Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Panorama sur la représentation des genres dans la presse française

Une nouvelle approche basée sur l'Entity Linking

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Elodie PHOMMADY

Directeur de mémoire :

Damien Nouvel

Encadrants :

Benoît Laurent

Guillaume Lechien

Année universitaire 2021/2022

Remerciements

Je tiens tout d'abord à témoigner ma gratitude envers mon directeur de mémoire, M. Damien NOUVEL, pour sa patience, sa bienveillance et ses précieux conseils durant toutes ces années, et particulièrement pendant le Master.

Je remercie mes tuteurs, M. Benoît LAURENT et M. Guillaume LECHIEN, pour leur temps, leur partage et leur gaieté. Je suis extrêmement reconnaissante de leur soutien moral.

Je voudrais également remercier tous les employés d'Aday pour leur accueil chaleureux et pour m'avoir permis de réaliser mon stage et mon mémoire dans d'excellentes conditions.

Je remercie l'équipe enseignante PluriTAL de m'avoir accompagné tout au long de mon parcours, depuis la licence LLCER pour certains, et de m'avoir fait découvrir le monde du TAL.

À mes amis et mes camarades, merci pour leur soutien et pour leur bonne humeur communicative.

Enfin, je tiens à remercier infiniment ma famille et surtout mes parents, pour tous leurs sacrifices et tout ce qu'ils ont fait pour que je reçoive une bonne éducation.

Préambule

Nous souhaitons mettre en évidence le fait que nous ne sommes pas experts en sociologie, et particulièrement en étude de genres. Certaines terminologies, constructions de phrases ou exemples peuvent paraître et/ou être offensants, notamment envers les personnes transgenres ou envers celleux qui ne se reconnaissent pas dans la perception binaire du genre. En aucun cas ce travail a pour intention de porter atteinte à tout individu. Nous souhaitons, au contraire, aider les personnes qui ne sont pas touchées par ces problématiques à mieux comprendre les enjeux que soulèvent l'effacement et la sous-représentation de certains genres.

Résumé

De nos jours, les médias ont indéniablement une influence sur nos sociétés, et inversement. De nombreux travaux ont mis l'accent sur les inégalités de genre, en particulier au sujet de la sous-représentation médiatique des femmes par rapport aux hommes, en abordant le genre comme un concept binaire. Les autres identités de genre sont encore très peu incluses dans les problématiques étudiées, notamment en TAL. Nous proposons dans ce mémoire une méthode d'analyse de la représentation d'un éventail plus large de genres (féminin, masculin, non-binaire, *gender queer*, *gender fluid*, trans) dans un vaste corpus de presse française, basée sur de la liaison d'entités mentionnées dans les articles (*Entity Linking*). Cette démarche présente des avantages, elle permet par exemple de traiter automatiquement un volume de données important, ou encore de prendre en compte cette diversité de genres. Elle fait cependant face à la complexité de la problématique, autant d'un point de vue sociologique que TAL. Les résultats obtenus confirment une sous-représentation des genres minoritaires, dans la presse comme dans les méthodes et ressources (bases de connaissance), face à laquelle quelques pistes de travail sont proposées.

Mots-clés : entity linking, dbpedia spotlight, identité de genre, représentation des genres, presse, français

Table des matières

Liste des figures

Liste des tableaux

Introduction

Les médias (presse, télévision, radio, réseaux sociaux,...) se sont démultipliés et diversifiés avec l'apparition des nouveaux supports en ligne. Ils sont devenus les plateformes de référence pour le partage d'informations. Avec ce progrès, nous constatons une augmentation de l'influence des médias, impactant inévitablement la société et notre perception des réalités sociales, ce qui a fait émerger de nouvelles problématiques. Au cours des années, nous avons observé une exposition plus ou moins importante de certaines personnes et de certaines identités de genres dans les médias. De nombreux travaux sur les rapports entre les genres ont mis en exergue la sous-représentation des femmes dans la presse. En plus des inégalités hommes-femmes, le combat des personnes transgenres, entre autres, a remué et redéfini la question de l'égalité des genres dans notre société. La vision binaire du genre s'est progressivement estompée pour faire place à un spectre plus large et diversifié, incluant des identités non binaires.

En parallèle, le domaine du Traitement Automatique des Langues (TAL), ou en anglais *Natural Language Processing* (NLP), a connu un essor considérable. L'accès aux informations et aux connaissances se faisant de plus en plus sur des supports numériques, le TAL peut être utilisé pour traiter de manière automatique cette abondance d'informations. Grâce à sa capacité d'extraire des informations dans des données non structurées, le TAL est utile dans de nombreux champs d'applications comme, par exemple, pour les systèmes de questions-réponses.

Réalisé dans le cadre du master TAL proposé à l'Institut National des Langues et Civilisations Orientales (Inalco), ce mémoire est le fruit d'une étude sur les méthodes de calcul et d'analyse de la représentation des genres dans un corpus de presse française moderne. Nous proposons des éléments de réflexion sur la pertinence du TAL, et plus particulièrement de l'*Entity Linking*, sur l'étude de la représentation des genres, en termes de visibilité et de stéréotypisation. Est-ce que l'*Entity Linking* permet de dresser une cartographie objective des genres dans les médias ? À quel degré le TAL apporte-t-il une contribution dans l'étude des déséquilibres de classes sociales dans les médias ? Quelle est la part de visibilité de chaque genre dans la presse française ? Les genres non binaires sont-ils représentés ? Quelles perceptions de la société sur les différents genres le TAL permet-il de faire remonter ? Nous tentons de répondre à ces questions par le biais de ce travail de recherche.

Dans un premier temps, nous remettrons le travail dans son contexte en reprenant ce qu'est le genre. Nous expliquerons également le principe et l'objectif de la tâche d'*Entity Linking*. Nous nous attarderons ensuite sur la constitution de notre corpus de travail, à savoir un ensemble d'articles de presse récupérés depuis l'immense base de données de l'entreprise de veille média Aday. Nous expliciterons les méthodes appliquées dans ce mémoire pour annoter notre jeu de données et, par la suite, nous analyserons les résultats obtenus concernant la représentation des genres à travers notre corpus. Nous terminerons en évoquant les limites et les possibilités d'amélioration de notre approche.

CHAPITRE 1

LA NOTION DE GENRE

Sommaire

Introduction

La problématique de ce mémoire repose non seulement sur des méthodes de TAL mais aussi sur des concepts sociologiques. Pour prétendre faire de l'analyse sur la représentation et la distribution des genres dans la presse, il important de poser quelques définitions sur les notions sociologiques mentionnées dans ce travail.

Dans ce chapitre, nous allons reprendre dans un premier temps la notion d'identité de genre : sa définition, la différence entre « sexe » et « genre », qui n'est souvent pas très claire pour tout le monde, et le caractère complexe et évolutif des identités de genre. Dans un second temps, nous évoquerons les tenants et les aboutissants de précédents travaux portant sur les identités de genre dans les médias d'une part, et en TAL d'autre part, avec une inclusion nuancée des genres non binaires [1].

1.1 Mise en contexte sociologique

Le concept de genre, et plus précisément d'identité de genre, ne cesse de se développer et suscite un intérêt grandissant, en particulier dans une société où l'on cherche à atteindre l'égalité et l'équité à toutes les échelles. En revanche, la notion de genre est encore perçue dans la plupart des cas sous un point de vue binaire. Qu'est-ce que le genre ? Comment est étudiée la question du genre ?

1. Fait allusion ici à tous les genres autres que masculin et féminin. Pas uniquement les personnes non-binaires, mais aussi trans, etc.

1.1.1 Définition de l'identité de genre

Dans la langue française, le terme « genre » est extrêmement polysémique. Le CNRTL (*Centre National de Ressources Textuelles et Lexicales*) en propose moult définitions qui déclinent toutes plus ou moins d'une description globale : « *Ensemble d'êtres ou d'objets ayant la même origine ou liés par la similitude d'un ou de plusieurs caractères* ». Le mot « genre » peut être utilisé pour parler d'une multitude de choses telles qu'une attitude ou une manière d'être, des préférences personnelles, un genre grammatical, un genre littéraire ou encore un genre musical. De nos jours, il est même employé sous la forme d'un adverbe, particulièrement à l'oral. Il est donc compliqué d'entreprendre quelconque conversation ou recherche centrée sur le genre sans en préciser le sens.

En anglais, la notion de genre sociologique est représentée simplement par le mot « *gender* ». Ce terme américain fut progressivement traduit dans les écrits francophones par « genre », qui apporte avec lui de nombreuses confusions quant à son interprétation et sa définition. Dans notre cadre de recherche, nous restons à un niveau de surface des définitions et ne nous plongeons pas dans une analyse approfondie du genre, nécessitant des connaissances solides en sciences sociales (sociologie, anthropologie, psychologie, études de genre,...).

Le genre d'un individu, comme employé dans le contexte social actuel, renvoie en réalité à son **identité de genre**, terme introduit aux États-Unis sous l'appellation « *gender identity* » à la suite d'une étude menée par le psychanalyste Robert Stroller dans les années 1960 opposant le genre à l'orientation sexuelle [Stoller, 1968]. Selon le site officiel du gouvernement du Canada [2], l'identité de genre est « *l'expérience intérieure et personnelle que chaque personne a de son genre. Il s'agit du sentiment d'être une femme, un homme, les deux, ni l'un ni l'autre, ou d'être à un autre point dans le continuum des genres.* ». Par souci de lisibilité, « genre » et « identité de genre » sont employés de manière interchangeable dans ce mémoire, même si l'OMS (*Organisation Mondiale de la Santé*) [3] et un article de QuestionSexualité [4] précisent que ces deux termes font référence à des concepts différents [5].

La mention de l'identité de genre est souvent accompagnée de celle de l'« expression de genre », signifiant la manière dont une personne extériorise et dévoile son genre aux autres. L'expression de genre peut être liée ou non à l'identité de genre d'un individu. Une personne s'identifiant au genre féminin peut se présenter de manière masculine aux yeux des autres, de par son style vestimentaire, par exemple. Tout comme il existe une différence entre expression et identité de genre, une distinction importante est à souligner entre « sexe » et « genre ».

1.1.2 Sexe et genre

Le genre, sous-entendu « identité de genre », et le sexe, sous-entendu « sexe biologique », sont des concepts totalement différents, bien qu'ils soient intimement liés. Le sexe d'un individu est assigné conformément à ses caractéristiques biologiques, que ce soit son anatomie (attributs génitaux, gonades,...) ou autres (chromosomes,

2. https://www.canada.ca/fr/ministere-justice/nouvelles/2016/05/identite-de-genre-et-expression-de-genre.html
3. https://www.who.int/health-topics/gender#tab=tab_1
4. https://questionsexualite.fr/connaitre-son-corps-et-sa-sexualite/la-diversite-de-genre/qu-est-ce-que-l-identite-de-genre
5. Le genre est exposé comme une construction sociale entraînant des normes, des rôles et des rapports humains particuliers, tandis que l'identité de genre se réfère à un ressenti intime.

production hormonale,...). Il n'est pas nécessairement binaire ni figé. Les personnes intersexes présentent des caractères sexuels qui ne se conforment pas aux définitions traditionnelles du sexe masculin ou féminin. Il est également possible de changer de sexe.

Les termes « sexe » et « genre » sont couramment utilisés de manière interchangeable car il est commun d'attribuer un genre à la naissance en relation directe avec le sexe biologique. Néanmoins, un individu peut s'identifier à un genre qui n'est pas traditionnellement conforme à son sexe, comme dans le cas des personnes transgenres.

Si l'identité de genre est fréquemment confondue avec le sexe, elle n'est pas non plus équivalente à l'orientation sexuelle. Ce sont des notions indépendantes. L'orientation sexuelle s'inscrit dans le concept de sexualité et désigne simplement l'attirance émotionnelle ou sexuelle qu'une personne a envers une autre. Toutes ces dissemblances ne sont pas encore bien ancrées dans le conscient collectif car ces notions sont relativement nouvelles et les terminologies employées sont sujettes à de nombreux changements.

1.1.3 Un concept complexe en perpétuelle évolution

Anciennement perçue comme une réalité binaire, la prise de parole de la communauté LGBTQIA+, entre autres, a mis en exergue une représentation plus vaste du genre, en soulevant notamment des questions sur la transidentité. L'identité de genre s'inscrit dans un spectre non figé, d'une certaine profondeur et complexité.

En fonction du sexe assigné à la naissance, un individu peut se placer dans une certaine dimension du spectre. Une personne se présentant comme **cisgenre** (ou « cis ») est une personne dont l'identité de genre est en adéquation avec le sexe (et parfois, par la même occasion, le genre) assigné à la naissance. À l'inverse, une personne peut être **transgenre** (simplifié par « trans »), auquel cas le sexe et/ou le genre assigné à la naissance ne correspond pas à son identité de genre.

L'identité de genre étant relative à chacun, il n'y a pas réellement de définitions exactes pour chaque genre et chaque individu peut en avoir sa propre version. Cependant, nous exposons tout de même ci-dessous quelques descriptions qui sont généralement liées à des identités de genre retrouvées dans la société occidentale :

- **agenre** : personne ne s'identifiant à aucun genre,
- **féminin** : personne en accord avec les propriétés féminines déterminées par la société,
- **masculin** : personne en accord avec les propriétés masculines déterminées par la société,
- **non-binaire** : terme générique pour désigner une personne qui ne s'identifie pas entièrement comme masculine, ni féminine,
- ***gender queer*** : personne dont le genre n'est pas normé (ni masculin ni féminin, les deux ou un mélange des deux,...),
- ***gender fluid*** : personne dont l'identité et l'expression de genre fluctuent selon le temps ou les circonstances.

Ces définitions sont à prendre avec du recul et cette liste ne rend pas compte des différentes dimensions du spectre de genre. La complexité du spectre entraîne beaucoup de confusions quant à la signification et l'usage de certains termes, tels que « non-binaire » et « *gender queer* » qui sont souvent utilisés comme synonymes. Il est impossible d'établir une liste exhaustive de toutes les identités de genre car, comme

l'affirme QuestionSexualité.fr, « *il existe autant de nuances et d'identités que de personnes* ».

La notion de genre varie aussi bien à l'échelle individuelle, avec le développement personnel de son identité de genre au cours de sa vie, qu'à l'échelle collective, avec la prise de conscience générale de la diversité des genres. La conceptualisation moderne du genre prend de plus en plus d'importance, et ce, à plusieurs niveaux : dans les relations humaines avec l'usage de pronoms adéquats pour s'adresser à quelqu'un [Lauscher et al., 2022], via les inégalités d'accès aux droits, via les discriminations, etc. La variété des genres est d'autant plus large lorsque l'on considère non seulement les genres occidentaux, mais aussi les genres originaires d'autres régions et d'autres cultures [Bahadurdesai, 2018]. La question de l'identité de genre a suscité une attention particulière ces dernières années auprès des nouvelles générations ainsi qu'auprès des chercheurs.

1.2 Représentation des genres

Le phénomène d'inégalité des genres est observé depuis plusieurs décennies en sciences sociales et en TAL. Il présente malgré tout un lien fort avec d'autres critères sociodémographiques (âge, profession,...). Il peut être abordé sous divers approches : quantitatives (« *gender imbalance* »), qualitatives (« *gender stereotyping* »), etc. De plus, l'éventail des identités de genre considéré varie d'un travail à un autre : conception binaire du genre, inclusion des personnes transgenres, ouverture vers un spectre plus large du genre, et bien d'autres.

1.2.1 Dans les médias

L'étude de la représentation de genres dans les médias n'est pas une tâche récente [Tuchman, 1979] et reste toujours d'actualité.

Les analyses quantitatives des genres dans les médias sont intéressantes dans le sens où nous pouvons faire varier l'échelle de comparaison (internationale, nationale,...) et elles permettent d'observer facilement l'évolution de la répartition des genres en fonction du temps. De manière globale, les écarts de représentation médiatique entre le genre masculin et féminin sont élevés. Au cours de ces vingts dernières années, le pourcentage de femmes représentées dans les médias a graduellement augmenté [6] mais ne permet pas d'invalider le phénomène d'« annihilation symbolique » [7] des femmes, soulevé par Gaye Tuchman [Tuchman, 1979]. Les situations de crises, comme la pandémie de Covid-19, ont fait reculer la progression vis-à-vis de l'exposition médiatique des femmes [Calvez, 2020].

Les analyses qualitatives contribuent, quant à elles, à l'étude des stéréotypes de genre. À l'encontre des analyses quantitatives, celles-ci dépendent fortement de la langue de travail et les stéréotypes relevés sont décomposés en observations sociolinguistiques. Les chercheurs ont souligné que, contrairement à la gent masculine, les femmes sont communément présentées à travers leurs attributs physiques, leur statut matrimonial et sont dépeintes comme « *victimes et non comme actrices des problèmes sociaux* » [Richard et al., 2022].

6. Constatation appuyée par le projet GMMP (*Global Media Monitoring Project*).
7. Effacement ou sous-représentation d'un groupe de personnes, maintenant une inégalité sociale.

Récemment, plusieurs travaux ont traité cette problématique de représentation en tirant parti des apports fournis par le TAL, notamment avec la prédiction de genre et l'extraction de citations [Richard et al., 2022, Asr et al., 2021].

1.2.2 Dans le TAL

[Richard et al., 2022] rend compte de l'écart de représentation des genres dans la presse française en considérant deux mesures générées via des approches computationnelles : le taux de masculinité des mentions et la proportion des hommes cités. Ces indicateurs quantitatifs, calculés de manière automatique sur une fenêtre hebdomadaire et affichés sur leur site [8], permettent de se rendre compte de l'évolution globale des déséquilibres de genres dans la presse française.

De nos jours, nous dénotons une réelle utilité de posséder des informations sur le genre (et/ou de sexe) d'une personne dans plusieurs domaines et pour de nombreuses applications, notamment en médecine. Ces renseignements peuvent être obtenus de différentes manières : par les personnes directement concernées (par exemple à l'aide d'un formulaire), par attribution grâce à des données externes, par prédiction (avec des modèles de classification, entre autres), etc.

L'inférence de genre sur la base de prénoms ou noms est un exemple classique de prédiction de genre. Parmi les outils les plus connus, nous retrouvons NamSor, Gender API, genderize.io ou bien encore Wiki-Gendersort [Bérubé et al., 2020]. Cependant, cette approche simplifie le concept du genre et se confronte à plusieurs difficultés dont la prise en compte des dimensions temporelle et spatiale [Sebo, 2021]. En effet, le ou les genre(s) attribué(s) à un prénom varie(nt) en fonction du temps et de la région du monde, introduisant une culture et des coutumes propres à ces endroits.

Pour la plupart des travaux de TAL, la question du genre se traduit par le traitement des biais de genres existants dans des modèles de langues. [Kaneko and Bollegala, 2022] met en évidence l'encodage de biais sociaux, dont le genre et l'âge font partie, dans les plongements de mots (*word embeddings*) des modèles de langues masqués comme BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019] ou ALBERT [Lan et al., 2020]. Ces biais de genre se propagent et engendrent des erreurs dans des tâches plus spécifiques comme la résolution de coréférences [Rudinger et al., 2018], la reconnaissance d'entités nommées ou bien la traduction automatique [Dev et al., 2021]. Pour que les traitements soient indépendants des biais existants et produisent des résultats plus « justes », un effort notable est consacré ces dernières années à la réduction de l'impact de ces biais [Sun et al., 2019].

1.2.3 Évolution vers l'inclusion des genres non binaires

Dans la plupart des champs d'études, les identités de genres qui ne se conforment pas à la binarité "classique" ne sont pas intégrées dans la liste des genres étudiés. Il en va de même pour les études basées sur le sexe, où les personnes intersexes sont rarement prises en compte. Tous les travaux de TAL ne suivent pas les mêmes critères d'inclusion en matière de théorisation et représentation du genre [Devinney et al., 2022]. Le traitement exclusif des genres masculin et féminin peut faire du tort aux personnes touchées par cette problématique [Dev et al., 2021]. L'exclusivité accordée à la conception binaire des genres provoque un effacement des

8. https://gendered-news.imag.fr/genderednews/

identités de genres non binaires, qui sont donc perçues comme non valides ou qui sont réduites à une étiquette « mixte/indéfini/autres ». Elle peut apparaître dans un questionnaire n'indiquant que « M » ou « F » comme genres, dans des jeux de données annotés en *male*/*female*/*unknown*, dans des *embeddings* pour une prédiction binaire de genres, et beaucoup d'autres. Les modèles de langues encodent généralement des propriétés sociolinguistiques associées aux genres binaires. Les applications dépendantes de ces modèles sont elles aussi affectées par ces biais ou par l'effacement de genres insufflés dans ces encodages.

Un autre versant des préjudices à l'encontre des personnes non binaires concerne la mauvaise classification (*misclassification*) de genre et l'emploi de termes inappropriés par rapport à l'identité de genre de la personne à laquelle on s'adresse (*misgendering* ou mégenrage).

Peu à peu, certains travaux de TAL tentent d'accorder de l'importance à ces genres non binaires [Cao and Daumé III, 2020, Cao and Daumé, 2021, Havens et al., 2022]. Il existe aussi des corpus, comme celui d'ORTOLANG [9] où les données sont associées à des personnes de diverses identités de genres, et des ressources spécialisées sur des genres non binaires, tel que *Nonbinary Wiki* [10].

Il est incontestable que le meilleur moyen de connaître le genre d'une personne est de recueillir cette information directement auprès des individus concernés. Cela permettrait de respecter au maximum l'identité de cette personne et d'analyser de manière plus rigoureuse les résultats obtenus. Néanmoins, il est clair que cette technique de collecte de données est difficilement envisageable pour élaborer une base de connaissance au volume important. De plus, sachant que ces informations peuvent changer au cours du temps, l'entretien et la mise à jour de ces ressources demanderaient également un effort considérable.

Conclusion

L'identité de genre est une notion encore peu connue et faisant l'objet de lacunes auprès du grand public. Actuellement, la confusion entre « sexe », « genre » et « orientation sexuelle » reste commune mais les discussions soulevées au sein de notre société tendent à nous éduquer sur ce sujet. Les recherches sur les inégalités hommes-femmes ont mis en avant la question de visibilité des genres. Au-delà de la conception traditionnelle des genres, nous constatons également un manque d'inclusion, voire une mise sous silence, des genres non binaires, aussi bien dans la sphère médiatique qu'académique. Ces dernières restent en majorité centrées sur une approche binaire du genre.

Pour certaines recherches et applications, il s'avère essentiel de recueillir des informations sur le genre des individus. Le seul moyen d'obtenir une information fidèle sur le genre est de générer un corpus dont les genres sont indiqués par les personnes elles-mêmes impliquées. Cependant, l'élaboration d'un tel corpus est coûteuse en temps et en moyens humains. De plus, il pourrait ne pas être complètement fiable sur une longue période puisque l'identité de genre est également dépendante de l'espace et du temps.

Pour palier au maximum à tous ces problèmes, nous proposons d'exploiter une base de connaissance possédant un grand nombre de ressources, comme DBpedia ou Wikidata, pour attribuer un genre à une entité désambiguïsée par l'*Entity Linking*.

9. https://www.ortolang.fr/market/corpora/gender_spectrum_speech?lang=en
10. https://nonbinary.wiki/wiki/Main_Page

CHAPITRE 2

AMBIGUÏTÉ ET ENTITY LINKING

Sommaire

Introduction

Savoir lever une ambiguïté est crucial pour comprendre certaines informations. La tâche de désambiguïsation (pour connaître le sens d'un mot, l'identité d'une personne mentionnée,...) est relativement facile pour les humains, sachant notre connaissance externe du monde et notre capacité à interpréter des intentions non explicites à l'aide du contexte. En revanche, pour une machine, la tâche s'avère être plus compliquée et entraîne un effort de réflexion, de conception et de programmation considérable. En TAL, cette question est en partie traitée grâce à la désambiguïsation lexicale (*word sense disambiguation*) et à la désambiguïsation d'entités (*entity disambiguation*). Cette dernière est généralement incluse dans un travail plus global qu'est l'*Entity Linking*.

Ce chapitre définit plus précisément ce qu'est l'*Entity Linking*. Nous reprenons les grandes parties du traitement d'un modèle d'*Entity Linking*, à savoir la détection et la désambiguïsation d'entités nommées, et présentons différentes formes de modèles existants.

2.1 Principe de l'Entity Linking

L'***Entity Linking***, abrégé en « EL », consiste à identifier des entités nommées dans un texte non structuré, et les lier à des entrées uniques dans une base de connaissance après désambiguïsation. Ce lien permet de distinguer sans ambiguïté la nature ou l'identité de l'entité nommée mentionnée.

D'après le Thésaurus de l'activité gouvernementale du Québec [1], une entité nommée est une « *chose existant dans le monde réel et qui est une instance ou membre de classes décrites par des concepts sujet* ». En d'autres termes, c'est une expression

1. https://www.thesaurus.gouv.qc.ca/tag/terme.do?id=15592

faisant référence à une entité unique définie dans le monde réel. Elle peut être représentée sous différentes formes textuelles :

— un nom propre (par exemple « Joe Biden »),
— une description complète ou incomplète (« le 46e président des États-Unis », « le président des USA », « l'ancien candidat à la vice-présidence »),
— un pronom (par exemple « il »), qui reprend une entité.

Plusieurs expressions peuvent renvoyer au même objet ou à la même entité. En linguistique, on dit que ces signifiants ont une même référence : ce sont des coréférences. À l'inverse, une même forme de surface peut pointer vers des références distinctes, qui peuvent être des entités nommées ou non, en fonction du cadre d'énonciation. Par exemple, « *Apple* » peut se rapporter à divers objets, tels que le fruit (traduction de « pomme » en anglais), l'entreprise américaine *Apple Inc.* ou encore la ville de New York que l'on surnomme « *Big Apple* ». Pour avoir une compréhension totale de l'information, il y a un réel besoin de désambiguïser ces expressions.

La désambiguïsation des entités dépend en grande partie du contexte (social, culturel, géographique, temporel,...). Le syntagme nominal « le président » ne pointe pas vers la même personne qu'il soit déclaré de nos jours ou 10 ans auparavant, en France ou dans un autre pays, en parlant d'une nation ou d'une organisation comme une entreprise.

L'EL peut être décliné sous d'autres noms : NEL (*Named-Entity Linking*), NERD (*Named-Entity Recognition and Disambiguation*), NED (*Named-Entity Desambiguation*), NEN (*Named-Entity Normalization*), etc. Cette tâche permet d'enrichir des textes à l'échelle des entités et ces informations complémentaires sont exploitables pour d'autres applications, comme des recherches biomédicales [Zheng et al., 2015] ou des systèmes de Question Answering [Diomedi and Hogan, 2021]. Le procédé d'EL se divise en deux étapes classiques :

1. la **reconnaissance d'entités nommées**, qui détecte les occurrences d'entités nommées dans un texte et les inscrit dans des classes prédéfinies, ou la **détection de mentions**, qui extrait les entités susceptibles d'être désambiguïsées,
2. la **désambiguïsation d'entités**, qui établit le lien entre l'entité identifiée et une entrée dans une base de connaissance.

La figure 2.1 [2], extraite de la page anglaise de Wikipédia sur l'EL, montre le déroulement général du processus.

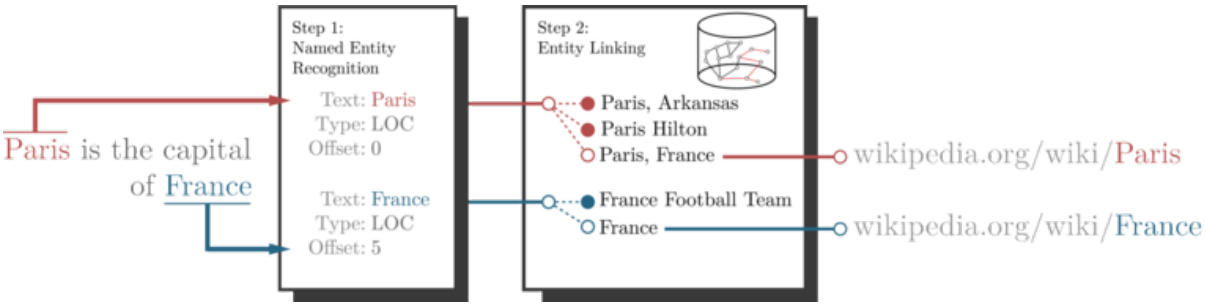

FIGURE 2.1 – Schéma général du processus d'Entity Linking

2. Crédit : By Aparravi - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=80106788

2.2 Décomposition du traitement en deux étapes

2.2.1 Reconnaissance d'entités nommées

Avec l'évolution de l'accès à l'information, notamment par le biais d'outils numériques, un besoin de compréhension s'est développé. Néanmoins, tenter de comprendre des textes peu voire non structurés dans leur ensemble s'avère être un objectif ambitieux compte tenu des nombreuses difficultés rencontrées, à la fois théoriques et pratiques. Une alternative s'est donc développée : l'extraction d'information. Elle cherche à extraire des informations pertinentes d'un texte pour un but donné.

La **reconnaissance d'entités nommées**, plus communément citée sous les abréviations REN, NER (*Named-Entity Recognition*) ou encore NERC *Named-Entity Recognition and Classification*), est une tâche qui a vu le jour avec le développement des systèmes d'extraction d'information et a été introduite par le biais des conférences MUC (*Message Understanding Conferences*), notamment au cours de la sixième édition. Bien que les études sur les entités nommées soient entreprises depuis longtemps, les recherches sur la REN restent tout de même très actuelles et sont encore étudiées de nos jours, comme dans les travaux de [Suárez et al., 2020] sur l'amélioration de modèles français.

[Suárez et al., 2020] définit la REN comme « *une tâche qui consiste à identifier des morceaux de textes désignant des entités nommées tel que des personnes, des lieux ou des organisations* » [3]. Étiqueter des entités nommées dans un texte non structuré implique deux objectifs : définir les frontières de chaque entité et les classer dans des catégories prédéfinies. Les premières classes introduites lors des conférences MUC [4] sont des éléments accompagnés d'un attribut qui permet de préciser la catégorie des entités nommées (tableau 2.1). Mais, en pratique, les classes peuvent être plus ou moins précises, sachant que la granularité des classes dépend entre autres des applications de la REN, comme pour une future tâche d'EL [Tedeschi et al., 2021].

Élement	**Attribut**	**Description**
ENAMEX (Identifieurs uniques : noms propres, acronymes,...)	ORGANIZATION	Entreprises, institutions gouvernementales ou autres organisations
	LOCATION	Lieux politiques (pays,...) ou géographiques (montagnes,...)
	PERSON	Noms de personnes ou familles
TIMEX (Expressions temporelles)	DATE	Dates complètes ou partielles
	TIME	Périodes de la journée complètes ou partielles
NUMEX (Expressions numériques)	MONEY	Montant d'argent
	PERCENT	Pourcentages

TABLE 2.1 – Catégories d'entités nommées au format MUC

Cependant, [Ehrmann, 2008] évoque en détail la difficulté de définir ce qu'est une entité nommée, en considérant à la fois l'aspect linguistique et l'aspect TAL. Ce sont des unités lexicales particulières qui posent de nombreuses questions et difficultés :

3. Traduit de : « *task consisting in identifying text spans that denote named entities such as person, location and organization names, to name the most important types* »

4. https://aclanthology.org/M95-1024/

— Elles peuvent se présenter sous plusieurs formes de surface (par exemple « Sir Lewis Hamilton », « L. Hamilton » ou « un pilote automobile britannique »).
— Elles peuvent être enchâssées (l'expression « le président de la France » désigne une personne mais inclus également une organisation [5]).
— Comment traiter les combinaisons de syntagmes ? (comme « Emmanuel et Brigitte Macron »)
— Comment délimiter les différentes formes que l'entité nommée peut avoir (multiples expressions définies,...) et leurs frontières textuelles ?
— etc.

Une même forme de surface peut également appartenir à plusieurs classes en fonction du contexte. Pour reprendre l'exemple d'« *Apple* », si elle est utilisée pour désigner le fruit, elle ne sera pas considérée comme entité nommée. Par contre, pour les deux autres références énoncées, il faut déterminer leur classe (organisation ou lieu) et circonscrire correctement les bornes de l'entité nommée. La difficulté d'annotation des entités nommées liée à la polysémie de leurs formes de surface doit nécessairement être surmontée par la désambiguïsation de ces dernières.

2.2.2 Désambiguïsation d'entités nommées

Les travaux sur la désambiguïsation d'entités s'inscrivent dans la continuité de la REN. Désambiguïser une entité identifiée dans un texte équivaut à l'associer avec une entrée unique dans une base de connaissance. Comme nous l'avons vu avec le traitement complexe des entités nommées, il n'est pas si simple pour une machine de faire le lien entre ces deux objets. Le processus conventionnel de désambiguïsation d'entités respecte l'ordre suivant :

1. **sélection de candidats** dans la base de connaissances pouvant correspondre à l'entité reconnue dans le texte,

2. **classement des candidats** en fonction des attributs qui permettent de rapprocher les candidats de la mention

Ne tenir compte que de la forme de surface de l'entité nommée ne suffit pas pour mener à bien chaque étape du traitement. Effectivement, nous avons constaté qu'une entité peut apparaître sous de multiples variations : expressions complètes ou incomplètes, formes abrégées, variations d'écritures selon la pays, etc. Par conséquent, la prise en compte à la fois de la mention et de son contexte est cruciale et représente une caractéristique importante dans la désambiguïsation des entités nommées.

L'encodage des informations est important, aussi bien que le choix des bases de connaissances employées, qui peut varier en fonction de l'objectif de l'EL. En effet, de nos jours, même si l'EL est majoritairement utilisé pour la presse et les réseaux sociaux, elle peut également utile dans d'autres domaines [Sowinski et al., 2022]. Les bases de connaissance pouvant être des graphes, tels que Wikidata, DBpedia, ou encore YAGO, nous observons l'existence de modèles avec des approches textuelles, qui peuvent aussi être combinées avec des approches à base de graphes [Mulang' et al., 2020].

5. Le syntagme « la France » indique dans ce cas-là une organisation et non un lieu car l'expression fait référence au gouvernement français.

2.3 Variété des approches et des modèles

Modèles *end-to-end*

Tous les modèles d'EL ne réalisent pas forcément la première étape du traitement. Étant donné que la détection d'entités nommées est une tâche qui a énormément été étudiée en TAL, la plupart des articles d'EL s'occupaient jusqu'à présent de la partie « désambiguïsation d'entités » seulement. C'est pour cette raison que NED et NEL se substituent parfois l'un à l'autre pour parler de la même tâche. Néanmoins, nous observons ces dernières années un essor des travaux qui abordent l'EL en *end-to-end*, réalisant donc la détection de mentions et la désambiguïsation d'entités au sein d'un même modèle [Kolitsas et al., 2018]. Mais, contrairement à la REN, la détection de mentions ne réalise pas forcément la classification d'entités nommées.

Pour les modèles qui se concentrent uniquement sur la désambiguïsation d'entités, nous supposons que les corpus dont ils font usage sont préalablement annotés en entités par des systèmes de REN. Pour ces données-là, l'usage des informations de classes d'entités peut s'avérer bénéfique pour la tâche de désambiguïsation [Martins et al., 2019].

Monolingue ou multilingue

Les modèles d'EL sont fortement liés à la langue traitée. Au fil du temps, les modèles multilingues se sont ajoutés aux modèles monolingues mais, comme dans la majeure partie des traitements de TAL, les avancements sont principalement réalisés en anglais et beaucoup moins dans les autres langues. Pour les langues peu dotées, les recherches sont encore très pauvres.

En ce qui concerne la langue française, l'EL reste une tâche sous-étudiée. La plupart du temps, la désambiguïsation d'entités en français est proposée dans les outils d'EL multilingues, comme DBpedia Spotlight [Daiber et al., 2013] ou Babelfy [Moro et al., 2014]. Le manque de données annotées limite l'entraînement et l'évaluation de modèles d'apprentissage automatique. Les rares corpus français étiquetés en entités nommées ne sont pas souvent libres d'accès, ni gratuits. Cela peut être expliqué par le coût important (en temps et en ressources) des annotations manuelles, mais aussi par les difficultés rencontrées pour étiqueter les entités nommées. [Ehrmann, 2008] souligne le fait que ces entités sont « *des objets difficiles à cerner et à traiter* ».

Approche multimodale

Le *Multimodal Entity Linking* (MEL) est une version de l'EL qui s'appuie sur des contextes de différentes natures pour désambiguïser des entités. Généralement, ce sont des contextes visuels (comme des images) qui, associés à du texte, permettent d'apporter des éléments d'informations supplémentaires et d'affiner la désambiguïsation [Adjali et al., 2020].

Conclusion

Le TAL a notamment permis d'automatiser et d'optimiser l'accès à la connaissance. Néanmoins, le caractère polysémique de la langue, sa complexité et ses variations mettent en évidence l'importance de désambiguïser des contenus textuels non

structurés. L'*Entity Linking* connaît donc une popularité grandissante au sein de la communauté mais ce n'est pas sans soulever d'autres questions liées à la difficulté du traitement des entités nommées.

L'essor que connaît l'*Entity Linking* fait avancer le progrès dans les autres applications de TAL (systèmes de questions-réponses, de recherche d'informations,...) et profite aussi à d'autres domaines d'études. Dans notre contexte de recherche, elle agit comme base de traitement pour nos observations sur les comportements et inégalités transmis à travers la presse française. Grâce à l'identification et la désambiguïsation des entités nommées effectuées par l'*Entity Linking*, nous pouvons joindre des personnes à leur genre en évitant de les mégenrer pour se rapprocher le plus fidèlement possible du portrait que dépeint la sphère médiatique envers les individus de genres différents.

CHAPITRE 3

ÉLABORATION DU CORPUS

Sommaire

Introduction

Afin de répondre à la problématique et d'étudier la représentation des genres dans la presse, il nous faut élaborer un corpus de travail qui donne une vue d'ensemble des articles que l'on retrouve dans les médias de presse.

Dans ce chapitre, nous introduisons dans un premier temps la ressource nous permettant d'établir un corpus d'articles de presse. Ensuite, nous abordons les points choisis pour filtrer les articles et la manière de les récupérer. Enfin, nous établissons une présentation générale du corpus constitué.

3.1 Source des données : Entreprise Aday

Avant de vouloir constituer un corpus sur la presse, il est nécessaire de savoir où récupérer ces données et de connaître les caractéristiques du support de données. Dans notre cas, les données sont collectées depuis les ressources proposées par la société Aday.

Aday est une entreprise de veille et d'analyses de contenus médias. Créée en 1980 sous le nom « Edd », elle s'est récemment ouverte à un panorama international et multilingue. Elle propose de plus en plus de services destinés aux entreprises. Leur plateforme de veille-médias, Tagaday [1], met à disposition du contenu actualisé chaque jour. Plus de 1,5 million de nouvelles données sont traitées quotidiennement, que ce soit des articles de presse nationale, presse régionale, presse magazine, presse spécialisée ou professionnelle, presse en ligne, des flux de presse audio-visuelle (télévision, radio,...) ou des publications sur les réseaux sociaux.

Avec plus de 400 millions d'articles, de dépêches et autres supports informationnels, Aday possède les ressources et les moyens suffisants pour pouvoir constituer un corpus représentatif de la presse française (plus particulièrement, de la presse francophone).

1. https://www.aday.fr/plateforme-tagaday/

3.2 Filtrage des articles

Afin d'obtenir un corpus représentatif de la presse moderne francophone, il y a plusieurs critères à prendre en compte. Dans un premier temps, nous considérons la dimension temporelle. Les questions autour de la parité hommes-femmes et de la tolérance envers la communauté LGBTQIA+ étant des problématiques émergentes, nous avons pour ambition d'analyser la représentation actuelle des genres dans les médias. Nous avons donc ciblé des articles dans une période donnée : entre le 1er Janvier 2021 et le 15 Octobre 2022 inclus.

Dans un deuxième temps, nous nous attardons sur les types d'articles choisis pour constituer le corpus. Nous nous limitons ici à de la presse écrite dans la langue française. Les ressources textuelles issues du traitement *Speech-to-Text* de médias audio-visuels (radios, émissions télévisées,...) sont donc exclues du corpus, pour en limiter son volume et pour avoir un jeu de données homogène sur de la presse écrite, entre autres. Nous décidons également de rester sur de la presse générale traitant de sujets nationaux ou internationaux :

- articles de presse quotidienne nationale,
- articles de presse magazine,
- articles de presse en ligne (flux internet),
- articles des certaines agences de presse : AFP (*Agence France Presse*) et Reuters (en français uniquement).

Les documents provenant de la presse régionale et la presse professionnelle, ainsi que les communiqués de presse, sont donc écartés du corpus pour éviter de sélectionner un type de presse extrêmement ciblé pour un public en fonction de leurs caractéristiques sociodémographiques. De tels articles pourraient introduire des entités nommées plus spécifiques et moins connues, qui ne sont pas forcément référencées dans des bases de connaissances libres comme DBpedia et Wikidata. À plus petite échelle, il y a certains documents que nous souhaitons éviter, tels que :

- des sommaires et publicités dans les magazines,
- des pages de jeux, caractéristiques du format magazine : mots croisés, mots fléchés, etc.
- des plannings (programmes télévisés, sorties cinéma,...),
- des prévisions météorologiques, astrologiques (horoscope), et autres,
- des articles sous forme de PDF, d'images, etc.

Dans le but de filtrer la plupart des articles indésirables, nous avons réfléchi à plusieurs critères de requête pouvant influencer la récupération des articles. Se baser sur un nombre minimum de mots par article est une solution envisagée, en émettant l'hypothèse qu'un article très court ne mentionne pas de personnes et qu'il ne contribue pas à la représentation des genres dans les médias. Néanmoins, nous ne pouvons pas nous fier entièrement à ce paramètre pour éviter l'ensemble des articles indésirables. Certains documents comme les sommaires peuvent avoir un nombre de mots suffisamment conséquent. En considérant cela, il est logique de vouloir fixer une valeur élevée pour le nombre de mot minimum. Mais dans ce cas-là, des petits articles pertinents pourraient être omis.

Pour remédier à cela, nous avons également pris en compte l'usage de certains mots dans les titres d'articles. En effet, ceux-ci peuvent laisser entrevoir le contenu de l'article, comme pour un article titré « SOMMAIRE ». Le lexique des contenus d'articles étant incroyablement large, nous ne recherchons pas directement la présence de ces mots au sein de l'article au risque d'éliminer une multitude de documents.

Suite à ces réflexions, nous avons décidé d'exclure les articles :

— bruts (sous forme de page magazine, de PDF,...), audio-videos ou biographiques,
— n'étant pas rédigé dans la langue française,
— de moins de 150 mots (sachant que nous nous basons sur le calcul de mots fourni dans les métadonnées de l'article),
— contenant les expressions « sommaire », « mots fléchés », « mots croisés », « *advertisement* » et « *advertisements* » dans le titre.

Les termes anglais « *advertisement* » et « *advertisements* » sont employés ici, à défaut de « publicité ». Ce dernier est un nom commun couramment utilisé dans la langue française. De plus, après avoir jeté un œil aux documents fournis par l'outil de recherche d'articles Aday, les titres contenant le mot « *advertisement* », contrairement à son équivalent français, pointaient la plupart du temps vers des vignettes publicitaires ou des articles illisibles.

Si tous ces choix visent à récupérer un corpus de presse avec le moins de bruit possible, ils introduisent par la même occasion un biais dans le corpus en écartant ou non, selon des conditions pré-définies, des documents pouvant être pertinents.

3.3 Récupération des données

La recherche de documents dans la base de données Aday est composée de plusieurs paramètres :

— une requête respectant la syntaxe interne à l'entreprise (eQuery) et passant par Elasticsearch,
— une date de début et de fin déterminant la période de recherche d'articles en fonction de leur date de publication,
— un pas de requête, pour préciser à quelle intervalle de temps la requête est lancée et les articles sont récupérés (heures, jours, mois ou années),
— un nombre maximal de documents attendus,
— les métadonnées et champs à récupérer pour chaque document (identifiant, titre, contenu de l'article,...).

Ces configurations permettent une recherche et une extraction d'articles plus précise et ciblée. Il est également possible de spécifier les éditeurs ou les groupes d'éditeurs [2] dont nous voulons obtenir les articles. Comme énoncé précédemment, nous avons fait le choix de rester dans la sphère de la presse écrite générale. Pour cela, nous avons sélectionné des groupes d'éditeurs précis, recensés dans le tableau A.1 en Annexe.

Pour chaque document, nous prélevons le corps de l'article et des métadonnées associées, choisies pour leur éventuelle utilité dans l'étude de la représentation genrée. Celles-ci sont précisées dans le tableau 3.1.

2. Les groupes d'éditeurs sont élaborés par Aday en fonction de la liste d'éditeurs qui composent leur banque de données.

Champ	Description
identifier	Identifiant unique Aday
uri	URI du document source dans la base de données Aday
publisher	Code à 4 caractères désignant l'éditeur
title	Titre du document
content	Contenu principal du document
word_count	Nombre de mots du document
language	Langue du document
schema	Type de document XML[3] : *press*, *wire*, *web*, *audiovideo*, *biography*, *raw*
doc_type	Nature du document : *article*, *cover*, *toc*,...
media_tag	Liste du ou des thème(s) traité(s) par l'éditeur[4]
theme	Thème du document

TABLE 3.1 – Sélection des métadonnées

Certains champs sont récupérés inévitablement, de manière automatique :
— l'identifiant, l'index et le score Elasticsearch,
— le type de document XML, qui est identique au champ *schema*,
— *page*, qui correspond au nombre de la première page de l'article pour les documents de type *press* et *raw*,
— *highlighted_words*, qui correspondent aux mots trouvés concordant avec la requête.

La requête est appliquée pour chaque jour de la période de recherche saisie et parmi les articles répondant à tous les critères mentionnés, un nombre limité d'articles aléatoires est récupéré pour composer le corpus final. Le nombre d'articles récupéré par jour est déterminé en divisant la taille maximale du corpus (information entrée par l'utilisateur) par le nombre de jours dans la période définie. Étant donné que nous sommes limités par les ressources à notre disposition (nombre d'articles disponibles dans la période donnée) et par le temps de calcul, nous restreignons la taille du corpus à 100 000 articles maximum.

3.4 Distribution des données

Au total, 92 560 articles sous la forme de fichiers JSON ont été recueillis pour ce corpus de presse française moderne. Malgré le filtrage, il y a quand même des articles non souhaités qui ont été récupérés :
— 5 articles ont un contenu vide, mais ont été tout de même collectés car un nombre de mots est renseigné dans les métadonnées,
— des articles de listes de produits à recommander et/ou à vendre,
— etc.

En considérant les jours comme classes, le corpus obtenu est équilibré dans le temps. Néanmoins, pour le nombre d'articles collectés en fonction des groupes d'édi-

3. S'apparente à la plateforme de diffusion de l'article.
4. Cette métadonnée correspond aux thèmes traités par l'ensemble des articles de l'éditeur, mais ne dit pas le ou les thème(s) exactement traité(s) dans l'article en question.

teurs et des éditeurs en eux-même, le corpus est très déséquilibré, ce qui est un point non négligeable à souligner lors des analyses des résultats car certaines classes auront donc un poids beaucoup plus important que d'autres. La distribution des articles dans le corpus est contenue dans le tableau 3.2.

Code du groupe d'éditeurs	Type d'éditeurs	Nombre d'articles
PURE	Sites web	20 104
SITESMAG	Sites web	17 158
SITESPQN	Sites web	16 020
PRESSEPQNACTUGENE	Journaux	8 802
PRESSEMAGCULTURE	Magazines	8 691
AFP	Agences de presse	7 761
PRESSEMAGACTUGENE	Magazines	3 953
PRESSEPQNSPORT	Journaux	2 079
PRESSEPQNECO	Journaux	1 738
PRESSEMAGECO	Magazines	1 132
PRESSEMAGFEM	Magazines	1 087
PRESSEMAGAUTO	Magazines	1 071
PRESSEMAGFINANCE	Magazines	826
PRESSEMAGSPORT	Magazines	679
PRESSEAGENCES	Agences de presse	408
PRESSEMAGPEOPLE	Magazines	319
PRESSEMAGSANTE	Magazines	276
PRESSEMAGNTIC	Magazines	171
PRESSEMAGINTERNATIONAL	Magazines	156
PRESSEMAGEDUC	Magazines	124

TABLE 3.2 – Distribution générale des articles du corpus selon les groupes d'éditeurs

Parmi les 1 707 éditeurs inclus dans le jeu de données, les 5 éditeurs ayant le plus d'articles dans le corpus sont :

— le site des *Echos* avec 3 310 articles,
— *AFP Fil Général*, avec 2 858 articles,
— le site du *Figaro*, avec 2 199 articles,
— *Le Monde*, avec 2 150 articles,
— le site de *Mediapart*, avec 2 051 articles.

Et inversement, nous comptons 24 éditeurs comptabilisant un seul article et qui, pour cette raison, ont un impact moindre sur le corpus en général, tels que *santemagazine.fr*, *rocknfolk.com*, *snowsurf.com*, *carnetsdeleconomie.fr*, *terre-futur.com*, *abcfeminin.com*, *Marie Claire Style*, *Zen et bien dans ma vie*, etc.

Conclusion

La constitution d'un corpus de presse amène à se poser de nombreuses questions pour lesquelles des décisions sont requises et doivent être explicitées. Pourquoi décidons-nous de sélectionner tel article plutôt qu'un autre ? Dans la plupart des cas, ces choix sont justifiés par l'objectif final. Ils induisent toutefois, de manière volontaire ou involontaire, des biais à prendre en compte.

À l'issue de cette procédure, nous disposons d'un corpus d'articles de presse se voulant être représentatif de la sphère médiatique actuelle en France, et qui présente par conséquent un fort déséquilibre des classes (expliqué en partie par le fait que les éditeurs ne publient pas à la même fréquence). Seule une analyse temporelle propose des classes équilibrées. Néanmoins, nous pouvons désormais procéder à l'annotation automatique du corpus pour pouvoir entreprendre une étude des écarts genrés de représentation dans les médias.

CHAPITRE 4

MÉTHODES

Sommaire

Introduction

Avec le corpus de presse en notre possession, nous entamons l'étape d'annotation automatique du corpus nécessaire à l'analyse de la représentation de genres dans les médias. Elle débute par une succession de pré-traitements, notamment pour enrichir les métadonnées des articles. Ensuite, nous expliciterons les processus permettant l'annotation et désambiguïsation des documents en entités de type *personne*. Dans un dernier temps, nous procédons à l'attribution des identités de genre aux personnes mentionnées dans le corpus à l'aide des ressources de DBpedia FR et Wikidata.

L'entièreté du processus et de la démarche globale appliquée dans ce mémoire est reprise dans le schéma A.1.

4.1 Pré-traitements

4.1.1 Récupération du contenu brut des articles

La méthode de récupération de documents avec Elasticsearch permet d'obtenir le corps des articles. Cependant, après analyse, ces contenus ont été normalisés. Entre autres, les guillemets droits ont été supprimés. Les guillemets français ont, pour la plupart, été gardés.

Néanmoins, la suppression de certains types de guillemets entraîne une perte d'information, sachant que les citations incarnent l'importance de la part de parole

accordée à quelqu'un [Richard et al., 2022]. La version de l'article adaptée pour les recherches Elasticsearch n'incluant pas les guillemets droits, nous devons donc passer par la version XML du document pour avoir accès à la version la moins retravaillée de l'article chez Aday et pour laquelle est tout de même associée des métadonnées.

Pour rappel, parmi les champs recueillis pour chaque article, nous avons également collecté l'URI du document source. Grâce à cet URI et la librairie Python *requests*, nous avons accès au document XML, qui contient la version la moins retravaillée du corps de l'article, depuis la *warehouse* d'Aday. Nous considérons que le texte principal est contenu dans les balises `<p>`. Si le texte est inclus dans d'autres balises (par exemple `<div>`), il n'est malheureusement pas récupéré. Dans le cas où aucun texte n'est trouvé, nous retournons la valeur `None`. Enfin, nous conservons le résultat dans le fichier JSON de l'article sous le champ *raw_content*.

```
"[...] C'est un privilège d'occuper ce poste, ce n'est jamais quelque
chose que l'on tient pour acquis , a-t-il souligné. [...]"

"[...] \"C'est un privilège d'occuper ce poste, ce n'est jamais
quelque chose que l'on tient pour acquis\", a-t-il souligné. [...]"
```

FIGURE 4.1 – Extrait d'un article récupéré par Elasticsearch lors de l'élaboration du corpus et de la version brute collectée après coup (AFP, 17 Août 2022)

4.1.2 Classification par thème

En nous avançant sur les futures analyses possibles du corpus, nous voulons comparer la répartition des genres dans les médias à plusieurs niveaux, notamment selon le thème de l'article.

Parmi les métadonnées collectées, *media_tag* et *theme* (tableau 3.1) pourraient convenir pour séparer les articles en classes de thème. Cependant, ces métadonnées présentent des inconvénients considérables :

- beaucoup de documents ont ces champs non renseignés (ayant la valeur `null`) : 59902 articles pour *media_tag* et 58194 articles pour *theme*,
- la métadonnée *media_tag* correspond aux thèmes d'un éditeur. Un éditeur pouvant traiter divers thèmes, cette métadonnée est soit une liste de valeurs, soit `null`. De plus, nous considérons ici les thèmes d'un éditeur. Il est moins précis de se baser sur cette métadonnée plutôt que sur le thème de l'article, même s'il est possible que plusieurs sujets soient traités dans un seul article.
- le *theme* de l'article étant renseigné par les éditeurs eux-mêmes, les étiquettes ne respectent pas un ensemble de valeurs en particulier. De ce fait, nous obtenons un nombre élevé d'étiquettes (5505 thèmes uniques) dont certaines peuvent éventuellement être regroupées sous une seule classe.

Ainsi, n'ayant pas de classification consistante sur les articles, nous décidons de procéder nous-même à la catégorisation des articles à l'aide d'un transformer pour la classification en thème (*topic classification*) qui dérive de l'utilisation du corpus MLSUM (*MultiLingual SUMmarization dataset*) [Scialom et al., 2020] : *flaubert-mlsum-topic-classification* [1]. Comme mentionné sur leur page Hugging Face, ce modèle de

1. https://huggingface.co/lincoln/flaubert-mlsum-topic-classification

détection de *topic* sur les articles de presse se base sur une version du modèle FlauBERT (*flaubert_base_cased*) [Le et al., 2020], un BERT entraîné sur un large corpus français, et est *fine-tuné* à partir d'un corpus d'articles de la base de données MLSUM. Les thèmes ont été extraits à partir des URLs des articles et ils procèdent ensuite à un nettoyage et regroupement de thèmes. Les classes retenues sont des thèmes ayant plus de 100 articles associés : économie, opinion, politique, société, culture, sport, environnement, technologie, éducation et justice.

Le modèle présent sur Hugging Face réalise une détection de thème sur la base du corps de l'article. Pour les articles de notre corpus ne présentant pas de corps principal, nous ne réalisons pas la classification sur le titre du document car il ne convient pas aux bonnes pratiques du modèle mis à disposition.

Une matrice de confusion (figure 4.2) est mise à disposition sur la page Hugging Face pour observer les performances du modèle. Sauf pour la classe « justice » où le modèle ne parvient pas à prédire le bon thème dans la majorité des cas, la prédiction des autres classes est correcte plus de la moitié du temps, avec un maximum de 92% des articles correctement associés à la catégorie « sport ». Nous notons la difficulté du modèle à reconnaître les articles de « justice », qui les confond avec des articles de « société » (à 60%). Seuls 33% des documents de justice sont bien prédits.

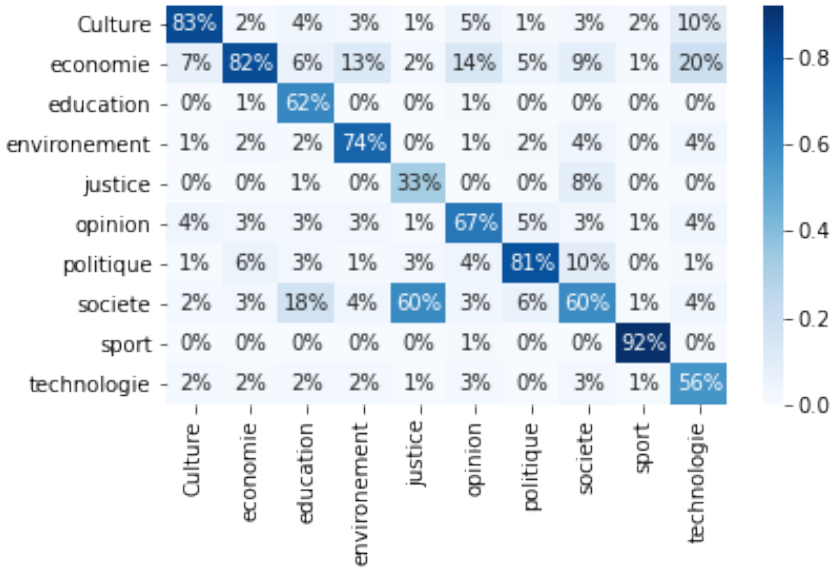

FIGURE 4.2 – Matrice de confusion du transformer *flaubert-mlsum-topic-classification* (reprise de la page Hugging Face du projet)

Il est précisé que les résultats ne sont pas garantis sur le long terme, sachant que les articles de presse dépendent d'événements spécifiques dans une période temporelle donnée. Les analyses de représentation de genres en fonction des thèmes abordés dépendront également des prédictions de ce *transformer*. Les labels et les scores associés aux thèmes prédits sont mémorisés dans les champs *topic_label* et *topic_score* du fichier JSON pour chaque article.

4.1.3 Mise à jour des métadonnées

En plus des nouvelles métadonnées extraites, nous pouvons obtenir et ajouter des métadonnées supplémentaires aux articles à partir de celles déjà existantes : les dates de publication des articles, le nom de l'éditeur et les groupes d'éditeurs auxquels ils appartiennent, définis en interne chez Aday [2].

En ce qui concerne les dates, elles sont incluses dans plusieurs métadonnées existantes, par exemple *identifier* ou *uri*. Afin d'éviter de relancer des instructions sur ces métadonnées pour extraire la date de publication, nous la prélevons lors de ces pré-traitements et la sauvegardons sous la clé *yyyymmdd*.

Pour les noms et groupes d'éditeurs, nous nous servons de fichiers CSV qui nous ont également été utiles lors de l'élaboration du corpus (chapitre 3). Le fichier *pub_group.csv* contient sur chaque ligne le code de l'éditeur, son nom complet et la liste des groupes auxquels il appartient, ordonné du plus précis au plus général. Pour la métadonnée *group_publisher*, nous retenons le groupe d'éditeur le plus précis, qui correspond aux groupes mentionnés dans le tableau 3.2, pour chaque article du corpus.

Après avoir mis à jour toutes les données extraites pendant les prétraitements, nous obtenons au final des articles restés au format JSON avec des informations supplémentaires ajoutées.

4.1.4 Division en lots

Nous proposons également la possibilité de séparer le corpus en plusieurs lots, qui correspondent simplement à des regroupements aléatoires d'un nombre défini de documents. Cette option est mise à disposition pour le bon fonctionnement des traitements suivants, notamment pour éviter de saturer le serveur local utilisé avec DBpedia Spotlight.

Dans notre cas, nous avons choisi de faire des regroupements de 5000 articles, ce qui divise notre corpus en 18 lots dont un est composé uniquement de 2560 documents. Avec un nombre plus élevé de documents par lots, il semblait avoir une surcharge de traitements dans le terminal qui faisait interrompre brutalement le processus, sans avoir enregistré les fichiers déjà traités. Pour l'analyse de la représentation des genres, nous reprenons bien en compte le corpus complet, indépendamment des lots.

4.2 Annotation automatique des entités

Pour traiter le corps principal des articles en entrée, certaines étapes sont indispensables, comme celle de la tokénisation.

4.2.1 Utilisation d'un modèle de langue spaCy

Pour effectuer la tokénisation, nous utilisons un *tokenizer* proposé dans un modèle de langue de spaCy. Bien qu'il existe d'autres outils sur le marché, nous avons fait le choix d'utiliser le framework proposé spaCy pour diverses raisons. C'est une bibliothèque Python *open-source*, gratuite et moderne, fréquemment utilisée dans les projets de Traitement Automatique des Langues. Elle propose des modèles de

2. Les groupes sont organisés de manière hiérarchique, c'est-à-dire que nous avons des groupes d'éditeurs généraux qui incluent plusieurs groupes d'éditeurs plus précis.

langues récents et pré-entraînés avec plusieurs composants utilisables pour multiples types d'analyses (morphologiques, syntaxiques,...) et la prise en main des modèles de langues mis à disposition est relativement facile.

Par contre, il y a des inconvénients à souligner, comme le fait qu'il soit plus compliqué d'entraîner nos propres modèles sur spaCy et que les modèles de langues en français n'ont pas forcément les meilleures performances comparés à d'autres langues.

spaCy propose tout de même quatre modèles de langues français :

— *fr_core_news_sm*,
— *fr_core_news_md*,
— *fr_core_news_lg*,
— *fr_dep_news_trf*, la version *transformer* basée sur *camembert-base*.

Ces modèles sont entraînés sur des textes écrits de presse française (*news and media*) et utilisent tous le même *tokenizer* contenant les mêmes règles pour la tokénisation des textes en langue française. Par ailleurs, nous avons ajouté quelques règles personnalisées liées notamment à la gestion de toutes les formes de guillemets qui sont collées à des ponctuations fortes ou bien des apostrophes.

Nous comparons ensuite les performances de leurs composants énoncées sur https://spacy.io/models/fr.

Scores	**Modèles spaCy**			
	sm	***md***	***lg***	***trf***
POS_ACC	96.21	97.20	97.30	98.69
MORPH_ACC	95.26	96.09	96.78	97.94
MORPH_MICRO_F	97.26	98.02	98.45	99.23
SENTS_F	87.20	90.97	89.64	94.47
DEP_UAS	87.20	89.99	89.66	94.64
DEP_LAS	82.96	85.97	85.81	92.50
TAG_ACC	93.34	94.33	94.42	95.78
LEMMA_ACC	90.29	90.76	90.71	91.22
ENTS_F	81.15	83.33	84.12	-
Taille du modèle	15 MB	43 MB	545 MB	382 MB

TABLE 4.1 – Informations des modèles français spaCy

Les performances des modèles classiques sont assez équivalentes, notamment entre *fr_core_news_md* et *fr_core_news_lg*. En considérant les résultats des composants qui nous intéressent (comme le *sentencizer* éventuellement) et la taille du modèle, *fr_core_news_md* est le modèle classique le plus adapté pour ce projet. Le transformer *fr_dep_news_trf* peut aussi être une option. Il est plus performant que les modèles classiques mais d'autres points nous dissuadent de travailler avec ce modèle :

- Il est plus lourd que *fr_core_news_md*.
- L'absence de composant pour la reconnaissance d'entité nommée n'est pas un problème en soi car il ne va pas être utilisé. Néanmoins, bien qu'il y ait des scores pour la segmentation en phrases, il n'y a pas de composant propre pour celle-ci. Une segmentation en phrases est possible avec le *dependency parser* mais nous n'avons pas plus d'idées sur sa qualité.
- La taille des documents traités par ce modèle est limité[3]. Envoyer des docu-

3. Message d'avertissement obtenu lors d'un essai de traitement : "Token indices sequence length is

ments trop grands ne va pas cesser brutalement le traitement mais les annotations ne vont pas être réalisées sur l'entièreté du document.

En prenant en compte toutes ces caractéristiques, notre choix de modèle s'est porté sur *fr_core_news_md*.

4.2.2 Entity Linking avec DBpedia Spotlight

Pour intégrer d'autres genres que ceux binaires et éviter de causer du tort aux personnes touchées en cas d'erreurs de classifications (*misclassifications*), nous récupérons les informations de genres depuis une base de connaissance après avoir lié les entités présentes dans nos données textuelles.

Parmi les modèles et outils d'*Entity Linking* gratuits et libres, nous avons choisi DBpedia Spotlight [Daiber et al., 2013] pour de multiples raisons. Il réalise de l'*Entity Linking* en français, ce que tous les outils ne proposent pas forcément. Il est également possible de le manipuler par plusieurs moyens (API REST ou modèle déployé en local). L'avantage principal dans notre cas est qu'il est facile de combiner son utilisation avec celle de spaCy. En effet, nous utilisons ce modèle d'*Entity Linking* sous la forme de *wrapper* spaCy, appelé *spacy-dbpedia-spotlight* [4].

Il est possible de le télécharger avec `pip` ou via le dépôt git. Ainsi, ce *wrapper* est utilisable en Python avec spaCy en intégrant le composant `dbpedia_spotlight` directement dans un modèle de langue vide ou dans un pipeline de modèles de langue existants. Ce composant réalise la reconnaissance et la désambiguïsation d'entités sur la base de connaissance DBpedia, indépendamment des autres composants du pipeline spaCy. Les résultats de l'*Entity Linking* sont sauvegardés en tant que `Span` dans l'attribut `ents` d'un document spaCy, normalement utilisé par le composant de reconnaissance d'entités nommées de spaCy. Il est possible de configurer des paramètres du modèle, tels que :

- `dbpedia_rest_endpoint` : permet de réaliser l'Entity Linking soit via l'API REST, hébergé sur un serveur distant, soit via le modèle déployé localement,
- `process` : pour choisir jusqu'à quelle étape le traitement est effectué (détection de mentions (*spot*), génération de candidats pour chaque mention (*candidates*) ou désambiguïsation (*annotate*)),
- `confidence` : possibilité de choisir la valeur du seuil de confiance pour la désambiguïsation,
- `types` : permet de filtrer l'Entity Linking pour qu'elle ne soit réalisée seulement sur des types d'entités précis et recensés dans DBpedia (par ex. *DBpedia :Place*),
- Et d'autres paramètres (filtrage `sparql`,...).

Compte tenu du nombre de requêtes qu'il aurait fallu exécuter pour notre corpus et souhaitant respecter l'usage des autres utilisateurs de l'API public de DBpedia Spotlight, nous optons pour l'option consistant à déployer le modèle sur un serveur local. Cette méthode présente également d'autres avantages, comme l'accès à plus de langues pour l'*Entity Linking* et un temps de calcul plus rapide.

En prenant en compte notre objectif principal qui est d'obtenir une représentation des genres dans la presse française, l'annotation des entités nommées de type *personne* est primordiale. L'*Entity Linking* sur les autres entités peut être complé-

longer than the specified maximum sequence length for this model (4539 > 512). Running this sequence through the model will result in indexing errors".

4. https://github.com/MartinoMensio/spacy-dbpedia-spotlight

mentaire mais n'est pas nécessaire. Par conséquent, nous laissons le choix à l'utilisateur d'appliquer ou non l'*Entity Linking* sur tous les types d'entités ou uniquement sur les entités de type *DBpedia :Person*, pour aussi réduire le temps de calcul, ce qui est un élément à prendre en compte dans un contexte industriel et également académique.

Malgré les commodités permettant une utilisation plus personnalisée du modèle, DBpedia Spotlight présente tout de même des inconvénients. Bien qu'il soit utilisé dans un pipeline spaCy, le modèle a recours à sa propre tokénisation et sa propre reconnaissance d'entités nommées (ou plutôt la détection de syntagmes pouvant être liées à une entrée dans DBpedia, comme défini par [Daiber et al., 2013]). Il n'est pas possible de contourner ou de remplacer ces étapes du processus car *spacy-dbpedia-spotlight* se présente comme un modèle d'*Entity Linking* en *end-to-end*. Par conséquent, nous ne pouvons pas utiliser séparément le désambiguïsation de DBpedia Spotlight avec un modèle plus performant de reconnaissance d'entités, par exemple. Il n'est donc pas utile de laisser actif le composant *ner* du pipeline dans le modèle spaCy, puisque les résultats ne seront pas exploités par DBpedia Spotlight. Après avoir été identifiées et désambiguïsées, les entités nommées du texte sont liées à des entités dans la version française de DBpedia, qui ne contiennent pas les informations recherchées, c'est-à-dire l'identité de genre des personnes. Il est ainsi indispensable de passer par une autre étape pour attribuer un genre à l'entité liée.

4.3 Attribution de genre via des bases de connaissances

Nous ne passons pas par de l'inférence de genre, mais bien par une attribution de genre à l'aide de bases de connaissances, étant donné que les modèles de prédictions ne classent généralement qu'en genre binaire. Les bases de connaissances nous permettent d'augmenter les informations en genres.

4.3.1 Extraction de données sur DBpedia FR et Wikidata

DBpedia ne fournit pas d'information sur le genre, contrairement à Wikidata [5]. Wikidata est l'une des plus grande base de connaissance libre, gratuite, multilingue et collective. Ces données sont structurées et servent également de support pour les autres projets Wikimedia, comme pour les mises à jour des encadrés informatifs dans les pages Wikipédia. Pour récupérer le genre de l'entité désambiguïsée sur DBpedia, il faut donc passer par l'entité Wikidata correspondante.

DBpedia Spotlight ne nous donne que peu d'informations sur l'entité désambiguïsée comme l'URI de la ressource, qui est intimement lié à l'article de Wikipédia de l'entité, ou bien leurs types, qui sont les classes dont l'entité est l'instance. Pour faire le lien entre DBpedia FR et Wikidata et pour que l'entité désambiguïsée pointe bien sur la même personne, nous extrayons l'identifiant Wikidata contenu dans le triplet ayant comme prédicat `owl:sameAs` et comme objet le préfixe `dbpedia-wikidata`. Pour cela, nous décidons de passer par du *web scraping* avec la bibliothèque *Beautiful Soup*.

Ensuite, avec l'identifiant Wikidata en main, nous utilisons un client Wikidata afin de récupérer des propriétés d'une entité, notamment celle sur son genre. Dans notre cas, nous recherchons l'attribut de la propriété « *sex or gender* » d'une entité Wikidata. Comme vous pouvez le constater, la propriété ne fait pas seulement mention

5. https://www.wikidata.org/wiki/Wikidata:Main_Page

du genre mais aussi du sexe, ce qui n'est pas l'objet de notre analyse. Il faut en tenir compte dans les extractions obtenues.

Il se peut néanmoins que le genre n'arrive pas à être extrait pour diverses causes :

— l'URI de l'entité pointe sur une page de désambiguïsation,
— l'entité typée comme *personne* dans DBpedia FR n'est pas en réalité une personne (ex. http://fr.dbpedia.org/resource/Attentat_du_19_décembre_2016_à_Berlin)
— l'URI pointe sur un groupe de personne ou un duo,
— l'entité dans Wikidata ne possède pas la propriété recherchée (entité faisant référence à un duo, propriété simplement absente),
— etc.

4.3.2 Filtrage dépendant du cadre d'expérience

Pour éviter de chercher à nouveau des genres déjà identifiés, nous conservons les informations récoltées dans un dictionnaire. Et, en gardant en mémoire la visée de cette étude, nous faisons en sorte d'attribuer un genre aux entités qui sont identifiées comme *personne*. Nous pourrions réaliser l'attribution de genre à toutes les entités, quel que soit leur type, sachant que certaines n'ont pas de type renseigné. Cependant, cela prend énormément de temps, ce qui n'est pas raisonnable au vu des objectifs de notre travail.

Après avoir effectué un test sur un extrait du corpus actuel, le traitement prend environ 1 minute par article. Si la vitesse du programme restait constante, pour un volume de données un peu plus grand, par exemple 1000 articles, cela prendrait 16 heures et 40 minutes. Sachant qu'il existe un nombre très conséquent d'entités pouvant être reconnues et désambiguïsées, la technique du dictionnaire ne permettrait pas de baisser de manière drastique le temps de calcul, sauf si les articles mentionnent souvent les mêmes entités. Néanmoins, cela présente tout de même un avantage. Chaque entité serait liée à son identifiant Wikidata, ce qui peut être utile si nous voulons effectuer d'autres travaux par la suite, comme des comparaisons sur la base d'un autre critère sociodémographique.

Conclusion

Afin de réaliser une étude de la représentation des genres dans la presse française, il est indispensable de disposer de données portant des informations relatives aux genres. Nous avons donc annoté automatiquement l'ensemble des articles du corpus en personne et en identité de genre. Pour cela, nous nous sommes aidés du modèle de DBpedia Spotlight et de Wikidata, base de connaissance constamment alimentée et mise à jour. Si nous considérons que les bases de connaissances sont approvisionnées d'informations fondées et non déduites, l'attribution de genres à des entités désambiguïsées par l'*Entity Linking* permet une annotation plus juste et plus respectueuse des individus mentionnés. Le corpus annoté nous permet à présent de pouvoir extraire des éléments de travail quantitatifs et qualitatifs dans le but de réaliser nos analyses sur les discriminations de genres dans la presse.

CHAPITRE 5

RÉSULTATS

Sommaire

Introduction

L'application de notre modèle nous permet de disposer d'un corpus volumineux annoté automatiquement en personnes avec leur genre. À partir de là, nous pouvons proposer des analyses quantitatives et qualitatives de la représentation des genres dans la presse française. Nous entreprenons dans un premier temps une analyse sur l'intégralité des données. Ensuite, nous spécifions l'analyse en fonction de plusieurs paramètres tels que le thème des articles et la date de parution.

Nous tenons à préciser un point avant d'entreprendre les analyses. Vu que les annotations de genre dépendent de la propriété « *sex or gender* » de Wikidata, des étiquettes ne correspondant pas à des identités de genre ont été récupérées : *intersex* et *travesti*. Le premier fait référence à un sexe. Le second décrit une personne qui change de façon épisodique ou permanente son expression de genre. Dans le cadre de notre étude sur la représentation des identités de genres, nous écartons de nos analyses toute personne annotée comme intersexuée ou travestie.

5.1 Représentation à l'échelle du corpus entier

Sur notre corpus composé de 92 560 documents, 32 474 ne contiennent pas d'entités désambiguïsées de type *personne*, ce qui représente plus d'un tiers du jeu de données. Néanmoins, nous rappelons que cette estimation dépend de la fiabilité des traitements effectués précédemment. Les articles contenant des personnes désambiguïsées répertorient en moyenne 6,5 occurrences de personnes, avec un minimum d'une seule mention et un maximum de 234 occurrences de personnes.

5.1.1 Analyse quantitative

Concernant l'analyse quantitative du contenu pour la représentation de genres, nous souhaitons savoir combien de fois des personnes d'un certain genre sont mentionnées. Nous observons un écart imposant d'une part entre les genres masculins et féminins, et d'autre part de manière frappante entre les genres binaires et non binaires. Ces différences engendrent un manque de visibilité pour les genres en sous-représentation.

Nous comptabilisons un total de 391 793 mentions de personnes dans l'ensemble du corpus. En revanche, le nombre de personnes mentionnées, désambiguïsées par DBpedia Spotligbht et pour lesquelles nous avons pu attribuer un genre s'élève à 390 867. Les mentions de personnes de genre masculin représentent 80,18% du nombre total de mentions, soit 313 386 occurrences (figure 5.1). En partant du postulat que la société est composée d'un nombre équilibré d'hommes et de femmes [1], les individus s'identifiant au genre féminin sont en sous-représentation (19,74% du nombre total de personnes mentionnées) face aux individus de genre masculin. Que ce soit en comptabilisant les occurrences uniques ou non des mentions de personnes, la présence d'hommes est environ quatre fois plus importante que celle des femmes. Ces chiffres renforcent l'existence du phénomène d'annihilation symbolique des femmes dans les médias, déjà révélé par Tuchman depuis 1979 [Tuchman, 1979].

La sous-représentation dans les données est encore plus marquante pour les identités de genres non binaires, bien que nous ne sachions pas objectivement quelles sont leurs proportions dans la société. Représentant moins de 0,09% des personnes nommées, nous notons une absence quasi-totale des personnes *genderqueer* (0,005%), non-binaires (0,03%), *genderfluid* (0,02%) et transgenres (femmes trans à 0,03% et hommes trans à 0,0005%) dans le corpus de presse.

Nous tenons à mentionner de nouveau que ces résultats sont à prendre avec du recul vu qu'ils dépendent de la performance des outils utilisés, qui n'ont peut-être pas bien repéré ni lié les personnes dans les articles.

Identité de genre	**Nombre de mentions de personnes genrées**	**Nombre de personnes uniques mentionnées**
MALE	313 386	24 747
FEMALE	77 147	6 200
TRANS WOMAN	130	21
NON-BINARY	109	18
GENDERFLUID	74	3
GENDERQUEER	19	3
TRANS MAN	2	2

TABLE 5.1 – Occurrences des mentions en fonction des genres

1. Hypothèse s'appuyant sur les données renseignées par la Banque mondiale (homme/femme), l'Insee (*Institut national de la statistique et des études économiques*) et l'Ined (*Institut national d'études démographiques*) pour les années 2021/2022 sur la distribution démographique (mondiale et française) en sexe. À notre connaissance, les rendus disponibles sur la répartition de la population en France sont calculés par rapport au sexe ou à l'orientation sexuelle, mais pas encore en fonction des identités de genre (ou seulement centrés sur certaines comme pour les personnes transgenres).

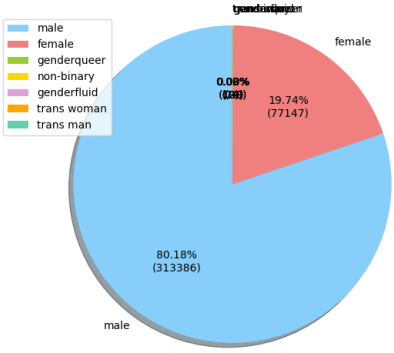

FIGURE 5.1 – Nombre total d'occurrences des personnes selon leur genre

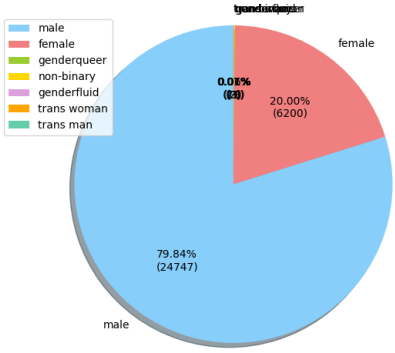

FIGURE 5.2 – Nombre de mentions uniques des personnes selon leur genre

Nous avons également calculé la proportion des genres à l'échelle d'un article (formule 5.1). Cette mesure correspond aux fréquences relatives de personnes genrées sur le nombre total de personnes du document. Pour calculer le taux moyen des mentions genrées par article à l'échelle du corpus entier (formule 5.2), nous avons effectué une simple moyenne des fréquences relatives. Par conséquent, le taux moyen pour l'ensemble du corpus ne repose pas sur le nombre total d'articles du corpus, mais le nombre d'articles contenant au moins une personne.

$$\text{Taux(G)}_{Article} = \frac{\text{Nombre de mentions de personnes de genre G}}{\text{Nombre total de personnes mentionnées dans le document}} \tag{5.1}$$

$$\text{TauxMoyen(G)}_{Corpus} = \frac{\text{Somme des Taux(G)}_{Article}}{\text{Nombre d'articles contenant au moins une personne}} \tag{5.2}$$

Concernant le *Taux(G)*$_{Article}$, chaque identité de genre obtient un taux maximal égal à 1, à l'exception d'un genre : *genderqueer*. Cela signifie que, sur l'ensemble de notre corpus, au moins un article mentionne uniquement des personnes du même genre sauf pour les *genderqueer*. Aucun article ne contient exclusivement des individus *genderqueer*. Avec un pourcentage maximal de 25% sur le nombre total de personnes mentionnées dans un article, les mentions de personnes *genderqueer* sont accompagnées, au mieux, de trois fois plus de personnes d'autres genres (et/ou de personnes dont l'identité de genre n'a pas pu être attribuée) dans un même document.

Le ratio moyen sur l'ensemble du corpus des mentions de personnes genrées dans un article (correspondant à *TauxMoyen(G)*$_{Corpus}$) suit la même tendance que la proportion des mentions de personnes genrées sur le corpus global, avec 81,32% des personnes sont des hommes, 18,29% des femmes et moins de 0,1% de personnes non binaires (figure 5.3)[2].

2. La somme des pourcentages n'est pas parfaitement égale à 1 car nous rappelons que les personnes sans genre associé ne sont pas prises en compte.

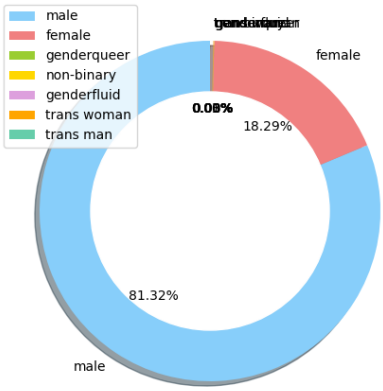

FIGURE 5.3 – Proportions moyennes des mentions de personnes genrées dans un article

5.1.2 Analyse qualitative

En plus d'entreprendre une analyse quantitative, nous abordons également l'étude des écarts de représentation sous une approche qualitative. Nous cherchons à déterminer quels sont les termes les plus employés et les plus spécifiques à un genre et si cela peut refléter des propriétés sociolinguistiques liées à une identité de genre. Pour cela, nous utilisons l'outil de textométrie TXM [Heiden et al., 2010] pour pouvoir identifier l'emploi caractéristique d'unités textuelles pour un genre plutôt qu'un autre grâce à leur indice de spécificité.

Chaque fichier au format TXT est associé à une identité de genre et contient sur chaque ligne uniquement le contexte local de l'entité genrée, après avoir soustrait l'entité elle-même du segment textuel car nous souhaitons analyser les termes accompagnant l'entité. La taille des contextes gauches et droits a été définie arbitrairement et fixée dans notre cas à 20 tokens, générant un contexte local de 40 tokens. Nous aurions pu considérer la phrase comme contexte de l'entité. Toutefois, cela signifiait que nous nous basions sur la segmentation en phrases réalisée par spaCy, qui n'est pas très fiable. Malheureusement, dans la plupart des cas, les composants pré-entraînés de spaCy ne semblent pas correctement gérer certaines ponctuations, tels que les guillemets français. De plus, les unités textuelles particulières comme les citations peuvent comporter des ponctuations fortes et, par conséquent, peuvent être mal segmentées. Par exemple, « *, avait lancé Nathalie Delon après le diagnostic de sa maladie.* » a été identifiée comme une phrase par spaCy. Or, il peut être intéressant de considérer les citations pour rendre compte de discriminations genrées [Richard et al., 2022].

Nous avons ensuite chargé ces fichiers dans un même corpus sous TXM et créé une partition en considérant chaque genre comme une sous-catégorie. Avec cette partition en genre, nous avons pu appliquer un calcul de spécificités qui fait ressortir les mots qui sont sur-représentés (ou sur-employés, avec un indice de spécificité positif) et ceux qui sont sous-représentés (ou sous-employés, avec un indice de spécificité négatif) dans une sous-catégorie considérée. L'analyse des contextes locaux sous TXM nous a permis d'observer plusieurs phénomènes.

Les individus féminins sont mentionnés avec d'autres personnes ou sont présentés relativement à une autre personne. Parmi les unités textuelles spécifiques au genre

féminin, il y a une prédominance de noms propres (« Anne », « Harry », « Meghan », « Markle », « Johnny »,...). Une grande partie de ces noms propres est liée à une famille et un événement particulier : les membres de la famille royale britannique, suite au décès de la reine Elisabeth II. Les femmes semblent être désignées en fonction de leur statut matrimonial ou familial : « fille », « mère », « couple », « maman » ou encore « mari ». Nous aurions pu inclure le mot « femme » dans la liste mais, compte tenu de son caractère polysémique, il peut aussi bien faire référence à un statut marital qu'un être humain de sexe ou de genre féminin, contrairement à sa déclinaison au pluriel qui a plus tendance à désigner le second sens, en considérant le fait que le corpus se place dans une société majoritairement monogame. La mention de personnes de genre féminin peut aussi être accompagnée d'un titre ou d'une profession (« actrice », « reine », « présidente », « chanteuse »,...). Le déterminant « la » et le pronom personnel « elle » sont particulièrement spécifiques au genre féminin, au même titre que « le » et « il » pour le genre masculin. Cela témoigne, comme attendu, du lien intrinsèque entre l'identité de genre et les genres grammaticaux dans une langue genrée, comme le français.

À propos du genre masculin, nous constatons une kyrielle de termes appartenant au champ lexical du sport, et plus précisément du football (« joueurs », « Mbappé », « entraîneur », « club », « match », « PSG », « attaquant », « Messi », « Neymar », « but », « OM »,...).

Concernant les « *trans woman* » (130 mentions de 21 personnes uniques), nous soulignons une mise en exergue des sujets autour de la transidentité et de la communauté LGBTQIA+ (« trans », « transidentité », « prides », « coming-out », formes dérivant du lemme « transgenre »,...). Il y a également un aspect temporel lié au parcours des personnes trans : « transition« , « réassignement » et l'usage de leurs morinoms ou *deadnames* [3] (comme « Bruce » pour Caitlyn Jenner).

5.2 Observations ciblées

Grâce aux métadonnées de chaque article, nous pouvons mettre en évidence différents volets de la discrimination genrée dans la presse. Nous avons pu constater précédemment l'impact de certains événements ou thèmes sur la représentation de genres.

5.2.1 Approche thématique

Nous souhaitons savoir si les représentations de genres varient en fonction des thèmes abordés par les articles de presse. Nous avons donc pris en compte les thèmes prédit par le transformer *flaubert-mlsum-topic-classification* et lorsque le *topic_score* est en dessous du seuil de 0,5, nous présumons que la classification en thème n'est pas vraiment motivée. Plusieurs justifications potentielles peuvent être mises en avant : résultat dû au hasard, thématique principale de l'article ne correspondant pas à une des classes prédéfinies,... Ces articles sont alors placés dans la classe "Indéfini". Les résultats suivants sont à remettre en contexte en considérant le fait que les classes ne sont pas équilibrées (tableau A.2).

Nous avons antérieurement observé un phénomène de sous-représentation sévère des genres non binaires. La figure 5.4 permet de faire un gros plan sur la distribution

3. Termes employés pour faire référence au prénom de naissance (ou le précédent nom) d'une personne qui a effectué un changement de prénom suite à sa transition de genre.

des identités de genre non binaires par rapport au thématique des articles. La majorité des personnes du corpus ne s'identifiant pas à un genre binaire est évoquée dans les articles de culture. Ce sont pour la plupart des célébrités évoluant dans le milieu artistique (chanteur·se, acteur·trice,...) comme Ezra Miller, Miley Cyrus ou encore Andréa Furet.

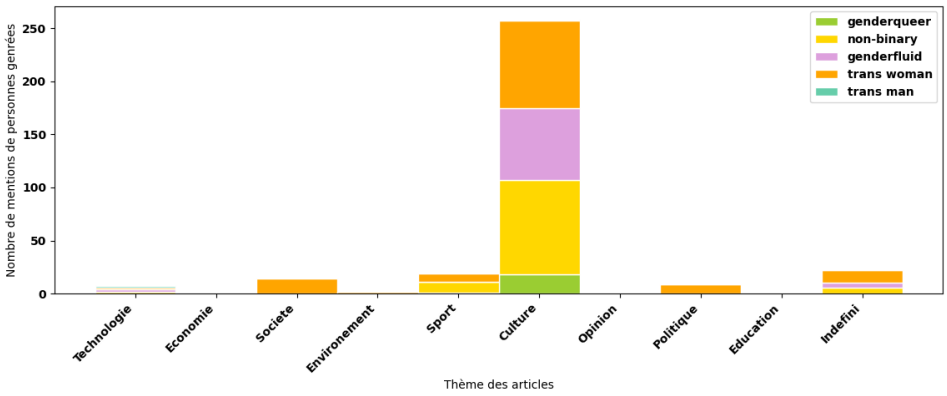

FIGURE 5.4 – Nombre de mentions de personnes genrées (*male* et *female* exclus) en fonction du thème principal des articles

Sur l'ensemble des thèmes, les personnes de genre féminin sont davantage mentionnées, de manière absolue et relative, dans les articles de culture (32 549 mentions qui représentent 30,31%), comme l'avait souligné [Richard et al., 2022]. Nous notons également que le taux de masculinité par article est le plus élevé dans les articles de sport, avec une moyenne arrondie à 0,90. La sur-représentation des hommes exacerbe *a contrario* la sous-représentation importante des six autres genres puisqu'ils représentent moins de 0,10 du ratio des personnes mentionnées dans un document traitant de sport. Par ailleurs, aucun individu *genderfluid* ni *trans man* n'est cité dans un article sportif. Des événements sportifs ponctuels tels que la Coupe du monde féminine de rugby 2021 [4], le Championnat d'Europe féminin de football 2022 ou la saison 2022 de tennis féminin (WTA Tour 2022) peuvent probablement réduire la discrimination dans la presse en faveur du genre féminin.

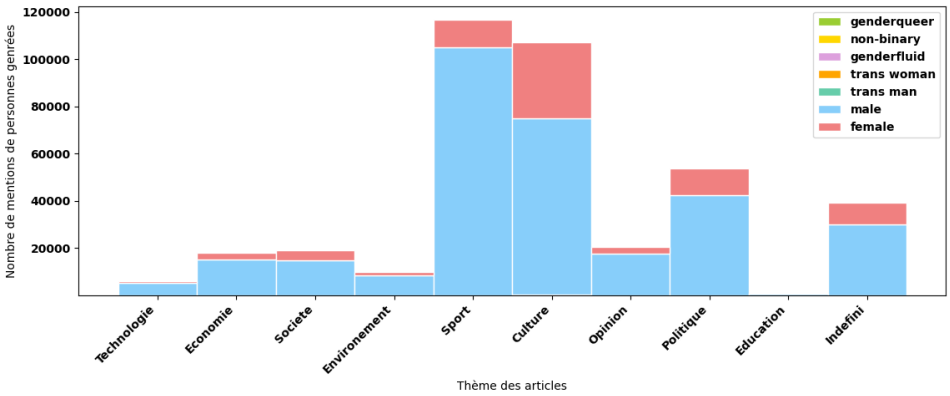

FIGURE 5.5 – Nombre de mentions de personnes genrées en fonction du thème principal des articles

4. Reportée à la fin de l'année 2022 pour cause de Covid

5.2.2 Approche diachronique

Étant donné le système utilisé pour la récupération d'articles, les mois sont relativement équilibrés en nombre d'articles sauf pour le dernier, vu que le corpus a été récupéré du 1 Janvier 2021 jusqu'au 15 Octobre 2022.

La figure 5.6 montre un rapport assez constant du taux de femmes par rapport aux hommes. Notre corpus s'inscrivant dans une période après la pandémie du Covid-19, nous constatons la faible présence des femmes dans la presse qui s'est renforcée suite à leur effacement de la scène médiatique engendré par la crise sanitaire [Calvez, 2020], avec une moyenne de 80,27% pour la gent masculine, 19,65% pour la gent féminine et 0,08% pour le regroupement des genres non normatifs.

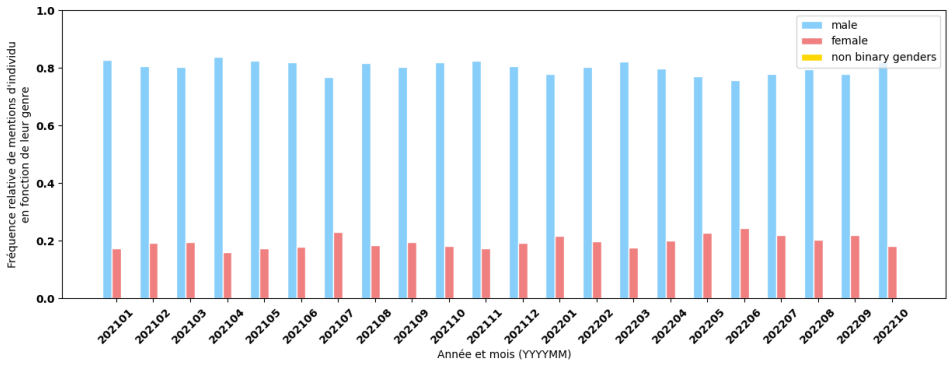

FIGURE 5.6 – Fréquence relative des personnes genrées en fonction du mois

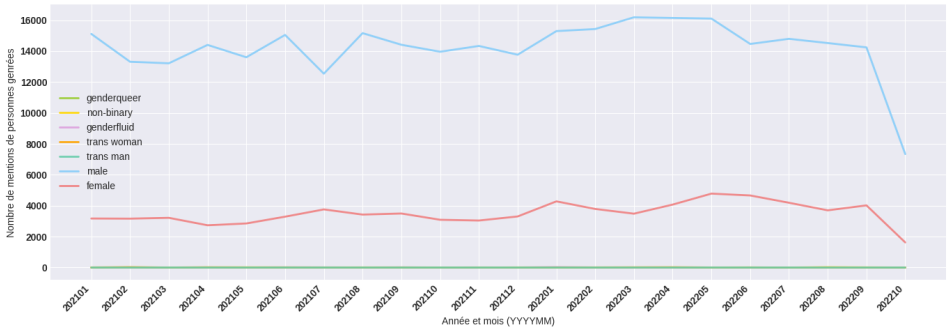

FIGURE 5.7 – Distribution des mentions de personnes genrées en fonction du mois

Nous corroborons avec les figures 5.6 et 5.7 notre précédente observation faisant état que les genres non binaires sont en sous-représentation écrasante et sont victimes d'une annihilation symbolique dans les médias. S'il y a bien une période de l'année où la société vise à mettre en avant de l'inclusion des personnes de la communauté LGBTQIA+ et faire accroître leur visibilité, ce serait pendant le mois des fiertés qui se déroule tous les ans, généralement en Juin. Cependant, ce n'est pas un phénomène qui semble se refléter dans la représentation des genres observée dans notre corpus.

Nous constatons quelques pics de mentions dans la figure 5.8 sur des mois particuliers. En revanche, d'après nos analyses qualitatives et quantitatives, nous ne

notons pas d'événements spécifiques pouvant être à l'origine de ces pics d'apparitions. Lorsque nous remettons en contexte, il y a en fait tellement peu de mentions de genres non binaires dans le corpus que ces augmentations accrues de mentions dans un mois ne sont pas dues à un événement mais simplement au fait d'avoir récupéré un article à un certain jour [5] contenant beaucoup de mentions de ces personnes. Lorsque nous comparons le nombre total de mentions et le nombre unique de personnes mentionnées à ces périodes-là, nous constatons en effet que ces pics correspondent à une seule personne. Pour illustrer ce propos, en Février 2022, nous comptons 23 occurrences de femmes trans dont 22 pointent vers Caitlyn Jenner uniquement dans un seul article le 18 Février 2022. En Janvier 2022 et plus précisément le 9 du mois, le même phénomène se reproduit pour Miley Cyrus avec 25 occurrences, traçant le pic rose. En remontant à l'origine de l'article, il est issu du magazine *Closer* accompagné de nombreuses photos dont les légendes mentionnent la star en question, ce qui explique en partie le nombre élevé de mentions dans un seul article.

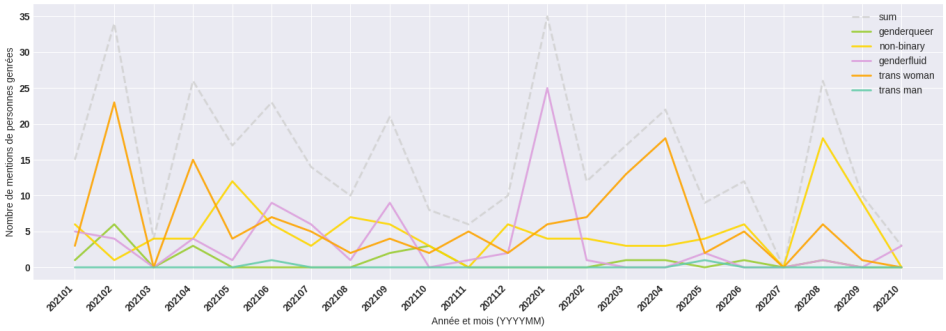

FIGURE 5.8 – Distribution des mentions de personnes non binaires en fonction du mois

Conclusion

Les analyses réalisées sur notre corpus s'inscrivent dans la continuité des travaux traitant de la même problématique de recherche.

L'exposition médiatique des individus masculins est écrasante et met sous silence la représentation des femmes dans les médias. Ajouté à cela, même si nous ne savons pas encore bien quelle est la proportion d'individus non binaires dans la société, les identités de genres non binaires appartiennent à une catégorie minoritaire victime d'un effacement de représentation dans la presse française encore plus important que celui observé pour le genre féminin. Les inégalités de représentation dans les médias sont affectées par de nombreux paramètres externes au genre (thèmes, critères sociodémographiques,...).

Les stéréotypes de genres sont également impactés par ces facteurs. La langue et ses propriétés sociolinguistiques jouent un rôle crucial dans l'interprétation des stéréotypes de genres.

5. Observable sur la figure A.3 en Annexe

CHAPITRE 6

DISCUSSIONS ET PERSPECTIVES

Sommaire

Introduction

Nous avons pu analyser la représentation à la fois quantitative et qualitative des écarts de genres sur un corpus de grande taille. Néanmoins, notre méthode n'étant pas infaillible, nous pouvons mettre en évidence des limites auxquelles elle s'est confrontée. Autant sur la méthode que sur les données utilisées, il y a plusieurs points qui pourraient être pris en considération concernant les observations obtenues sur la représentation des genres. En premier lieu, nous mettrons l'accent sur les performances insuffisantes de DBpedia Spotlight. Puis, nous évoquerons les défauts des bases de connaissances utilisées. Ensuite, nous nous questionnerons sur le point de départ de la sous-représentation de genres. Pour finir, nous suggérerons des perspectives d'améliorations pour l'approche adoptée dans le cadre de ce mémoire.

6.1 Performances de DBpedia Spotlight

6.1.1 Manque de robustesse

Pour réaliser l'*Entity Linking*, nous avons sélectionné DBpedia Spotlight pour plusieurs motifs mentionnés dans le chapitre 4. Considérant le faible nombre d'outils de désambiguïsation d'entités gratuits et en libre accès, nous avons négligé les performances de ce modèle qui s'avèrent être un obstacle conséquent pour notre étude de représentation des genres. DBpedia Spotlight n'est pas l'outil d'*Entity Linking* le plus robuste.

Le modèle actuel de DBpedia Spotlight découle de l'étude de [Daiber et al., 2013], fournissant une version améliorée du modèle initial reposant sur Lu-

cene [1] [Mendes et al., 2011]. D'après les résultats fournis par [Daiber et al., 2013] sur l'évaluation de DBpedia Spotlight dans plusieurs langues, les performances sur le français ne sont pas les meilleures. En français, nous obtenons une *f-mesure* de 45,02 pour la reconnaissance d'entités, résultant d'une précision et d'un rappel similaires, et une exactitude (*accuracy*) de 0,789 quant à la désambiguïsation des entités. Les performances dans les autres langues sont à peu près du même ordre, quelles que soient la taille et la langue du corpus en entrée.

Dans d'autres papiers, les performances de DBpedia Spotlight ne semblent pas énormément varier malgré la diversité des corpus utilisés. À notre connaissance, les *f-mesures* de la tâche globale d'*Entity Linking* en anglais avoisinent les 0,5 et ne dépassent pas le score de 0,68 [Kolitsas et al., 2018, Brasoveanu et al., 2020, Asgari-Bidhendi et al., 2021]. Nous n'avons pas pu trouver de scores ciblés sur la désambiguïsation de personnes pour le modèle actuel de DBpedia Spotlight. Pour évaluer les résultats du modèle sur notre corpus de presse et particulièrement sur les entités de type *personne*, il aurait fallu que nous procédions à une annotation manuelle des articles. Malheureusement, par manque de temps, cela n'a pas pu aboutir.

En revanche, nous avons examiné manuellement quelques sorties du modèle et avons observé de nombreuses erreurs ponctuelles de la part de DBpedia Spotlight, dont voici quelques exemples :

1. **Erreurs concernant la détection de mention de personnes**
   - Les entités nommées sous la forme de « prénom + nom », « titre de civilité + nom de famille », nom de scène, surnom, pronom, descriptions nominales, et autres se sont pas détectées.
   - Les noms complets de personnes sont partiellement reconnus.
2. **Erreurs concernant la désambiguïsation de personnes**
   - Les entités nommées partiellement reconnues ne sont pas liées à leur bonne entrée dans la base de connaissances.
   - Il y a des entités désambiguïsées qui pointent sur une entrée correspondant à une page de désambiguïsation.

6.1.2 Des traitements impactés par un biais de genre ?

Connaissant l'existence de biais de genre dans les traitements de TAL, nous nous questionnons sur leur éventuel impact sur l'*Entity Linking* de DBpedia Spotlight. Les conséquences de ces biais sont fréquemment illustrées par les analogies entre genres et professions qui sont encodées dans les *embeddings*. Par exemple, les modèles de langues ont tendance à considérer un médecin comme étant une personne de genre masculin et une nourrice de genre féminin. Par ces biais sont mis en avant les traitements stéréotypiques des genres binaires. Les genres non binaires sont, quant à eux, touchés différemment par cette problématique.

Le manque de représentation, qualitative et/ou quantitative, des genres non binaires ne permet pas une modélisation correcte de ces genres. D'après nos résultats, il y a apparemment peu d'entités nommées faisant référence à des personnes non binaires. De plus, leur identité de genre peut induire plusieurs particularités qui se traduisent quelquefois en caractéristiques textuelles : diverses appellations dues à des changements de nom, néopronoms, etc. Est-ce que ces termes se différencient assez des représentations binaires pour pouvoir être correctement encodés ? Ces der-

1. http://lucene.apache.org

niers ne sont hypothétiquement pas pris en considération dans l'encodage des données par les modèles. Le problème résiderait majoritairement dans la capacité ou non à encoder de l'information en relation avec ces genres.

```
Derrière            case         derrière            ADP
le                  det          le                  DET
pseudonyme          nsubj        pseudonyme          NOUN
de                  case         de                  ADP
Mykki               nmod         Mykki               PROPN     DBPEDIA_ENT http://fr.dbpedia.org/resource/Mykki_Blanco
Blanco              flat:name    Blanco              PROPN     DBPEDIA_ENT http://fr.dbpedia.org/resource/Mykki_Blanco
se                  expl:comp    se                  PRON
cache               ROOT         cache               VERB
l'                  det          le                  DET
artiste             obj          artiste             NOUN
pluridisciplinaire  amod         pluridisciplinaire  ADJ
Michael             obl:arg      Michael             PROPN
David               flat:name    David               PROPN
Quattlebaum         flat:name    Quattlebaum         PROPN
,                   punct        ,                   PUNCT
Jr                  nmod         jr                  NOUN
.                   punct        .                   PUNCT
,                   punct        ,                   PUNCT
né                  advcl        naître              VERB
le                  det          le                  DET
2                   obl:mod      2                   NUM
avril               nmod         avril               NOUN
1986                nmod         1986                NUM
```

FIGURE 6.1 – Exemple d'une annotation spaCy et DBpedia Spotlight pour l'artiste transgenre Mykki Blanco

Dans l'hypothèse où la liaison d'entités serait effectivement influencée par des biais de genre, le passage sous silence des personnes féminines et non binaires peut avoir été exacerbé. La réelle représentation de genres dans la presse est donc potentiellement différente des scores obtenus dans ce mémoire, qui seraient en défaveur des genres autres que masculin.

6.2 Difficultés liées à la base de connaissance

6.2.1 Entités inexistantes dans DBpedia

Une étude de la représentation des genres reposant sur la sortie d'un modèle d'Entity Linking ne dépend pas uniquement de la performance de ce dernier mais aussi de la base de connaissance utilisée.

La base de connaissance recense uniquement des personnalités appartenant à la sphère publique, mais pas toutes. Nous pouvons utiliser conjointement d'autres bases de connaissances définies sur des données plus spécifiques (par exemple, une base de connaissances de journalistes). Par contre, pour les instances occasionnelles d'individus anonymes ou inconnus du grand public dans les articles de presse, il est fortement probable que ces personnes n'aient pas d'entrées correspondantes dans aucune base de connaissance.

Pour les entités absentes de la base de connaissance, nous n'avons aucun moyen d'affirmer avec certitude leur identité de genre, mis à part si cette information est renseignée dans l'article en lui-même. Nous perdons des renseignements vitaux pour l'étude entreprise dans ce cadre de recherche.

Pour surmonter ces problèmes, des modèles sont actuellement développés pour générer et/ou alimenter automatiquement des graphes de connaissances à partir de données textuelles non structurées.

6.2.2 Absence d'informations pour une entrée

Un nombre assez conséquent d'entités désambiguïsées par DBpedia Spotlight ne possèdent pas d'informations complètes, notamment sur le type de l'entité. Pour rappel, dans l'objectif de gagner du temps de calcul, nous attribuons un genre exclusivement aux entités qui sont typés comme *personne* dans DBpedia FR, et plus précisément dans la réponse spaCy de *spacy-dbpedia-spotlight*. Or, il se peut que le champ `@types` du `span` spaCy, associé à l'entité désambiguïsée, soit vide, bien que la page en ligne liée à la ressource DBpedia fournit bien les informations recherchées.

De nombreuses personnes ne sont donc pas étiquetées en genre. De plus, il serait intéressant que les entités DBpedia dispose d'une propriété de genre, pour éviter de faire des allers-retours entre une multitude de bases de connaissance, à savoir Wikidata dans notre cas.

Nous avons également observé chez les entités désambiguïsées de type *personne* auxquelles nous n'avons pas pu conférer une identité de genre que Wikidata ne fournit pas systématiquement une information de genre à toutes les personnes de la base. Notre méthode doit donc être prête à s'adapter aux mises à jour de la base de connaissance.

6.3 Sous-représentation des genres au-delà de la presse

Comme l'analyse quantitative l'a souligné, comparé au genre masculin, tous les autres genres sont en sous-représentation, surtout les genres non binaires (ni masculin, ni féminin). Ces inégalités sont observables non seulement dans notre corpus de presse, mais aussi dans les bases de connaissances.

Identité de genre	**Nombre de personnes recensées**
MALE	142 498
FEMALE	42 725
TRANS WOMAN	34
NON-BINARY	13
TRANS MAN	13
GENDERQUEER	3
GENDERFLUID	1
AGENDER	1

TABLE 6.1 – Effectif des êtres humains de nationalité française nés après le 1e Janvier 1900 recensées dans Wikidata

Le tableau 6.1 montre toutes les étiquettes de la propriété « *sex or gender* » (correspondant bien à un genre) et leurs effectifs dans Wikidata après avoir appliqué des filtres sur la date de naissance et la nationalité des personnes, pour ne pas atteindre le temps limite de la requête. En prenant en compte le fait que la requête exclut les personnalités internationales, la représentation des genres dans Wikidata est également déséquilibrée, le genre masculin étant toujours en sur-représentation. Parmi les genres référencés dans Wikidata, certains sont même des sous-classes de « minorité de genre » (*gender minority*).

6.4 Perspectives d'améliorations

Remaniement du corpus

Il serait intéressant de voir si faire varier la taille du corpus change les résultats constatés. Est-ce que nous obtenons des écarts de représentation similaires avec un corpus beaucoup plus petit ? Est-ce qu'un nombre plus important d'articles reflète une sur-représentation moins forte des hommes ?

De plus, notre corpus pourrait bénéficier d'un nettoyage et d'une normalisation plus approfondis. Selon le point de vue adopté, certaines portions textuelles, respectant un format stylistique particulier, peuvent engendrer du bruit : les légendes de photos ou vidéos accompagnant l'article, les informations personnelles du journaliste qui a rédigé le papier, etc. Comment et quand est-ce que nous considérons ces unités textuelles comme pertinentes pour l'analyse ? Cette problématique pose des difficultés supplémentaires sur la détection des morceaux de textes à exclure de l'analyse.

Dans un autre cas, quelques articles peuvent également avoir un contenu identique à la lettre près, ce qui implique un biais dans l'analyse objective des représentations de genre dans la presse.

Il semble également y avoir des fautes d'étiquetage qui dépendent du contenu même des données. Dans notre jeu de données, nous avons pu constater sur certains documents qu'il y avait des fautes de frappe provenant de l'éditeur (absence d'une espace entre deux mots,...), ce qui a pu introduire des erreurs lors de la tokénisation et les traitements ultérieurs, en particulier la reconnaissance d'entités.

Marge de progrès sur l'Entity Linking

Au vu des performances de DBpedia Spotlight, il serait intéressant d'utiliser un autre modèle d'*Entity Linking*, que ce soit un modèle entraîné par nos soins ou un modèle existant (Babelfy, mGENRE [De Cao et al., 2021],...), dans l'objectif d'améliorer les résultats et obtenir une représentation de genre plus fiable.

Nous pouvons *fine-tuner* des modèles d'Entity Linking ou de désambiguïsation d'entités sur un jeu de données de presse française, annoté en conséquence. Le manque d'accès à ce genre de corpus préétablis nous amène à envisager une campagne d'annotation.

Ajout de traitements pour les coréférences et les citations

L'*Entity Linking* effectue de la désambiguïsation d'entités nommées ou de syntagmes susceptibles de correspondre à une entrée dans une base de connaissance. De ce fait, elle écarte une partie des coréférences, particulièrement les pronoms, de la représentation des genres. Une résolution de coréférences, préférablement sans biais de genre (comme énoncée dans le chapitre 1), permettrait à la fois de prendre en compte toutes ces coréférences non traités par l'*Entity Linking* et, si souhaité, d'assigner en une seule fois une identité de genre à une liste de coréférences associée à une seule et unique personne.

Ajouter une extraction de citations aux traitements serait également pertinent, sachant que les citations sont considérées comme étant des indicateurs forts d'accès à la parole dans les médias. Cela permettrait d'aborder la représentation des genres du point de vue des citations :

— Est-ce que la presse rapporte plus les propos de personnes masculines que d'autres genres ?
— Selon les identités de genre, y a-t-il des différences dans la reprise du propos ?
— La manière d'introduire le locuteur d'une citation (verbes de parole, prépositions,...) suit-elle des stéréotypes de genres ?
— Pour un thème en particulier, la presse cite-t-elle davantage les propos d'une certaine catégorie de genre ?
— Les médias privilégient-ils des individus d'un genre particulier pour les paroles d'experts ?
— etc.

Les travaux de [Richard et al., 2022] proposent quelques éléments de réponses à ces questions. Afin d'améliorer leur mesure calculée avec un modèle à base de règles, ces chercheurs travaillent sur l'élaboration de modèles de réseaux de neurones pour l'extraction de citations.

Optimisation de l'attribution de genre

Il serait intéressant dans de futurs travaux de reprendre plus en détail la partie sur l'attribution d'un genre à une entité. En effet, il y a plusieurs circonstances qui n'ont pas été prises en compte dans le système actuel d'attribution de genre.

Il paraît évident que l'attribution de genre ne peut se faire que sur les entités faisant référence à une seule et unique personne, et non à un groupe ou un duo. Par exemple, l'entité nommée « Anthony Russo » a été rattaché à l'entrée des frères Russo. Une identité de genre n'a pas pu lui être attribuée. Néanmoins, nous tenons à souligner le fait qu'il n'existe pas encore d'entrée dans DBpedia FR pour Anthony Russo seul.

Il existe aussi des entités Wikidata pour lesquelles la propriété « *sex or gender* » contient plusieurs valeurs. Néanmoins, le client Wikidata utilisé dans les scripts ne renvoie pour le moment qu'une seule de ces valeurs. Une réflexion additionnelle s'ajoute donc à la problématique de recherche existante : comment gérer les entités qui présentent cette singularité ?

Conclusion

Bien que la démarche proposée parte de la volonté de respecter les individus et leur identité, nous rencontrons plusieurs limites à la fois théoriques et pratiques. Leur résolution peut dépasser notre portée, comme pour le cas de la surreprésentation des hommes dans les ressources utilisées. La complexité de la problématique suscite de nombreux éléments de discussions.

Cependant, nous en tirons également une multitude de pistes d'améliorations envisageables pour obtenir des résultats plus fidèles à la représentation des genres dans les médias. Nous pourrons alors multiplier les analyses et explorer cette question sous différentes approches. Quel est le rôle de la presse vis-à-vis des inégalités de genres et de leur propagation ? Quelles en sont les conséquences ? À quel degré sont-elles nuisibles ? En termes d'inégalités de représentation, pouvons-nous établir des liens entre le genre et d'autres critères sociaux ? Des nouvelles réponses pourraient être apportées dans de futurs travaux.

Conclusion générale

Dans ce mémoire, nous avons entrepris une analyse de la représentation des genres dans un corpus de presse, élaboré à partir des ressources proposées par Aday, en adoptant une démarche fondée sur du Traitement Automatique des Langues. Le traitement de cette question a fait remonter des problématiques à la fois d'un point du vue sociologique et aussi en TAL.

Les précédentes études sur la représentation des genres dans les médias se sont davantage tournées vers la question de l'évolution de la place des femmes, en observant une sous-représentation perpétuelle des femmes sur ces plateformes. La définition du genre et les travaux sur ce sujet sont complexes, à l'image de l'évolution du spectre du genre. Le genre (ou l'identité de genre, dans notre cas) est encore très majoritairement étudié comme étant un concept binaire, notamment dans les travaux de TAL. Habituellement, les différentes identités de genres considérées sont « masculin », « féminin », et dans certains cas une classe « mixte » ou « indéterminé » vient s'ajouter aux genres binaires. Cela tend à invalider ou à effacer l'existence des genres non binaires. C'est pourquoi ce travail de recherche a pour volonté d'inclure les personnes ne s'identifiant pas à un genre binaire et de mettre en lumière la diversité des genres.

Les travaux sur la représentation en genre se basent généralement sur une prédiction de genre. Ici, l'utilisation de l'*Entity Linking* est une approche inédite par rapport à ce qui a pu être proposé auparavant pour la résolution de cette problématique. Cette démarche présente des avantages, comme l'inclusion des genres non binaires et la réduction du mégenrage des personnes impliquées. Néanmoins, la difficulté de la problématique a révélé des défauts dans l'approche adoptée, particulièrement sur le choix du modèle de désambiguïsation. Malgré tout, au vu des gros volumes de données que nous avons collectés, un gain en temps considérable a été permis avec l'emploi du TAL, au prix de la fiabilité.

Les mesures quantitatives, facilement compréhensibles par le grand public, et les facteurs qualitatifs, marquant des phénomènes sociolinguistiques, proposent une analyse sous plusieurs angles. Nous constatons que la représentation genrée observable dans la presse ne reflète pas la distribution réelle des genres dans la société. Un déséquilibre est frappant, notamment avec une sur-représentation écrasante du genre masculin et un effacement quasi-total des genres non binaires. Une représentation stéréotypée des femmes est également observée à travers les médias.

Ce travail de recherche a permis de souligner l'utilité des méthodes de TAL dans le traitement de sujets plus vastes, comme l'étude des genres, et de nouvelles données, comme l'émergence des représentations non binaires. Nous mettons également en exergue une multitude de points d'améliorations, ce qui laisse une large marge de manœuvre pour de futurs travaux souhaitant approfondir cette approche. Il serait intéressant de continuer à mobiliser et interroger les outils de TAL pour traiter des problématiques de société.

BIBLIOGRAPHIE

[Adjali et al., 2020] Adjali, O., Besançon, R., Ferret, O., Borgne, H. L., and Grau, B. (2020). Multimodal Entity Linking for Tweets. volume 12035, pages 463–478. arXiv :2104.03236 [cs]. – Cité page 21.

[Asgari-Bidhendi et al., 2021] Asgari-Bidhendi, M., Janfada, B., Havangi, A., Hossayni, S. A., and Minaei-Bidgoli, B. (2021). An Unsupervised Language-Independent Entity Disambiguation Method and its Evaluation on the English and Persian Languages. arXiv :2102.00395 [cs]. – Cité page 46.

[Asr et al., 2021] Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., and Taboada, M. (2021). The Gender Gap Tracker : Using Natural Language Processing to measure gender bias in media. *PLOS ONE*, 16(1) :e0245533. Publisher : Public Library of Science. – Cité page 15.

[Bahadurdesai, 2018] Bahadurdesai, B. R. (2018). Karnataka's Jogappas can now live a gender-fluid life. *The Hindu*. – Cité page 14.

[Brasoveanu et al., 2020] Brasoveanu, A. M., Weichselbraun, A., and Nixon, L. (2020). In Media Res : A Corpus for Evaluating Named Entity Linking with Creative Works. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 355–364, Online. Association for Computational Linguistics. – Cité page 46.

[Bérubé et al., 2020] Bérubé, N., Ghiasi, G., Sainte-Marie, M., and Larivière, V. (2020). Wiki-Gendersort : Automatic gender detection using first names in Wikipedia. – Cité page 15.

[Calvez, 2020] Calvez, C. (2020). Rapport sur la place des femmes dans les médias en temps de crise. – Cité pages 14 et 43.

[Cao and Daumé, 2021] Cao, Y. T. and Daumé, H. (2021). Toward Gender-Inclusive Coreference Resolution : An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics*, 47(3) :615–661. – Cité page 16.

[Cao and Daumé III, 2020] Cao, Y. T. and Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics. – Cité page 16.

[Daiber et al., 2013] Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems - I-SEMANTICS '13*, page 121, Graz, Austria. ACM Press. – Cité pages 21, 34, 35, 45 et 46.

[De Cao et al., 2021] De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., and Petroni, F. (2021). Multilingual Autoregressive Entity Linking. arXiv :2103.12528 [cs, stat]. – Cité page 49.

[Dev et al., 2021] Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., and Chang, K.-W. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. arXiv :2108.12084 [cs]. – Cité page 15.

[Devinney et al., 2022] Devinney, H., Björklund, J., and Björklund, H. (2022). Theories of "Gender" in NLP Bias Research. arXiv :2205.02526 [cs] version : 1. – Cité page 15.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. – Cité page 15.

[Diomedi and Hogan, 2021] Diomedi, D. and Hogan, A. (2021). Question Answering over Knowledge Graphs with Neural Machine Translation and Entity Linking. arXiv :2107.02865 [cs]. – Cité page 18.

[Ehrmann, 2008] Ehrmann, M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Theses, Paris Diderot University. – Cité pages 19 et 21.

[Havens et al., 2022] Havens, L., Alex, B., Bach, B., and Terras, M. (2022). Uncertainty and Inclusivity in Gender Bias Annotation : An Annotation Taxonomy and Annotated Datasets of British English Text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, Seattle, Washington. Association for Computational Linguistics. – Cité page 16.

[Heiden et al., 2010] Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. volume 2, page 1021. Edizioni Universitarie di Lettere Economia Diritto. Issue : 3. – Cité page 40.

[Kaneko and Bollegala, 2022] Kaneko, M. and Bollegala, D. (2022). Unmasking the Mask – Evaluating Social Biases in Masked Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11) :11954–11962. Number : 11. – Cité page 15.

[Kolitsas et al., 2018] Kolitsas, N., Ganea, O.-E., and Hofmann, T. (2018). End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics. – Cité pages 21 et 46.

[Lan et al., 2020] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. arXiv :1909.11942 [cs]. – Cité page 15.

[Lauscher et al., 2022] Lauscher, A., Crowley, A., and Hovy, D. (2022). Welcome to the Modern World of Pronouns : Identity-Inclusive Natural Language Processing beyond Gender. arXiv :2202.11923 [cs]. – Cité page 14.

[Le et al., 2020] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT : des

modèles de langue contextualisés pré-entraînés pour le français. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278, Nancy, France. ATALA. – Cité page 31.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv :1907.11692 [cs]. – Cité page 15.

[Martins et al., 2019] Martins, P. H., Marinho, Z., and Martins, A. F. T. (2019). Joint Learning of Named Entity Recognition and Entity Linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics. – Cité page 21.

[Mendes et al., 2011] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia spotlight : shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*, pages 1–8, Graz, Austria. ACM Press. – Cité page 46.

[Moro et al., 2014] Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation : a Unified Approach. *Transactions of the Association for Computational Linguistics*, 2 :231–244. – Cité page 21.

[Mulang' et al., 2020] Mulang', I. O., Singh, K., Prabhu, C., Nadgeri, A., Hoffart, J., and Lehmann, J. (2020). Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models. arXiv :2008.05190 [cs]. – Cité page 20.

[Richard et al., 2022] Richard, A., Bastin, G., and Portet, F. (2022). GenderedNews : Une approche computationnelle des \'ecarts de repr\'esentation des genres dans la presse fran\c{c}aise. arXiv :2202.05682 [cs]. – Cité pages 14, 15, 30, 40, 42 et 50.

[Rudinger et al., 2018] Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics. – Cité page 15.

[Scialom et al., 2020] Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020). MLSUM : The Multilingual Summarization Corpus. arXiv :2004.14900 [cs]. – Cité page 30.

[Sebo, 2021] Sebo, P. (2021). Performance of gender detection tools : a comparative study of name-to-gender inference services. *Journal of the Medical Library Association*, 109(3). – Cité page 15.

[Sowinski et al., 2022] Sowinski, P., Wasielewska-Michniewska, K., Ganzha, M., and Paprzycki, M. (2022). Topical Classification of Food Safety Publications with a Knowledge Base. arXiv :2201.00374 [cs]. – Cité page 20.

[Stoller, 1968] Stoller, R. J. (1968). A further contribution to the study of gender identity. *The International Journal of Psycho-Analysis*, 49(2) :364–369. – Cité page 12.

[Sun et al., 2019] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing : Literature Review. arXiv :1906.08976 [cs]. – Cité page 15.

[Suárez et al., 2020] Suárez, P. J. O., Dupont, Y., Muller, B., Romary, L., and Sagot, B. (2020). Establishing a New State-of-the-Art for French Named Entity Recognition. – Cité page 19.

[Tedeschi et al., 2021] Tedeschi, S., Conia, S., Cecconi, F., and Navigli, R. (2021). Named Entity Recognition for Entity Linking : What Works and What's Next. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics. – Cité page 19.

[Tuchman, 1979] Tuchman, G. (1979). Women's Depiction by the Mass Media. *Signs*, 4(3) :528–542. Publisher : University of Chicago Press. – Cité pages 14 et 38.

[Zheng et al., 2015] Zheng, J. G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., and Ji, H. (2015). Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(1) :S4. – Cité page 18.

ANNEXE A

ANNEXE

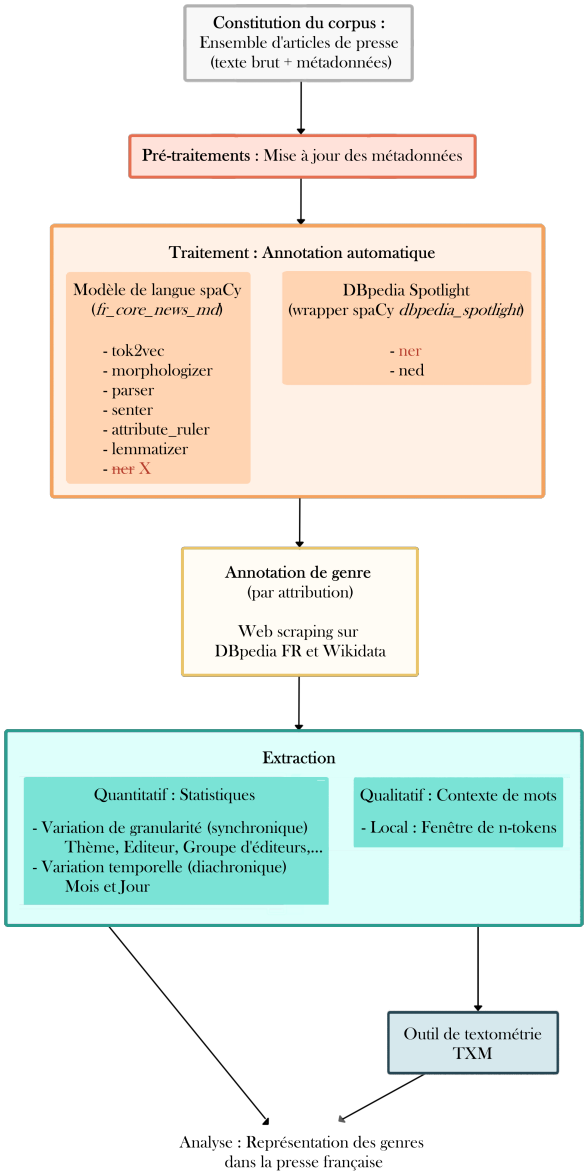

FIGURE A.1 – Schéma du processus général de la méthode adoptée dans ce travail pour l'étude de représentation des genres

Code du groupe	**Description**	**Exemples d'éditeurs**
PRESSEPQNACTUGENE	Presse quotidienne nationale : Actualité générale	*Le Monde*, *20 Minutes*, *La Croix*, *Le Figaro*, *Aujourd'hui en France*, *L'Humanité*,...
PRESSEPQNECO	Presse quotidienne nationale : Économie	*Les Echos*, *La Tribune*, *Aujourd'hui en France Economie*, *jdheditions.fr*,...
PRESSEPQNSPORT	Presse quotidienne nationale : Sport	*L'Equipe*, *Midi Olympique Rouge*, *Midi Olympique Vert*, *Paris-Turf*...
PRESSEMAGACTUGENE	Presse magazine : Actualité générale	*Marianne*, *Le Monde - Magazine*, *Okapi*, *Le Journal du Dimanche*,...
PRESSEMAGAUTO	Presse magazine : Automobile	*Moto Magazine*, *L'Auto-Journal*, *Gazoline*,...
PRESSEMAGCULTURE	Presse magazine : Culture - Loisirs	*Music Story*, *tetu.com*, *Connaissance de la Chasse*, *joursdefrance.lefigaro.fr*,...
PRESSEMAGECO	Presse magazine : Économie	*Alternatives Economiques*, *Le Parisien Economie*, *Journal des Entreprises Le Brief*,...
PRESSEMAGEDUC	Presse magazine : Éducation	*Parenthèse*, *X Passion*, *Le Journal des Grandes Ecoles*,...
PRESSEMAGFEM	Presse magazine : Féminin	*ELLE*, *Femme Actuelle*, *Marie Claire*,...
PRESSEMAGFINANCE	Presse magazine : Finances	*Investir Hebdo*, *Option Finance - Hebdo*, *Mieux Vivre Votre Argent*,...
PRESSEMAGINTERNATIONAL	Presse magazine : International	*Courrier International*, *Courrier de l'Atlas*, *Afrique Magazine*,...
PRESSEMAGNTIC	Presse magazine : Nouvelles technologies	*Micro Pratique*, *Modèle Magazine*,...

Suite sur la prochaine page

Code du groupe	**Description**	**Exemples d'éditeurs**
PRESSEMAGPEOPLE	Presse magazine : People	*Gala*, *Voici*, *Public*, *Closer*,...
PRESSEMAGSANTE	Presse magazine : Santé - Médecine	*Top Santé*, *Egora.fr*, *Vivre Bio*, *Psychologies*,...
PRESSEMAGSPORT	Presse magazine : Sport	*L'Equipe Mag*, *L'Officiel du Cycle*, *Sport Foot Magazine*,...
AFP	Agence France Presse	*AFP Fil Général*, *AFP Multimédia*, *AFP Fil Economique*, *AFP Fil International*,...
PRESSEAGENCES	Autres agences de presse	*Reuters Fil Général (français)*,...
PURE	Médias en ligne : Pure Players	*mediapart.fr*, *zonebourse.com*, *melty.fr*, *footmercato.net*,...
SITESPQN	Médias en ligne : Sites de presse quotidienne nationale	*lesechos.fr*, *lefigaro.fr*, *20minutes.fr*, *lemonde.fr*,...
SITESMAG	Médias en ligne : Sites de magazines	*voici.fr*, *courrierinternational.com*, *lepoint.fr*,...

TABLE A.1 – Liste des groupes d'éditeurs sélectionnés dans le corpus

Thème	**Nombre d'articles associés**	
	Contenant des personnes	***Total***
CULTURE	14 737	19 113
SPORT	14 443	16 531
ECONOMIE	5 964	14 901
INDEFINI	7 197	11 981
SOCIETE	4 698	7 922
TECHNOLOGIE	2 014	6 397
ENVIRONNEMENT	3 085	6 252
POLITIQUE	4 504	4 599
OPINION	3 182	4 365
EDUCATION	257	494

TABLE A.2 – Nombre de documents en fonction des thèmes

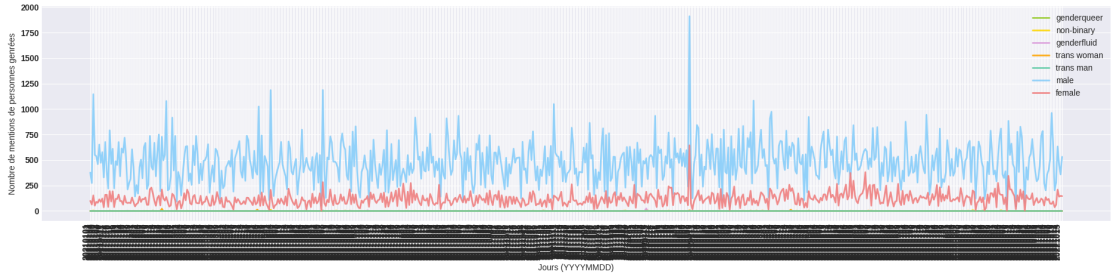

FIGURE A.2 – Effectifs des genres en fonction des jours de parution des articles

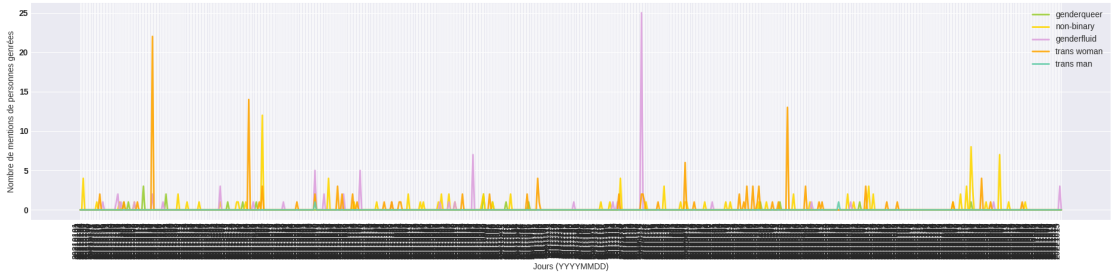

FIGURE A.3 – Effectifs des genres non binaires en fonction des jours de parution des articles

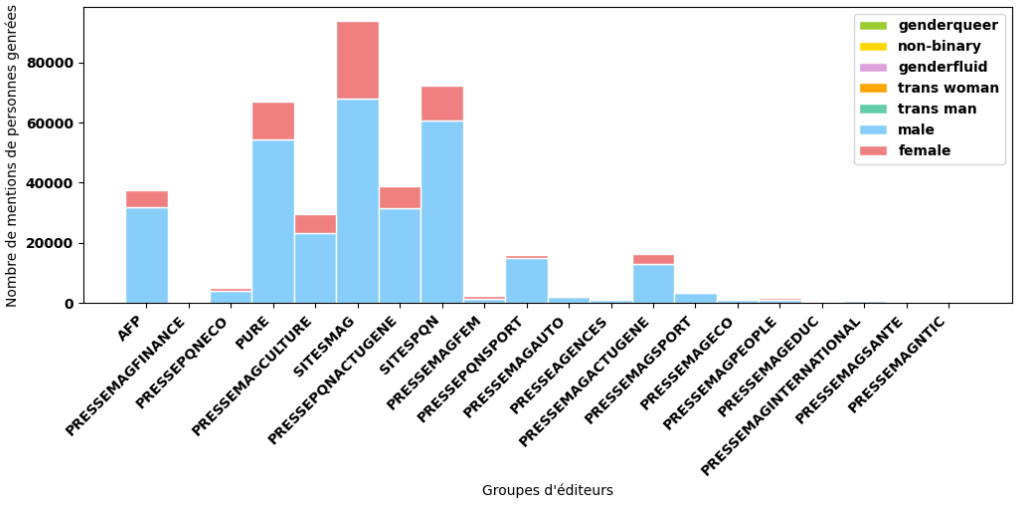

FIGURE A.4 – Effectifs des mentions de personnes genrées en fonction des groupes d'éditeurs

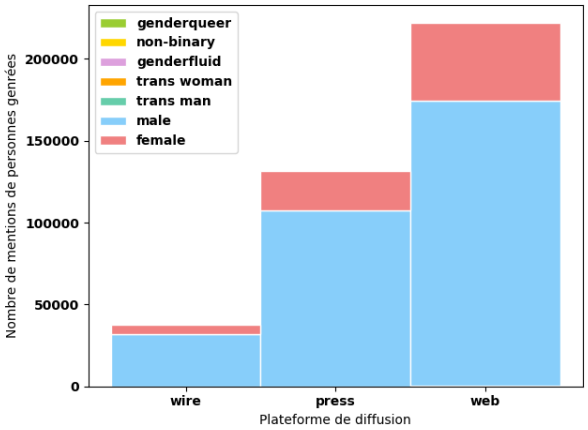

FIGURE A.5 – Effectifs des mentions de personnes genrées en fonction des plateformes de diffusion

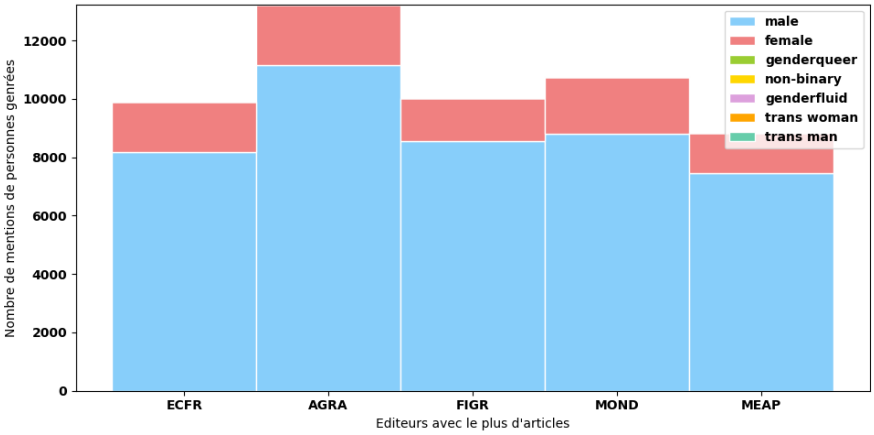

FIGURE A.6 – Effectifs des mentions de personnes genrées pour les 5 éditeurs ayant le plus d'articles dans le corpus