer sur des modèles de plus en plus grands, mais bien sur des données de meilleure qualité (*Ibid*.).

ANNEXE A

DÉTAIL DE L'ÉCHANTILLON INITIAL SÉLECTIONNÉ

1. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. et Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. Dans *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (p. 1630-1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

2. Blodgett, S. L., Barocas, S., Daumé Iii, H. et Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5454-5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

3. Bender, E. M. et Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604. https://doi.org/10.1162/tacl_a_00041

4. Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5248-5264). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

5. Lalor, J., Yang, Y., Smith, K., Forsgren, N. et Abbasi, A. (2022). Benchmarking Intersectional Biases in NLP. Dans *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 3598-3609). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.263

6. Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5248-5264). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

7. Ruder, S., Vulić, I. et Søgaard, A. (2022). Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. Dans *Findings of the Association for Computational Linguistics: ACL 2022* (p. 2340-2354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.184

8. Costa-jussà, M. R., Hardmeier, C., Radford, W. et Webster, K. (dir.). (2020). *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.0

9. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y. et Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5491-5501). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.487

10. Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G. et Gonen, H. (dir.). (2022). *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics. https://aclanthology.org/2022.gebnlp-1.0

11. Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N. et Chang, K.-W. (2022, 13 octobre). *On Measures of Biases and Harms in NLP*. arXiv. Récupéré le 15 août 2023 de http://arxiv.org/abs/2108.03362

12. Rudinger, R., May, C. et Van Durme, B. (2017). Social Bias in Elicited Natural Language Inferences. Dans *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (p. 74-79). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1609

13. Hassan, S., Huenerfauth, M. et Alm, C. O. (2021). Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. Dans *Findings of the Association for Computational Linguistics: EMNLP 2021* (p. 3116-3123). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.267

14. Webster, K., Costa-jussà, M. R., Hardmeier, C. et Radford, W. (2019). Gendered Ambiguous Pronoun (GAP) Shared Task at the Gender Bias in NLP Workshop 2019. Dans *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (p. 1-7). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3801

15. Schick, T., Udupa, S. et Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408-1424. https://doi.org/10.1162/tacl_a_00434

16. Costa-jussa, M., Gonen, H., Hardmeier, C. et Webster, K. (dir.). (2021). *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2021.gebnlp-1.0

17. Li, Y., Zhang, G., Yang, B., Lin, C., Ragni, A., Wang, S. et Fu, J. (2022). HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models. Dans *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (p. 334-346). Association for Computational Linguistics. https://aclanthology.org/2022.findings-aacl.32

18. Ahn, J. et Oh, A. (2021). Mitigating Language-Dependent Ethnic Bias in BERT. Dans *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (p. 533-549). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.42

19. Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T., Njie, M. et Matthews, J. (2021). Gender Bias in Natural Language Processing Across Human Languages. Dans *Proceedings of the First Workshop on Trustworthy Natural Language Processing* (p. 45-54). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.trustnlp-1.6

20. Costa-jussà, M. R., Hardmeier, C., Radford, W. et Webster, K. (dir.). (2019). *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/W19-3800

ANNEXE B

DÉTAIL DU CORPUS OFFICIEL SÉLECTIONNÉ

1. Blodgett, S. L., Barocas, S., Daumé Iii, H. et Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5454-5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

2. Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5248-5264). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

3. Lalor, J., Yang, Y., Smith, K., Forsgren, N. et Abbasi, A. (2022). Benchmarking Intersectional Biases in NLP. Dans *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 3598-3609). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.263

4. Ruder, S., Vulić, I. et Søgaard, A. (2022). Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. Dans *Findings of the Association for Computational Linguistics: ACL 2022* (p. 2340-2354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.184

5. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y. et Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5491-5501). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.487

6. Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N. et Chang, K.-W. (2022, 13 octobre). *On Measures of Biases and Harms in NLP*. arXiv. Récupéré le 15 août 2023 de http://arxiv.org/abs/2108.03362

7. Schick, T., Udupa, S. et Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408-1424. https://doi.org/10.1162/tacl_a_00434

8. Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T., Njie, M. et Matthews, J. (2021). Gender Bias in Natural Language Processing Across Human Languages. Dans *Proceedings of the First Workshop on Trustworthy Natural Language Processing* (p. 45-54). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.trustnlp-1.6

ANNEXE C

GRILLE D'ANALYSE DU CORPUS

Grille d'analyse du corpus							
Titre : Language (Technology) is Power: A Critical Survey of "Bias" in NLP, 2020							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptualisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative	Compréhension des biais, performance du système	Explicite, non spécifiée
Attributions	Conceptualisé	Explicite	Représentation et allocation	Explicite	Qualitative	Compréhension des biais	Non spécifiée
Citations des sources originales	Indeed, the term "bias" (or "gender bias" or "racial bias") is used to describe a wide range of system behaviors, even though they may be harmful in different ways, to different groups, or for different reasons.	Turning first to (R1), we argue that work analyzing "bias" in NLP systems will paint a much fuller picture if it engages with the relevant literature outside of NLP that explores the relationships between language and social hierarchies. Many disciplines, including sociolinguistics, linguistic anthropology, sociology, and social psychology, study how language takes on social meaning and the role that language plays in maintaining social hierarchies. [...] As a result, many groups have sought to bring about social changes through changes in language, disrupting patterns of oppression and marginalization via so-called "gender-fair" language (Sczesny et al., 2016; Menegatti and Rubini, 2017), language that is more inclusive to people with disabilities (ADA, 2018), and language that is less dehumanizing (e.g., abandoning the use of the term "illegal" in everyday discourse on immigration in the U.S. (Rosa, 2019)).	We used a previously developed taxonomy of harms for this categorization, which differentiates between so-called allocational and representational harms (Barocas et al., 2017; Crawford, 2017). Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to dif- ferent social groups; representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether. [...] Adapting and extending this taxonomy, we categorized the 146 papers' motivations and techniques into the following categories: Allocational harms. Representational harms: Stereotyping that propagates negative generalizations about particular social groups. Differences in system performance for different social groups, language that misrepresents the distribution of different social groups in the population, or language that is denigrating to particular social groups.	We provide a list of all 146 papers in the appendix. [...] Once identified, we then read each of the 146 papers with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating "bias." [...] We then describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. We argue that such work should examine the relationships between language and social hi- erarchies; we call on researchers and practitioners conducting such work to articulate their conceptualizations of "bias" in order to enable conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why; and we recommend deeper engagements between technologists and communities affected by NLP systems. We also provide several concrete research questions that are implied by each of our recommendations.	Our survey includes all papers known to us analyzing "bias" in NLP systems—146 papers in total. We omitted papers about speech, restricting our survey to papers about written text only. To identify the 146 papers, we first searched the ACL Anthology1 for all papers with the keywords "bias" or "fairness" that were made available prior to May 2020. [...] Once identified, we then read each of the 146 papers with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating "bias." We used a previously developed taxonomy of harms for this categorization, which differentiates between so-called allocational and representational harms (Barocas et al., 2017; Crawford, 2017). [...] We then describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems.	By surveying 146 papers analyzing "bias" in NLP systems, we found that (a) their motivations are often vague, inconsistent, and lacking in norma- tive reasoning; and (b) their proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not en- gage with the relevant literature outside of NLP. To help researchers and practitioners avoid these pitfalls, we proposed three recommendations that should guide work analyzing "bias" in NLP sys- tems, and, for each, provided several concrete re- search questions. These recommendations rest on a greater recognition of the relationships between language and social hierarchies—a step that we see as paramount to establishing a path forward.	-

Grille d'analyse du corpus							
Titre : Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview, 2020							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative	Compréhension des biais, performance du système	Explicite, non spécifiée
Attributions	Conceptalisé	Vague	Représentation et allocation	Explicite	Mixte	Compréhension des biais	Non spécifiée
Citations des sources originales	In essence, biases are priors that inform our decisions (a dialogue system designed for elders might work differently than one for teenagers). Still, undetected and unaddressed, biases can lead to negative consequences: There are aggregate effects for demographic groups, which combine to produce predictive bias. [...] We identify four points within the standard supervised NLP pipeline where bias may originate: (1) the training labels (label bias), (2) the samples used as observations — for training or testing (selection bias), (3) the representation of data (semantic bias), or (4) due to the fit method itself (overamplification). Our definition of predictive bias in NLP builds on its definition within the literature on standardized testing (i.e., SAT, GRE, etc.) Specifically, Swinton (1981) states: By "predictive bias," we refer to a situation in which a predictive model is used to predict a specific criterion for a particular population, and is found to give systematically different predictions for subgroups of this population who are in fact identical on that specific criterion. [...] Our framework focuses on NLP, but it follows Glymour and Herington (2019) in providing probabilistic based definitions of bias.	The general phenomenon of biased predictive models in NLP is not recent. The community has long worked on the domain adaptation problem (Jiang and Zhai, 2007; Daume III, 2007): models fit on newswire data do not perform well on social media and other text types. This problem arises from the tendency of statistical models to pick up on non-generalizable signals during the training process. In the case of domains, these non-generalizations are words, phrases, or senses that occur in one text type, but not another. However, this kind of variation is not just restricted to text domains: it is a fundamental property of human-generated language: we talk differently than our parents or people from a different part of our country, etc. (Pennebaker and Stone, 2003; Eisenstein et al., 2010; Kern et al., 2016). In other words, language reflects the diverse demographics, backgrounds, and personalities of the people who use it.	Predictive models in NLP are sensitive to a variety of (often unintended) biases throughout the development process. As a result, fitted models do not generalize well, incurring performance and reliability losses on unseen data. They also have socially undesirable effects by systematically underserving or mispredicting certain user groups. [...] Still, undetected and unaddressed, biases can lead to negative consequences: There are aggregate effects for demographic groups, which combine to produce predictive bias. While it is essential to address the effects of bias, it can leave the fundamental origin unchanged (Gonen and Goldberg, 2019), requiring researchers to rediscover the issue over and over.	[...] much work has focused on bias effects and symptoms rather than their origins. While it is essential to address the effects of bias, it can leave the fundamental origin unchanged (Go- nen and Goldberg, 2019), requiring researchers to rediscover the issue over and over. The "bias" discussed in one paper may, therefore, be quite different than that in another. A shared definition and framework of predictive bias can unify these efforts, provide a common terminology, help to identify underlying causes, and allow coordination of countermeasures (Sun et al., 2019). However, such a general framework had yet to be proposed within the NLP community. To address these problems, we suggest a joint conceptual framework, depicted in Figure 1, outlining and relating the different origins of bias. [...] We hope this paper will help researchers spot, com- pare, and address bias in all its various forms.	We base our framework on an extensive survey of the relevant NLP literature, informed by selected works in social science and adjacent fields. [...] Our primary contributions include: (1) a conceptual framework for identifying and quantifying predictive bias and its origins within a standard NLP pipeline, (2) a survey of biases identified in NLP models, and (3) a survey of methods for countering bias in NLP organized within our conceptual framework.	We present a comprehensive overview of the recent literature on predictive bias in NLP. Based on this survey, we develop a unifying conceptual framework to describe bias sources and their ef- fects (rather than just their effects). This framework allows us to group and compare works on countermeasures. Rather than giving the impression that bias is a growing problem, we would like to point out that bias is not necessarily something gone awry, but rather something nearly inevitable in statistical models.	-

Grille d'analyse du corpus							
Titre : Social Biases in NLP Models as Barriers for Persons with Disabilities, 2020							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif/motivation**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, non spécifiée
Attributions	Non conceptualisé	Explicite	Représentation et allocation	Vague	Quantitative	Performance du système	Explicite
Citations des sources originales	This is important because NLP models are increasingly being used for tasks such as fighting online abuse (Jigsaw, 2017), measuring brand sentiment (Mostafa, 2013), and matching job applicants to job opportunities (De-Arteaga et al., 2019). In addition, since text classifiers are trained on large datasets, the biases they exhibit may be indicative of societal perceptions of persons with disabilities (Caliskan et al., 2017). If models inappropriately condition on mentions of disability, this could impact people writing, reading, or seekiing information about a disability. Depending on how such models are deployed, this could potentially result in reduced autonomy, reduced freedom [...] bias with respect to different disability groups has been relatively under-explored. However, over one billion individuals (about 15% of the world's population) are persons with disabilities, and disability is sometimes the subject of strong negative social biases. [...] We have so far worked in a purely technical framing of model biases—i.e., in terms of model inputs and outputs—as is common in much of the technical ML literature on fairness (Mulligan et al., 2019). However, normative and social justifications should be considered when applying a statistical definition of fairness.	Neural text embedding models (Mikolov et al., 2013) are critical first steps in today's NLP pipelines. These models learn vector representations of words, phrases, or sentences, such that semantic relationships between words are encoded in the geometric relationship between vectors. Text embedding models capture some of the complexities and nuances of human language. However, these models may also encode undesirable correlations in the data that reflect harmful social biases. [...] It is important to recognize that social norms around language are contextual and differ across groups. [...] models for detecting abuse can be used to nudge writers to rethink comments which might be interpreted as toxic (Jurgens et al., 2019). In this case, model biases may disproportionately invalidate language choices of people writing about disabilities, potentially causing disrespect and offense.	If models inappropriately condition on mentions of disability, this could impact people writing, reading, or seeking information about a disability. Depending on how such models are deployed, this could potentially result in reduced autonomy, reduced freedom of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals.[...] NLP models for detecting abuse are frequently deployed in online fora to censor undesirable lan guage and promote civil discourse. Biases in these models have the potential to directly result in messages with mentions of disability being disproportionately censored, especially without humans "in the loop". Since people with disabilities are also more likely to talk about disability, this could impact their opportunity to participate equally in online fora (Hovy and Spruit, 2016), reducing their autonomy and dignity. Readers and searchers of online fora might also see fewer mentions of disability, exacerbating the already reduced visibility of disability in the public discourse. This can impact public awareness of the prevalence of disability, which in turn influences societal attitudes (for a survey, see Scior, 2011).	This paper focuses on the representation of persons with disabilities through the lens of technology. Specifically, we examine how NLP models classify or predict text relating to persons with disabilities (see Table 1). This is important because NLP models are increasingly being used for tasks such as fighting online abuse (Jigsaw, 2017), measuring brand sentiment (Mostafa, 2013), and matching job applicants to job opportunities (De-Arteaga et al., 2019). In addition, since text classifiers are trained on large datasets, the biases they exhibit may be indicative of societal perceptions of persons with disabilities. [...] Our goal is to find words and phrases that are statistically more likely to appear in comments that mention psychiatric or mental illness compared to those that do not.	Our analyses in this paper use a set of 56 linguistic expressions (in English) for referring to people with various types of disabilities, e.g. a deaf person. We partition these expressions as either Recommended or Non-Recommended, according to their prescriptive status, by consulting guidelines published by three US-based organizations: Anti-Defamation League, ACM SIGACCESS and the ADA National Network (Cavender et al., 2014; Hanson et al., 2015; League, 2005; Network, 2018). Following (Garg et al., 2019; Prabhakaran et al., 2019), we use the notion of perturbation, whereby the phrases for referring to people with disabilities, described above, are all inserted into the same slots in sentence templates. We start by first retrieving a set of naturally-occurring sentences that contain the pronouns he or she. We then select a pronoun in each sentence, and "perturb" the sentence by replacing this pronoun with the phrases described above. Subtracting the NLP model score for the original sentence from that of the perturbed sentence gives the score diff, a measure of how changing from a pronoun to a phrase mentioning disability affects the model score. We perform this method on a set of 1000 sentences extracted at random from the Reddit subcorpus of (Voigt et al., 2018). Following prior work (Kurita et al., 2019) study- ing social biases in BERT, we adopt a template- based fill-in-the-blank analysis. Given a query sentence with a missing word, BERT predicts a ranked list of words to fill in the blank. [...] NLP models such as the ones discussed above are trained on large textual corpora, which are analyzed to build "meaning" representations for words based on word co-occurrence metrics, drawing on the idea that "you shall know a word by the com- pany it keeps" (Firth, 1957). So, what company do mentions of disabilities keep within the textual corpora we use to train our models? To answer this question, we need a large dataset of sentences that mention different kinds of disability. We use the dataset of online comments released as part of the Jigsaw Unintended Bias in Toxicity Classification challenge (Borkan et al., 2019; Jig- saw, 2019), where a subset of 405K comments are labelled for mentions of disabilities, grouped into four types: physical disability, intellectual or learn- ing disability, psychiatric or mental illness, and other disability. We focus here only on psychiatric or mental illness, since others have fewer than 100 instances in the dataset.	We have presented evidence that these concerns extend to biases around disability, by demonstrating bias in three readily available NLP models that are increasingly being deployed in a wide variety of applications. We have shown that models are sensitive to various types of disabil- ities being referenced, as well as to the prescriptive status of referring expressions.	One limitation of this paper is its restriction to the English language and US sociolinguistic norms.

Grille d'analyse du corpus							
Titre : Gender Bias in Natural Language Processing Across Human Languages, 2021							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, vague, non spécifiée
Attributions	Non conceptualisé	Vague	Allocation	Vague	Quantitative	Performance du système	Explicite
Citations des sources originales	Bolukbasi et al. developed a method for measuring gender bias using word embedding systems like Word2vec. Specifically, they defined a set of highly gendered word pairs such as ("he", "she") and used the difference between these word pairs to define a gendered vector space. They then evaluated the relationship of profession words like doctor, nurse or teacher relative to this gendered vector space. Ideally, profession words would not reflect a strong gender bias. However, in practice, they often do. According to such a metric, doctor might be male biased or nurse female biased based on how these words are used in the corpora from which the word embedding model was produced. Thus, this gender bias metric of profession words as calculated from the Word2Vec model can be used as a measure of the gender bias learned from corpora of natural language.	Gender bias in NLP has been well studied in English, but has been less studied in other languages. In this paper, a team including speakers of 9 languages - Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof - reports and analyzes measurements of gender bias in the Wikipedia corpora for these 9 languages. We develop extensions to profession- level and corpus-level gender bias metric calculations originally designed for English and apply them to 8 other languages, including languages that have grammatically gendered nouns including different feminine, masculine, and neuter profession words.	[...] relationship of profession words like doctor, nurse, or teacher relative to this gendered vector space. They demonstrated that word embedding software trained on a corpus of Google news could associate men with the profession computer programmer and women with the profession homemaker. Systems based on such models, trained even with "representative text" like Google news, could lead to biased hiring practices if used to, for example, parse resumes and suggest matches for a computer programming job.	While our goal in this study was to identify a defining set and profession set that could more easily be used across many languages and for which the T-test results indicated no statistically significant difference in results over the English Wikipedia corpus, it would be interesting to repeat this analysis with additional variations in the defining set and profession set.	[...] we applied Bolukbasi et al.'s methodology to computing and comparing corpus-level gender bias metrics across different corpora of the English text (Babaeianjelodar 2020). [...] Here we build on the work of Bolukbasi et al. and our own earlier work to extend these important techniques in gender bias measurement and analysis beyond English. [...] We translate and modify Bolukbasi et al.'s defining sets and profession sets in English for 8 additional languages and develop extensions to the profession-level and corpus-level gender bias metric calculations for languages with grammatically gendered nouns. We use this methodology to analyze the gender bias in Wikipedia corpora for Chinese (Mandarin Chinese), Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof.	We have extended an influential method for comput- ing gender bias from Bolukbasi et al., a technique that had only been applied in English. We made key modi- fications that allowed us to extend the methodology to 8 additional languages, including languages with grammatically gendered nouns. With this, we quantified how gender bias varies across the Wikipedia corpora of 9 languages [...]	One important limitation to note is that for many laguages, if a word is expressed with a multi-word phrase (e.g. astronomer in Arabic), the word count reported by Word2Vec for this phrase will be zero. For each language, there is a tokenizer that identifies the words or phrases to be tracked. In many cases, the tokenizer identifies words as being separated by a space. The Chinese tokenizer however attempts to recognize when multiple characters that are separated with spaces should be tracked as a multi-character word or concept. This involves looking up a string of characters in a dictionary.

Grille d'analyse du corpus							
Titre : Benchmarking Intersectional Biases in NLP, 2022							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisés, non conceptualisés	Explicite, vague, non spécifiée	Représentation, allocation, non spécifiés	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, non spécifiée
Attributions	Non conceptualisé	Vague	Allocation	Vague	Quantitative	Performance du système	Explicite
Citations des sources originales	Accordingly, in this study we perform a broad benchmark analysis of intersectional bias encompassing the following key characteristics: [...] Inclusion of five demographic dimensions: gender, race, age, education, and income. Having three or more dimensions on many of the tasks affords opportunities to examine bias for various demographic intersection subgroups in a more in-depth manner. [...] Intersectional biases arising as a result of interacting demographics have been studied in the broader machine learning literature, either from a theoretical perspective (Kearns et al., 2018; Yang et al., 2020), or in the context of facial recognition (Buolamwini and Gebru, 2018). [...] We build on the emergent literature on intersectional biases by assessing datasets encompassing up to five demographic dimensions, in conjunction with state-of-the-art word embeddings and debiasing methods, on downstream tasks where biased predictions can lead to allocational harm (§3.1).	Other recent studies have empirically shown that the biases inherent in language models for gender and race intersections might exceed those observed for gender and race alone (Tan and Celis, 2019), and that only debiasing along a single dimension can be problematic (Subramanian et al., 2021).	[...] allocational harms - these "arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups." Allocational harm is often aligned with downstream tasks/interventions guided by the NLP model. [...] Biases in predictions for healthcare-related variables (Psychometrics), or personality type variables (MBTI, FIPI) can affect an individual's health care plan, personalized interventions, job prospects, etc. Biased predictions for drug rating sentiment can affect which drugs a future user chooses to take. [...] Collectively, the results underscore the allocational harm implications of NLP models on several downstream tasks - ones that even well-designed and well-intentioned debiasing strategies cannot overcome. This can be problematic in the era of personalized marketing and precision health, with NLP-based persona-generation playing a bigger role. For tasks like numeracy and literacy, this can affect how a patient is treated by a medical staff during a hospital visit (i.e., a false positive high literacy prediction for a person who has trouble understanding his or her medical record). For the personality indicators, inconsistent predictions may lead to biased decisions in the workplace (e.g., a manager looking to form a team of extroverts). [...] We build on the emergent literature on intersectional biases by assessing datasets encompassing up to five demographic dimensions, in conjunction with state-of- the-art word embeddings and debiasing methods, on downstream tasks where biased predictions can lead to allocational harm (§3.1).	There is a need for a more systematic analysis of how current state-of- the-art language models and mitigation strategies perform with regards to intersectional bias in down-stream tasks.	[...] we benchmark multiple NLP models with regards to their fairness and predictive performance across a variety of NLP tasks. In particular, we assess intersectional bias - fairness across multiple demographic dimensions. In this study we perform a broad benchmark analysis of intersectional bias (Figure 1) encompassing the following key characteristics: • Benchmark analysis on ten downstream sequence classification tasks related to five datasets that span common modes of user- generated content: Twitter, forums, Reddit, and survey responses. For these tasks, we also note the allocational harm implications of disparate impact, namely the harm associated with biased NLP-guided interventions. • Inclusion of five demographic dimensions: gender, race, age, education, and income. Having three or more dimensions on many of the tasks affords opportunities to examine bias for various demographic intersection subgroups in a more in-depth manner. On four of the datasets, these demographics are selfreported as opposed to being algorithmically or heuristically inferred - an important consideration for debiasing research. • Evaluation of three prominent word embeddings, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GloVe (Pennington et al., 2014), and four state-of-the-art model debiasing methods (Ravfogel et al., 2020; Kaneko and Bollegala, 2021; Zmigrod et al., 2019; Webster et al., 2020). This allows us to draw empirical insights regarding the ef- fectiveness of mitigation strategies for down- stream tasks.	Our benchmark evaluation offers empirical evidence that the concerns voiced in recent critical surveys about too much emphasis on representational debi- asing devoid of explicit normative goals (Blodgett et al., 2020), relative to mitigation of downstream allocational harm, are well-founded. [...] We also look at known debiasing methods for these models and show that while the debiased versions maintain predictive performance (as expected), they do not help with mitigating bi- ases. While most models are relatively fair when looking at a single demographic characteristic, accounting for intersectional groups leads to less fair models and wider ranges of bias because of the combinatorial considerations of the intersectional groups.	In this work, our scope is debiasing embeddings, not debiasing classifiers. While there is much work in the area of debiasing classifiers, here we restrict our focus to the debiasing of embeddings.

Grille d'analyse du corpus							
Titre : Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP, 2021							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifié	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, non spécifiée
Attributions	Non conceptualisé	Vague	Non spécifié	Vague	Mixte	Performance du système	Explicite
Citations des sources originales	With model sizes continually increasing (Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020; Fedus et al., 2021), ever-larger pretraining datasets are necessary both to prevent overfitting and to provide access to as much world knowledge as possible. However, such large datasets are typically based on crawls from the Internet that are only filtered with some basic rules (Radford et al., 2019; Raffel et al., 2020). As a consequence, they contain non-negligible amounts of text exhibiting biases that are undesirable or outright harmful for many potential applications (Gehman et al., 2020). Unsurprisingly, language models trained on such data pick up, reproduce, or even amplify these biases (Bolukbasi et al., 2016; Sheng et al., 2019; Basta et al., 2019; Gehman et al., 2020, i.a.). [...] If the model is told which biases are unde- sired—and it is able to discern their presence—it should be able to avoid them even if they are present in some of the texts it has been trained on. [...] Unfortunately, Perspective API covers only a limited set of emotional concepts and does not explicitly measure many relevant biases known to be found in pretrained language models [...].	When trained on large, unfiltered crawls from the Internet, language models pick up and reproduce all kinds of undesirable biases that can be found in the data: They often generate racist, sexist, violent, or otherwise toxic language.	As a consequence, they contain non-negligible amounts of text exhibiting biases that are undesirable or outright harmful for many potential applications.	Building training datasets with more care and deliberation, an alternative solution discussed by Bender et al. (2021), is important, especially for improving linguistic and cultural diversity in online and other forms of communication. However, for large language models that are available for common global languages, it is desirable to also have other mechanisms to address bias because dataset curation and documentation is extremely resource intensive, given the amount of data required. It can also necessitate building different training sets and, accordingly, training different models for each desired behavior, which can result in high environmental impact (Strubell et al., 2019). In this paper, we therefore propose an approach that, instead of trusting that a model will implicitly learn desired behaviors from the training data, makes explicit how we expect it to behave at test time: If the model is told which biases are undesired—and it is able to discern their presence—it should be able to avoid them even if they are present in some of the texts it has been trained on.	[...] we first explore whether language models are able to detect when their own outputs exhibit undesirable attributes, based only on their internal knowledge—a process to which we refer as self-diagnosis. We then investigate whether this ability can be used to perform self-debiasing, that is, whether language models can use this knowl- edge to discard undesired behaviors in a fully unsupervised fashion. To this end, we propose a decoding algorithm that reduces the probability of a model producing biased text, requiring nothing more than a textual description of the undesired behavior, which can be as simple as a single keyword (e.g., ''sexist'', ''racist'', ''homophobic'', or ''violent'' in Figure 1; see §4 for details). [...] To evaluate the self-diagnosis capabilities of current language models, we follow Gehman et al. (2020) and consider all emotional concepts covered by Perspective API4 as attributes (toxicity, severe toxicity, sexually explicit, threat, profanity, identity attack); we investigate whether language models are able to detect when their outputs exhibit one of these attributes. [...] As Perspective API only covers a limited set of attributes, we are unable to test the effectiveness of our method for many relevant biases (e.g., gender bias) using only RealToxicityPrompts. Therefore, we additionally evaluate self-debiasing on CrowS-Pairs (Nangia et al., 2020), a dataset that measures the degree to which nine different types of social bias are present in MLMs (race/color, gender, occupation, nationality, religion, age, sexual orientation, phyical appearance and disability.) [...] For complementing our automatic evaluation with human judgments, we randomly select 100 prompts from the challenging subset of RealToxicityPrompts. For these prompts, we use Amazon Mechanical Turk to collect human annotations for continuations generated with both regular GPT2-XL and GPT2-XL with self-debiasing ($\lambda = 100$). Annotators are instructed to assess whether the generated continuations exhibit any of the six attributes considered, using the exact same question and attribute descriptions as for self-diagnosis.	[...] we first demonstrate a surprising finding: Pretrained language models recognize, to a considerable degree, their undesirable biases and the toxi- city of the content they produce. We refer to this capability as self-diagnosis. Based on this finding, we then propose a decoding algorithm that, given only a textual description of the undesired behavior, reduces the probability of a language model producing problematic text. We refer to this approach as self-debiasing. Self-debiasing does not rely on manually cu- rated word lists, nor does it require any training data or changes to the model's parameters. While we by no means eliminate the issue of language models generating biased text, we believe our approach to be an important step in this direction.	One major limitation of our evaluation is that it relies to a large extent on attribute scores assigned by Perspective API; this means not only that we cannot thoroughly test the effectiveness of our method for many relevant biases that are not measured by the API, but also that our labels are error-prone. For example, Perspective API may fail to detect more subtle forms of bias and be overreliant on lexical cues (Gehman et al., 2020). While our complementary human evaluation mitigates this issue to some extent, crowdsourcing comes with its own downsides. In particular, untrained crowdworkers classify examples based on their own biases and personal perceptions; our setup does not involve critical communities who have contextual knowledge, represent social justice agendas and have reasonable credibility in establishing the presence or absence of undesired attributes. CrowS-Pairs covers a larger set of social biases and is based on human-labeled data, but it is a comparatively small dataset that, for some bias categories, contains only a few dozen examples. [...] As for the limitations of self-diagnosis and self-debiasing, both algorithms rely on simple templates and attribute descriptions; as our ex- periments in §3.3 show, modifying templates and descriptions can—in some cases—result in quite different self-diagnosis performance. In addition, finding descriptions that are well understood by current generations of language models may be inherently difficult for some forms of bias. We also find that the proposed self-debiasing algo- rithm is often overly aggressive in filtering out harmless words that do not really contribute to undesired bias in the generated sentence.

Grille d'analyse du corpus							
Titre : On Measures of Biases and Harms in NLP, 2022							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisé, non conceptualisé	Explicite, vague, non spécifiée	Représentation, allocation, non spécifiés	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, vague, non spécifié
Attributions	Conceptualisé	Explicite	Représentation et allocation	Explicite	Qualitative	Compréhension des biais	Explicite
Citations des sources originales	Bias in language models is commonly defined as "skew that produces a type of harm" (Crawford, 2017) towards different social groups, though it is a complex notion that is often not well-defined in existing literature (Blodgett et al., 2020; Delobelle et al., 2022; Talat et al., 2022).	Dehumanizing language uses techniques such as moral disgust, denial of agency, or likening members of a target group to non-human entities (Markowitz and Slovic, 2020) to reinforce normative identities—often as indication of a biological hierarchy of 'species' within humankind. [...] A single instance of language may represent/cause multiple forms of harm (e.g., some Stereotyping harms may also be Dehumanization harms). Does the measure provide a method for measuring multiple harms separately as well as in aggregate (e.g., are subsets of the underlying data tagged along multiple axes)? What language and culture is the bias and measure most relevant to? [...] Language data for bias measures is sourced primarily in two ways: by extracting from existing textual data or by generating from specific templates. While the first has the advantage of being more similar to "real samples" that models see, the latter has the advantage of testing for specific artifacts by construct.	The relevant harms can be subdivided into representational or allocational harms, depending on whether there is a generalization of harmful representations of groups or if there is a tangible, disparate distribution of resources between groups, respectively (Crawford, 2017).	This paper is motivated by two main goals. The first goal is to define a practical framework for harms that is both theoretically-motivated and empirically useful for describing bias measures. [...] The second goal is to define a collection of documentation questions around bias measures that helps others capture measure limitations and align operationalizations of "biases" to harms.	To achieve these goals, we organize a practical framework of harms, a tagged collection of 43 existing bias measures and the associated harms, a set of documentation questions, and a collection of case studies.	we organize a framework to define and distinguish between different types of harms—presented through heuristics and documentation questions—to guide more inten- tional development of bias measures. Our proposed documentation template also facilitates combining, comparing, and utilizing different bias measures, and continuously re-visiting them to update limi- tations and comparative understanding with other measures.	We acknowledge that our framework of harms has been created from a US-centric perspective and has been influenced by the Social Dominance Theory (Sidanius and Pratto, 2001), which can be limiting from a global perspective and does not include cultural harms. While some definitions and operationalizations of harms in our framework (e.g., Stereotyping, Disparagement) may be applicable to other cultural perspectives, we note that there may be some that require cultural context-specific updates and also that there are other harms that we did not cover. There are also other bias measures in this rapidly growing space that we may not have covered and tagged with harms measured. Additionally, we do not focus on specific downstream applications where each measure might be used and encourage further analysis on these applications.

Grille d'analyse du corpus							
Titre : Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold, 2022							
Catégories à l'étude	**Biais**	**Biais et langage**	**Type de préjudices**	**Objectif**	**Méthodologie**	**Résultats**	**Limitation**
Classifications possibles	Conceptalisés, non conceptualisés	Explicite, vague, non spécifiée	Représentation, allocation, non spécifiés	Explicite, vague, non spécifié	Qualitative, quantitative, mixte	Compréhension des biais, performance du système	Explicite, vague, non spécifiée
Attributions	Non conceptualisé	Vague	Non spécifiés	Vague	Qualitative	Compréhension des biais	Non spécifiée
Citations des sources originales	We observe that NLP research often goes beyond the square one setup, e.g, focusing not only on accuracy, but also on fairness or interpretability, but typically only along a single dimension. Most work targeting multilinguality, for example, considers only accuracy; most work on fairness or interpretability considers only English; and so on. Such one-dimensionality of most research means we are only exploring a fraction of the NLP research search space. We provide historical and recent examples of how the square one bias has led researchers to draw false conclusions or make unwise choices, point to promising yet unexplored directions on the research manifold, and make practical recommendations to enable more multi-dimensional research. [...] We will refer to this prototype as NLP's SQUARE ONE—and to the bias that follows from it, as the SQUARE ONE BIAS. We argue this bias manifests in a particular way: Since research is a creative endeavor, and researchers aim to push the research horizon, most research papers in NLP go beyond this prototype, but only along a single dimension at a time. Such dimensions might include multilinguality, efficiency, fairness, and interpretability, among others. The effect of the SQUARE ONE BIAS is to baseline novel research contributions, rewarding work that differs from the prototype in a concise, one-dimensional way.	Overall, almost 70% of papers evaluate only on English, clearly highlighting a lack of language diversity in NLP (Bender, 2011; Joshi et al., 2020). [...] While studies have demonstrated the ability of word embeddings to capture linguistic information in English, it remains unclear whether they capture the information needed for processing morphologically rich languages (Tsarfaty et al., 2020). [...] Another bias in our models relates to word order. In order for n-gram models to capture inter- word dependencies, words need to appear in the n-gram window. This will occur more frequently in languages with relatively fixed word order compared to languages with relatively free word order (Bender, 2011).	We argue that the SQUARE ONE BIAS has several negative effects, most of which amount to the study of one of the above dimensions being biased by ignoring the others. Specifically, by focusing only on exploring the edges of the manifold, we are not able to identify the non-linear interactions between different research dimensions. [...] We discuss the impact of this prototype on our research community, and the bias it introduces. We then discuss the negative effects of this bias.	We provide historical and recent examples of how the square one bias has led researchers to draw false conclusions or make unwise choices, point to promising yet unexplored directions on the research manifold, and make practical recommendations to enable more multi-dimensional research.	We annotate the 461 papers that were presented orally at ACL 2021, a representative cross-section of the 779 papers accepted to the main conference. The general statistics from our classification of ACL 2021 papers are presented in Table 1. In addition, we highlight the statistics for the conference areas (tracks) corresponding to 3 of the 4 dimensions4, as well as for the top 5 areas with the most papers. We show statistics for the remaining areas in Appendix A.2. We additionally visualize their distribution in Figure 1. Overall, almost 70% of papers evaluate only on English, clearly highlighting a lack of language diversity in NLP (Bender, 2011; Joshi et al., 2020). [...] We discuss the impact of this prototype on our research community, and the bias it introduces. We then discuss the negative effects of this bias. We also list work that has taken steps to overcome the bias.	We highlighted the associated SQUARE ONE BIAS, which encourages research to go beyond the proto- type in a single dimension. We discussed the problems resulting from this bias, by studying the area statistics of a recent NLP conference as well as by discussing historic and recent examples. We finally pointed to under-explored research directions and made practical recommendations to inspire more multi-dimensional research in NLP.	-

BIBLIOGRAPHIE

ACL Anthology. (2021). *Requesting Corrections - ACL Anthology*. https://aclanthology.org/info/corrections/

ACL Anthology. (2023). *What is the ACL and what is Computational Linguistics? | ACL Member Portal*. https://www.aclweb.org/portal/what-is-cl

Ahn, J. et Oh, A. (2021). Mitigating Language-Dependent Ethnic Bias in BERT. Dans *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (p. 533-549). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.42

Andrejevic, M. (2020). *Automated media* (Taylor&Francis Group). Routledge.

Angwin, J. et Paglen, T. (2019, 23 avril). Interviewé par R. Harmanci. Forget about "privacy": Julia Angwin and Trevor Paglen on our data crisis. https://www.fastcompany.com/90337954/who-cares-about-liberty-julia-angwin-and-trevor-paglen-on-privacy-surveillance-and-the-mess-were-in

Antoniak, M. et Mimno, D. (2021). Bad Seeds: Evaluating Lexical Methods for Bias Measurement. Dans *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (p. 1889-1904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.148

Arendt, H. (1998). *The Human Condition* (The University of Chicago Press).

Aznar, J., Bender, E. M. et Lascarides, A. (2020). Linguistic Fundamentals for Natural Language Processing II : 100 Essentials from Semantics and Pragmatics. *Morgan & Claypool publishers.*, *62*, 250 pages.

Bansal, R. (2022, 6 mars). *A Survey on Bias and Fairness in Natural Language Processing*. arXiv. Récupéré le 15 août 2023 de http://arxiv.org/abs/2204.09591

Beaucher, V. et Jutras, F. (2007). Étude comparative de la métasynthèseet de la méta-analyse qualitative. *Recherches qualitatives*, *27*(2), 58. https://doi.org/10.7202/1086786ar

Bender, E. M. (2022). *On NYT Magazine on AI: Resist the Urge to be Impressed*. Medium. https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd

Bender, E. M. (2023). Interviewé par D. Hirning. Emily M. Bender Q&A | UW ChatGPT, Ethics & Careers in Computational Linguistics. https://www.compling.uw.edu/academic-experience/faculty/emily-bender-q-a

Bender, E. M. et Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A. et Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (p. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Bender, E. M. et Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5185-5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M. et Mitchell, M. (2021). Interviewé par S. Charrington. Can Language Models Be Too Big? 🦜 with Emily Bender and Margaret Mitchell on Apple Podcasts [The TWIML AI Podcast]. https://podcasts.apple.com/ca/podcast/can-language-models-be-too-big-with-emily-bender/id1116303051?i=1000514394149

Bérubé, F. S. (2007). Le principe responsabilité de Hans Jonas et la responsabilité sociale. *Mémoire de maîtrise*. https://archipel.uqam.ca/3268/1/M9722.pdf

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns (New York, N.Y.)*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Blodgett, S. L., Barocas, S., Daumé Iii, H. et Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5454-5476). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.acl-main.485

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. et Kalai, A. (2016, 21 juillet). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. arXiv. https://doi.org/10.48550/arXiv.1607.06520

Bontems, V. (2017, 26 octobre). *L'éthique des techniques selon Simondon par Vincent Bontems*. https://www.youtube.com/watch?v=EGssBngeYBA

Bujokas, E. (2020). *Text Classification Using Word Embeddings and Deep Learning in Python — Classifying Tweets from Twitter*. Towards Data Science. https://towardsdatascience.com/text-classification-using-word-embeddings-and-deep-learning-in-python-classifying-tweets-from-6fe644fcfc81

Castello, M. et Lajeunesse, Y. (2019). *Traitement automatique du langage naturel, un domaine de l'IA payant pour les affaires*. Novipro. https://www.novipro.com/blogue/fr/webinaire-les-affaires-traitement-automatique-du-langage-naturel-un-domaine-de-lia-payant-pour-les-affaires

Cheng, L., Ge, S. et Liu, H. (2022). *Toward Understanding Bias Correlations for Mitigation in NLP*. https://doi.org/10.48550/ARXIV.2205.12391

Chowdhury, M. Z. I. et Turin, T. (2019). *Synthesizing Quantitative and Qualitative Studies in Systematic Reviews: The Basics of Meta-analysis and Meta-synthesis*. ResearchGate. https://www.researchgate.net/publication/337926243_Synthesizing_Quantitative_and_Qualitative_Studies_in_Systematic_Reviews_The_Basics_of_Meta-analysis_and_Meta-synthesis?enrichId=rgreq-bd6709efe2e11939749a200d1f593e80-XXX&enrichSource=Y292ZXJQYWdlOzMzNzkyNjI0MztBUzo4MzU3MzExMzYyMTI5OTNAMTU3NjI2NTIzODQyNw%3D%3D&el=1_x_3&_esc=publicationCoverPdf

Corbyn, Z. (2021, 6 juin). Microsoft's Kate Crawford: 'AI is neither artificial nor intelligent'. *The Observer*, Technology. https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent

Costa-jussa, M., Gonen, H., Hardmeier, C. et Webster, K. (dir.). (2021). *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2021.gebnlp-1.0

Costa-jussà, M. R., Hardmeier, C., Radford, W. et Webster, K. (dir.). (2019). *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/W19-3800

Costa-jussà, M. R., Hardmeier, C., Radford, W. et Webster, K. (dir.). (2020). *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.0

Crawford, K. (2009). Adult responsibility in insecure times. *Soundings*, *41*(41), 45-55. https://doi.org/10.3898/136266209787778939

Crawford, K. (2017, 10 décembre). *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017*. https://www.youtube.com/watch?v=fMym_BKWQzk

Dean, J. (2020). *About Google's approach to research publication - Google Docs*. https://docs.google.com/document/d/1f2kYWDXwhzYnq8ebVtuk9CqQqz7ScqxhSIxeYGrWjK0/preview?pru=AAABdlOOKBs*gTzLnuI53B2IS2BISVcgAQ#heading=h.aplcvu32myqt

Delouvée, S. (2018). 5. Perception sociale, stéréotypes et préjugés. Dans *Manuel visuel de psychologie sociale* (p. 87-110). Dunod. https://doi.org/10.3917/dunod.delou.2018.01.0087

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N. et Chang, K.-W. (2022, 13 octobre). *On Measures of Biases and Harms in NLP*. arXiv. Récupéré le 15 août 2023 de http://arxiv.org/abs/2108.03362

Ellul, J. (2012). *Le bluff technologique* (Éditions Fayard).

Encyclopædia Universalis. (2023). *INFORMATIQUE - Principes, Automates et langages formels*. https://www.universalis.fr/encyclopedie/informatique-principes/3-automates-et-langages-formels/

Estrach, M. (2022, 27 février). *Artificial Intelligence 101: Risks on AI Implementation (Part II)*. Arcadia. https://www.byarcadia.org/post/artificial-intelligence-101-risks-on-ai-implementation-part-ii

Feenberg, A. (2005). *Critical Theory of Technology: An Overview*.

Finfgeld, D. L. (2003). Metasynthesis: The State of the Art—So Far. *Qualitative Health Research*, *13*(7), 893-904. https://doi.org/10.1177/1049732303253462

Gebru, T., Margaret, M. et Bender, E. M. (2023). *Stochastic Parrots Day: A Retrospective With « Stochastic Parrots » Authors*. DAIR-Tube. https://peertube.dair-institute.org/w/vW2AoH552jgH7Swh8ePjQV

Golumbia, D. (2009). *The Cultural Logic of Computation*. Harvard University Press. https://doi.org/10.4159/9780674053885

Guo, Y., Yang, Y. et Abbasi, A. (2022). Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. Dans *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 1012-1023). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.72

Guy, G. R. et Labov, W. (dir.). (1997). *Social interaction and discourse structures*. J. Benjamins.

Halpern, O. (2014). Beautiful Data: A History of Vision and Reason since 1945. *Duke University Press*, *121*(2), 537-538. https://doi.org/10.1093/ahr/121.2.537

Hao, K. (2020, 4 décembre). *We read the paper that forced Timnit Gebru out of Google. Here's what it says.* MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Hao, K. (2023). *The Hidden Workforce That Helped Filter Violence and Abuse Out of ChatGPT - The Journal. - WSJ Podcasts*. Wall Street Journal. https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413

Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G. et Gonen, H. (dir.). (2022). *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics. https://aclanthology.org/2022.gebnlp-1.0

Hassan, S., Huenerfauth, M. et Alm, C. O. (2021). Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. Dans *Findings of the Association for Computational Linguistics: EMNLP 2021* (p. 3116-3123). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.267

Hayles, N. K. (1999). *How we became posthuman: virtual bodies in cybernetics, literature, and informatics* (Nachdr.). Univ. of Chicago Press.

Heidegger, M. (1998). Traditional Language and Technological Language: *Journal of Philosophical Research*, *23*, 129-145. https://doi.org/10.5840/jpr_1998_16

Hern, A. (2018, 12 janvier). Google's solution to accidental algorithmic racism: ban gorillas. *The Guardian*, Technology. https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people

Hern, A. et Milmo, D. (2023, 24 février). Everything you wanted to know about AI – but were afraid to ask. *The Guardian*, Technology. https://www.theguardian.com/technology/2023/feb/24/ai-artificial-intelligence-chatbots-to-deepfakes

Hessenthaler, M., Strubell, E., Hovy, D. et Lauscher, A. (2022). Bridging Fairness and Environmental Sustainability in Natural Language Processing. Dans *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (p. 7817-7836). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.533

Hughes, A. (2023). *ChatGPT: Everything you need to know about OpenAI's GPT-4 tool*. https://www.sciencefocus.com/future-technology/gpt-3

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y. et Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5491-5501). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.487

Jean, A. (2019). *De l'autre côté de la Machine* (Éditions de l'Observatoire).

Jonas, H. (1992). *Le principe responsabilité : une éthique pour la civilisation technologique* (CERF).

Kumaraswamy, A. (2017, 8 décembre). *20 lessons on bias in machine learning systems from NIPS 2017 Keynote*. Packt Hub. https://hub.packtpub.com/20-lessons-bias-machine-learning-systems-nips-2017/

Lalor, J., Yang, Y., Smith, K., Forsgren, N. et Abbasi, A. (2022). Benchmarking Intersectional Biases in

NLP. Dans *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 3598-3609). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.263

Laugée, F. (2015). Solutionnisme. https://la-rem.eu/2015/04/solutionnisme/

Li, Y., Zhang, G., Yang, B., Lin, C., Ragni, A., Wang, S. et Fu, J. (2022). HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models. Dans *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (p. 334-346). Association for Computational Linguistics. https://aclanthology.org/2022.findings-aacl.32

Mallat, S. (2019). *L'apprentissage par réseaux de neurones profonds*. https://www.youtube.com/watch?v=WYZp2IZFLcE

Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T., Njie, M. et Matthews, J. (2021). Gender Bias in Natural Language Processing Across Human Languages. Dans *Proceedings of the First Workshop on Trustworthy Natural Language Processing* (p. 45-54). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.trustnlp-1.6

Milmo, D. (2023, 10 février). Google v Microsoft: who will win the AI chatbot race? *The Guardian*, Technology. https://www.theguardian.com/technology/2023/feb/10/google-v-microsoft-who-will-win-the-ai-chatbot-race-bard-chatgpt

Morozov, E. (2014). *To save everything, click here : the folly of technological solutionism* (PublicAffairs). Paperback first published.

Naughton, J. (2023, 4 février). ChatGPT isn't a great leap forward, it's an expensive deal with the devil. *The Observer*, Opinion. https://www.theguardian.com/commentisfree/2023/feb/04/chatgpt-isnt-a-great-leap-forward-its-an-expensive-deal-with-the-devil

Newton, C. (2020). *The withering email that got an ethical AI researcher fired at Google*. https://www.platformer.news/p/the-withering-email-that-got-an-ethical

Noble, S. (2018, 26 février). Interviewé par J. Snow. Bias already exists in search engine results, and it's only going to get worse [MIT Technology Review].

https://www.technologyreview.com/2018/02/26/3299/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. et Dean, J. (2021, 23 avril). *Carbon Emissions and Large Neural Network Training*. arXiv. https://doi.org/10.48550/arXiv.2104.10350

Perrigo, B. (2023). *OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Pitron, G. (2021). *L'enfer numérique : Voyage au bout d'un like* (Éditions Les Liens qui libèrent).

Pitron, G. (2022, 26 mars). Interviewé par P.-L. Poyau. Guillaume Pitron : « Le numérique n'a pas été conçu pour être vert ou régler un quelconque problème environnemental ». https://lvsl.fr/guillaume-pitron-le-numerique-na-pas-ete-concu-pour-etre-vert/

Pommier, É. (2014, 26 décembre). Interviewé par G. Mosna-Savoye. Hans Jonas. Pour une éthique de la société technologique. https://www.radiofrance.fr/franceculture/podcasts/les-chemins-de-la-philosophie/hans-jonas-pour-une-ethique-de-la-societe-technologique-9272708

Prabhakaran, S. (2020). *T Test (Students T Test) - Understanding the math and how it works*. Machine Learning Plus. https://www.machinelearningplus.com/statistics/t-test-students-understanding-the-math-and-how-it-works/

Rouvroy, A. (2018, 6 mars). *Rencontre avec Antoinette Rouvroy : gouvernementalité algorithmique et idéologie des big data*. https://www.youtube.com/watch?v=cQCeAe8wPKU

Ruder, S., Vulić, I. et Søgaard, A. (2022). Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. Dans *Findings of the Association for Computational Linguistics: ACL 2022* (p. 2340-2354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.184

Rudinger, R., May, C. et Van Durme, B. (2017). Social Bias in Elicited Natural Language Inferences. Dans *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (p. 74-79). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1609

Sandelowski, M. et Barroso, J. (2007). *Handbook for Synthesizing Qualitative Research*. Springer Publishing Company, Inc.

Schaul, K., Chen, S. Y. et Tiku, N. (2023). *Inside the secret list of websites that make AI like ChatGPT sound smart*. Washington Post. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Schick, T., Udupa, S. et Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408-1424. https://doi.org/10.1162/tacl_a_00434

Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5248-5264). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

Sherzer, J. (2012). Langage et culture : une approche centrée sur le discours. *Langage et société*, *139*(1), 21-45. https://doi.org/10.3917/ls.139.0021

Singh, M. (2020). *Google workers demand reinstatement and apology for fired Black AI ethics researcher | Alphabet | The Guardian*. The Guardian. https://www.theguardian.com/technology/2020/dec/16/google-timnit-gebru-fired-letter-reinstated-diversity

Skopeliti, C. et Milmo, D. (2023, 8 février). ChatGPT needs a huge amount of editing : users views mixed on AI chatbot. *The Guardian*, Technology. https://www.theguardian.com/technology/2023/feb/08/chatgpt-users-views-ai-chatbot-essays-emails

Strubell, E., Ganesh, A. et McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Dans *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (p. 3645-3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. et Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. Dans *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (p. 1630-1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Tucker, I. (2017, 28 mai). « A white mask worked better »: why algorithms are not colour blind. *The Observer*, Technology. https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when-algorithms-are-racist-facial-recognition-bias

van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W. et Schulz, K. (2022). *Undesirable biases in NLP: Averting a crisis of measurement*. https://doi.org/10.48550/ARXIV.2211.13709

Vincent, J. (2023). *Google announces ChatGPT rival Bard, with wider availability in 'coming weeks'*. The Verge. https://www.theverge.com/2023/2/6/23588033/google-chatgpt-rival-bard-testing-rollout-features

Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D. et Tsvetkov, Y. (2018). RtGender: A Corpus for Studying Differential Responses to Gender. Dans *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1445

Vu, K. (2022, 18 mai). *GPT-2 (GPT2) vs. GPT-3 (GPT3): The OpenAI Showdown - DZone*. dzone.com. https://blog.exxactcorp.com/gpt2-vs-gpt3-the-openai-showdown/

Webster, K., Costa-jussà, M. R., Hardmeier, C. et Radford, W. (2019). Gendered Ambiguous Pronoun (GAP) Shared Task at the Gender Bias in NLP Workshop 2019. Dans *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (p. 1-7). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3801