

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

« I'M GAY, BUT GG » - TOXIQUE, BIAISÉ, OU LES DEUX ? : DESCRIPTION DES BIAIS IDENTITAIRES
DES MODÈLES DE DÉTECTION AUTOMATIQUE DE TOXICITÉ DANS LES CLAVARDAGES DE JEUX
VIDÉO

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

MAITRISE EN LINGUISTIQUE

PAR

JOSIANE VAN DORPE

JANVIER 2025

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je dis souvent que si la théorie du Karma est vraie, alors j'ai dû sauver un pays entier de la destruction sans même le savoir. Il m'est difficile de voir d'autres raisons pour lesquelles j'aurais la chance inouïe d'être entourée de gens tous plus incroyables les uns que les autres, qui m'ont donné des outils, m'ont permis de grandir, d'apprendre, de me tromper, et m'ont donné des opportunités qui ont changé ma vie. J'espère évidemment avoir un impact positif sur leur vie également, et je continue de faire de mon mieux pour leur rendre la pareille autant que possible.

Je tiens tout d'abord à exprimer ma reconnaissance infinie à mon directeur, Grégoire Winterstein. Son support, ses conseils et sa confiance en moi dans les moments plus difficiles m'ont permis de mener ce projet à terme. Surtout, grâce à lui, j'ai découvert le monde passionnant de la linguistique informatique, intégré un laboratoire de recherche, publié des articles à l'international et effectué un stage idéal qui m'a d'ailleurs permis d'obtenir un emploi que je ne m'imaginai pas même dans mes rêves les plus fous.

Merci, merci à mon fiancé, Frédérick, qui a été un support moral essentiel à mon retour aux études, et ce depuis le tout début de mon baccalauréat. Sans sa patience, son amour, son écoute et son calme, je ne me serais pas rendue à cette étape. J'espère pouvoir être un support aussi solide pour son propre retour aux études.

Merci aux membres de ma famille qui m'ont soutenue tout au long de l'écriture : ma mère, la meilleure maman du monde, une personne essentielle à ma réussite qui m'a supportée dans les (très) hauts et les (très) bas, et ce, à toute heure du jour et de la nuit ; mon père, un fort soutien calmant et rassurant dans des moments de grand stress ; ma sœur, qui, malgré sa vie bien remplie de maman extraordinaire, m'a encouragée, a toujours été à mon écoute, et a toujours apporté son aide. Merci également à mes tantes et ma cousine pour leur soutien.

Merci à mes ami.es pour leur patience et leurs encouragements, et d'avoir accepté ma version « hologramme » aux nombreuses soirées manquées pour les études : Sab, Raph, Chloé, Auré, Luke, Wyatt, Jeep, Samanta, Deeps, Kiki. Merci à tout le SLIC et la SALLE pour de beaux moments à discuter de linguistique et pour tout le gossip. Un merci spécial à Samuel, qui a réussi à me motiver et à me rassurer à chaque conversation.

Merci aux étudiants et employés de La Forge-Ubisoft et PIXEL, notamment Doriane, Zachary, Andrea et Amanda. Dans un monde de sciences informatiques et de génie logiciel où je me sentais peu à ma place, leur support et leurs conseils m'ont été indispensables pour mener ma recherche à terme. Merci à mon superviseur, Nicolas Grenon-Godbout, qui en plus de m'avoir guidée pendant mon stage, m'a creusé une place au sein de son équipe de scientifiques de données. Je remercie d'ailleurs tous les membres de cette équipe incroyable, Gabriel, Bettina, Simon et Jean-Michel, qui n'ont jamais hésité à m'apporter leur aide et qui m'ont permis de découvrir et de comprendre des outils et des méthodes qui m'ont servi pour ce mémoire.

Finalement, je remercie les organismes qui m'ont offert un soutien financier essentiel pour passer au travers de mes études à la maîtrise sans avoir à me soucier de manquer de sous pour payer mes sessions et survivre : merci au CRSH, au FRQSC, à Mitacs Accélération, ainsi qu'au CRLEC pour ces bourses.

TABLE DES MATIÈRES

REMERCIEMENTS	ii
LISTE DES FIGURES.....	vi
LISTE DES TABLEAUX	viii
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES.....	ix
RÉSUMÉ.....	x
ABSTRACT	xi
INTRODUCTION	1
CHAPITRE 1 LES MODÈLES DE LANGUE ET ÉMERGENCES LINGUISTIQUES.....	3
1.1 Entraînement et utilisation des modèles.....	3
1.2 Représentation de la langue	5
1.3 Information linguistique contenue dans les plongements	7
1.3.1 Distances entre les plongements et couches cachées.....	7
1.3.2 Plus que le contexte : la <i>story</i>	9
1.3.3 Contextes dynamiques et la substitution lexicale	10
1.3.4 Plongements et leurs <i>word stories</i>	11
1.4 Conclusion.....	14
CHAPITRE 2 TOXICITÉ : DÉFINITION ET DÉTECTION	15
2.1 Définir la toxicité.....	15
2.2 En ligne et dans les jeux vidéo	15
2.3 Méthodes de détection de toxicité.....	17
2.4 Information linguistique : toxicité dans les plongements.....	18
2.5 Limitations.....	19
2.6 Conclusion.....	21
CHAPITRE 3 BIAIS : MODÈLES DE LANGUE DÉTECTION DE TOXICITÉ.....	22
3.1 Définir les biais.....	22
3.2 Information linguistique : biais dans les plongements	23
3.3 Biais en détection de toxicité.....	28
3.4 Atténuation de biais et difficultés associées.....	29
3.4.1 Techniques utilisées.....	29
3.4.2 Problèmes rencontrés.....	31

3.4.3 Neutraliser la langue pour atténuer les biais.....	31
3.5 Conclusion.....	33
CHAPITRE 4 QUESTION DE RECHERCHE	34
CHAPITRE 5 MÉTHODE	36
5.1 Collecte de substituts.....	36
5.1.1 Terme cible	36
5.1.2 Jeu de données	37
5.1.3 Recrutement de participant·es	39
5.1.4 Création et déroulement de la tâche de substitution	40
5.2 Analyses qualitatives et analyses des plongements	44
5.2.1 Analyses qualitatives.....	44
5.2.2 Analyses des plongements.....	44
5.2.2.1 Récupération des plongements.....	45
5.2.2.2 Calculs de similarité	46
5.2.2.3 Méthode de regroupement.....	46
5.3 Conclusion.....	49
CHAPITRE 6 RÉSULTATS ET DISCUSSION	50
6.1 Analyses qualitatives : perception du sens de <i>gay</i> toxique et non toxique.....	50
6.1.1 Catégorie « homosexualité » et le cas de <i>homo</i>	53
6.1.2 Scores de similarité	55
6.1.3 Catégories positives	55
6.1.4 Autres catégories et observations	55
6.1.5 Conclusion : <i>Stories</i> de <i>gay</i> dans la communauté de joueur·euses.....	56
6.2 Analyse de regroupements des plongements	57
6.2.1 Regroupements des substituts par ligne cible.....	57
6.2.2 Regroupement des phrases cibles	63
6.3 Conclusion : <i>Story</i> de <i>gay</i> détectée dans les modèles	66
CONCLUSION.....	68
6.4 Limites de la recherche	68
ANNEXE A APPROBATION ÉTHIQUE.....	70
APPENDICE A FORMULAIRE DE CONSENTEMENT ÉTHIQUE	71
APPENDICE B INSTRUCTIONS POUR LA TÂCHE DE SUBSTITUTION	75
APPENDICE C JEU DE DONNÉES ET SUBSTITUTS OBTENUS.....	76
APPENDICE D REGROUPEMENTS OBTENUS.....	82
RÉFÉRENCES	89

LISTE DES FIGURES

Figure 1.1 Visualisation d'un réseau de neurones avec trois couches cachées (DeepAI, 2019).....	4
Figure 1.2 Projections en trois dimensions de représentations vectorielles fictives indiquant les relations sémantiques ou morphosyntaxiques (Dolphin, 2023).	8
Figure 1.3 Exemple de regroupements de phrases avec les termes <i>madhouse</i> et <i>asylum</i> (Chronis & Erk, 2020).	13
Figure 2.1 Délimitations des concepts de <i>Dark Participation</i> , <i>Toxicité</i> et <i>Trolling</i> dans les jeux vidéo en ligne (Kowert, 2020).	17
Figure 3.1 Traduction des mots <i>friend</i> , <i>doctor</i> et <i>nurse</i> sans flexion de genre en anglais à ami/médecin — au masculin et amie/infirmière — au féminin.	24
Figure 3.2 Autres choix pour la traduction de la deuxième phrase : malgré le changement de genre du terme <i>ami</i> , le mot <i>infirmière</i> reste au féminin.	24
Figure 3.3 Les réponses de ChatGPT montrent la présence d'une association engineer-homme, secretary-femme.....	25
Figure 3.4 Les réponses obtenues dans l'application ChatGPT montrent les biais en français également. 26	
Figure 3.5 Retirer le caractère biaisé ou toxique d'un terme revient à déchirer une page de sa <i>story</i> , qui contient autant des éléments appartenant aux biais qu'à la toxicité.	32
Figure 5.1 Image annonçant le recrutement de participant-es pour la tâche de substitution.....	40
Figure 5.2 Une interaction pour laquelle les participant-es doivent trouver un substitut.	41
Figure 5.3 Exemple donné aux participants.es avant de commencer la tâche.....	43
Figure 5.4 Un exemple de l'utilisation de l'algorithme <i>k-moyens</i> : des données sont regroupées ensemble selon leurs propriétés.	48
Figure 5.5 Exemple de la méthode du coude pour repérer la meilleure valeur de <i>k</i>	48
Figure 6.1 Visualisation 2D des groupements des plongements des substituts de la phrase « bruh im black ». Les plongements sont issus du modèle BERT.....	58
Figure 6.2 Visualisation 3D des plongements des substituts de la phrase #5 « and you are gay now ». Les plongements sont issus du modèle BERT.	59
Figure 6.3 Visualisation 2D des plongements des substituts de la phrase #5 « and you are gay now ». Les plongements sont issus du modèle RoBERTa préentraîné sur des clavardages de jeux vidéo.	60

Figure 6.4 Visualisation 2D des plongements des substituts de la phrase #17 « ill accept being gay :) ». Les plongements sont issus du modèle BERT.....	61
Figure 6.5 Visualisation 2D des plongements des substituts de la phrase « ill accept being gay :) ». Les plongements sont issus du modèle RoBERTa.....	62
Figure 6.6 Visualisation 2D des plongements des substituts de la phrase #12 « why so mad we got gay banners ». Les plongements sont issus du modèle BERT.....	63
Figure 6.7 Visualisation 2D des regroupements des phrases cibles avec les plongements de BERT – les losanges sont des phrases toxiques et les cercles sont des phrases non toxiques.....	64
Figure 6.8 Visualisation 2D des regroupement des phrases cibles avec les plongements de RoBERTa – les losanges sont des phrases toxiques et les cercles sont des phrases non toxiques.....	65

LISTE DES TABLEAUX

Tableau 1.1 Choix de substituts lexicaux pour le nom « charge », dans le sens d'une personne aux soins d'une autre. Les mots en italiques se retrouvent dans WordNet, mais ne sont pas reliés au terme cible « charge » (Kremer et al., 2014).....	11
Tableau 5.1 Exemples de phrases du jeu de données, accompagnés de leur contexte et annotation de toxicité (1 = toxique et 0 = non toxique)	39
Tableau 6.1 Catégories identifiées parmi les substituts des lignes cibles toxiques	51
Tableau 6.2 Catégories identifiées parmi les substituts des lignes cibles non toxiques.....	53
Tableau 6.3 Regroupements des phrases cibles avec les plongements de BERT	64
Tableau 6.4 Regroupements des phrases cibles avec les plongements de RoBERTa.	65
Tableau 6.5 Tableau complet du jeu de données, avec les substituts obtenus.....	76
Tableau 6.6 Tous les regroupements obtenus avec les plongements (BERT) des substituts.	82
Tableau 6.7 Tous les regroupements obtenus avec les plongements (RoBERTa) des substituts.	85

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

GML : Grands modèles de langue

TAL : Traitement automatique des langues

RÉSUMÉ

Alors que les communications en ligne occupent une place grandissante dans les interactions sociales, garantir des environnements surs et inclusifs pour les groupes marginalisés devient crucial. Dans le contexte des discussions en ligne dans les jeux vidéo, des modèles de langue sont utilisés pour détecter et modérer les contenus toxiques. Cependant, ces modèles peuvent pénaliser de manière disproportionnée les communautés marginalisées lorsqu'elles tentent d'affirmer leur identité ou de discuter de leurs communautés. Ce mémoire explore les interactions entre la toxicité et les biais identitaires dans les clavardages de jeux vidéo multijoueurs en ligne. L'étude analyse les relations sémantiques encodées dans les plongements linguistiques afin de mieux comprendre les compromis nécessaires entre l'atténuation des biais et la perte de performance. Une collecte de substituts lexicaux au terme « gay » dans des contextes toxiques et non toxiques, ainsi qu'une analyse des regroupements dans l'espace des plongements, met en lumière la relation profonde entre biais identitaires et toxicité, compliquant ainsi la séparation des deux concepts. En démontrant l'interaction sémantique entre biais et toxicité et comment elle est encodée dans les plongements, cette recherche plaide pour une compréhension approfondie des représentations linguistiques, afin de concevoir des modèles plus inclusifs, tout en soulignant la nécessité de prudence dans l'application des modèles de détection automatique.

Mots clés : TAL, biais, toxicité, jeux vidéo, modèles de langue

ABSTRACT

As online communications increasingly play a central role in social interactions, ensuring safe and inclusive environments for marginalized groups becomes crucial. In the context of online multiplayer game chats, language models are used to detect and moderate toxic content. However, these models may disproportionately penalize marginalized communities when they attempt to assert their identity or discuss their communities. This thesis explores the interactions between toxicity and identity bias in online multiplayer game chats. The study analyzes the semantic relationships encoded in language embeddings to better understand the necessary trade-offs between bias mitigation and performance loss. A collection of lexical substitutes for the term "gay" in toxic and non-toxic contexts, as well as an analysis of clusters in the embedding space, highlights the deep relationship between identity bias and toxicity, making the separation of the two concepts difficult. By demonstrating the semantic interaction between bias and toxicity and how it is encoded in embeddings, this research advocates for a deeper understanding of linguistic representations to design more inclusive models, while emphasizing the need for caution in the application of automatic detection models.

Keywords : toxicity, biases, nlp, language models, videogames

INTRODUCTION

Les interactions en ligne occupent de plus en plus une place centrale dans la manière dont les individus communiquent et échangent. Avec cela s'impose la nécessité de créer des environnements plus inclusifs et accueillants pour les groupes marginalisés. Des lois et réglementations sont mises en place graduellement à l'échelle globale pour assurer que ces espaces soient sécuritaires pour tous (Loi sur les préjudices en ligne, 2021; Single market for digital services (Digital Services Act), 2022). Dans le contexte des jeux vidéo en ligne multijoueurs compétitifs, ces réglementations sont particulièrement essentielles pour garantir un espace où compétition et respect peuvent coexister. Pour agir en accord avec ces lois et pour rejoindre un public diversifié, des entreprises productrices de jeux vidéo reconnaissent désormais l'importance de modérer les discours et d'encourager des comportements responsables afin de maximiser l'engagement des utilisateurs tout en maintenant un climat de compétition sain (Miller, 2019; Unity, 2021). Cette dynamique est toutefois complexe à gérer, car elle exige de trouver un équilibre entre la liberté d'expression et la modération des contenus toxiques. Ce débat s'inscrit dans une transformation culturelle plus large, marquée par un changement des normes sociales quant à ce qui est jugé acceptable dans l'espace public. Par exemple, des sujets autrefois considérés comme tolérables, tels que les blagues sur des violences sexuelles ou autres blagues misogynes, sont devenus largement inacceptables, notamment depuis le mouvement #MeToo en 2016.

Aujourd'hui, des outils accessibles tels que ChatGPT et Copilot, fonctionnant à l'aide de modèles de langue entraînés sur de volumineux corpus de texte en ligne, capturent ces changements culturels (Davani et al., 2023; Ziems et al., 2024). Ces modèles deviennent ainsi des miroirs déformants, révélant des aspects de notre propre culture numérique. Dans le cas de modèles de langue qui sont créés avec le but de détecter les discours haineux et toxiques dans des endroits d'échange en ligne comme les clavardages de jeux vidéo, la fonctionnalité de détection de ces discours repose sur les biais et les représentations sociales présents dans les données sur lesquelles ils sont entraînés (Bender et al., 2021; Davidson et al., 2017, 2019; Yang et al., 2023). Toutefois, ces modèles peuvent pénaliser de manière disproportionnée des communautés déjà marginalisées. En tentant de s'exprimer ou de s'identifier, ces groupes peuvent voir leurs discours automatiquement catégorisés comme toxiques, ce qui amplifie leur invisibilité et leur sentiment d'exclusion (Davidson et al., 2019; Dixon et al., 2018; Sap et al., 2019). Il semble important d'atténuer ces biais identitaires, mais la tâche est complexe. En analysant les représentations linguistiques encodées dans les modèles, nous avons l'opportunité non seulement de mieux comprendre leur fonctionnement, mais

aussi de découvrir comment sont encodés les liens entre la toxicité et les biais identitaires ainsi que les éléments qui reflètent ces deux concepts.

Le présent mémoire propose d'explorer ces problématiques. Premièrement, dans le premier chapitre, nous examinerons comment les modèles de langue actuels sont entraînés et comment il est possible d'obtenir des représentations de la langue par leur entraînement. Deuxièmement, au chapitre suivant, nous nous pencherons sur la question de la toxicité en ligne, en particulier dans les espaces de jeux vidéo, et sur les méthodes employées pour la détecter et la modérer. Ce chapitre abordera également les limites actuelles de ces approches, notamment les biais identitaires s'y glissant. Enfin, dans le chapitre 3, nous analyserons plus en profondeur les biais présents dans les modèles de langue et leurs impacts sur la détection de la toxicité, avant d'explorer les méthodes d'atténuation de ces biais. Le chapitre 4 permettra de formuler les questions guidant la recherche sur l'interaction entre toxicité et biais identitaires dans les modèles de langue dans une optique d'atténuation de biais, et les chapitres 5 et 6 permettront de proposer une approche méthodologique et d'analyser les résultats obtenus afin de discuter des implications de ces biais dans les contextes des jeux vidéo en ligne.

CHAPITRE 1

LES MODÈLES DE LANGUE ET ÉMERGENCES LINGUISTIQUES

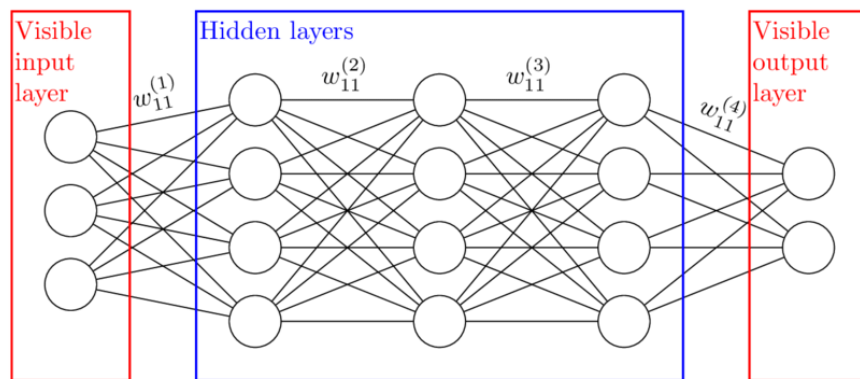
Les progrès fulgurants dans le domaine de l'apprentissage automatique observés ces dernières années ouvrent la voie à l'automatisation de nombreuses tâches complexes. Des techniques de traitement automatique des langues (TAL), plus précisément celles qui sont issues de l'intelligence artificielle, sont utilisées pour entraîner des outils polyvalents appelés des modèles de langues. Ces modèles servent de base à la création de systèmes permettant la traduction automatique, la classification de texte, l'analyse de sentiment, la révision de texte, l'assistance à l'écriture, la création d'un robot conversationnel et plus encore. La première étape pour accomplir ces tâches consiste à utiliser un modèle de langue préentraîné. Ce chapitre permettra de mieux comprendre l'entraînement et l'utilisation des modèles, le fonctionnement de l'encodage d'une représentation de la langue, et les informations linguistiques contenues dans ces représentations.

1.1 Entraînement et utilisation des modèles

La phase de préentraînement d'un modèle correspond à une phase initiale d'« apprentissage » des régularités linguistiques présentes dans le langage. Cet « apprentissage » se déroule dans l'architecture du modèle. Pendant cette étape, d'énormes volumes de texte sont donnés au modèle pour identifier des motifs sémantiques, syntaxiques, et contextuels. Un modèle de base repose sur une architecture qui s'inspire de façon approximative du fonctionnement du cerveau humain : il est composé d'un réseau de neurones artificiels à plusieurs couches (voir Figure 1.1). Ces couches successives forment une hiérarchie dans laquelle les connexions entre les neurones contiennent des poids ajustables représentant une information, appelés paramètres, qui sont transformés et raffinés à chaque étape. Les couches d'entrée du réseau reçoivent des données textuelles brutes, comme des séquences de mots. L'objectif primaire de l'utilisation d'un modèle est de prédire le prochain mot, la prochaine phrase, ou un terme masqué dans une phrase (p. ex. : je vais me chercher un MASK pour écrire). À travers le réseau, chaque couche de neurones est mise à jour de manière à extraire des caractéristiques générales du texte en se concentrant sur des relations contextuelles entre les mots. Les couches internes du modèle, appelées « couches cachées », sont responsables de l'analyse des interactions entre les termes du texte. Enfin, le résultat final est une phrase cohérente ou un ensemble de prédictions qui respectent les règles et structures du langage naturel, apprises de manière implicite à travers l'exposition massive aux données textuelles.

Parmi les modèles préentraînés plus récents et fréquemment utilisés, on retrouve les modèles de type Transformers comme BERT (Devlin et al., 2019), GPT (Radford et al., 2019), LLAMA2 (Touvron et al., 2023) et Gemini (Team et al., 2023). Ces modèles sont catégorisés comme étant des grands modèles de langage (GML) parce que leur architecture est plus complexe que les réseaux de neurones comme ceux de la Figure 1.1, ils contiennent des millions ou milliards de paramètres avec plusieurs couches cachées. Ce sont les outils derrière les suggestions de texte dans des moteurs de recherche et la génération de contenu et plusieurs robots conversationnels comme ChatGPT, Copilot et Gemini.

Figure 1.1 Visualisation d'un réseau de neurones avec trois couches cachées (DeepAI, 2019).



Les modèles préentraînés capturent ainsi des représentations du langage qui « imite » la compétence langagière humaine de produire du texte cohérent et approprié selon le contexte. Pour exploiter pleinement ces outils puissants et les représentations apprises, il est possible de les adapter pour effectuer des tâches plus précises, comme celles énumérées un peu plus tôt. Pour ce faire, un second processus d'« apprentissage » appelé *fine-tuning* ou ajustement fin est enclenché, cette fois avec un corpus étiqueté qui contient des paires de données d'entrée et de sorties. Prenons un exemple de classification de texte, où nous détenons une large quantité de récits dans un corpus, chacun accompagné de leur type (aventure, policier, fantastique, science-fiction, etc.) clairement identifié. L'étape d'ajustement fin sert à modifier les paramètres du modèle préentraîné de façon à ce que les régularités apprises soient reliées à la nouvelle tâche, et que les prédictions résultantes du modèle soient les étiquettes présentes dans le corpus étiqueté — dans cet exemple, on s'attend à obtenir un type de récit. Une fois l'étape terminée, il est ainsi possible de présenter un nouveau récit au modèle pour obtenir le type correspondant.

Si l’ajustement fin nécessite un large volume de données, il reste bien moindre que celui exigé par la phase de préentraînement des larges modèles de langue. Ces exigences sont nécessaires pour traiter les vastes ensembles de données et effectuer les calculs complexes requis pour obtenir des paramètres de haute qualité qui représentent les régularités linguistiques. Ce sont ces représentations qui donnent aux modèles de langue leur utilité remarquable pour effectuer des tâches linguistiques complexes. Dans la prochaine partie, nous explorerons en détail les caractéristiques et le fonctionnement des représentations de la langue obtenues à la suite d’un préentraînement ou d’un ajustement fin, en mettant en lumière leur rôle crucial dans le domaine du traitement du langage naturel.

1.2 Représentation de la langue

La création d’un modèle de langue peut se faire de diverses manières, conduisant à des modèles de différentes complexités et tailles. Comme vu à la section précédente, les modèles sont créés en ajustant un nombre élevé de paramètres selon des régularités linguistiques observées dans le corpus. Ces paramètres permettent de générer des vecteurs, représentations numériques des données textuelles. Les vecteurs sont obtenus en associant chaque mot (appelés tokens) à une représentation qui est ensuite raffinée à travers les couches du modèle. Les embeddings peuvent être calculés à différents niveaux, comme au niveau des tokens, des phrases ou même des documents entiers issus du corpus d’entraînement, selon le contexte d’utilisation. Les vecteurs sont appelés *embeddings* en anglais, et *plongements* ou *représentation* en français. Ces deux derniers seront utilisés dans ce texte. Il existe deux types de plongements principaux : les plongements statiques et les plongements dynamiques.

Les représentations contenues dans un modèle appelé *bag-of-words*, par exemple, sont des représentations dites *statiques* qui représentent le document complet, et sont produites à partir de la fréquence d’apparition de chaque terme dans le texte. Un document ayant simplement la phrase de l’exemple (1) aurait alors la représentation en (1).

- (1) a. « il aime les patates et les tomates, mais il n’aime pas les céleris et les avocats »
b. {« il » : 2, « aime » : 2, « les » : 4, « patates » : 1, « et » : 2, « tomates » : 1, « mais » : 1, « n’ » : 1, « pas » : 1, « céleris » : 1, « avocats » : 1}

D’autres modèles comme *Word2Vec* (Mikolov et al., 2013) ou *GloVe* (Pennington et al., 2014) sont construits à partir des principes de la sémantique distributionnelle, selon laquelle le sens d’un mot est défini par sa distribution dans l’usage de la langue, ce qui l’accompagne (Firth, 1957; Gastaldi, 2021;

Wittgenstein, 2014). En suivant ces principes, il est possible de concevoir le sens d'un mot en observant ce qui l'entoure. En lisant l'exemple (2), par exemple, même sans savoir ce qu'est un *mosswine*¹, le contexte est familier :

- (2) a. Les **mosswines** fouillent le sol avec leur truffe pour trouver de la nourriture.
- b. La viande des **mosswines** est un mets de choix. Dociles mais chargent si on les énerve.
- c. La peau épaisse du **mosswine** le protège des attaques des prédateurs et des intempéries.
- d. Les **mosswines** sont généralement situés dans des zones qui contiennent des champignons.

Truffe, fouiller le sol, peau épaisse, viande étant un met de choix, champignons... Ces termes indiquent que le *mosswine* est une créature semblable au cochon ou au sanglier. Le principe de sémantique vectorielle, une approche standard de la représentation du sens lexical en TAL, se base sur la sémantique distributionnelle pour représenter un terme en tant que vecteur, un point dans un espace multidimensionnel. Pour `Word2Vec` et pour `GloVe`, les représentations obtenues sont statiques, ce qui signifie qu'un terme possède une seule représentation même s'il est polysémique. Ainsi, chaque terme du vocabulaire d'un corpus est associé à un vecteur dans un espace multidimensionnel. Les valeurs des dimensions sont établies par le modèle pendant le processus d'apprentissage et peuvent être interprétées comme des axes de sens différents. Dans la phrase (1), chaque terme sera ainsi représenté par une suite de nombres : le terme « les » sera associé à une suite de nombres de taille n où n est le nombre de dimensions: $[x_1, x_2, x_3 \dots x_n]$.

Les GML mentionnés à la section précédente, les Transformers comme BERT (Devlin et al., 2019) et GPT (Radford et al., 2019), ont une architecture qui inclut des réseaux de neurones dont les couches contiennent un mécanisme d'attention (Vaswani et al., 2017), permettant une analyse simultanée de toutes les positions d'un terme dans une séquence de texte. Les représentations issues des modèles Transformers sont dites *dynamiques* : la représentation d'un terme dépend du contexte dans lequel on le retrouve. Lorsqu'un texte est passé à BERT, par exemple, plusieurs étapes se déroulent pour produire les plongements. D'abord, le texte est séparé en tokens – en mots. Bien que le préentraînement d'un GML

¹ Les mosswines des créatures fictives, retrouvées dans les jeux de la série Monster Hunter. Les phrases de l'exemple (2) proviennent en partie des descriptions de la créature dans les différents jeux, retrouvées sur la page suivante : <https://mogapedia.fandom.com/fr/wiki/Mosswine>

contienne une grande quantité de mots, certains termes sont rares et il est difficile d'identifier les motifs dans lesquels ils se retrouvent. Pour pallier ce problème, le processus de tokénisation se fait en mots ou en sous-mots. Le terme *disloyal*, par exemple, est divisé ainsi : « di », « ##sl », « ##oya », « ##l ». Les sous-mots sont aussi appelés WordPieces (Wu et al., 2016) et consistent en des pièces de mots qui permettent de diviser les mots inconnus en petites parties et leur assigner une représentation. Comme le montre l'exemple de *disloyal*, les sous-mots ne sont pas nécessairement des morphèmes, et n'ont donc pas nécessairement de sens lexical ou grammatical. Il a été démontré que cette méthode de tokenisation est très efficace, car elle permet d'obtenir des représentations pour tous les mots d'un texte même dans le cas où ce mot n'était pas présent dans les données d'entraînement du modèle. Une fois les tokens obtenus, ils sont convertis en identifiants numériques. En accord avec ce qui a été mentionné à la section précédente, chaque couche permet d'affiner les représentations contextuelles des tokens, en capturant des informations de plus en plus abstraites à mesure que l'on progresse dans les couches. Alors qu'un modèle comme Word2vec ne contient qu'une seule couche cachée, le modèle BERT en compte 12. Chaque couche de BERT ou autre modèle Transformers utilise un mécanisme d'attention qui permet de mesurer l'importance relative des tokens les uns par rapport aux autres. Cela favorise, entre autres, la capture de relations à longue distance entre les mots et l'amélioration de la modélisation syntaxique et sémantique (Jawahar et al., 2019). Ce processus permet ainsi de générer des représentations contextuelles riches, contenant des informations essentielles sur chaque token. Par conséquent, cette approche se traduit par de meilleures performances dans des tâches linguistiques comme la traduction automatique et la classification de texte, en encodant diverses informations linguistiques à travers les plongements créés.

1.3 Information linguistique contenue dans les plongements

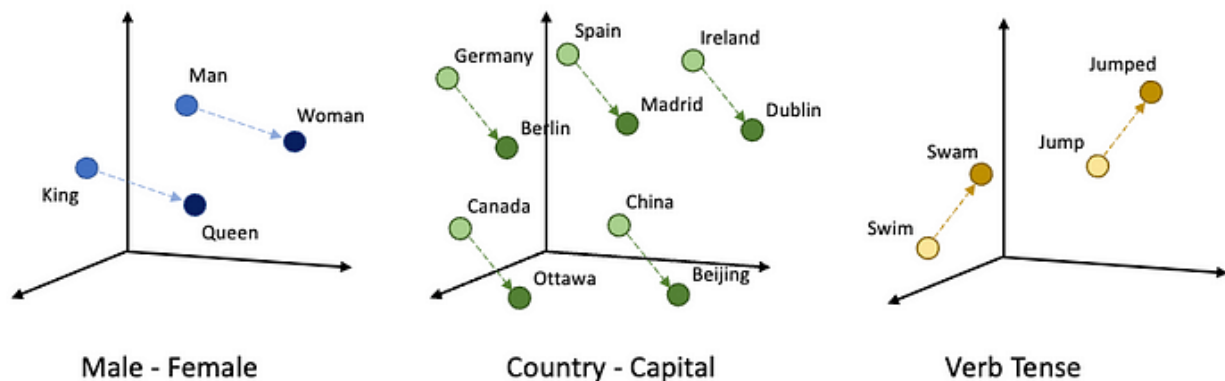
1.3.1 Distances entre les plongements et couches cachées

Les représentations résultantes de l'entraînement d'un large modèle de langue sont des vecteurs de centaines de dimensions et sont ainsi en elles-mêmes ininterprétables par l'humain. Bien que le vecteur d'un token donné ne soit pas interprétable, il est tout de même possible de mesurer les distances entre les vecteurs dans l'espace multidimensionnel produit pour observer les relations linguistiques qui émergent de ces représentations. Des analyses de ces relations ont permis d'identifier que les régularités distributionnelles représentées dans les plongements laissent émerger, par exemple, des relations sémantiques et pragmatiques — comme des implicatures et présuppositions (Jeretic et al., 2020), ou des relations d'hyponymie et de synonymie (Padó & Lapata, 2003; Sahlgren, 2006) — et morphosyntaxiques,

et ce même dans les plongements statiques (Mikolov et al., 2013; Pennington et al., 2014; Caliskan et al., 2017; Jawahar et al., 2019; Dolphin, 2023).

La Figure 1.2 permet de visualiser le concept en utilisant des plongements statiques fictifs. Les vecteurs ayant plusieurs dimensions, des projections en deux ou trois dimensions sont utilisées pour mettre en évidence certaines relations. On peut observer que la première projection met en évidence une relation de genre (masculin ou féminin) entre les termes, puisque la distance entre *Man* et *Woman* est similaire à celle entre *King* et *Queen*. La deuxième projection exemplifie la relation sémantique pays-capitale. Finalement, la troisième permet d'illustrer la relation de flexion entre un même verbe au présent et au passé. De même, en analysant les écarts de distance entre les occurrences d'un terme dans un corpus au fil du temps, il est possible de réaliser des études diachroniques. Ces analyses permettent d'observer les similitudes entre les termes d'une époque à une autre et d'identifier ainsi les changements de connotation qu'ils ont pu subir (Hamilton et al., 2016; Kozlowski et al., 2019).

Figure 1.2 Projections en trois dimensions de représentations vectorielles fictives indiquant les relations sémantiques ou morphosyntaxiques (Dolphin, 2023).



Contrairement aux modèles comme *Word2vec* et *GloVe*, les représentations dynamiques générées par les Transformers s'adaptent au contexte spécifique de chaque mot, offrant ainsi une meilleure compréhension contextuelle du texte. La taille des modèles Transformers est plus importante, non seulement en ce qui concerne le nombre de dimensions dans les plongements, mais aussi en nombre de plongements pour chaque terme : comme mentionné dans les sections précédentes, les réseaux détiennent plusieurs couches « cachées », et chaque couche correspond à un plongement qui joue son rôle dans l'encodage d'informations linguistiques. La structure de BERT, par exemple, compte habituellement 11 couches de neurones cachés. Plusieurs études ont été menées pour identifier avec

une plus grande précision les informations linguistiques encodées dans chaque couche (Geiger et al., 2022; Hoover et al., 2019; Jawahar et al., 2019; Merchant et al., 2020). Parmi ces travaux, Jawahar et al. (2019) soulignent des parallèles entre la structure interne de BERT et l'analyse syntaxique traditionnelle, notamment en ce qui concerne sa modélisation compositionnelle. Dans le cadre de cette recherche, nous nous intéressons ainsi aux représentations dynamiques, en particulier celles des modèles de la famille BERT (Hewitt & Manning, 2019).

1.3.2 Plus que le contexte : la *story*

Erk et Chronis (2022) discutent de la richesse des informations contenues dans les plongements des mots d'un modèle comme BERT. Leur travail permet de mettre en lumière comment les plongements de mots peuvent être considérés comme des *word stories embeddings*², révélant ainsi la manière dont les informations sont encapsulées dans les représentations dynamiques. Nous connaissons tous le sens d'un mot comme *restaurant*, un « établissement où l'on sert des repas moyennant paiement » (Le Robert, s. d.) selon sa définition dans le dictionnaire. Ces définitions sont intéressantes et pratiques, mais ne contiennent toutefois pas toute l'information reliée au mot et laissent de côté une partie du sens, car elles se concentrent sur l'information objective et référentielle. Les *stories* contiennent des informations non seulement liées au sens et au contexte textuel d'un mot, mais elles se rapportent aussi aux connotations, aux éléments de connaissances préalables, aux scènes et événements prototypiques avec les émotions et jugements qui leur sont reliés, ainsi qu'aux contextes culturels du mot. Elles sont des manières de connecter le mot et sa représentation et tous ces éléments, et ainsi d'encoder dans les plongements le point de vue humain sur le monde, ce qu'ils trouvent important, utile ou inutile. Inévitablement, un mot évoque la *story* qui lui est liée lexicalement, mais il évoque aussi les *stories* qui ne lui sont pas liées lexicalement. Le mot « restaurant », par exemple, réfère à plus que le simple bâtiment où de la nourriture est servie à des clients. La *story* du mot peut aussi faire référence à l'ambiance romantique/familiale/conviviale, à l'action de sortir de chez soi, prendre un véhicule pour s'y rendre, considérer l'argent nécessaire pour le paiement, le type de nourriture d'un restaurant en particulier et les émotions ou souvenirs qui y sont rattachés, etc. Les *stories* sont dynamiques et peuvent changer selon les informations contenues dans la phrase. De façon similaire à la polysémie, qui met en évidence les multiples

² En français, « histoire de mots ». Les termes « word story/stories » sont conservés et non traduits dans ce texte en raison de ses connotations spécifiques qui enrichissent la compréhension du concept. En effet, cette expression évoque une dimension narrative, qui met de l'avant la connexion entre les mots et tout ce qui s'y rattache. Cette nuance est difficile à transmettre par une traduction française directe. De plus, l'utilisation du terme original permet de reconnaître l'apport d'Erk et Chronis (2022) à la clarification de ce concept.

sens qu'un seul mot peut posséder, les *stories* plongent plus profondément dans les nuances contextuelles qui entourent l'interprétation d'un mot. Par exemple, le mot « banc » peut évoquer plusieurs émotions et images; s'il se réfère à un banc d'école, il pourrait être associé à l'apprentissage; un banc de parc évoque des images de nature et d'air frais; un banc de sable rappelle des baignades dans une rivière ou un sentiment de relaxation sur la plage. Ce concept se retrouve ainsi au-delà de relations issues de la sémantique lexicale comme les hyperonymes et hyponymes et considère qu'un mot évoque nécessairement sa *story* complète.

L'idée de *word stories* est parallèle à une autre notion en sémantique appelée *frames*, telle que décrite par Fillmore (1982). Comme les *stories* chaque mot est intrinsèquement lié à une *frame*. Un mot évoque automatiquement ses *frames*, un ensemble d'expériences et de connaissances qui lui sont associées, organisées en catégories conceptuelles spécifiques, ce qui permet de structurer et d'interpréter ces expériences de manière cohérente. On retrouve des éléments similaires à ce qui est contenu dans une *word story* : les participant·es impliqué·es, les connaissances culturelles, les informations connexes, les émotions et les jugements. La différence principale avec les *stories* est que les *frames* sont statiques : les scènes et les événements évoqués sont prototypiques et offrent une stabilité quant à la catégorisation conceptuelle, peu importe le contexte (Erk & Chronis, 2022). Le mot « restaurant » évoque automatiquement certaines *frames* spécifiques comme les rôles des clients·es, serveurs·es, chefs·es, le menu, le paiement, et l'action de manger. Les *stories* capturent tout cela, mais offrent aussi une grande variation selon le contexte d'utilisation de « restaurant », comme expliqué ci-dessus. La perspective d'un passage d'une notion statique à dynamique ouvre la voie à des analyses plus approfondies sur la manière dont la *story* des mots évolue à travers différents contextes, ou même qu'une *story* évolue dans un même contexte, mais avec différents termes.

1.3.3 Contextes dynamiques et la substitution lexicale

Explorer des méthodes empiriques pour analyser les variations du sens des mots en fonction de leur contexte d'utilisation permet de mieux comprendre la richesse des *stories*. Avec cette analyse comme objectif, Kremer et al. (2014) créent un large corpus de substitution lexicale, CoInCo, en demandant à des annotateurs et annotatrices de fournir des termes pour remplacer chaque mot d'une phrase en fonction du contexte d'apparition. Iels comparent leur corpus à WordNet (Fellbaum, 1998), une ressource lexicale couramment utilisée pour établir des relations lexicales qui sont déterminées par des relations sémantiques en dehors de tout contexte, comme les relations d'hyperonymie et d'hyponymie. Les

substituts obtenus pour les différents contextes ne se retrouvent pas toujours dans les synonymes retrouvés dans WordNet : pourtant, ils sont jugés comme étant adéquats par les participant-es. La diversité des substituts obtenus pour un même mot dans des phrases différentes rend compte de l'importance du contexte dans la manière dont les mots sont utilisés et interprétés. En effet, Erk et Chronis (2022) discutent des résultats de Kremer et al. (2014) en mentionnant que les termes proposés par les annotateurs sont adaptés au contexte et à la *story* plutôt qu'aux relations présentes dans les synonymes de WordNet. Le Tableau 1.1 illustre cela avec le terme *charge* dans le sens d'une personne aux soins d'une autre. Même si le sens est identique dans les deux phrases, les substituts de la première phrase contrastent avec ceux de la deuxième phrase : l'un semble faire allusion à une personne à charge comme une personne en apprentissage qui accompagne un mentor, alors que l'autre renvoie plutôt à une scène de combat où les personnes à charge sont de rang inférieur à la personne décrite.

Tableau 1.1 Choix de substituts lexicaux pour le nom « charge », dans le sens d'une personne aux soins d'une autre. Les mots en italiques se retrouvent dans WordNet, mais ne sont pas reliés au terme cible « charge » (Kremer et al., 2014).

Phrases	Substituts
Now, how can I help the elegantly mannered friend of my Nephys and his surprising young <u>charge</u> ?	dependent, person, <i>task</i> , <i>lass</i> , <i>protégé</i> , <i>effort</i> , <i>companion</i>
The distinctive whuffle of pleasure rippled through the betas on the bridge, and Rakal let loose a small growl, as if to caution his <u>charges</u> against false hope.	dependent, <i>private</i> , <i>companion</i> , <i>follower</i> , <i>subordinate</i> , <i>prisoner</i> , <i>teammate</i> , <i>ward</i> , <i>junior</i> , <i>underling</i> , <i>enemy</i> , <i>group</i> , <i>crew</i> , <i>squad</i> , <i>troop</i> , <i>team</i> , <i>kid</i>

La tâche de substitution lexicale est une méthode efficace pour voir comment les mots sont utilisés dans des contextes spécifiques pour ensuite identifier des liens de relations et de *story* entre différents termes. Dans le cadre de cette recherche, nous utiliserons une tâche similaire pour faire émerger les relations entre la toxicité et les biais. Avant d'entrer en détail dans cette partie de la méthode, discutons d'abord de la place des *stories* dans les plongements dynamiques.

1.3.4 Plongements et leurs *word stories*

Avec un changement de sens aussi subtile pour deux mots identiques, il est à se demander si les plongements capturent vraiment ces différences sémantiques d'apparence minimes. Toujours selon Erk et Chronis (2022), il est possible d'observer l'effet de la *story* sur les plongements dynamiques des modèles, permettant ainsi d'obtenir plus d'information sur les *stories*. Pour ce faire, elles se basent sur des relations

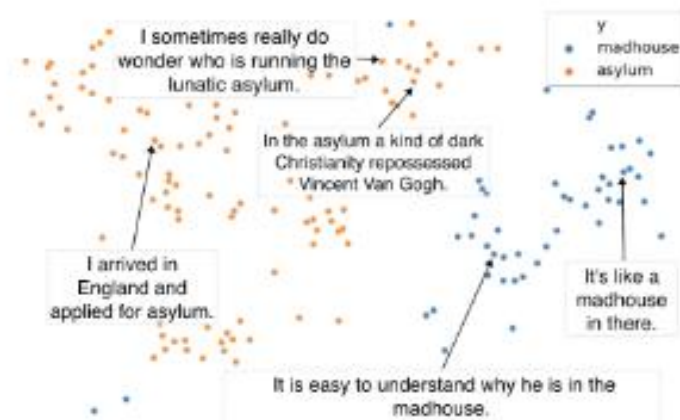
sémantiques à partir d'éléments lexicaux, notamment la similarité et la relation (*relatedness* en anglais) : des termes *similaires* ont des propriétés en commun, alors que des termes *reliés* sont liés aux mêmes sujets. Deux termes similaires ont une plus grande propension à se retrouver à la même place dans une phrase. Par exemple, le terme « avion » est **similaire** au terme « hélicoptère » puisque les deux sont des véhicules pouvant s'élever dans les airs; dans la phrase « l'avion/hélicoptère s'élève dans les airs », les deux prennent la même place. En revanche, « livre » est **relié** à « bibliothèque », car on retrouve des livres à la bibliothèque, on peut y lire un livre en silence, on peut emprunter un livre, etc.; ces termes sont plus propices à se suivre, comme dans la phrase « ce livre est disponible à la bibliothèque ». C'est le lien de relation qui est le plus représentatif de la notion de *story*, puisque les connexions entre les termes dépendent des interactions contextuelles entre eux. Ceci complète la théorie des *frames* en introduisant la dimension dynamique des *stories*. Prenons l'exemple de « femme » et « femoid », un terme péjoratif utilisé par certains groupes en ligne, tels que les incels³, pour désigner les femmes de manière dégradante. Alors que « femme » et « femoid » désignent tous les deux une personne du genre féminin, les *frames* associés à ces termes diffèrent considérablement. « Femme » active une *frame* qui inclut des rôles sociaux, des traits humains et des interactions quotidiennes. En revanche, « femoid » active une péjorative qui dénote une perception déshumanisante, comparant les femmes à des androïdes dépourvus d'émotions (Jaki et al., 2019). Les *frames* se chevauchent malgré tout, puisque « femoid » désigne une personne du genre féminin. Ainsi, les *stories* complètent les *frames* en montrant comment les termes, bien qu'ils partagent une dénotation et des *frames*, peuvent véhiculer des narrations différentes et évoluer selon les contextes et les intentions des utilisateurs.

Chronis et Erk (2020) s'intéressent à la présence du lien de relation dans les différents plongements associés à des tokens. Pour obtenir plus d'information sur comment ce lien est encodé dans les plongements, elles collectent d'abord des termes cibles en contexte, par exemple le terme *river* dans les phrases « How can he get all three safely over the *river*? » et « The *river* Sol is the southernmost of the Empire's rivers. » Elle procèdent par la suite à l'extraction des plongements des termes cibles, puis appliquent des algorithmes permettant de créer des groupes de phrases à partir des plongements du terme (voir Figure 1.3 pour un exemple de visualisation de regroupements avec les termes *madhouse* et *asylum*). En observant comment les phrases sont regroupées, elles déterminent si les différents sens du terme sont bien identifiés par le modèle : avec l'exemple de *river*, la première phrase est groupée avec

³ Les *incels* sont les membres majoritairement masculins d'une communauté en ligne qui exprime une haine et un mépris envers les femmes (Farrell et al., 2019; Jaki et al., 2019; Van Dorpe & Sénéchal, 2022).

d'autres phrases dans lesquelles *river* désigne un cours d'eau naturel, alors que la deuxième phrase est groupée avec d'autres phrases indiquant *river X*, *X* étant le nom d'une rivière. Elles effectuent également une seconde analyse de regroupement permettant de vérifier à quelle étape les représentations de BERT encodent les liens de *relation*. Pour cette analyse, quatre jeu de données qui contiennent des paires de termes ayant un lien de relation sont utilisés et regroupés. Ainsi, elles découvrent que les plongements des couches cachées finales de BERT, en particulier dans la 11^e couche, sont les plus utiles et appropriées pour accomplir efficacement des tâches en lien avec les relations entre des termes.

Figure 1.3 Exemple de regroupements de phrases avec les termes *madhouse* et *asylum* (Chronis & Erk, 2020).



Une grande partie des informations relatives aux liens de relations (et par extension, aux *stories*) se retrouvent donc dans la couche 11. Identifier la couche la plus pertinente pour des analyses nécessitant de faire ressortir la relation entre les termes est essentiel pour les analyses effectuées dans la recherche décrite dans ce mémoire. Par conséquent, nous nous concentrerons sur cette couche spécifique. La méthode de regroupement utilisée par Chronis et Erk (2020), utilisée également par Erk et Chronis (2022) pour vérifier l'effet des *stories* sur les plongements, sera également utilisée aux fins d'analyses. Tel que décrit à la section précédente, une tâche de substitution par des humains permet d'obtenir des termes qui partagent une *story*, un lien de relation. En regroupant des termes substitués obtenus pendant une telle tâche et en identifiant les thèmes de chaque groupe, nous pouvons ainsi en connaître plus sur la *story* du terme substitué et vérifier si la *story* est encodée adéquatement dans les plongements. Les prochains chapitres de ce mémoire permettront de mieux comprendre la pertinence de ce type d'analyse pour des problématiques telles que la détection de toxicité et des biais dans les plongements, en commençant par une exploration de la définition de la toxicité.

1.4 Conclusion

Ce chapitre a permis de poser les bases théoriques nécessaires pour comprendre le fonctionnement élémentaire des modèles de langues, et les mécanismes sous-jacents à leur utilisation pour automatiser des tâches complexes. La section 1.2 a permis de rendre compte de l'évolution rapide des technologies qui sous-tendent ces modèles, résultant des progrès réalisés dans les méthodes d'encodage des représentations linguistiques.

À la suite de la lecture de la section 1.3, il est plus clair que les relations contextuelles encodées dans les plongements contiennent entre autres des notions de *frames* et de *stories*. Des termes ayant des *frames* partagées divergent en *stories* selon les connotations sociales et les contextes d'utilisation, comme l'illustrent les exemples des termes « femme » et « femoid ». En s'appuyant sur les travaux d'Erk et Chronis (2022), ce chapitre conclut que la capture des relations sémantiques — en particulier des *stories* — dans les plongements est essentielle pour que l'utilisation d'un modèle soit efficace, en tenant compte des interactions entre les termes. La couche 11 d'un modèle basé sur BERT a été identifiée comme étant celle contenant le plus d'informations sur les *stories*. Les chapitres suivants permettront d'explorer ce que la présence des *stories* dans les plongements implique dans l'utilisation des modèles pour la détection de toxicité automatique, notamment en compliquant la dissociation des biais identitaires. Le CHAPITRE 2 se concentrera plus précisément sur la question de la détection de la toxicité.

CHAPITRE 2

TOXICITÉ : DÉFINITION ET DÉTECTION

Ce chapitre permet d'abord d'explorer une définition générale de la toxicité, avant de suggérer une définition qui s'accorde mieux avec le type de toxicité retrouvée en ligne et dans les jeux vidéo. Il sera ensuite question de survoler les méthodes de détection de toxicité, notamment avec les modèles de langues, et de comprendre où se retrouvent les informations liées à la toxicité dans les plongements d'un modèle. Finalement, les limitations des méthodes de détection actuelles seront abordées.

2.1 Définir la toxicité

À première vue, la définition de *toxicité* est une étiquette générale pour décrire tout comportement, substance ou produit qui a un effet nocif sur les éléments avec lesquels il interagit. En biologie ou en sciences de l'environnement, par exemple, on parle de toxicité pour désigner la propriété et la mesure de la propriété d'une substance toxique à empoisonner un organisme ou avoir un effet néfaste sur un environnement et sa biodiversité (Canada, 2017; Usito, s. d.). La définition change d'un domaine et d'un contexte à l'autre, mais la notion de conséquence indésirable reste. Une façon assez directe de mesurer les effets toxiques d'une substance lorsqu'ingérée par un humain est d'observer le déclin (ou non) de sa santé physique. La toxicité dans les communications étant un phénomène social, et de fait, subjectif, la mesure effectuée est plus complexe. La prochaine section permet de mieux définir la toxicité telle que retrouvée en ligne.

2.2 En ligne et dans les jeux vidéo

Un impact majeur de la toxicité se retrouve dans la santé des communautés en ligne et de leur diversité : avoir une présence en ligne expose l'utilisateur ou l'utilisatrice à la réception de menaces, de harcèlement et d'autres comportements toxiques, les incitant ainsi à quitter cet espace (Aroyo et al., 2019; Carillo & Marsan, 2016). Les communautés de joueurs et joueuses sont aussi affectées par la toxicité : iels sont vus comme perpétuant une culture toxique. Malgré la diversité des individus de la communauté, iels sont reconnus pour être souvent fermé-es à la diversité en refusant d'accepter dans leur groupe les joueurs ou joueuses ayant des traits ou une personnalité qui ne correspond pas au modèle préétabli du groupe (Kowert, 2020). Dans le cadre de ce travail, nous nous intéressons plus précisément à la toxicité présente

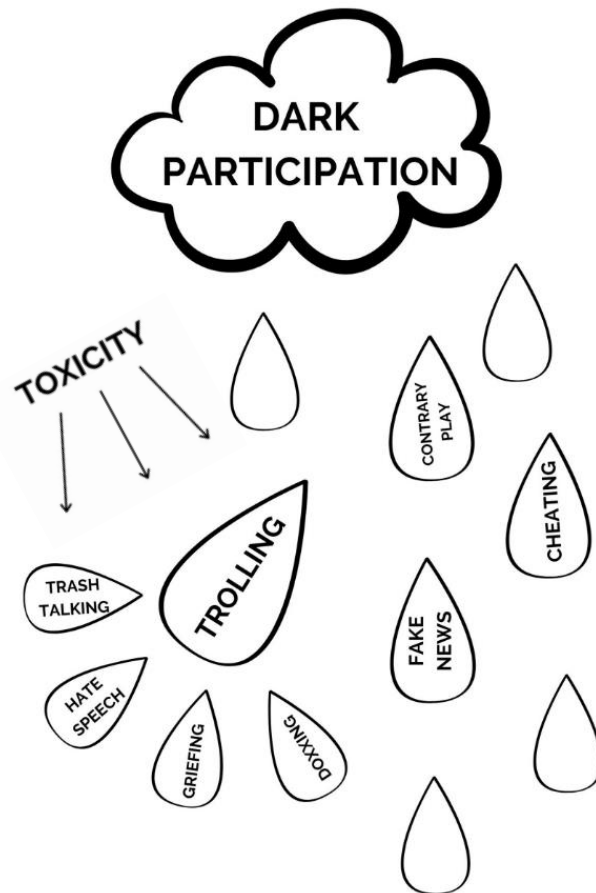
dans les échanges écrits dans les jeux vidéo multijoueurs en ligne, où des textes écrits ont un impact indésirable sur ceux qui les lisent.

Qu'est-ce qui détermine si un énoncé est toxique ou non ? La réponse dépend du contexte textuel, mais aussi du climat sociopolitique et du vécu de l'individu qui écrit ou consomme le contenu (Garg et al., 2023). Dans le contexte du jeu vidéo en ligne, on retrouve des éléments similaires aux autres médias sociaux qui permettent des publications, commentaires, ou autres discussions. Afin d'approfondir la compréhension du problème, de nombreux chercheurs et chercheuses tentent de définir la notion de toxicité en ligne. On la retrouve sous plusieurs noms : « toxicité » est évidemment souvent rencontrée (Aroyo et al., 2019; Dixon et al., 2018; Wulczyn et al., 2017), il est parfois question de « discours haineux » (Anderson & Barnes, 2022; Ciftci et al., 2017; Davani et al., 2023; Davidson et al., 2017; Sap et al., 2019; Watanabe et al., 2018), ou bien de « haine et harcèlement » (ADL, 2022; Ingersoll, 2019; Thomas et al., 2021). Chaque autrice et auteur attribue une définition selon les besoins de la recherche. On retrouve souvent une taxonomie se basant sur les façons de causer du préjudice à une cible, avec une intention. La taxonomie de Thomas et al. (2021) utilise en effet ces concepts pour identifier les attaques possibles dans le contenu toxique. On y retrouve, par exemple, de l'intimidation, des comportements visant à troller (tentative délibérée de provoquer ou de contrarier quelqu'un), du discours haineux, du contenu offensant ou profane, etc. Alors que le discours haineux (*hate speech*) est parfois considéré comme une sous-catégorie de la toxicité, une distinction claire est à faire pour Davidson et al. (2017). Un discours haineux est défini comme un « langage utilisé pour exprimer de la haine envers des groupes cibles, ou bien pour intentionnellement être désobligeant, humilier, ou insulter les membres du groupe » [Ma traduction], et du contenu offensant englobe tout texte potentiellement offensant qui ne correspond pas à la définition de discours haineux.

Les recherches sur la toxicité dans les jeux vidéo portent le plus souvent sur les comportements généraux des joueuses et joueurs. L'Anti-Defamation League (ADL, 2020, 2022) rapporte les résultats d'un sondage auprès des joueur-euses qui les questionnait sur leur expérience avec des « comportements perturbateurs », incluant du *trolling*, des insultes, des menaces de violence physique, de l'harcèlement sexuel, etc. Un comportement perturbateur peut aussi être relié à des actes plus concrets ou en dehors du clavardage, comme le *swatting* (un faux signalement à des services d'urgence pour que la personne reçoive l'intervention d'une équipe SWAT), le *stalking* (conduite obsessionnelle et intrusive visant à traquer ou harceler une personne) ou bien d'être intentionnellement mauvais au jeu pour ruiner la partie pour les autres joueur-euses (Kwak et al., 2015). Kowert (2020) propose une terminologie commune sur le sujet en

suggérant l'utilisation de l'expression « Dark participation » comme terme générique englobant tous les comportements déviants rencontrés en ligne (voir Figure 2.1). Les comportements toxiques représentent une forme spécifique de ces actions ; ils se caractérisent par le fait de causer du tort à la santé ou au bien-être d'une autre personne. Cette distinction est significative, car de tels comportements sont souvent définis de manière culturelle (Kowert, 2020; Kwak et al., 2015).

Figure 2.1 Délimitations des concepts de *Dark Participation*, *Toxicité* et *Trolling* dans les jeux vidéo en ligne (Kowert, 2020).



En considérant la diversité des termes pour désigner le phénomène, les termes *toxique* et *toxicité* seront ici employés pour désigner tout comportement perturbateur (propos haineux, insultes, préjugices, etc.) dans les clavardages des jeux.

2.3 Méthodes de détection de toxicité

Il est difficile de faire état de la situation ou de prendre des mesures pour contrer le problème de toxicité si on ne peut pas identifier les cas où elle se manifeste. Les approches actuelles comprennent la

modération par des êtres humains et la détection de mots-clés ; cependant, ces méthodes présentent plusieurs défis propres. Assurer une modération efficace à grande échelle avec des modérateurs humains est un processus chronophage avec un grand coût humain puisque les modérateurs et modératrices consomment un très grand nombre de contenus difficile à traiter sans nécessairement avoir un soutien psychologique adéquat (Arshat & Etcovich, 2018). La détection de mots-clés seule ne capture pas toute la complexité et la subtilité des comportements toxiques en ligne, et il est difficile de garder les dictionnaires de mots-clés à jour avec l'évolution rapide des termes utilisés par les joueur·euses. Le contexte joue également un rôle crucial puisque le jugement de toxicité porté sur un texte peut varier en fonction de son environnement contextuel (Pavlopoulos et al., 2020). Des techniques de Traitement Automatique du Langage (TAL) permettent de diminuer l'implication humaine dans le processus, de considérer le contexte, et d'identifier rapidement le contenu toxique.

En TAL, la détection de la toxicité constitue essentiellement une tâche de classification de texte, présentée au CHAPITRE 1. Dans cette tâche, un modèle de langue préentraîné est finement ajusté pour pouvoir obtenir des prédictions sur le caractère toxique ou non d'un texte. Cette adaptation du modèle repose sur un ensemble de données annotées où les exemples sont étiquetés comme toxiques ou non toxiques. En entraînant le modèle sur ces données, les caractéristiques des textes toxiques sont encodées dans les plongements, ce qui permet ensuite de généraliser et de détecter efficacement la toxicité dans de nouveaux textes. La classification du texte toxique est souvent effectuée à l'aide d'un score de toxicité entre 0 et 1, où le 0 indique du texte ne comportant aucune toxicité et 1 étant le plus toxique. La prédiction contient aussi généralement des scores pour différents attributs, ou catégories, possibles de toxicité (p. ex. : insultes, menaces, contenu sexuellement explicite, discours haineux, etc.) (Lees et al., 2022; Prabhakaran et al., 2019; Yang et al., 2023). Un modèle comme *Perspective API* (Lees et al., 2022) peut donc indiquer qu'un texte est considéré comme étant toxique (1), et qu'il a un score plus haut dans la catégorie « menaces », ce qui signifie qu'il est probable que la toxicité reconnue soit due au caractère menaçant du texte.

2.4 Information linguistique : toxicité dans les plongements

Comme nous l'avons détaillé dans le présent chapitre, la toxicité est un concept subjectif dont la perception est influencée par divers facteurs tels que l'environnement socioculturel de chaque individu ou de chaque communauté. Cela signifie que, même avec un énoncé qui ne contient que des termes toxiques, par exemple une seule insulte directe, on ne peut quantifier et généraliser le degré d'impact ou

le contenu offensant présent. Tous les éléments qui permettent d’attribuer les termes lexicaux à leur caractère toxique ou non toxique se rapportent directement aux notions de *frame* et de *story*, telles que décrites à la section 1.3.2 de ce mémoire.

Avec une meilleure compréhension de l’information linguistique contenue dans les plongements, comme vu à la section 1.3, nous émettons l’hypothèse que ces informations recoupent ce qui constitue les *stories* de termes ou de phrases toxiques. On peut ainsi s’attendre à ce qu’un grand nombre d’informations sur la toxicité se retrouve dans la couche numéro 11 de BERT. Cela confirme le choix énoncé précédemment d’utiliser les plongements de cette couche spécifique pour les analyses subséquentes de la recherche actuelle. Malgré la présence des éléments constituant la toxicité dans les plongements, il existe plusieurs enjeux et défis reliés à la détection de toxicité à l’aide de modèles de langue.

2.5 Limitations

Les enjeux et défis qui compliquent la tâche de détection reposent d’abord et avant tout sur le fait qu’il existe une volonté de développer des systèmes de détection universels, pouvant s’appliquer à plusieurs plateformes, peu importe la langue parlée et la géographie des utilisateurs. Certains défis se présentent au niveau technique, par exemple dans le choix de type et de quantité de contexte à ajouter au corpus d’entraînement. En effet, certains chercheurs et chercheuses ont tenté d’ajouter des métadonnées d’utilisateurs et utilisatrices des plateformes évaluées (Fehn Unsvåg & Gambäck, 2018) en incorporant des modèles spécialisés pour les conversations à plusieurs échanges en plus de données sur le public utilisant la plateforme (Lu et al., 2020; Yang et al., 2023). D’autres défis concernent les enjeux d’annotation de jeu de données, comme la diversité des participantes et participants et la qualité des annotations obtenues. D’autres défis concernent plutôt des aspects d’une analyse linguistique avancée, comme le contenu implicitement toxique, le sarcasme et l’ironie, les métaphores et les références implicites (van Aken et al., 2018). Pour ce projet, nous nous concentrerons principalement sur les enjeux sociolinguistiques.

Obtenir un système de détection universel n’est pas envisageable, d’abord parce que, comme mentionné à la section 2.1, il existe actuellement un manque de consensus sur la définition du contenu toxique : certains le regroupent avec le discours haineux, tandis que d’autres le considèrent comme une sous-catégorie ou une catégorie principale d’autres comportements perturbateurs. Cette diversité de perspectives rend difficile l’établissement de normes rigides pour définir la toxicité en ligne et créer des systèmes inclusifs. En effet, Davidson et al. (2017) distinguent le discours haineux et le contenu offensant

précisément en raison de la subjectivité de la perception de l'offense. Ce qui peut être perçu comme offensant par un groupe de personnes peut ne pas l'être pour d'autres. Un exemple frappant de ce phénomène est l'utilisation de *slurs*. Davis et McCready (2020) définissent le *slur*⁴ comme un terme qui, entre autres, évoque un contenu expressif offensant qui peut être utilisé de façon dénigrante auprès d'un groupe en particulier, et ce groupe est défini par une caractéristique qui lui est intrinsèque, telle que la race (mot en *N*, *Chink*), le genre (salope, chienne), la sexualité (fif, tapette) etc. Le contenu expressif est le contenu qui rend un *slur* offensant de façon inhérente. C'est un contenu complexe qui contient les faits historiques, les stéréotypes et les attitudes sociales envers le groupe visé. Un *slur*, lorsque énoncé, évoque nécessairement tout le bagage du contenu expressif qui lui est associé sans égard pour l'intention de l'individu qui prononce, écrit ou signe le terme. L'impact du contenu expressif varie alors seulement selon le contexte au moment de l'énonciation : les intentions, les individus présents ou visés, le climat sociopolitique actuel, etc. Le défi de détection apparaît donc dans l'exemple de l'utilisation du mot en « n » pour désigner la communauté noire : bien que de nombreux systèmes de détection automatique le considèrent comme une manifestation de discours haineux, il est utilisé quotidiennement par de nombreuses communautés noires sans connotation haineuse (Davidson et al., 2017, 2019). Ce qui peut être considéré comme toxique ou non est donc largement variable d'une communauté à une autre, même lorsqu'il s'agit de termes qui, intrinsèquement, sont associés à du contenu offensant.

Un autre défi de taille en détection de toxicité et de TAL en général réside dans le fait que la plupart des modèles de langue sont entraînés sur l'anglais, principalement en raison de la disponibilité abondante de contenu en ligne dans cette langue, et une rareté de contenu dans d'autres langues. Même lorsque les modèles de détection sont entraînés sur plusieurs langues, comme Perspective API qui couvre 12 langues (Lees et al., 2022), les données d'entraînement présentent souvent des déséquilibres de distribution, avec une prédominance d'exemples en anglais. Cela met en lumière une fois de plus le problème de la subjectivité : lorsque des locuteurs natifs sont sollicités pour annoter les données, la localisation géographique des annotateurs peut avoir un impact significatif sur la perception de la toxicité (Davidson et al., 2017; Kowert, 2020; Kwak et al., 2015; Kwak & Blackburn, 2015). Pour contrer minimalement l'effet du contexte culturel et social de la société et de la présence de l'anglais, plusieurs modèles de détection

⁴ Nous choisissons à nouveau de ne pas traduire le terme *slur*, dont la description sémantique et pragmatique est détaillée dans le texte de Davis et McCready (2020). Proposer une traduction implique une analyse pour s'assurer que les éléments sémantiques se retrouvent adéquatement dans le terme proposé, ce qui n'est pas l'objet de la recherche actuelle.

automatique de toxicité multilingues ont été développés (De Smedt et al., 2018; Gevers et al., 2022; Lees et al., 2022), parfois avec un point de mire sur une langue ou un groupe de langue (Gevers et al., 2022; Jhaveri et al., 2022; Leite et al., 2020; Mubarak et al., 2017).

Finalement, en apprenant une représentation de la langue sur un nombre très large de données humaines, les prédictions issues des modèles de langue laissent émerger des biais humains (Bolukbasi et al., 2016; Caliskan et al., 2017; Davani et al., 2023; Davidson et al., 2019, 2019; Garg et al., 2023). Dans la détection de toxicité, cela est problématique puisque certains termes liés à l'identité d'une personne peuvent être identifiés de façon erronée comme étant toxiques. Ce sujet sera développé plus en profondeur dans le CHAPITRE 3 de ce mémoire.

2.6 Conclusion

Bien qu'on ne puisse offrir de définition générale objective de la toxicité, ce chapitre a permis d'explorer plus en profondeur le concept, en proposant une définition adaptée au contexte du jeu vidéo en ligne, où les clavardages entre joueurs et joueuses sont fréquemment utilisés. Le terme « toxicité » désigne, pour ce mémoire, tout comportement perturbateur dans les clavardages des jeux. Même lorsqu'assignée à cette définition, la toxicité reste un phénomène subjectif, influencé par des facteurs socioculturels et contextuels.

La section 2.3 présente les méthodes actuelles de détection de la toxicité, en se concentrant surtout sur les approches basées sur les modèles de langue. Comme discuté dans la section 2.4, la détection automatique de la toxicité à l'aide de ces méthodes est intrinsèquement liée aux informations contextuelles et connotations sociales encodées dans les plongements, notamment aux *stories* introduites au CHAPITRE 1. Ainsi, le caractère toxique d'un texte peut être plus efficacement observé dans les plongements en se concentrant sur la 11^e couche d'un modèle basé sur BERT.

Une limitation majeure de ces systèmes, détaillée à la section 2.5, est que les modèles sont généralement entraînés sur des données qui reflètent souvent des déséquilibres géographiques et sociaux. Par exemple, l'utilisation de termes comme le mot en « n » illustre la complexité de la détection de la toxicité, où un même mot peut être perçu comme offensant ou neutre selon le contexte et la communauté qui l'utilise. Le chapitre suivant se penche sur un type de biais introduit par cette limitation et qui, comme la toxicité, est fortement lié à la notion de *story*.

CHAPITRE 3

BIAIS : MODÈLES DE LANGUE DÉTECTION DE TOXICITÉ

Ce chapitre débute un peu de la même manière que le chapitre précédent : après une proposition de définition générale des biais, il sera question de restreindre cette définition aux modèles de langues et à un biais social plus spécifique à la recherche actuelle. Le chapitre se continuera ensuite en abordant comment ces biais se présentent spécifiquement dans la détection de toxicité. Finalement, nous survolerons les méthodes actuelles d'atténuation de biais afin de mieux comprendre ce qu'elles impliquent et les problèmes possibles de leur utilisation.

3.1 Définir les biais

Similairement à la toxicité, il n'y a pas de consensus universel sur la définition de biais. On retrouve la notion dans plusieurs domaines, par exemple dans l'étude de la cognition humaine et de la psychologie (Gilovich et al., 2002; Gratton & Gagnon-St-Pierre, 2020; MacCoun, 1998) ou en santé (Webster et al., 2022), et la définition peut varier selon son utilisation.

Une définition trop générale des biais serait insuffisante pour une discussion concrète ; il est donc nécessaire d'en restreindre le champ d'application. Dans le cas de la recherche décrite dans ce projet de mémoire, il est question des biais manifestés par des humains, des biais sociaux. Les biais sociaux se réfèrent aux stéréotypes ou aux préjugés qui influencent la perception, la cognition et le comportement des individus dans les interactions sociales (Doucerein, 2020; Webster et al., 2022). Encore une fois comme la toxicité, les biais sociaux se manifestent et sont perçus différemment selon le vécu, la culture et le climat socioculturel de chaque individu. Il convient d'ailleurs de noter que même la classification « biais sociaux » peut englober une diversité de mécanismes et de contextes. Un autre type de biais sera mentionné à quelques reprises au fil de ce texte, le biais d'association (Greenwald et al., 1998). Ce biais fait référence à la tendance des humains à relier plus facilement certains concepts entre eux, plutôt qu'à d'autres, souvent en raison de stéréotypes ou d'expériences préexistantes. Un exemple courant est l'association des insectes avec des émotions négatives, telles que le dégoût, la peur ou l'inconfort, tandis que les fleurs sont plus facilement liées à des sentiments positifs comme la tranquillité, la joie ou la douceur. Cet exemple est généralement inoffensif, toutefois nous retrouvons aussi des biais d'associations avec une implication sociale importante, comme l'association de prénoms féminins avec des termes en lien avec la

famille (parent, enfant, maison, etc.), alors que le genre masculin est associé à des termes en lien avec une carrière (science, professionnel, bureau, etc.)

Dans les sous-sections suivantes, en discutant des biais présents dans les modèles de langue et dans la détection de toxicité, nous affinerons davantage le champ d'application et la définition en décrivant spécifiquement le type de biais qui nous intéresse pour cette recherche.

3.2 Information linguistique : biais dans les plongements

Les représentations issues de modèles de langue sont obtenues en utilisant tous les textes d'entraînement donnés au modèle, et une multitude d'informations sémantiques y sont encodées. Déjà pour les plongements statiques, qui requièrent habituellement moins de données, Caliskan et al. (2017) ont identifié que des biais humains d'association émergeaient des propriétés latentes encodées dans les vecteurs de GloVe. En plus de retrouver les biais « inoffensifs » identifiés à la section précédente tels que *insectes-émotions négatives* et *fleurs-émotions positives*, on y retrouve les biais d'associations sociaux comme *femme-famille* et *homme-métiers*.

Pour un modèle Transformers comme BERT, il est question de corpus de plusieurs milliards de termes (Devlin et al., 2019). Les données proviennent entre autres du Common Crawl⁵, une ressource massive qui compile des informations provenant de diverses sources en ligne telles que des livres numérisés, des articles de presse, des publications, des commentaires et publications sur les réseaux sociaux, et bien plus encore. La vaste étendue de ces données est cruciale pour obtenir les performances remarquables des modèles de traitement de la langue. Cependant, contrairement à ce que l'on pourrait supposer, cette abondance de texte ne garantit pas nécessairement une diversité et une représentativité adéquates. En fait, comme souligné par Bender et al. (2021) dans une étude sur la taille impressionnante de ces modèles de langues, les données peuvent être biaisées et ne pas refléter de manière équilibrée la diversité de la population utilisant Internet. La Figure 3.1 offre une illustration du phénomène dans la traduction automatique, en prenant pour exemple la traduction des phrases anglaises « *My friend is a doctor. My other friend is a nurse* » vers le français à l'aide de Google Traduction⁶. Les termes *friend*, *doctor* et *nurse* en anglais ne portent pas de marque de genre et rien n'indique dans le contexte le genre des personnes

⁵ <https://commoncrawl.org/>

⁶ L'exercice de traduction a été effectué le 05 juin 2024 sur <https://translate.google.com/>

référéées. Néanmoins, dans la traduction en français, la première phrase est au masculin avec *ami* et *médecin* alors que la deuxième phrase est au féminin avec *amie* et *infirmière*. Pour ajouter au problème, la plateforme complète de Google Traduction offre la possibilité de sélectionner une phrase traduite pour voir les autres options de traduction. Toutefois, les biais sont toujours présents : en sélectionnant la phrase « Mon autre amie est infirmière », les choix sont ceux observés dans la Figure 3.2, avec le genre de *ami* pouvant changer au masculin, mais le genre d'*infirmière* est inchangé. Le biais d'association, voulant que les femmes soient plus facilement associées au métier d'infirmier et les hommes au métier de médecin, est présent dans les données et se retrouve mis en évidence dans les résultats de l'application du modèle. Comme l'exemple le démontre, les applications des modèles de langue doivent être faites avec prudence et en avisant la population des biais qui vont nécessairement affecter les performances et les résultats.

Figure 3.1 Traduction des mots *friend*, *doctor* et *nurse* sans flexion de genre en anglais à ami/médecin — au masculin et amie/infirmière — au féminin.

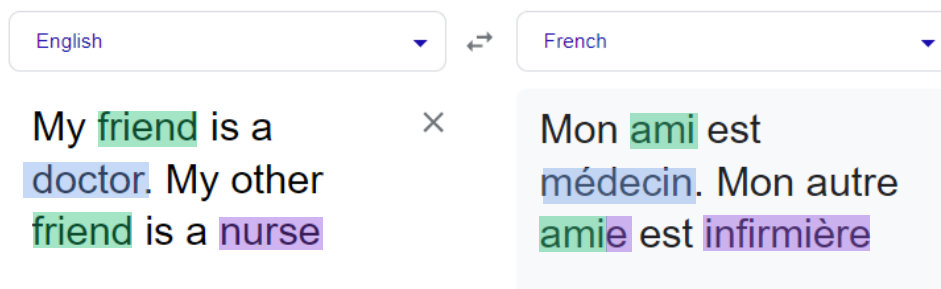


Figure 3.2 Autres choix pour la traduction de la deuxième phrase : malgré le changement de genre du terme *ami*, le mot *infirmière* reste au féminin.

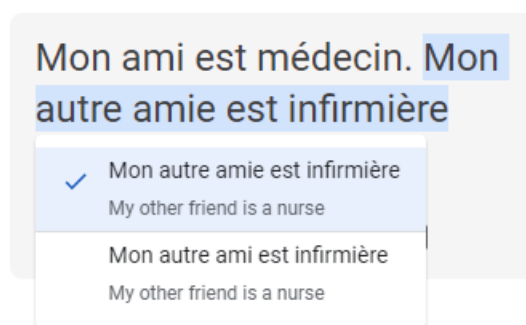
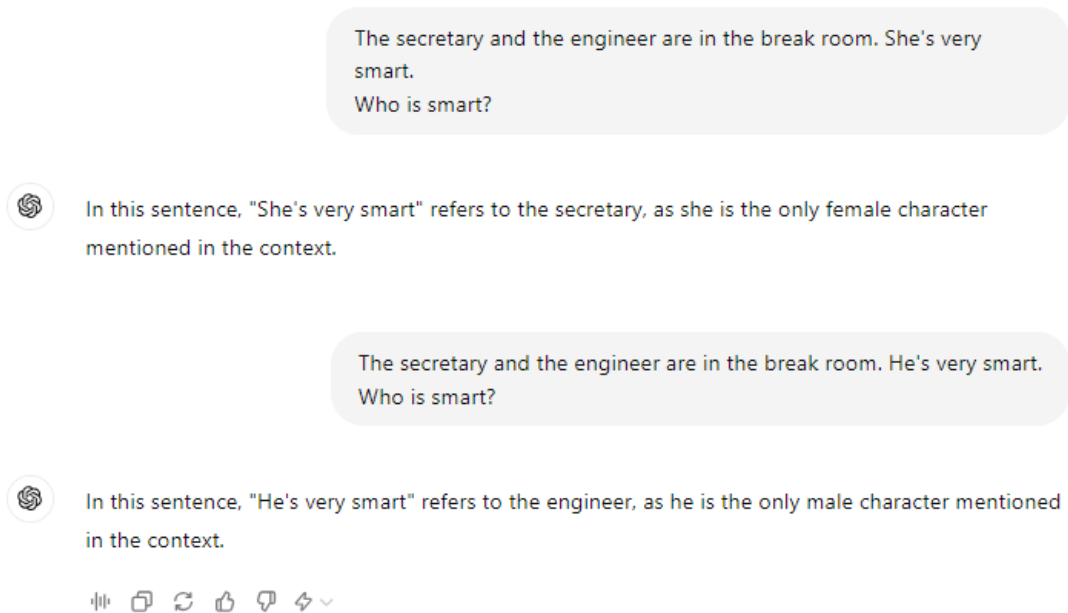
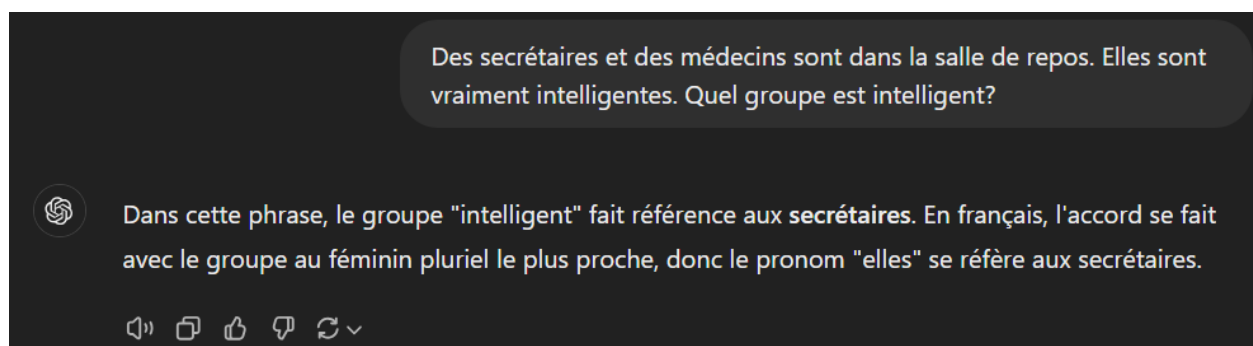


Figure 3.3 Les réponses de ChatGPT montrent la présence d'une association engineer-homme, secretary-femme.



Le test peut également être réalisé avec l'application ChatGPT, dans la version gratuite 4o-mini (2024). Cette fois-ci, les termes « engineer » et « secretary » sont utilisés pour voir l'association présente dans les plongements de ChatGPT. La Figure 3.3 montre les réponses obtenues dans l'application, qui indique que « secretary » fait référence à un personnage féminin, et « engineer » à un personnage masculin, même si aucune information sur le genre des personnages n'a été mentionnée. Le test est fait en anglais pour la simple raison que plusieurs termes liés aux emplois en français varient selon le genre de la personne. Dans la Figure 3.4, des métiers épicènes sont mis au pluriel pour éviter l'utilisation de déterminants qui pourraient révéler le genre. Avec le pronom « Elles », la réponse obtenue est claire que les secrétaires sont intelligentes. La même question en utilisant « Ils » est différente, puisque « Ils » peut se désigner les deux groupes à la fois.

Figure 3.4 Les réponses obtenues dans l'application ChatGPT montrent les biais en français également.



Rappelons que ces biais ne sont pas présents que dans les plongements du modèle derrière ChatGPT, ils se retrouvent d'abord chez les humains, d'où leur présence dans les GML. Les biais sociaux ne sont pas seulement reproduits et renforcés dans les applications des modèles de langue par leur présence dans les applications, ils sont également amplifiés. Les plongements issus de ces modèles capturent les associations entre des termes à partir de statistiques distributionnelles, mais le signal des associations est amplifié (Caliskan et al., 2017). Shah et al. (2020) expliquent en effet le phénomène de suramplification des biais lors de la création des plongements, dans les phases d'« apprentissage » comme le préentraînement ou l'ajustement fin. Une différence, même minime, entre des attributs humains (p. ex. : âge, origines, personnalité) dans les données est exacerbée dans les prédictions issues de l'application du modèle. Nul besoin de se plonger dans les grands modèles de langues récents pour observer le phénomène : en 2014, la compagnie Amazon souhaitait améliorer son processus de recrutement à l'aide d'un système issu de l'IA qui permettrait d'attribuer des scores aux nombreux CV reçus selon leur contenu (Dastin, 2018). Les données d'entraînement du modèle consistaient en une grande quantité de CV de développeurs en logiciel reçus par la compagnie dans les 10 dernières années. L'utilisation du mot « développeurs » au masculin est volontaire ici, puisque la majorité des CV utilisés provenaient d'hommes. Les scores attribués automatiquement aux CV étaient ainsi plus favorable si le CV de l'employé en attente d'embauche contenait des termes souvent utilisés par des ingénieurs masculins (p. ex. : *executed* ou *captured*). On retrouve ainsi un biais humain — les hommes étant plus souvent associés à des emplois dans les domaines technologiques — exacerbé dans l'application d'un modèle. Le phénomène d'amplification a été observé à plusieurs reprises, autant dans des modèles de langue (Angwin et al., 2016; Bordia & Bowman, 2019; Kurita et al., 2019) que des modèles de reconnaissance visuelle (Zhao et al., 2017) et de génération d'image (Sun et al., 2023).

Blodgett et al. (2020) résumant les différentes approches pour appréhender les biais en TAL, en mettant en lumière l'importance de définir précisément le concept de biais, d'identifier les personnes potentiellement affectées par ces biais, ainsi que les mécanismes impliqués. Pour se conformer à ces suggestions, il est essentiel de premièrement comprendre les sources de biais pertinentes. Suresh et Gutttag (2021) proposent une taxonomie exhaustive des types et sources de biais qui peuvent survenir à chaque phase de création d'un système utilisant l'apprentissage machine. Ces sources de biais peuvent être regroupées en deux catégories principales : la génération de données et la conception et l'implémentation d'un modèle. Cette dernière catégorie ne sera pas traitée dans cette recherche. Un exemple notable de biais survenant dans la génération de données est le biais historique. Il se manifeste lorsque les données (telles que des livres, des journaux, etc.) sont produites par des êtres humains à une certaine époque, puis deviennent désuètes à l'époque actuelle, causant du tort à certaines communautés (p. ex. : les stéréotypes impliquant les femmes au foyer). Un autre type de biais, connu sous le nom de biais de représentation, se produit lorsque les données utilisées pour l'entraînement sont sélectionnées de manière non représentative, comme pour l'entraînement de larges modèles de langues sur les données du Common Crawl. Cela peut entraîner une distorsion dans les résultats du modèle, car il ne capture pas adéquatement la diversité ou la réalité du domaine qu'il est censé représenter.

Le biais de représentation, en tant que biais social, provoque un déséquilibre dans les données, donnant surtout la parole aux opinions et communautés dominantes. Cela se traduit par une prédominance de certains groupes sociaux, qui expriment également leurs opinions de manière plus significative. Par exemple, sur les forums en ligne, nous observons une surreprésentation des hommes anglophones résidents aux États-Unis ou au Royaume-Uni, âgés de 18 à 29 ans, par rapport à d'autres groupes démographiques (Bender et al., 2021). Pour ajouter à la situation, des corpus dits « propres » sont créés, comme The Colossal Clean Crawled Corpus (Raffel et al., 2020), en filtrant des mots potentiellement liés au discours haineux. Cependant, les termes retirés font parfois référence à des thèmes jugés obscènes par quelques individus, mais importants pour certaines communautés. Par exemple, des termes injurieux qui ont été revendiqués par ces mêmes communautés, tels que « queer » et « twink » pour certaines personnes LGBTQ+ (Bender et al., 2021). Cet exemple de biais de représentation explique en partie les biais particuliers qu'on peut retrouver dans la détection de toxicité, abordé dans la section suivante.

3.3 Biais en détection de toxicité

Les biais identitaires font partie des biais de représentation et se retrouvent dans les modèles de détection de toxicité (Blodgett et al., 2020; Davidson et al., 2019; Garg et al., 2023; Kiritchenko & Mohammad, 2018; Prabhakaran et al., 2019; Sap et al., 2019). On retrouve ces biais puisque les données utilisées pour l’ajustement fin de ces modèles contiennent avec une surabondance de certains termes liés à des caractéristiques identitaires — telles que la couleur de peau (noir, blanc, brun), le genre (femme, homme, trans, non-binaire), l’orientation sexuelle, et ainsi de suite — étiquetés comme toxiques en raison de leur utilisation préjudiciable par les voix dominantes. En conséquence, lorsque des modèles de détection automatiques sont également utilisés pour cacher et effacer les textes identifiés comme étant toxiques, les biais sont reproduits et amplifiés : une personne homosexuelle pourrait ne pas pouvoir discuter de sujet concernant son identité, par exemple le mariage gai, puisque les termes « gai » ou « lesbienne » sont identifiés comme étant toxiques (Wang et al., 2014; Zhou et al., 2021). Ce phénomène souligne l’importance cruciale d’une annotation équilibrée et sensible ainsi que d’une surveillance constante des biais lors de la conception et de l’entraînement de tels modèles. De plus, ce type de biais est souvent abordé dans les différents travaux comme étant binaire : on évalue les biais de genre « homme » VS « femme », les biais de race « blanc » VS « noir » (Davidson et al., 2019; Garg et al., 2023; Kiritchenko & Mohammad, 2018; Park et al., 2018; Sap et al., 2019). Dès le départ, cela force à mettre de côté des communautés complètes, par exemple les personnes non-binaires, où les personnes qui ne sont ni « noirs » ni « blancs ». Pour résumer en utilisant les recommandations de Blodgett et al. (2020), les biais observés dans les modèles de toxicités sont des biais identitaires. Ils peuvent affecter les communautés qui sont parfois déjà marginalisées en les invisibilisant ou en promouvant des opinions d’apparence dominantes.

Dans le cadre d’un stage auprès de la compagnie de développement de jeux vidéo Ubisoft, Van Dorpe et al. (2023) évaluent les biais du modèle de détection de toxicité basé sur BERT, nommé Toxbuster (Yang et al., 2023), développé au sein du département de recherche et développement de la compagnie. L’évaluation visait à examiner la « réactivité » du modèle aux termes liés à l’identité, qui peuvent potentiellement véhiculer des biais. La réactivité est analysée en vérifiant si les prédictions issues de Toxbuster étaient trop agressives ou trop passives quant à la toxicité de ces termes. La recherche a révélé, par exemple, que des termes tels que « asian », « lesbian », « trans », « mexican », « gay » et « black » étaient excessivement associés à l’étiquette « toxique », tandis que le terme « yellow » était insuffisamment associé à cette étiquette (Van Dorpe et al., 2023). Il a aussi été identifié que ces biais

provenaient principalement d'une sur ou sous-représentation de ces termes en tant que toxiques dans les ensembles de données utilisés pour l'ajustement fin. Les données d'entraînement présentaient un nombre beaucoup plus élevé de lignes toxiques que de lignes non toxiques qui contenaient les termes biaisés identifiés. Pendant la phase d'apprentissage du modèle, la présence fréquente du terme dans les phrases toxiques a été encodée comme étant un motif : une ligne a plus de chance d'être toxique si le terme est dans la phrase que s'il n'y est pas. Compte tenu du fait que ces termes sont souvent utilisés dans des contextes toxiques et peuvent véhiculer des biais identitaires, il peut être difficile de justifier une réduction de la présence de biais déjà existants.

3.4 Atténuation de biais et difficultés associées

La présence de biais dans les modèles de langue en général et dans les modèles de détection de toxicité pose un problème important, surtout lorsque les modèles sont déployés et utilisés par des industries ou par des utilisateurs qui ne connaissent pas les risques associés. Dans une démarche visant à accroître l'inclusivité des modèles, diverses techniques sont mises en œuvre pour atténuer les biais. L'étude des biais et de leur atténuation est souvent retrouvée sous le nom de *Fairness* dans les différents articles. De façon similaire à la source des biais, les méthodes diffèrent selon leur type d'approche et selon l'étape à laquelle elles sont appliquées, notamment les phases avant, durant ou après l'entraînement d'un modèle (Garg et al., 2023; Hort et al., 2024). Dans tous les cas, les approches sont majoritairement réactives : l'atténuation est faite une fois les biais identifiés.

3.4.1 Techniques utilisées

Les recherches visant à atténuer les biais avant la création du modèle tentent généralement de prévenir les biais d'échantillonnage lors de la collecte des données et des étiquettes nécessaires à l'entraînement (Sap et al., 2019; Zhou et al., 2021). Il est important de noter que ces méthodes sont parfois appliquées après l'entraînement, en réaction à une évaluation des biais identifiés. Parmi les méthodes de modification d'échantillonnage utilisées, on retrouve l'ajustement de l'importance de chaque échantillon (Du & Wu, 2021; Yu et al., 2023), ainsi que le retrait ou l'ajout de données. Souvent, l'ajout de données consiste à intégrer davantage de données correspondant à des groupes minoritaires et sous-représentés dans l'ensemble de données. Dixon et al. (2018), par exemple, tentent d'atténuer les biais dans un modèle de détection de toxicité en équilibrant leur jeu de données. Ils extraient des exemples non toxiques de termes identitaires à partir d'articles de Wikipédia, en supposant que ces articles sont non toxiques en raison de leur présence sur le site, indiquant que la publication a été approuvée. Cependant, la question permettant

de savoir si de nouveaux biais ont été introduits à la suite de ces interventions n'a pas été abordée (Garg et al., 2023). Zhao et al. (2017) ont utilisé une technique d'augmentation de données pour atténuer un biais de genre en inversant les termes genrés dans des phrases contenant des termes associés aux femmes ou aux hommes. Ils créent donc des exemples où le mot « cooking » est retrouvé plus souvent avec des termes masculins, pour diminuer l'association présente avec les termes féminins.

D'autres chercheurs se concentrent sur la diminution des biais à l'étape d'entraînement du modèle. Ces techniques impliquent des régularisations et des modifications des algorithmes d'entraînement. Des travaux passent par l'optimisation d'hyperparamètres (paramètres de configurations) de leur modèle pour augmenter leur niveau de *fairness* (Cruz, 2020; Perrone et al., 2021; Tizpaz-Niari et al., 2022). Zhao et al. (2018) altèrent la fonction de perte d'un modèle GloVe (Pennington et al., 2014) pour que, entre autres, les vecteurs de termes qui ne sont habituellement pas associés à un genre en particulier soient orthogonaux à la direction de vecteurs de genre.

Certaines méthodes d'atténuation de biais sont appliquées une fois l'entraînement des modèles terminé. Plusieurs de ces approches tentent de corriger les prédictions, notamment en appliquant diverses façons de gérer le seuil d'activation du modèle. Kamiran et al. (2012, 2017) utilisent une méthode qui permet de modifier la prédiction du modèle si elle est près du seuil. L'ajustement est appliqué de cette façon : la prédiction est favorable si le texte réfère à un individu qui appartient à un groupe discriminé, mais défavorable si l'individu appartient à un groupe privilégié. Bolukbasi et al. (2016) et Park et al. (2018), quant à eux, approchent l'atténuation de biais en manipulant les plongements issus de l'entraînement d'un modèle Word2vec. Comme Zhao et al. (2018), ils considèrent qu'un biais de genre est atténué si la distance entre des termes étant associés aux femmes (p. ex. : termes reliés à la famille ou aux arts) et les termes plus souvent associés aux hommes (p. ex. : termes reliés à la carrière ou à la science) est réduite dans l'espace vectoriel. L'objet principal de ces techniques est l'ensemble des termes problématiques qui véhiculent un biais choisi : pour un biais de genre, les plongements modifiés sont ceux des termes qui sont associés à des groupes selon leurs propriétés intrinsèques. Comme mentionné ci-haut, plusieurs techniques utilisées avant l'entraînement sont aussi utilisées après, une fois les biais identifiés, pour tenter de réentraîner le modèle.

3.4.2 Problèmes rencontrés

L'atténuation des biais sur les performances des modèles implique une baisse de performance dans des modèles de classification de texte, et donc affecte les prédictions obtenues par l'utilisation du modèle (Berk et al., 2017; Haas, 2019; Hort et al., 2024). Garg et al. (2023) mentionnent qu'une combinaison de plusieurs techniques comme l'atténuation des biais dans des vecteurs statiques comme word2vec et inverser deux groupes dans les phrases aident à augmenter la performance, mais ne permet pas d'atténuer les biais de façon aussi puissante. Des solutions comme celles-ci tentent de diminuer l'effet du compromis entre performance et la présence de biais. Cependant, un choix doit toujours être fait entre les deux. L'atténuation cause donc une perte de l'information contenue dans les plongements, information essentielle pour le bon fonctionnement du modèle. Pourtant, plusieurs facteurs font de l'atténuation réactive de biais une tâche ardue et souvent superficiellement efficace. Gonen et Goldberg (2019), dans un article intitulé *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them*, démontrent entre autres que les techniques de correction de Bolukbasi et al. (2016) et Zhao et al. (2018) ne font que masquer le problème. Leur recherche permet de prouver que même si le modèle semble présenter moins de biais évidents dans leur performance, ils demeurent toujours encodés dans les plongements.

Parmi les techniques énoncées pour atténuer les biais, et ainsi enlever l'information comme quoi le terme est biaisé, on retrouve plusieurs exemples où il est tenté de rendre la langue et les termes qui véhiculent un biais plus « neutre ». Pour ce faire, un corpus « neutre » qui contient des contre-exemples aux biais est sélectionné pour égaliser les représentations, comme l'ont fait Zhao et al. (2017). Toutefois, ces approches nous confrontent aux problèmes décrits à la section 2.5, notamment le risque que l'application d'un filtre quelconque sur un corpus d'entraînement puisse avoir des effets néfastes sur des communautés déjà marginalisées. L'augmentation de données peut également introduire involontairement d'autres biais, qui ne sont pas toujours évalués par la suite.

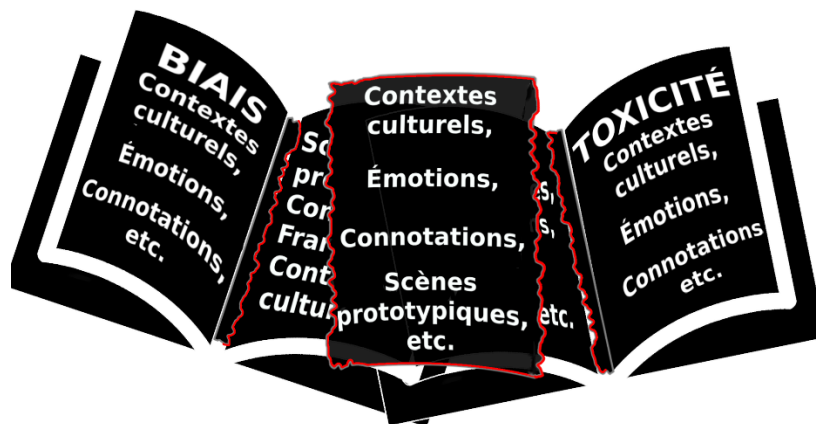
3.4.3 Neutraliser la langue pour atténuer les biais

L'intuition de neutraliser les termes vient de l'idée qu'il existe une contrepartie non toxique ou non biaisée à ces termes. Comme nous l'avons vu à la section 1.3.2 et 1.3.4, les *frames* sont une façon de conceptualiser des informations attachées à des mots. Nous avons noté que les termes « femme » et « femoid » partagent une *frame*, la référence à une personne de genre féminin, mais se distinguent par la connotation péjorative de la *frame* de « femoid ». La contrepartie « neutre » de « femoid » serait alors

logiquement « femme », car en éliminant la connotation toxique de « **femoid** », on revient à la frame de base qui désigne simplement une personne sans jugement négatif. Cependant, en prenant en compte les notions de sémantiques plus dynamiques comme les *stories*, il devient évident que cette approche néglige des éléments importants présents dans les plongements et les contextes des mots.

Erk et Chronis (2022) avancent l'idée que les biais sociaux se manifestent principalement au travers d'une forme de *story* impliquant des associations culturelles et des jugements. Comme mentionné à la section 1.3.4, les points de vue et les jugements des humains sur le monde, sur ce qu'ils déterminent utile et inutile, etc. sont tous des éléments encodés dans les plongements. Retirer le biais associé à un mot est donc une tentative de retirer et modifier sa *story*, dans un espoir de rendre la langue plus neutre et du même coup, plus inclusive. Dans le cas d'atténuation de biais dans un modèle de détection de toxicité, cette situation est particulièrement délicate, car les biais et la toxicité peuvent partager des éléments de la *story* d'un terme. Atténuer les biais et les rendre « neutres » retire la connotation négative du terme, ce qui entraîne alors une perte d'information qui affecte également la capacité du modèle à identifier la toxicité. En d'autres termes, comme l'illustre la Figure 3.5 si l'on réduit les biais associés à certaines identités ou groupes, on risque de diminuer la sensibilité du modèle à l'utilisation toxique de ces termes, car les *stories* de la toxicité et des biais peuvent se chevaucher. Cependant, selon Beaver et Stanley (2023), la neutralité de la langue n'est pas atteignable. Ils affirment que les mots sont imbriqués dans la pratique de la langue et, à ce titre, véhiculent une idéologie. Il n'existe donc pas de mot neutre dans une langue humaine. L'utilisation d'un terme ne peut en effet pas être neutre, car il évoque inévitablement sa *story*.

Figure 3.5 Retirer le caractère biaisé ou toxique d'un terme revient à déchirer une page de sa *story*, qui contient autant des éléments appartenant aux biais qu'à la toxicité.



Les *slurs* illustrent bien la difficulté de trouver une contrepartie « neutre » à des termes ayant un contenu expressif aussi complexe. La tâche ne se limite pas à remplacer le terme offensant par une désignation neutre du groupe visé, comme utiliser « noir » pour le *slur* du mot en « n », ou bien, dans le cas de biais par un Ashwell (2016) mentionne que les *slurs* sont souvent trop attachés à des normes sociales indiquant comment les gens devraient se comporter, ainsi il n'existe simplement pas d'équivalent neutre. Prenons le *slur* anglais *slut*. Ce terme désigne une femme qui a plusieurs partenaires sexuels, ou du moins qui semble avoir plusieurs partenaires, et se rattache à une norme sociale qui veut qu'il ne soit pas acceptable qu'une femme agisse ainsi. Selon la définition d'un *slur* de Davis et McCready (2020), l'information comme quoi une femme recevant l'injure appartient à la description de *slut* n'est pas nécessairement offensante, mais le contenu expressif qui lui est rattaché contient la norme, le jugement que les personnes répondant à la description devraient subir. Or, ces normes sont évoquées automatiquement lors de l'énonciation du terme, et ne sont pas dissociables. De plus, du point de vue de la sémantique formelle, il n'existe pas de terme qui ait une simple dénotation, dénuée de toute connotation. Ainsi, il semble futile de tenter de dissocier complètement un terme de son contexte, de sa *story*, et de rendre ce terme totalement neutre.

3.5 Conclusion

Ce troisième chapitre a permis de définir et comprendre les biais dans les modèles de langue, en se concentrant particulièrement sur les biais identitaires et leur manifestation dans la détection de toxicité. Ces biais affectent de manière disproportionnée les communautés marginalisées. Ils proviennent et sont renforcés par des données d'entraînement déséquilibrées.

La section 3.4 aborde les méthodes d'atténuation des biais, soulignant leurs limites et les compromis entre la performance des modèles et la réduction des biais. Les tentatives de neutralisation de la langue se heurtent aux *stories* déjà encodées dans les mots et les plongements.

Le chapitre a montré que les biais dans la détection de toxicité exigent une approche nuancée. Les concepts de biais identitaires et de toxicité sont fortement liés au concept de *story* : les connotations, jugements et contextes encodés dans les plongements s'entremêlent pour former une part essentielle de ce qui rend un mot à la fois identitaire et potentiellement toxique. Ces interactions sont au cœur du chapitre suivant, qui présente la formulation des questions de recherche.

CHAPITRE 4

QUESTION DE RECHERCHE

Les espaces en ligne, notamment les clavardages de jeux vidéo multijoueurs, sont fréquemment confrontés à au moins deux problèmes majeurs : la toxicité et les biais sociaux. Ces deux phénomènes nuisent aux communautés qui utilisent ces espaces et créent un environnement hostile. Lutter contre l'un de ces problèmes peut involontairement exacerber l'autre : détecter la toxicité pour la diminuer implique que la possibilité d'introduire des biais identitaires augmente, alors qu'éviter les biais implique de ne pas tenter de diminuer la toxicité avec les technologies actuelles efficaces. C'est un cercle vicieux indésirable qui rend la gestion de ces espaces particulièrement difficile. C'est pourquoi des techniques d'atténuation des biais ont été développées pour tenter de trouver un équilibre entre ces deux maux.

Un dilemme additionnel est introduit en utilisant les techniques d'atténuation de biais, car l'efficacité de la détection de toxicité diminue à un rythme comparable à la diminution de la présence des biais. Le réflexe des solutions actuelles d'atténuation de biais est de neutraliser les termes biaisés, et ainsi de retirer ou diminuer la connotation négative des termes biaisés. Les concepts explorés progressivement au fil des chapitres de ce mémoire démontrent pourtant que ceci paraît être un exercice futile. Sans avoir comme but la neutralisation d'un terme biaisé, existe-t-il des éléments qui pourraient être retirés ou modifiés dans les représentations des termes afin de réduire les biais identitaires tout en minimisant l'impact sur les informations relatives à la toxicité ? Obtenir des éléments de réponse à cette question nécessite une meilleure compréhension des interactions toxicité-biais identitaires dans les plongements.

Avant même de commencer à observer les interactions toxicité-biais dans les plongements, il est nécessaire de comprendre cette relation comme elle se présente dans le contexte des clavardages de jeux vidéo. Comme vu dans les chapitres précédents, observer une telle relation est possible à l'aide d'une tâche de substitution : pour un terme possiblement biaisé relié à l'identité d'une personne, nous cherchons des termes substitués qui partagent un lien de relation avec le terme biaisé, et donc qui peuvent être utilisés dans un même contexte sans en changer le sens. Afin de qualifier la relation entre le terme biaisé et la toxicité, nous devons obtenir des substitués pour des occurrences du terme biaisé autant dans des lignes de clavardages toxiques que des lignes non toxiques. À partir des substitués obtenus, comme l'ont fait Kremer et al. (2014) ainsi que Chronis et Erk (2020), il sera possible d'identifier les sens possibles du terme biaisé et de comparer ces sens dans des contextes toxiques VS non toxiques.

Comme exploré précédemment dans ce mémoire, cette relation entre termes cibles (termes biaisés) et termes substitués en contextes toxiques ou non devrait se manifester dans les plongements sous forme de *story* : un ensemble dynamique de connotations, de jugements, de contextes culturels, etc. Les sens identifiés suite à une tâche de substitution permettent de caractériser les relations entre les termes biaisés et leurs substitués, et ainsi nous permettent de connaître ce qui devrait se retrouver dans une *story* pour vérifier si celle-ci est bien encodée. Nous cherchons ainsi à répondre aux questions suivantes :

1. Les *stories* associées aux termes véhiculant un biais identitaire dans les clavardages de jeux vidéo et à leurs substitués, caractérisées par les représentations qui sont extraites d'un modèle de langue, se distinguent-elles significativement des *stories* dérivées des substitués ?
2. Les plongements reflètent-ils les différences entre *stories* toxiques et les *stories* non toxiques ?

Répondre à ces questions fournira une compréhension plus précise des liens entre la toxicité et les biais identitaires, en clarifiant ce qui compose la *story* des termes biaisés. Cela permettra aussi de déterminer si les éléments associés aux biais – par exemple, le sens relié à l'identité - peuvent être dissociés de ceux qui caractérisent une phrase toxique.

Si la réponse à la première question montre que les *stories* dérivées des substitués ne diffèrent pas de celles encodées dans les plongements, cela signifiera que ces dernières capturent bien la *story* des termes biaisés, et que certains éléments ou significations peuvent être isolés. En revanche, si les différences sont marquées, cela pourrait indiquer l'émergence d'une nouvelle *story* dans les plongements, nécessitant une analyse plus approfondie pour comprendre son recoupement avec celle des substitués. Quant à la deuxième question, si les plongements ne reflètent pas les différences entre *stories* toxiques et non toxiques observées dans les substitués, cela suggérerait que la relation entre toxicité et biais est trop étroitement liée dans les plongements pour permettre une neutralisation des biais sans compromettre la détection de la toxicité. Pour obtenir une réponse à ces questions, nous détaillons la méthode utilisée dans le prochain chapitre.

CHAPITRE 5

MÉTHODE

Pour identifier et qualifier les *stories* associées à des termes biaisés et répondre aux deux questions énoncées, nous collectons d’abord des substituts lexicaux dans des contextes variés, incluant à la fois des phrases toxiques et non toxiques. Une fois les substituts collectés, nous procédons à des analyses qualitatives pour recueillir une première compréhension des *stories* véhiculées par les termes possiblement biaisés en identifiant les différents sens. Par la suite, afin de vérifier si la *story* est encodée dans les représentations linguistiques des modèles, nous obtenons les plongements des substituts et appliquons une méthode de regroupement (*clustering*) pour vérifier si un modèle de langue capture bel et bien les subtilités de la *story* associée à ces termes. Pour ce chapitre, les termes « ligne » et « phrase » sont utilisés de façon interchangeable pour désigner une entrée de clavardage.

5.1 Collecte de substituts

En suivant la méthodologie de Kremer et al. (2014), nous collectons des substituts lexicaux pour des termes susceptibles de véhiculer un biais identitaire dans le contexte de clavardages de jeux vidéo. L’objectif est d’obtenir plusieurs substituts pour chaque terme cible. Pour ce faire, il est d’abord nécessaire de constituer un jeu de données composé de lignes de clavardages annotées comme toxiques ou non toxiques, afin de permettre une comparaison entre ces deux types de contextes.

En tant qu’auteurice de ce mémoire, bien que je parle régulièrement anglais et que je sois familière avec le contexte des jeux vidéo, le français est ma langue première et je ne suis ainsi pas la mieux placée pour identifier un grand nombre de substituts en anglais dans des situations aussi spécifiques. De plus, les jeux vidéo multijoueurs compétitifs ne sont pas dans mes intérêts principaux et je n’ai que peu d’expérience avec les clavardages de ces jeux, à part quelques exemples vus lors de mon stage. Afin de garantir une collecte complète et représentative, nous avons donc fait appel à des participant-es répondant à des critères spécifiques.

5.1.1 Terme cible

Les termes pour lesquels nous cherchons à obtenir des substituts sont ceux pouvant véhiculer un biais identitaire. Comme mentionné dans la section 3.3, Van Dorpe et al. (2023) ont développé un jeu de données permettant d’évaluer les termes pour lesquels un modèle de détection automatique de toxicité

est le plus réactif. Cette banque, constituée de 46 termes, a été établie en consultation avec les groupes ressources d'employés chez Ubisoft. Ces groupes rassemblent des individus partageant des caractéristiques ou des expériences de vie similaires (couleur de peau, origine, genre, etc.). La consultation des membres de plusieurs de ces groupes nous a permis d'obtenir un point de vue diversifié et informé pour confirmer la pertinence des termes et leur rapport à l'identité d'une communauté.

Cependant, pour que notre analyse soit vraiment pertinente, il est essentiel de s'assurer que ces termes peuvent effectivement véhiculer un biais identitaire dans des contextes toxiques ou non toxiques. Parmi la banque de termes, nous avons sélectionné « gay », dans le sens de « l'orientation sexuelle d'une personne attirée sexuellement ou romantiquement par des individus du même genre », comme terme cible en raison de sa fréquence dans les données décrites en 5.1.2, où il apparaît dans des contextes toxiques comme non-toxiques. Malgré une surreprésentation du terme dans un contexte toxique, nous avons identifié suffisamment de lignes non toxiques pour obtenir un jeu de données équilibré. Van Dorpe et al. (2023) ont également constaté un biais dans la détection de toxicité par un modèle automatique, qui marque plus souvent les lignes comme toxiques lorsqu'elles contiennent ce terme, même lorsqu'elles ne le sont pas.

Nous limitons le nombre de termes à l'étude à un seul afin d'éviter que les participants et participantes ne se brûlent à la tâche et que la qualité des annotations ne soit compromise par la fatigue cognitive. En choisissant un unique terme, nous pouvons aussi faire une analyse plus approfondie de ses utilisations et des substitutions obtenues.

5.1.2 Jeu de données

Dans le cadre de cette recherche, un jeu de données disponible en libre accès aurait été intéressant, car il aurait permis de bénéficier de données préalablement annotées et validées par la communauté scientifique, garantissant ainsi une certaine fiabilité et reproductibilité des résultats. Nous avons donc initialement récupéré le jeu de données CONDA (Weld et al., 2021), composé de lignes de clavardage du jeu DOTA 2 (Defense Of The Ancients 2) déjà annotées pour la toxicité. C'est, à notre connaissance, le seul jeu de donnée de cette nature en libre accès au moment d'effectuer cette recherche. Toutefois, ce jeu de données ne s'est pas avéré adéquat pour notre recherche. En effet, peu de lignes contenaient des termes possiblement biaisés, et les annotations se limitaient à des catégories telles que « toxicité implicite », « toxicité explicite », « action » ou « autre ». À un niveau plus granulaire, chaque terme avait sa propre

annotation précisant des catégories. Parmi les quelques lignes où « gay » apparaissait, les annotations manquaient de clarté sur le caractère toxique ou non de la ligne, rendant le contexte difficile à évaluer. Certaines annotées « autre » (et ainsi, possiblement non toxiques) ne contenaient que le terme cible, par exemple « gays? » avec une annotation supplémentaire au niveau du terme indiquant (T) pour toxique.

Après avoir recherché un jeu de données plus approprié, nous avons finalement obtenu un ensemble de données de clavardages de deux jeux multijoueurs compétitifs, gracieusement fourni par un producteur de jeux vidéo qui a demandé à rester anonyme. Les annotations de toxicité de ce jeu de données ont été recueillies à la fois auprès d'employés de la compagnie et de participant-es externes. Le recrutement des participant-es externes a été organisé de manière à maximiser la diversité, afin d'assurer une perspective aussi complète que possible. Chaque ligne a été annotée plusieurs fois, permettant ainsi de garder l'annotation la plus populaire.

Nous y avons extrait 20 lignes contenant le terme cible *gay* et ajouté, lorsque possible, 8 lignes avant et 8 lignes après chaque occurrence pour enrichir le contexte, tel qu'exemplifié dans le Tableau 5.1. Le choix d'une sélection de 20 lignes est entre autres expliqué à la section 5.1.4 de ce chapitre : la tâche effectuée à partir de ces lignes est répétitive, et il est nécessaire de limiter le nombre de lignes pour s'assurer d'obtenir des réponses pertinentes. 20 lignes permet d'obtenir 10 phrases toxiques et 10 non toxiques, ce qui offre une bonne balance entre monotonie de la tâche et obtenir suffisamment de points de comparaison entre toxique et non toxique. Parmi les phrases du jeu de données contenant le terme cible, nous avons extrait les premières qui contenaient quatre mots et plus afin d'avoir suffisamment de contexte. Encore une fois pour éviter la monotonie de la tâche, nous avons également ajouté quatre phrases contenant un autre terme, *black*, en tant que phrases leurres offrant une distraction. Ces phrases et leur contexte ont été choisis avec les mêmes critères que les phrases cibles contenant *gay*, notamment les premières qui contenaient au minimum quatre mots, avec une condition de plus : le terme devait être utilisé de façon à faire référence à la communauté noire plutôt que seulement la couleur (p. ex. : « a black gun » n'a pas été retenu). Le jeu de données original ayant un accent sur le terme *gay*, nous n'avons pas extrait les annotations de toxicité pour les phrases contenant *black*.

Pour conserver l'anonymat des joueur-euses, ainsi que pour l'anonymat de l'identité des jeux dont elles proviennent, les lignes ont été modifiées. Les pseudonymes des joueur-euses ont été remplacés par « Player_1 », le chiffre changeant selon l'ordre d'apparition. Lorsqu'un pseudonyme est mentionné dans

le clavardage, le terme correspondant a été utilisé. Les noms des personnages, d'armes, de gadgets, d'habiletés et d'endroits retrouvés dans les jeux ont été retirés et remplacés par des mots génériques entre accolades (p. ex. : {character} please come over at the {map location}). Au total, le jeu de données contient 24 lignes : 20 avec le terme cible (divisé en 10 lignes toxiques, 10 non toxiques) et 4 qui sont des éléments distrayeurs. Le jeu de données complet est présenté dans l'Appendice C.

Tableau 5.1 Exemples de phrases du jeu de données, accompagnés de leur contexte et annotation de toxicité (1 = toxique et 0 = non toxique)

Tox	Ligne-cible	Contexte précédent	Contexte suivant
1	ur kinda gay for that	[Player_2: Player_1 is mad already for some reason lol], [Player_1: MY INJURE], [Player_3: nice Player_9], [Player_1: LMAO], [Player_4: Player_10 afk], [Player_1: IMAO], [Player_1: hugp], [Player_1: FUCK YUOU Player_9]	[Player_1: I HOPE U DIE], [Player_2: he thinks i stole the kill on {character}], [Player_1: ONG U DID], [Player_2: damn], [Player_1: THAT WAS MY INJURE], [Player_1: I SWEAR ONG], [Player_2: boohoo], [Player_3: play time]
1	that was a gay move, running away like that	No lines before	[Player_2: ur whole team runs], [Player_3: rm?]
0	maybe he's gay?	[Player_2: fart in my mouth?], [Player_3: sure!!!], [Player_1: oh nooo], [Player_3: {female character} and {male character} should put a ring on eachother], [Player_4: they did irl], [Player_1: bro how do you know {male character} is a heterosexua?], [Player_4: but its not disability month anymore so they broke up]	[Player_4: no it was for pride month], [Player_4: hey gu7s==], [Player_4: wlcom to my epsode of fornte], [Player_2: {character} blow urself up], [Player_3: /Vigger], [Player_1: he's cheating], [Player_1: you should blow urself upo buddy], [Player_1: omg]
0	i do, im gay	[Player_2: loose asshole], [Player_1: f, died again like in the bible huh], [Player_2: why you so mad lol], [Player_2: cant take a joke], [Player_1: istn it normal to complain about bitch bois], [Player_2: bet you only take dick in the ass], [Player_2: EZ], [Player_3: ez]	[Player_2: bet you are], [Player_1: are you homophobic you sad asshole], [Player_2: your parents dont love you]

5.1.3 Recrutement de participant-es

L'étape finale de la collecte des substituts consistait à recruter des participant-es pour la tâche. Nous avons recruté un total de 18 personnes. Ceux-ci devaient répondre à plusieurs critères : avoir l'anglais pour langue d'usage, résider au Canada, avoir au moins 30 heures d'expérience dans un jeu vidéo multijoueur compétitif, et être âgés de 18 ans ou plus. Les critères de résidence et d'âge étaient principalement

motivés par des raisons éthiques et parce que les jeux d'où sont issues les données sont destinés à des joueurs et joueuses de 17 ans et plus. La langue première demandée était l'anglais, car les clavardages analysés étaient exclusivement en anglais et les substituts devaient également être fournis dans cette langue. De plus, les 30 heures d'expérience de jeu étaient nécessaires en raison du vocabulaire spécifique utilisé dans les clavardages (p. ex. : « gg » pour *good game*, « gn » pour *good night*, « wp » pour *well played*) sont fréquemment utilisés et peuvent rendre les lignes et leur contexte incompréhensibles pour une personne qui n'est pas familière avec ces termes.

Les personnes participantes ont été recrutées par une publication du lien vers l'expérience sur les réseaux sociaux Facebook et LinkedIn. La publication était accompagnée de l'image de la Figure 5.1, qui contient les informations générales sur la tâche à réaliser ainsi que les critères. Certains participant-es ont aussi entendu parler de la tâche par le bouche-à-oreille, et m'ont écrit personnellement pour que je leur envoie le lien.

Figure 5.1 Image annonçant le recrutement de participant-es pour la tâche de substitution

Seeking Participants


Linguistics Study on Bias in In-Game Chat Toxicity Detection

To participate : you must...

- Have **English** as your first language.
- Be **18 years old or over**.
- Be currently residing in **Canada**.
- Have a minimum of **30 hours of gameplay** experience in any online competitive multiplayer game.

What's Involved:


- **Read** game chat interactions that contain toxic language and complete a short task for each interaction.
- Fill the answers at your **own pace** (approx. **less than 35 minutes**).




Contains toxic and offensive language

How to Participate:


Interested in being a part of this study? Participate **from home** by accessing the **link** provided in the description, or scan this **QR code** to start.



Thank you to all of you who take the time to **share this post** or participate to the study.



For questions or more details, contact **Josiane Van Dorpe** on messenger or at the following email address:
van_dorpe.josiane@courrier.uqam.ca



5.1.4 Création et déroulement de la tâche de substitution

Les 24 phrases choisies étaient divisées en trois groupes de données pour l'expérience : toxiques (10 lignes), non toxiques (10 lignes) et éléments distracteurs (4 lignes). Un autre commentaire récurrent suite

aux prétests provient à nouveau du fait que la difficulté de la tâche augmente avec le nombre de phrases présentées à la personne participante. En effet, plus de 15 phrases créait une redondance, ce qui conduisait les participant-es à fournir des substituts moins pertinents vers la fin de la tâche. Iels avaient tendance à simplement répéter le même terme (p. Ex.: *homosexual* sans suffisamment tenir compte du contexte. Pour cette raison, nous avons créé une rotation de 5 groupes de participant-es, qui verraient chacun un total de 12 lignes : 4 de chaque groupe. Ainsi, chaque phrase contenant le terme cible « gay » a été présentée deux fois au cours du cycle des cinq groupes de participants. Les phrases contenant le terme « black » étaient présentées à chaque participant-e. L'ordre d'affichage des lignes était aléatoire pour chaque participant-e afin de garantir que les phrases n'étaient pas vues dans le même ordre par deux personnes du même groupe. Les lignes étaient affichées sous forme d'interactions, avec le contexte précédent et suivant clairement indiqué, comme illustré à la Figure 5.2.

Figure 5.2 Une interaction pour laquelle les participant-es doivent trouver un substitut.

Previous context :

Player2: stick your hand in your ass and itll warm up
Player3: nothing better to do
Player3: ill just use your ass like your dad
Player3: we got {map name}
Player2: No ban for the win
Player3: not a bad map
Player3: im just not good at it
Player4: pre ggs

Target interaction :

Player1: if u ban ur **gay**

Context after :

Player4: pre gn
Player2: no bad or your dad touches your asshole
Player4: u guys suck
Player1: gge
Player1: pre gn
Player2: DONT LISTEN TO SIMP HIS MENTAL IS CHALKED
Player3: call me elton John im banning
Player4: its over for yall

Please enter the substitutes here, separated by a comma :

Go to next interaction

La tâche de substitution a été construite à l'aide de PC IBEX Farm⁷ (Zehr & Schwarz, 2018), une plateforme de création d'expériences permettant de personnaliser et de diffuser une expérience en ligne gratuitement. L'utilisation de la plateforme nécessite de se familiariser avec la documentation, et quelques connaissances en programmation, notamment les langages de programmation CSS, HTML et Javascript, sont d'une grande aide. La tâche comprend 3 sections principales : le formulaire de consentement éthique, les instructions, puis la tâche. Les réponses sont enregistrées une fois la dernière question de la tâche remplie.

En accédant au lien de la tâche, les participant-es voient d'abord le formulaire de consentement éthique (voir Appendice A), qu'ils doivent lire et accepter. Ce formulaire décrit le contenu auquel ils seront exposés, qui peut être choquant et difficile à lire. Il est précisé qu'ils peuvent quitter le questionnaire à tout moment sans que leurs réponses ne soient enregistrées. Au besoin, un lien vers une banque de contacts pour urgence psychologique par province canadienne est également inclus dans le formulaire. Les coordonnées de mon superviseur et moi-même sont aussi disponibles pour toute question. À la fin du formulaire, les participant-es doivent indiquer leur prénom, nom, puis cocher les cases confirmant la prise de connaissance du formulaire, leur consentement à la tâche, et leur éligibilité. Il est impossible de continuer à la section suivante si ces champs obligatoires ne sont pas remplis. Ils ont également l'option d'indiquer leur courriel pour obtenir une copie du formulaire ainsi qu'un lien vers les résultats de la recherche lorsqu'ils seront disponibles.

La deuxième partie présente les instructions pour réaliser la tâche. Il est précisé que les réponses sont anonymes : bien que leur nom soit saisi dans le formulaire de consentement, les réponses à la tâche ne sont pas associées à ce formulaire. Les instructions complètes sont disponibles dans l'Appendice B. Les participant-es lisent une description générale de ce qu'ils verront pendant la tâche – des interactions tirées de clavardages de jeux vidéo, avec une ligne cible et le contexte précédent et suivant – et ce qu'ils doivent faire – repérer le terme cible surligné, puis indiquer des substituts dans la boîte réservée à cet effet. Des spécifications sont présentées sur le type de substitut attendu : il doit fonctionner avec l'intention et le sens de la ligne cible, il doit autant que possible contenir un seul mot, il doit prioriser l'intention plutôt que d'utiliser un synonyme direct, et il est possible de réutiliser le même substitut pour différentes interactions. Un dernier avertissement de contenu est présenté, au-dessus d'un bouton « I

⁷ <https://farm.pcibex.net/>

understand » qui permet de continuer l’affichage des instructions. Une fois le bouton cliqué, un exemple fictif apparait avec des réponses déjà fournies. Pour éviter d’influencer leurs réponses, le terme « girl » est utilisé en exemple, comme montré dans la Figure 5.3. Pour commencer la tâche, il faut appuyer sur le bouton « start ».

Figure 5.3 Exemple donné aux participant-es avant de commencer la tâche.

Here is an example of the task below :

Previous context :

- Player_3: Player_1 you suck
- Player_2: {battle tech} ruuuuules :3
- Player_3: lmao i keep carrying you noobs, get gud

Target interaction

- Player_1: this is such a **girl** play

Context After

- Player_2: gg
- Player_1: ez gn

Please enter the substitutes here, separated by a comma :

weak, dumb, inefficient, ridiculous

Note : Your answers will only be saved once all the questions are answered. You will see the message *'The results were successfully sent to the server. Thanks!'*.

La tâche elle-même est répétitive. Les participant-es voient un total de 12 interactions, dont 8 qui contiennent le terme cible. La Figure 5.2 offre un exemple d’interaction vu pendant la tâche. Les participant-es peuvent prendre autant de temps qu'ils le souhaitent pour chaque interaction. Une fois les substituts entrés dans la boîte, ils peuvent passer à l’interaction suivante. Il n’est pas obligatoire d’avoir entré une réponse pour continuer.

Il est important de noter qu'un compteur associé au numéro de groupe assigné à chaque personne participante est activé au moment de l'enregistrement des réponses. Cela signifie que si plusieurs personnes accèdent au lien dans un délai rapproché, elles recevront le même numéro de groupe. Lorsque ces personnes complètent la tâche, probablement également dans un délai rapproché, le compteur est activé plusieurs fois de façon successive, ce qui peut influencer le nombre de fois qu'un exemple est présenté aux participant-es. Pour cette raison, le groupe 2 a été le plus souvent assigné.

5.2 Analyses qualitatives et analyses des plongements

5.2.1 Analyses qualitatives

Après avoir collecté les substituts, nous réalisons une première analyse qualitative en examinant leur répartition entre les phrases toxiques et non toxiques. En suivant la méthode de Kremer et al. (2014), nous vérifions manuellement si le sens du terme cible *gay* varie selon la phrase et comment ces variations apparaissent. Pour ce faire, nous regroupons intuitivement les substituts en thèmes ou en catégories en fonction du contexte. Par exemple, des substituts comme *happy* et *sad* permettent d'identifier un thème lié aux émotions.

Ces observations fournissent déjà des indices sur les changements subtils de sens et de *story* du terme selon les contextes. En comparant ensuite les éléments de *story* entre phrases toxiques et non toxiques, nous cherchons à déterminer si les différences sont dues à la toxicité ou à d'autres facteurs contextuels. Ces variations de sens serviront ensuite de base de comparaison pour évaluer si les plongements capturent ces mêmes nuances, contribuant ainsi à répondre aux questions de recherche.

5.2.2 Analyses des plongements

Chronis et Erk (2020), pour explorer les liens de relation entre des termes, extraient les plongements de termes et appliquent des algorithmes de regroupement sur les différentes occurrences. Nous désirons approfondir les connaissances sur les *stories* en récupérant les plongements des substituts et en appliquant une méthode de regroupement pour vérifier si un modèle de langue contient bel et bien les subtilités de la *story* associée à ces termes.

Chronis et Erk (2020) ont mis à disposition du public le code de leur recherche, qui a permis d'étudier les liens de *relation*, et ainsi à la *story*, dans les plongements de modèles de langue. Cependant, les corpus utilisés pour leur analyse diffèrent des substituts collectés pour ce mémoire. Après avoir tenté d'adapter leur code, nous avons constaté que des ajustements substantiels étaient nécessaires pour l'adapter à notre contexte. Nous avons donc utilisé le fil conducteur de leur méthode pour l'extraction des plongements et les méthodes de regroupement, tout en ajustant les étapes spécifiques à notre jeu de données, pour écrire un code adéquat pour notre utilisation⁸.

⁸ Le code écrit est disponible en ligne à l'adresse suivante : https://github.com/josiane212/memoire_2024.git

5.2.2.1 Récupération des plongements

Comme Chronis et Erk (2020), nous utilisons le modèle BERT de type 'bert-base-uncased', accessible via la plateforme HuggingFace (Wolf et al., 2020), pour extraire les plongements des termes. Puisque nous nous intéressons au lien de *relation* entre les termes cibles et les termes substitués capturé par le modèle, c'est le plongement de la 11^e couche, et donc la dernière couche cachée, qui sera extraite pour chacun des termes.

À partir des phrases originales et des substitués collectés, nous générons une liste de phrases possibles : par exemple, en substituant *gay* par *homosexual* dans la phrase « i do, i'm gay », nous obtenons « i do, i'm homosexual ». Chaque phrase, accompagnée de son contexte précédent et suivant, est traitée par le modèle BERT pour produire les plongements. Comme mentionné à la section 1.2, le processus de tokenisation d'un texte par BERT est fait en segmentant le texte en sous-mots lorsque les mots complets ne sont pas reconnus. Le terme « gayest », par exemple, est divisé en « gay » et « #est ». Cela permet de prendre en compte tous les éléments d'un texte sans décider de leur importance avant d'effectuer les étapes suivantes de la création des plongements. Les tokens sont ensuite convertis en identifiants numériques avant d'être transformés en plongements à l'aide du modèle. La couche #11 est la dernière couche cachée du modèle.

Si un seul mot peut être divisé en plusieurs tokens avant d'être traité par BERT, cela implique que chaque partie du mot sera assignée à un plongement. Pour cette recherche, nous désirons n'avoir qu'un seul plongement par terme à analyser. Dans le cas de « gayest », sur quel token doit-on se concentrer ? Nous suivons à nouveau la méthode de Chronis et Erk (2020), qui elles-mêmes suivent le précédent de créer un nouveau plongement par la moyenne des plongements de tous les tokens qui composent le mot.

Dans le contexte de cette recherche, il peut être intéressant de porter les mêmes analyses sur un modèle ayant été pré-entraîné sur des données tirées de jeux vidéo multijoueurs en ligne. Nous utilisons un modèle, développé et obtenu par Yang et al. (2024), pré-entraîné sur les clavardages des jeux *Rainbow Six Siege* et *For Honor*. Ce modèle utilise la structure du modèle RoBERTa (Liu et al., 2019). La structure de BERT et de RoBERTa est sensiblement la même, ce qui signifie que le modèle peut être utilisé pour les analyses de cette recherche. Les différences principales entre les deux se retrouvent dans le pré-entraînement de base : le volume de données de pré-entraînement du modèle RoBERTa est généralement plus grand, et la source de données est différente (jeux vidéo pour ce modèle RoBERTa,

Common Crawl et autres pour BERT). La performance d'un modèle basé sur RoBERTa est aussi généralement meilleure que celle d'un modèle basé sur BERT.

5.2.2.2 Calculs de similarité

Les plongements, comme décrits à la section 1.3.1 de ce mémoire, notamment la Figure 1.2, reflètent une grande quantité d'informations sémantiques. Pour mieux décrire les relations entre substituts-terme cible et pour ainsi ajouter à l'analyse qualitative, nous comparons les termes cibles et leurs substituts via la similarité cosinus, qui mesure l'angle entre deux vecteurs de plongement. Un score proche de 1 indique une forte proximité sémantique, tandis qu'un score proche de 0 montre une différence marquée. Ces calculs permettent d'identifier les substituts les plus proches du terme cible dans l'espace sémantique, et nous permet d'observer si les termes ont un lien de similarité tels que décrit à la section 1.3.4 . Par exemple, pour *gay*, nous comparons sa similarité cosinus avec un terme substitut comme *homosexual* ou *queer*, notamment entre contextes toxiques et non toxiques.

Dans le cas où le score de similarité est plus élevé entre un terme substitut et le terme cible, cela signifie que les substituts sont utilisés de manière plus cohérente dans le contexte, suggérant un sens similaire. Cela pourrait ainsi indiquer que les substituts sont de bonne qualité et considérés comme interchangeables. Si la ligne cible est toxique et le score de similarité élevé, par exemple le terme *weak* dans la phrase « if u ban ur gay », cela signifie que les connotations péjoratives sont partagées entre les termes. À l'inverse, dans les contextes non toxiques, cela indique que les termes sont peut-être moins stigmatisants. Pour des termes qui ne sont pas du tout reliés au sens de base de *gay*, un haut score de similarité indique possiblement que le contexte prend plus d'importance que toute relation sémantique.

5.2.2.3 Méthode de regroupement

Nous réalisons deux types de regroupements pour les analyses : d'abord, les regroupements des plongements des substituts pour chaque ligne cible, puis les regroupements des plongements des lignes cibles elles-mêmes. Le premier type nous permettra de comparer les *stories* identifiées par l'analyse qualitative des substituts avec celles contenues dans les plongements. Le deuxième type nous aidera à déterminer si les plongements rendent compte les différences entre les *stories* de phrases toxiques et non toxiques. Pour les deux types de regroupements effectués, la méthode appliquée est la même.

Les plongements sont d'abord extraits – les plongements des substituts pour le premier type, puis pour le deuxième type les plongements du terme *gay* pour chaque ligne cible - puis regroupés par un algorithme *k-moyens*. Cette méthode de regroupement demande de déterminer au préalable le nombre de groupes voulus (*k*). La Figure 5.4 illustre rapidement le fonctionnement de l'algorithme : à partir de données non groupées, l'algorithme permet d'identifier les groupes qui ressortent. Sur la figure, il peut sembler évident que $k=3$ est optimal, mais cela est beaucoup plus difficile avec un grand volume de données. Nous employons donc la méthode dite du coude pour identifier le nombre optimal de groupes. Cette méthode consiste à tester différentes valeurs de *k* et à calculer l'inertie pour chaque valeur. L'inertie mesure la somme des distances au carré entre chaque point de données et le centre du groupe, indiquant ainsi à quel point les groupes sont compacts. Un graphique est produit avec l'inertie sur l'axe des ordonnées et les valeurs de *k* sur l'axe des abscisses, produisant un graphique ayant une forme de coude plié (voir Figure 5.5). Le « coude » de la courbe indique le *k* optimal, bien que la détermination du coude ne soit pas toujours évidente. En cas d'incertitude, plusieurs valeurs proches du coude sont testées. Pour chaque valeur de *k*, nous calculons le score Silhouette, qui évalue la cohésion des groupes et leur séparation par rapport aux autres groupes. En général, un score Silhouette entre 0.25 et 0.5 est faible, un score entre 0.5 et 0.7 est raisonnable, et un score au-dessus de 0.7 indique une très bonne cohésion. La valeur de *k* avec le plus grand score silhouette est celle sélectionnée pour les analyses subséquentes.

Une fois les regroupements obtenus, nous déterminons quel élément du groupe est le plus représentatif. Cela peut donner plus d'informations sur le « thème » du groupe, ou ce qui le définit le mieux. Pour obtenir cette information, il suffit de calculer la valeur d'un point central au groupe, appelé *centroïde*, puis d'identifier le point de donnée le plus près.

Figure 5.4 Un exemple de l'utilisation de l'algorithme *k-moyens* : des données sont regroupées ensemble selon leurs propriétés.

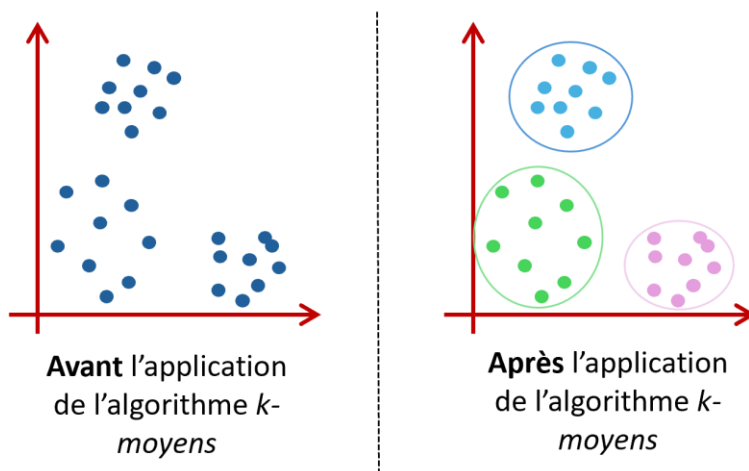
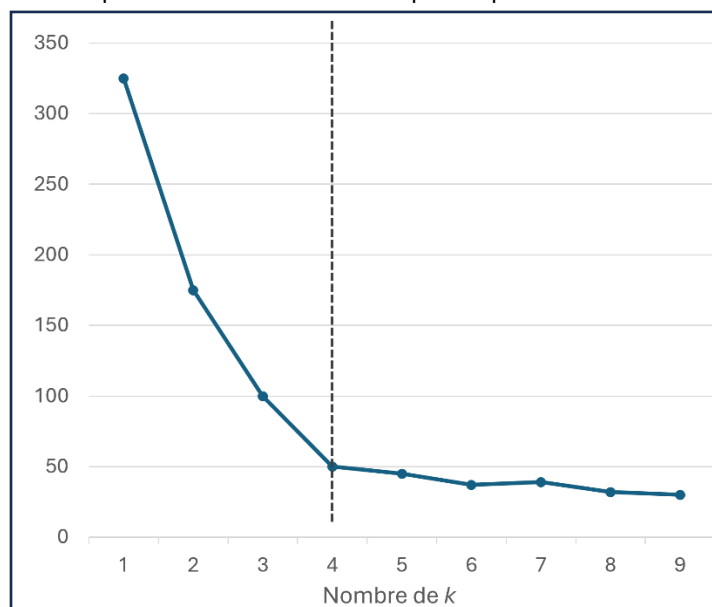


Figure 5.5 Exemple de la méthode du coude pour repérer la meilleure valeur de k .



Les plongements issus de l'utilisation d'un modèle comme 'BERT-base' ont par défaut 768 dimensions. Ce grand volume de dimensions rend difficile l'application efficace d'algorithme de regroupements comme le *k-moyens*. Nous utilisons la technique de réduction de dimensionnalité d'analyse en composantes principales (ACP) pour réduire les plongements à 3 dimensions, ce qui facilite du même coup la visualisation des termes dans un graphique tridimensionnel, tout en conservant les informations pertinentes et en améliorant l'efficacité du clustering. La réduction est effectuée sur un groupe de

plongements plutôt que sur un seul plongement à la fois, augmentant ainsi la pertinence des dimensions finales, car elle prend en compte le contexte global des termes.

5.3 Conclusion

Ce chapitre présente la méthode utilisée pour analyser les *stories* encodées dans les plongements, en s'appuyant sur la collecte et l'étude de substituts lexicaux dans des contextes toxiques et non toxiques. Le terme à l'étude est « gay » présents dans des contextes toxiques et non toxiques extraits de clavardages de jeux vidéo. En nous appuyant sur Erk et Chronis (2022), nous utilisons une méthode quantitative de regroupement des plongements de la couche 11 de deux modèles : 'bert-base-uncased' et un modèle RoBERTa préentraîné sur des données de clavardage de jeu vidéo. Les méthodes de regroupement permettent d'observer les relations entre les termes cibles et leurs substituts, et donc les *stories* encodées lors d'un préentraînement d'un modèle de langue.

Une analyse qualitative pour examiner les différences entre les *stories* en contextes toxiques et non toxiques est aussi décrite dans ce chapitre, en tenant compte de la dynamique des biais identitaires et de la toxicité. En combinant les analyses quantitatives et qualitatives, il est possible de vérifier si les plongements saisissent des distinctions pertinentes entre contextes toxiques et non toxiques et comment les connotations associées à certains termes se manifestent à travers les *stories*. Cette préparation de méthode ouvre naturellement la voie au chapitre suivant, qui présente les résultats obtenus et la discussion associée.

CHAPITRE 6

RÉSULTATS ET DISCUSSION

6.1 Analyses qualitatives : perception du sens de *gay* toxique et non toxique

À la suite de la collecte de substituts, nous avons accumulé un total de 410 termes substituts pour les 24 phrases et contextes présentés, pour un total de 186 termes uniques, 13.04 termes uniques par phrase, et une moyenne de 22 réponses par participant·e. Pour ce qui est des 20 phrases avec *gay*, nous avons obtenu 245 termes, dont 121 uniques, à 11.25 termes uniques par réponse en moyenne. Nous avons obtenu plus de réponses (126) pour les phrases avec *black*, puisque les quatre lignes ont été présentées à tous·tes les participant·es. Les substituts obtenus pour chaque phrase sont présentés dans l'Appendice C, accompagnés de leur fréquence d'apparition et du score de similarité associé.

Les substituts permettent d'abord une analyse qualitative, agrémentée de la mesure de similarité avec le terme cible, pour chaque phrase cible. Nous avons manuellement regroupé chaque substitut des catégories de sens, que nous appelons ici catégories ou thèmes, qui les englobent. Les catégories et regroupements ont été choisis selon le sens perçu des substituts en prenant en compte le contexte, et donc de façon intuitive. Le Tableau 6.1 pour les lignes toxiques, et le Tableau 6.2 pour les lignes non toxiques, détaillent les thèmes identifiés avec les substituts, qui sont placés dans leur catégorie en ordre décroissant de score de similarité. Ces thèmes sont spécifiques à la ligne cible et à son contexte. Il est important de spécifier que ce ne sont pas des *frames* ou *stories* à part entière, ce sont plutôt des sous-éléments de ces concepts. Par exemple, la catégorie « injuste » pourrait s'inscrire dans une *frame* d'évaluation de l'équité, qui inclut des rôles prédéterminés comme une action, un acteur, les partis affectés, etc. (Ruppenhofer et al., 2010)⁹. Obtenir ces catégories nous permet d'identifier les différentes utilisations et les différents sens qui composent la *story* de *gay*.

Le tableau contient des termes surlignés en jaune, qui sont les trois des termes les plus similaires à *gay* pour cette ligne. Tous les scores de similarité sont indiqués dans l'Appendice C, mais nous utilisons ici seulement les trois premiers afin de voir s'il y a des régularités évidentes qui ressortent. Le calcul de similarité a été effectué avec les plongements extraits de BERT. Parmi tous les substituts, le score de

⁹ On retrouve les différentes *frame* sur *framenet* (Ruppenhofer et al., 2010). L'entrée décrite ici se retrouve sur cette page : https://framenet.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Fairness_evaluation

similarité avec *gay* ne descend pas sous 0.70, ce qui est relativement élevé. Comme indiqué dans la section 5.2.2.2, cela s’explique par le fait que les contextes dans lesquels ces termes apparaissent sont identiques, et des liens de similarité fortes entre ces termes et le terme cible *gay* sont encodées dans les plongements. Malgré cette explication pour les scores élevés, il vaut tout de même la peine d’observer certaines catégories qui ont des termes similaires qui ne renvoient pas directement au sens de base de *gay*.

D’une ligne à l’autre, certains substituts reviennent, mais ne sont pas nécessairement assignés aux mêmes thèmes : par exemple, *bad* est parfois présent dans la catégorie « provocation » ou « négativité/insatisfaction » quand il est respectivement un attribut d’une personne ou d’une chose, parfois dans la catégorie « incompétence/inhabilité » quand il prend le sens de « éprouver des difficultés à faire quelque chose », par exemple. Les cellules surlignées seront expliquées au fil du texte.

Tableau 6.1 Catégories identifiées parmi les substituts des lignes cibles toxiques

#	Ligne cible	Catégorie	Substituts
1	gay for not letting us win	Incompétence/Inhabilité	weak, dumb , stupid, idiotic, noob
		Provocation	rude, bad
		Ennuie/Déplaisir	lame , loser, null, zero
		Injustice	cheap
		Absurdité	rubbish, trashy
2	gayest thing ive ever seen	Incompétence/Inhabilité	weakest, noobest , (most) stupid, dumb
		Négativité/Insatisfaction	worst, most disappointing, terrible, bad, most displeasing
		Ennuie/Déplaisir	lamest
		Injustice	most unfair, most broken, cheap, most unbalanced
		Positivité	best
3	ur kinda gay for that	Incompétence/Inhabilité	dumb , stupid, weak, idiot, idiotic
		Provocation	rude , bad
		Ennuie/Déplaisir	lame
		Malhonnêteté	thief
4	who puts exclamation marks at the end of their sentences that's gay	Incompétence/Inhabilité	stupid, dumb , retarded
		Ennuie/Déplaisir	lame
		Insignifiance	irrelevant, useless
5	and you are gay now	Incompétence/Inhabilité	weak, stupid, bad, noob
		Ennuie/Déplaisir	a loser
		Compassion	nice, caring, pacifist
		Caprice	whining

		Trahison	traitor, backstabber
		Homosexualité	[homosexual, queer, homo], [fag, faggot]
6	why r u so gay	Incompétence/Inhabilité	dumb , stupid, idiot, bad, idiotic
		Vulnérabilité	weak, pussy, overly sensitive
		Sévérité	terrible, mean
		Homosexualité	[queer, homo], [fag]
7	yall gay or somethin?	Incompétence/Inhabilité	dumb , stupid, bad, weak, new, terrible, noob, misunderstanding
		Compassion	affectionate
		Caprice	annoying, immature, childish
		Vulnérabilité	pussy, sensitive
		Trahison	unfair, ganging up, traitors, disloyal, cheaters, colliders
		Homosexualité	[homo, homosexual , queers], [fag, fags, faggots]
8	if u ban ur gay	Incompétence/Inhabilité	stupid, noob
		Ennuie/Déplaisir	lame, bad, un-fun, unfair, a loser
		Crainte	pussy, cowardly
		Fragilité émotionnelle	salty, overly sensitive, weak, compensating
		Immaturité	annoying, baby, whiny
		Homosexualité	[homosexual, queer], homo , [fag]
9	that's pretty gay of him	Incompétence/Inhabilité	weak , dumb, stupid
		Ennuie/Déplaisir	not fun, party-poopers
		Distinction	not like us
		Crainte	cowardly
		Injustice	unfair, bad balancing
		Provocation	rude, mean, disrespectful
		Féminité	feminin
		Homosexualité	[homosexual, queer], homo
10	that was a gay move, running away like that	Incompétence/Inhabilité	dumb , weak
		Lâcheté	coward, cowardly
		Ennuie/Déplaisir	lame , boring, loser
		Injustice	cheap , nasty, low, ungamedly
		Offense	nasty, mean, bad

Tableau 6.2 Catégories identifiées parmi les substituts des lignes cibles non toxiques.

#	Ligne cible	Catégorie	Substituts
11	maybe he's gay?	Homosexualité	homosexual, queer
		Incompétence/Inhabilité	weak , not good
12	why so mad we got gay banners	Couleur	colorful, flashy
		Inadéquation	ugly , dumb, useless, dumbass, unfitting, sucky
		Homosexualité	Homosexual, queer
13	and my parents know im gay	Homosexualité	Homosexual, lesbian, queer
14	Nothing wrong with being gay	Homosexualité	Homosexual, queer
15	gay people are normal tooo	Homosexualité	Homosexual, lgbt, queer , homo, lgbtq, lgbtq+, lgbtq+
16	yeah and I'm part of the gays	Homosexualité	homosexuals, homos, lgbt community , dick-lovers, lgbtq, homosexual, lgbtq+ community,
17	ill accept being gay :)	Homosexualité	Homosexual, lgbt, queer , homo, lgbtq+
		Distinction	different, odd
		Incompétence/Inhabilité	dumb, bad, noob
		Victoire	right, good, correct, decisive, the winner
18	im gay too :)	Homosexualité	Homosexual, lgbt, queer , homo, lgbtq+, homosexual
		Amitié	Friendly, team-player
19	Indeed, I am gay	Homosexualité	Homosexual, lesbian, homo
		Incompétence/Inhabilité	simple, noob, unskilled
		Ruse	witty, owning
		Nonchalance/Déplaisir	wicked, trolling, (doesn't gaf)
20	I do, I'm gay	Homosexualité	homosexual, homo
		Féminité	a woman
		Support positif	witty , proud, good, an ally to gays,
		Nuisance	A troll

6.1.1 Catégorie « homosexualité » et le cas de *homo*

Au premier regard, on constate rapidement une différence majeure entre les catégories de lignes toxiques et non toxiques. La catégorie « homosexualité », surlignée en gris dans le tableau pour un repérage plus aisé, est présente pour toutes les lignes cibles (10/10), alors qu'elle ne se présente que 5 fois sur 10 dans les lignes toxiques. Cette catégorie désigne le sens indiqué au CHAPITRE 5, *gay* dénotant une personne attirée par des individus du même genre.

Les termes inclus dans cette catégorie peuvent eux-mêmes être séparés en deux sous-catégories, entre crochets dans le tableau, qui divisent les termes utilisés par la communauté même¹⁰ et les termes injurieux (p. ex. : *fag* et *faggot*). Lorsque les termes injurieux sont présents dans la catégorie, la cellule est colorée en rose. Le terme *homo* est considéré un entre-deux de par son absence sur la liste de termes utilisés par la communauté, et son appartenance à l'expression « no homo ». « No homo » est souvent utilisé par des personnes pour affirmer leur non-homosexualité et peut perpétuer l'hétéronormativité, en suggérant que montrer des caractéristiques féminines implique l'homosexualité (Brown, 2011; Seal, 2021). Pour cette raison, cette expression et le terme *homo* sont moins acceptés au sein de la communauté. Toutefois, puisque nous ne connaissons pas l'orientation sexuelle des participant-es et leur statut d'appartenance à cette communauté, il est difficile d'affirmer s'il devrait être considéré hors contexte comme étant un terme injurieux ou non. Les quatre premières phrases non toxiques sont les seules qui n'ont pas le terme *homo*, ce qui pourrait indiquer que les participant-es ayant vu ces phrases ne considèrent pas le terme comme acceptable pour désigner quelqu'un de *gay*.

Sans inclure *homo* comme terme injurieux, toutes les lignes toxiques sauf une contiennent les substituts *fag* ou *faggot*, ce qui est un élément important de la *story* de *gay* toxique : dans un contexte toxique et lorsqu'on fait référence au sens de base de *gay*, ces termes semblent ancrés dans le sens du terme, et donc font partie de la *story*. Considérant cette information et le fait que la catégorie « homosexualité » apparaît dans tous les exemples non toxiques, il convient de rappeler que les techniques d'atténuation de biais visent à séparer le terme de son sens identitaire. Cependant, cette dissociation n'est pas simple. Même dans les contextes toxiques, où 5 exemples sur 10 sont rattachés à la catégorie « homosexualité », il est difficile de reconnaître quand le terme est utilisé de manière non biaisée (et donc, aucune attache avec la catégorie) ou dans un autre sens.

Avant d'affirmer que le biais existe ou non sur un terme et ensuite tenter de retirer le biais, il est donc essentiel de s'assurer que les autres sens, les autres catégories sont bien reconnus par le modèle. Le risque est qu'en essayant de neutraliser le lien avec l'identité, on retire aussi les autres thèmes cruciaux pour indiquer une utilisation toxique.

¹⁰ Du moins, au Canada, comme l'indique ce glossaire créé par le Gouvernement du Canada : <https://www.canada.ca/en/women-gender-equality/free-to-be-me/2slgbtqi-plus-glossary.html> (Government of Canada, 2022)

6.1.2 Scores de similarité

Lorsque seuls deux termes qui ont les plus hauts scores de similarité sont dans la catégorie « homosexualité », le troisième terme apparaît souvent dans les catégories « incompetence/inhabileté » ou « ennui/déplaisir » comme pour les lignes #6, #7 et #9. La relation de similarité est un lien sémantique identifié à l'aide de calculs réalisés sur les plongements de la 11e couche de BERT. Cela signifie que ces scores sont une partie de la *story* qui informe sur l'utilisation de *gay* dans ces contextes. Si les modèles encodent adéquatement la *story* d'un terme véhiculant un biais identitaire comme *gay*, ces relations de similarité devraient ressortir dans nos analyses subséquentes sur les regroupements.

6.1.3 Catégories positives

Quelques catégories, surlignées en vert dans le Tableau 6.1 et le Tableau 6.2, sont considérées intuitivement plus positives que les autres. Les thèmes de ces catégories sont « victoire », « support positif », « compassion », « amitié » et « positivité ». Il est intéressant de noter que la catégorie « compatissant » apparaît à deux reprises, mais exclusivement dans les lignes qualifiées de toxiques. Ce fait relève une information importante : même dans des contextes toxiques, des thèmes positifs persistent et sont liés à l'utilisation de *gay*. La présence de plusieurs catégories positives parmi les lignes est une information cruciale pour l'atténuation des biais. Même lorsque toxiques, des bouts de *story* positifs existent et sont attachés à l'utilisation du terme. Comme pour les informations reliées aux scores de similarité, il est nécessaire que l'information contenue dans les plongements du modèle permette de distinguer les catégories positives des catégories négatives pour affirmer qu'il est également possible d'isoler l'information biaisée.

6.1.4 Autres catégories et observations

D'autres catégories reviennent souvent parmi les thèmes, notamment « incompetence/inhabileté », « ennuyeux/déplaisant ». La présence et fréquence de ces catégories indiquent que certains thèmes restent relativement stables dans la *story* de *gay*.

Dans les phrases non toxiques, on retrouve généralement une moins grande quantité de catégories, et elles font souvent référence directement à l'identité d'un individu. Lorsque quelqu'un se désigne comme *gay*, cela est fréquemment non toxique, comme les phrases « *Indeed, I am gay* » ou « *i do, im gay* ». En revanche, lorsque le terme est utilisé pour désigner quelqu'un d'autre ou pour qualifier un objet, cela devient souvent toxique et considéré comme une insulte directe, comme « *if u ban ur gay* » ou « *yall gay* ».

or somethin? ». Cette observation suggère possiblement une méthode plus « simple », qui pourrait permettre de détecter la toxicité seulement sur la base d'une analyse grammaticale. Toutefois, déjà parmi les 20 phrases extraites des données de clavardage, on retrouve des exceptions à cette règle. Les lignes « *why so mad we got gay banners* » et « *Nothing wrong with being gay* » sont des lignes qui ne désignent pas la personne qui les écrit, mais qui sont tout de même considérées comme non toxiques. La différence semble provenir entre autres du fait que les lignes expriment une intention positive, un certain support aux communautés discutées, ou du moins qui remettent en doute la toxicité qui fait partie de la *story*. En effet, « *why so mad we got gay banners* » n'est pas nécessairement utilisé pour critiquer l'apparence des bannières, mais remet en doute l'attitude toxique des interlocuteurs qui eux expriment un mécontentement. Ces intentions semblent s'appliquer même lorsque des catégories « négatives » sont présentes dans les lignes non toxiques, où elles véhiculent plutôt de l'autodérision.

6.1.5 Conclusion : *Stories* de *gay* dans la communauté de joueur-euses

Selon les résultats obtenus, il existe des différences entre les usages toxiques et non toxiques du terme *gay* par les joueur-euses, notamment la fréquence de la présence de la catégorie « homosexualité » dans la *story* et l'intention véhiculée par le contexte, mais aussi de nombreuses similarités telles que les catégories positives et les termes ayant un score élevé de similarité avec le terme cible. D'une ligne à l'autre, toxiques ou non toxiques, les changements sont souvent subtils, avec différentes catégories de sens qui se répètent. Le biais identitaire se manifeste sous la catégorie « homosexualité » : bien qu'elle soit surtout présente dans les phrases non toxiques, elle se retrouve aussi dans des phrases toxiques, ce qui indique que le lien à l'identité peut se retrouver entremêlé dans les sens toxiques ou non toxiques du contenu de la *story*. Pour retirer le biais et ainsi isoler le sens identitaire du sens toxique, il est important d'effectuer la première étape de bien distinguer les éléments de la *story*.

À partir de la tâche de substitution par des humains et des analyses qualitatives effectuées, nous avons extrait les nombreux sens que peut prendre le terme *gay* et qui composent en partie sa *story*. Nous avons aussi identifié que les différences entre les sens toxiques et non toxiques ne sont pas aussi grandes qu'il le faudrait pour assurer qu'il est possible de distinguer les sens toxiques des sens non toxiques. Les liens de relations entre toxicité et biais sont ici assez forts. Pour répondre aux deux questions posées au CHAPITRE 4, il est nécessaire de comparer ces résultats avec ce qui est retrouvé dans les plongements issus de modèles de langue.

6.2 Analyse de regroupements des plongements

Nous avons extrait les plongements des substituts de la 11e couche des modèles préentraînés BERT et RoBERTa. Une réduction ACP a ensuite été appliquée pour réduire ces plongements à trois dimensions seulement. Il est important de noter que les phrases avec 3 substituts ou moins (#13 « and my parents know i'm gay », #14 « Nothing wrong with being gay ») n'ont pas subi la réduction et conséquemment les analyses de regroupement, puisque la méthode de réduction de plongements exige d'avoir au minimum autant d'exemples que de dimensions finales. Les deux lignes n'ont que 2 ou 3 substituts, tous appartenant à la catégorie « homosexualité ». Les regroupements de BERT ont en moyenne un score Silhouette de 0.49, ce qui indique une cohésion acceptable parmi les regroupements. Pour le modèle RoBERTa préentraîné sur les données de jeux vidéo, cette moyenne est de 0.43, ce qui se rapproche davantage d'un score faible.

Considérant la théorie explorée au CHAPITRE 1, il est entendu que les plongements issus des modèles préentraînés encodent un bon nombre d'informations linguistiques, pouvant être syntaxiques et sémantiques. Maintenant que nous avons obtenu les sens qui composent la *story* dérivée des substituts, nous pouvons vérifier si la subtilité des sens de *gay* est identifiée par les modèles, et si les modèles voient une distinction entre les *stories* des sens toxiques et non toxiques.

6.2.1 Regroupements des substituts par ligne cible

Pour chaque ligne cible, nous avons appliqué la méthode *k-moyens* sur les plongements des substituts, réduits à 3 dimensions. Les résultats issus de la méthode de regroupement se retrouvent à l'Appendice D, qui contient les lignes cibles, les regroupements identifiés et les scores moyens d'Inertie et de Silhouette. Seules quelques visualisations seront incluses dans le texte, aux fins de démonstration.

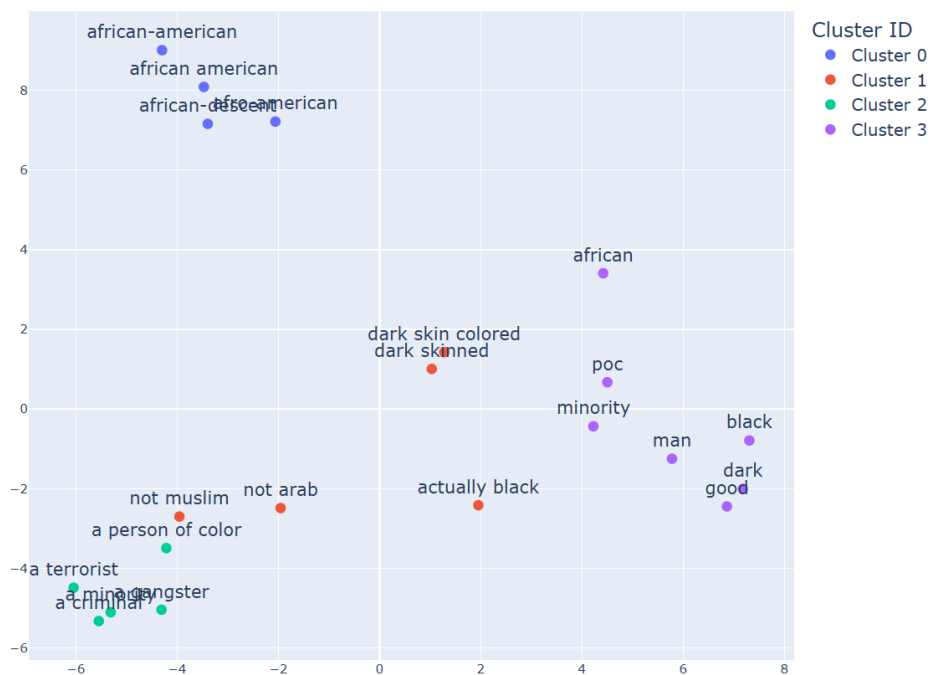
Les lignes cibles avec *black* ayant été présentées à tous les participant-es, nous avons collecté 126 substituts, dont 22 substituts uniques par ligne. Nous n'avons pas de distinction toxique VS non toxique pour ces lignes puisqu'elles n'étaient pas annotées dans le jeu de données fourni. Par contre, nous pouvons observer si certains concepts généraux des substituts ont été identifiés par les modèles. La Figure 6.1 permet de visualiser les groupements faits avec les plongements extraits du modèle BERT pour les substituts de la phrase « bruh im black ». Les visualisations sont créées à partir de plongements en trois dimensions, ce qui signifie qu'il est possible que certains termes, sur une projection en deux dimensions, semblent appartenir à un groupe différent. Une vue en 3D permettrait de mieux voir la distance entre ces

regroupements, toutefois il est parfois difficile d'obtenir un graphique en 3D où tous les éléments sont bien visibles.

Les groupements sont assez bien formés : on retrouve les différentes variations de « african american » regroupées ensemble. Le *cluster 1* semble référer à la couleur de peau, mais on peut également observer des substituts plutôt reliés à la religion musulmane ou l'origine arabe. Le *cluster 2* contient tous les termes qui débutent par le déterminant indéfini *a*, mais le lien sémantique semble moins crucial. Le *cluster 3* regroupe plusieurs termes qui auraient probablement pu subir un autre niveau de regroupement, par exemple pour distinguer *poc* et *minority* de *man*.

Figure 6.1 Visualisation 2D des groupements des plongements des substituts de la phrase « bruh im black ». Les plongements sont issus du modèle BERT.

2D BERT - bruh im black

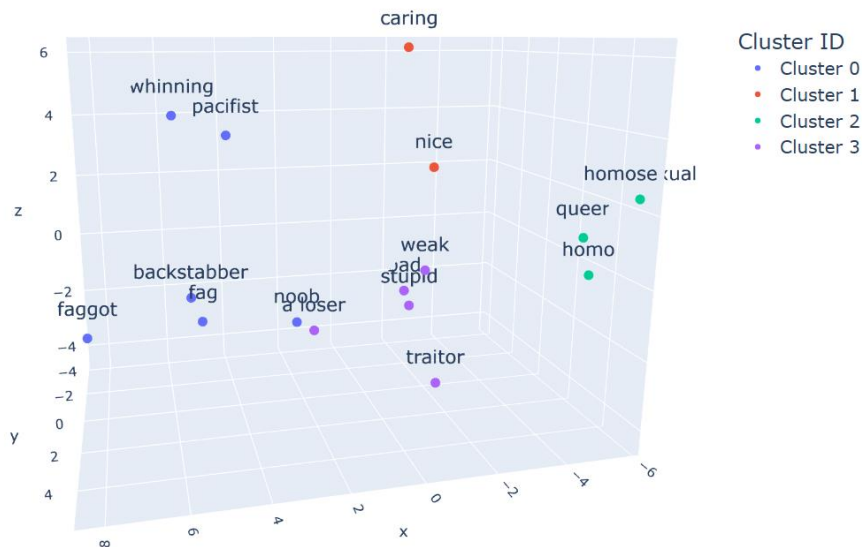


Ainsi, nous pouvons vérifier quelles catégories identifiées manuellement à la section 6.1 sont également identifiées par le modèle. Prenons quelques exemples de phrase toxiques et non toxiques qui ont plus d'une catégorie de substituts, et donc des *story* et des sens plus complexes. La Figure 6.2 montre la phrase #5 « and you are gay now », une phrase toxique. Les catégories de substituts pour cette phrase sont : incompetence/inaptitude, ennui/déplaisir, compassion, caprice, trahison et homosexualité. C'est l'une des rares phrases toxiques qui inclut une catégorie plus positive, « compassion ». Parmi les regroupements

de plongements issus de BERT pour cette phrase, peu semblent logiques ou représentatifs des catégories, à l'exception de *homosexual*, *queer* et *homo*, qui sont regroupés ensemble. Ces trois termes sont également les plus similaires à *gay* dans le contexte de cette phrase, indiquant possiblement que ce sens identitaire est bien encodé dans les plongements. Parmi les autres sens, celui relié à la trahison semble perdu : *backstabber* et *traitor* ne sont pas regroupés ensemble. On observe un regroupement qui se rapproche de la catégorie de compassion avec *caring* et *nice*. Le terme *pacifist*, par contre, se retrouve dans un regroupement qui ne semble pas avoir un thème particulier : *backstabber*, *faggot*, *fag*, *noob* et *whinning* en font tous partie. Le *cluster 3* ne semble pas non plus se distinguer par son sens, et on y retrouve des termes des catégories « ennui/déplaisir », « incompetence » et « trahison ». Les éléments les plus représentatifs de chaque regroupement, indiqués dans l'Appendice D, ne permettent pas non plus d'inférer un thème pour chacun des groupes.

Figure 6.2 Visualisation 3D des plongements des substituts de la phrase #5 « and you are gay now ». Les plongements sont issus du modèle BERT.

3D BERT - and you are gay now



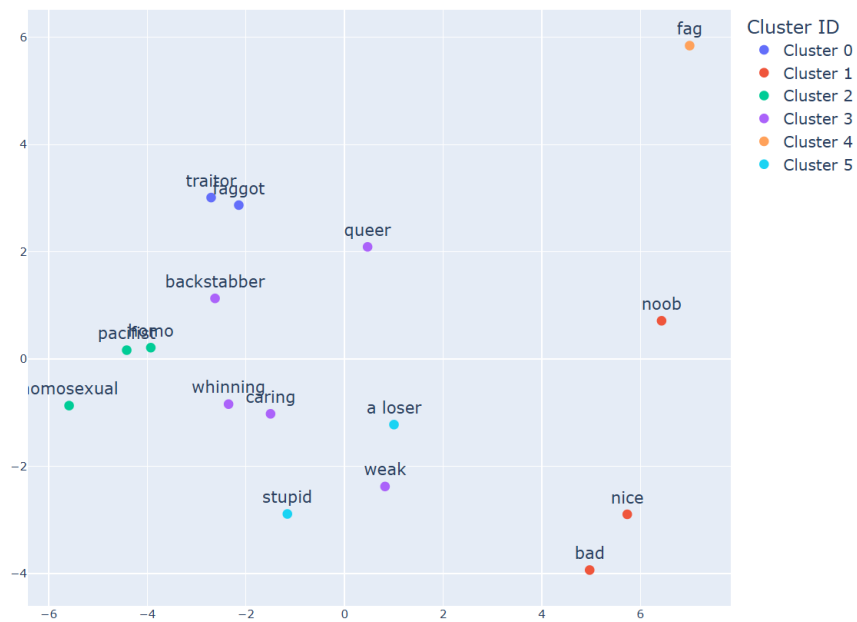
À notre connaissance, le préentraînement du modèle BERT contient peu ou pas de clavardages de jeux vidéo compétitifs multijoueurs. Peut-être qu'un modèle utilisé et préentraîné spécifiquement pour ce type de données pourrait mieux déceler les subtilités de la toxicité. La Figure 6.3 montre la même phrase et les mêmes substituts, mais avec les plongements extraits du modèle RoBERTa préentraîné sur des données tirées de jeux vidéo. Plus de regroupements ont été créés avec les plongements de ce modèle. Malgré tout, il est très difficile de reconnaître ne serait-ce qu'un seul thème d'un regroupement. Seuls *homosexual* et

homo se retrouvent dans un même groupe. *Queer* est plutôt regroupé avec des termes des catégories « compassion », « trahison », « incompetence » et « caprice ». Même les termes insultants liés à l’homosexualité, *fag* et *faggot*, se retrouvent dans des regroupements différents. Alors que les données d’entraînement de ce modèle incluent exactement ce genre de données, il est surprenant que les informations liées à la *story* semblent perdues.

À partir de ces informations, on peut émettre l’hypothèse que le sens initial du mot *gay*, sa *frame* principale, est encodée adéquatement dans les représentations de cette phrase et de ses substituts. Le reste de la *story* du terme est présent, mais n’est pas suffisamment bien défini pour confirmer que les informations reliées peuvent être extraites de façon aisée et précise.

Figure 6.3 Visualisation 2D des plongements des substituts de la phrase #5 « and you are gay now ». Les plongements sont issus du modèle RoBERTa préentraîné sur des clavardages de jeux vidéo.

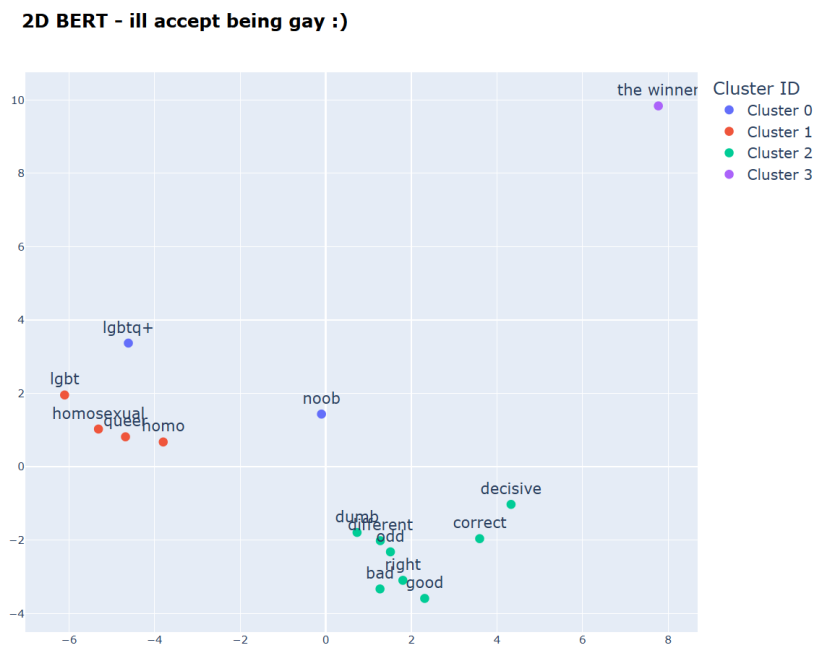
2D RoBERTa - and you are gay now



Prenons un exemple de phrase non toxique #17, « ill accept being gay :) ». Les regroupements construits à partir des plongements de BERT se retrouvent sur la Figure 6.4. Ici, on peut observer que les termes de la catégorie « homosexualité » sont majoritairement regroupés ensemble, à l’exception de *lgbtq+* qui se retrouve dans le *cluster 0*, tout comme *noob*, plutôt que dans le *cluster 1* avec les autres termes de la catégorie. Alors que nous retrouvons la catégorie « victoire » parmi les sens indiqués par les substituts, ces termes ne sont pas retrouvés dans un même regroupement, et se retrouvent plutôt avec tous les autres

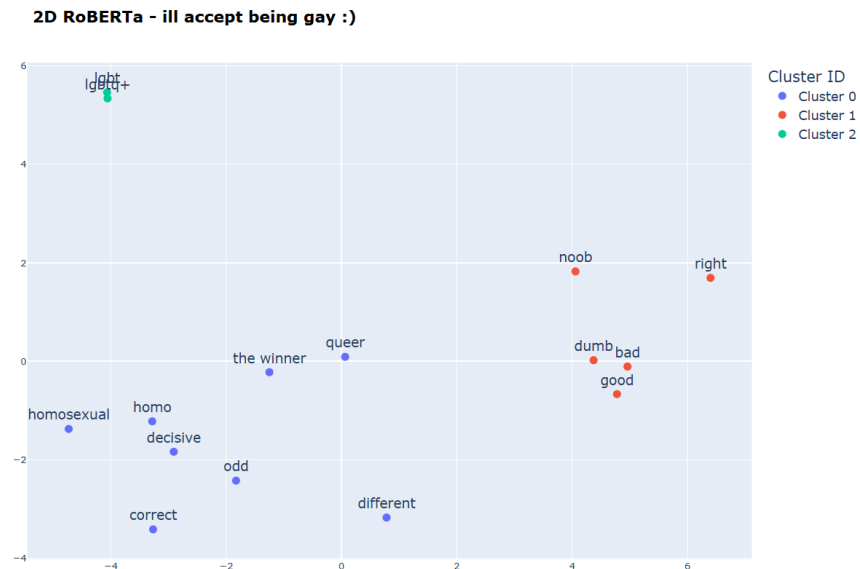
termes substitués, peu importe leur catégorie. Le substitut *the winner* semble être isolé. Il est possible que cela soit dû au fait qu'il est composé de deux mots plutôt qu'un seul, une différence syntaxique relevée qui expliquerait la séparation de ce substitut. À nouveau, même pour une phrase non toxique, les regroupements ne semblent pas être cohérents avec les éléments de *story* identifiés dans la section précédente.

Figure 6.4 Visualisation 2D des plongements des substituts de la phrase #17 « ill accept being gay :) ». Les plongements sont issus du modèle BERT.



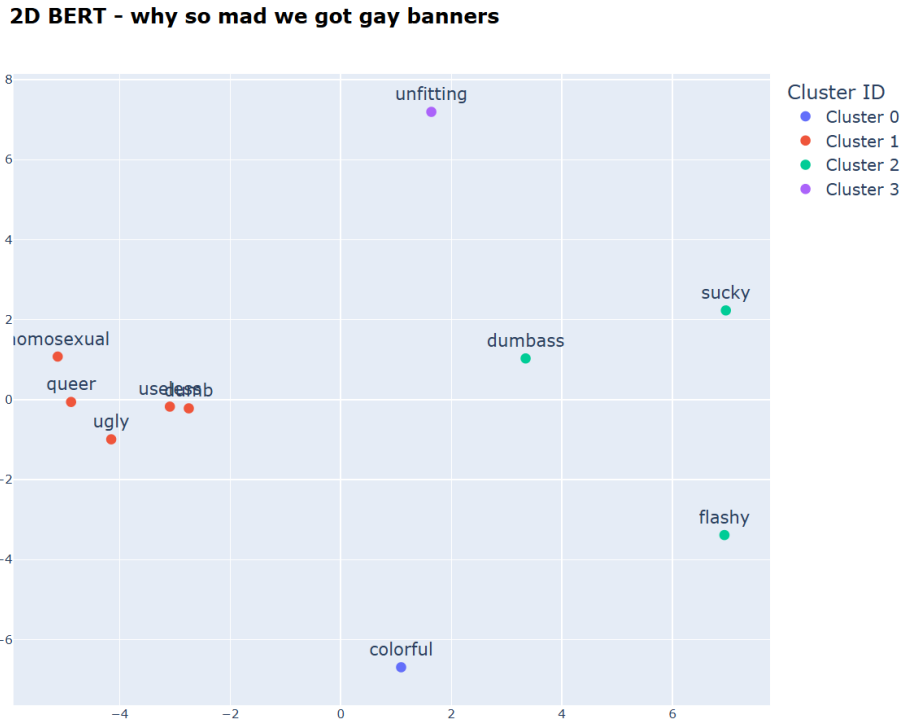
Les regroupements offerts par RoBERTa pour la même ligne cible, visualisés à la Figure 6.5, ne semblent pas non plus indiquer une distinction claire entre les sens. Les regroupements ne distinguent pas la catégorie « homosexualité », à l'exception des termes *lgbt* ainsi que *lgbtq+*. La catégorie positive n'est pas non plus identifiée.

Figure 6.5 Visualisation 2D des plongements des substituts de la phrase « ill accept being gay :) ». Les plongements sont issus du modèle RoBERTa.



Alors que la Figure 6.2 et la Figure 6.4 indiquent peut-être la possibilité d’obtenir une distinction claire de la catégorie du sens de base de *gay* dans les plongements, la phrase # 12, contredit cette hypothèse. En effet, la Figure 6.6 illustre les regroupements obtenus avec les plongements de BERT. Les termes de la catégorie « homosexualité » sont dans le même groupe que certains des termes de la catégorie « inadéquation », mais seulement quelques-uns. Les deux termes reliés aux couleurs, *flashy* et *colorful*, n’appartiennent pas non plus au même regroupement. Encore une fois, il ne semble pas y avoir de logique particulière à la formation des regroupements. Pour ce qui est du modèle RoBERTa pour la même phrase, il n’y a que deux regroupements, qui ne distinguent pas les différentes catégories.

Figure 6.6 Visualisation 2D des plongements des substituts de la phrase #12 « why so mad we got gay banners ». Les plongements sont issus du modèle BERT.



Les visualisations qui se sont retrouvées dans cette section permettent de rendre compte du fait que les modèles ne semblent pas en mesure de bien distinguer les différents sens dérivés des substituts, ce qui répond à notre première question : les *stories* associées aux termes véhiculant un biais identitaire dans les clavardages de jeux vidéo et à leurs substituts, caractérisées par les représentations qui sont extraites d'un modèle de langue, se distinguent significativement des *stories* dérivées des substituts. En plus de se distinguer, notre analyse ne nous a pas permis d'identifier de *story* cohérente qui émerge des regroupements obtenus.

6.2.2 Regroupement des phrases cibles

Pour répondre à la deuxième question, nous vérifions si les plongements encodent adéquatement des informations sur la toxicité de la phrase. Pour ce faire, les Figure 6.7 et Figure 6.8, accompagnées respectivement des Tableau 6.3 et Tableau 6.4 ci-dessous, indiquent que la distinction toxique-non toxique n'est pas effectuée dans les plongements. Les losanges représentent les phrases toxiques, et les cercles les phrases non toxiques.

Pour tenter d'identifier une pertinence dans les regroupements, nous vérifions si d'autres motifs sont présents parmi les catégories ou même parmi les termes similaires. Par exemple, le *cluster 1* du Tableau 6.3 regroupe #4 « who puts exclamation marks at the end opf their sentences thats gay » et #12 « why so mad we got gay banners ». Tous deux une classification de toxicité différente et des catégories qui ne se recoupent pas non plus. Le point commun possible de ces deux phrases est leur formulation en tant que question. Toutefois, la ligne la plus représentative du regroupement est #15 « gay people are normal tooo », qui n'est ni une question, ni toxique, et qui ne consiste qu'à une seule catégorie de sens (« homosexualité ») qui n'est pas présente dans tous les autres items du regroupement.

Figure 6.7 Visualisation 2D des regroupements des phrases cibles avec les plongements de BERT – les losanges sont des phrases toxiques et les cercles sont des phrases non toxiques.

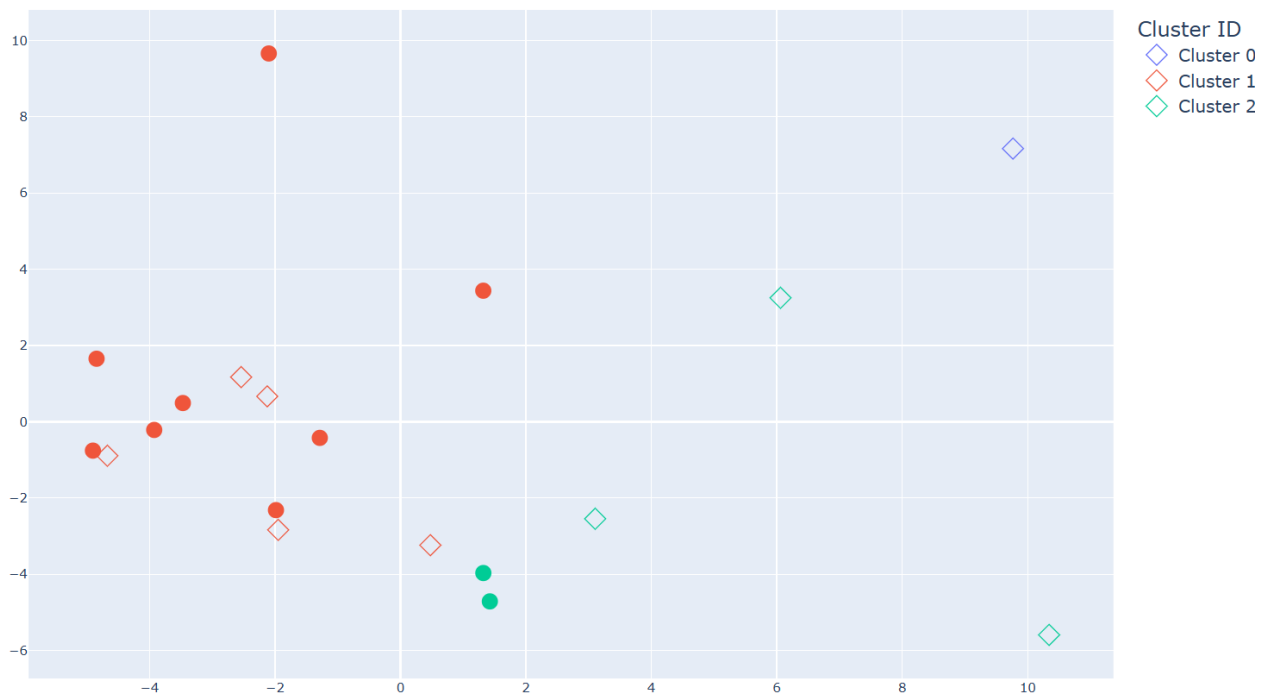


Tableau 6.3 Regroupements des phrases cibles avec les plongements de BERT

ID	Lignes cibles	Élément le plus représentatif
0	['that was a gay move, running away like that']	that was a gay move, running away like that
1	['who puts exclamation marks at the end opf their sentences thats gay', 'and you are gay now', 'why r u so gay', 'yall gay or somethin?', 'if u ban ur gay', "maybe he's gay?", 'why so mad we got gay banners', 'and my parents know im gay', 'Nothing wrong with being gay', 'gay people are normal tooo', 'ill accept being gay :)', 'Indeed, I am gay', 'i do, im gay']	gay people are normal tooo

2	['gay for not letting us win', 'ur kinda gay for that', "that's pretty gay of him", 'im gay too :)', 'youre gay arent you']	ur kinda gay for that
---	---	-----------------------

Pour le modèle RoBERTa, le même problème survient. Les regroupements ne distinguent aucunement les phrases toxiques des phrases non toxiques. Le *cluster 1*, par exemple, « that was a gay move, running away like that » et « why so mad we got gay banners » ne partagent ni termes similaires en commun, ni catégorie, ni classification de toxicité.

Figure 6.8 Visualisation 2D des regroupement des phrases cibles avec les plongements de RoBERTa – les losanges sont des phrases toxiques et les cercles sont des phrases non toxiques.

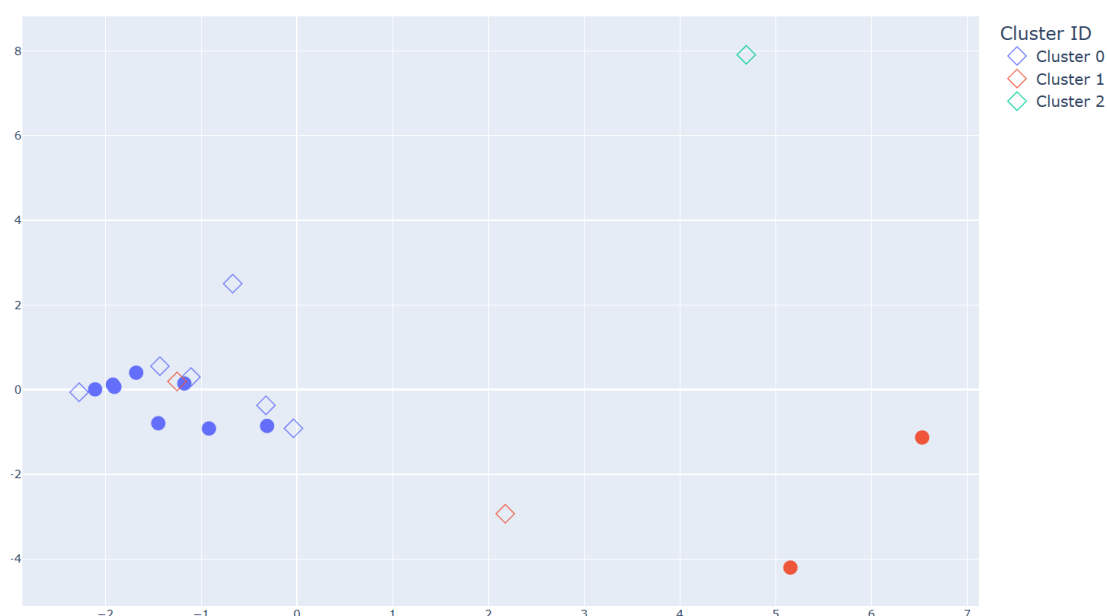


Tableau 6.4 Regroupements des phrases cibles avec les plongements de RoBERTa.

ID	Lignes cibles	Élément le plus représentatif
0	['ur kinda gay for that', 'and you are gay now', 'why r u so gay', 'yall gay or somethin?', 'if u ban ur gay', "that's pretty gay of him", "maybe he's gay?", 'and my parents know im gay', 'Nothing wrong with being gay', 'ill accept being gay :)', 'im gay too :)', 'Indeed, I am gay', 'i do, im gay', 'youre gay arent you']	ur kinda gay for that
1	['who puts exclamation marks at the end opf their sentences thats gay', 'that was a gay move, running away like that', 'why so mad we got gay banners', 'gay people are normal tooo']	that was a gay move, running away like that
2	['gay for not letting us win']	gay for not letting us win

À partir de ces résultats, il est clair que les deux modèles ne saisissent pas les informations liées à la *story* de toxicité, car les caractères toxiques et non toxiques des phrases ne sont pas des éléments encodés dans les plongements. Cela nous permet de répondre à notre deuxième question : les plongements ne reflètent pas les différences entre stories toxiques et les stories non toxiques.

6.3 Conclusion : *Story* de *gay* détectée dans les modèles

Au fil de ce chapitre, nous avons pu constater que les modèles ne semblent pas en mesure de distinguer clairement les différentes catégories qui composent en partie les *stories* de *gay*, et ne distinguent pas non plus les phrases à caractère toxique de celles non toxiques. Ces informations nous permettent de répondre à nos deux questions énoncées au CHAPITRE 4 :

1. Les *stories* associées aux termes véhiculant un biais identitaire dans les clavardages de jeux vidéo et à leurs substituts, caractérisées par les représentations qui sont extraites d'un modèle de langue, se distinguent significativement des stories dérivées des substituts.
2. Les plongements ne reflètent pas les différences entre stories toxiques et les stories non toxiques.

La réponse à la première question pourrait indiquer que les *stories* diffèrent, mais qu'il est tout de même possible d'identifier une *story* dans les plongements en retrouvant des thèmes parmi les regroupements obtenus. Cela offrirait ainsi une possibilité de commenter sur les différences entre les *stories* identifiées par des humains et celles encodées dans des plongements. Toutefois, les regroupements ne semblent pas cohérents sémantiquement, et aucun élément de sens ne ressort particulièrement. Certains regroupements indiquent des régularités syntaxiques (p. ex. : plusieurs mots, un terme en particulier, etc.), toutefois à nouveau de façon inconstante également.

Ces réponses sont des informations cruciales pour les efforts d'atténuation de biais : la *story* n'étant pas adéquatement encodée dans les plongements, on ne peut appliquer des méthodes qui tentent de retirer un élément déjà difficile à détecter dans l'information sémantique d'un terme. En retirant le biais, le risque de retirer également d'autres informations importantes pour déterminer si une phrase est toxique ou non est trop élevé. Grâce à ces résultats, il est clair qu'il est important de vérifier si les éléments de sens sont bien encodés avant d'effectuer toute manipulation des plongements. Il faut se demander, entre autres, si les sens toxiques et non toxiques se distinguent facilement. De plus, il est nécessaire de vérifier si le sens

de base est adéquatement encodé dans les représentations. Si c'est le cas, il faut vérifier qu'il soit bien isolé et facilement séparable.

CONCLUSION

Tout au long de ce mémoire, nous avons exploré les relations complexes entre toxicité et biais identitaire dans les clavardages de jeux vidéo en ligne en analysant les représentations linguistiques produites à l'aide des modèles de langue préentraînés. Nos analyses ont révélé des distinctions significatives entre les *stories* associées aux termes véhiculant un biais identitaire et celles dérivées des substituts obtenus par des participant-es, indiquant que les informations retrouvées dans les plongements ne reflètent pas les différences entre stories toxiques et non toxiques. Cela souligne une limite importante dans l'utilisation de modèles de langues pour la détection de toxicité.

En effet, ces résultats mettent en évidence les défis liés à l'atténuation des biais dans les systèmes de détection de toxicité. Il est crucial de faire preuve de prudence lorsque l'utilisation de ces modèles est nécessaire. En effet, bien que ces outils puissent être pratiques et performants pour gérer les discours en ligne, il est impératif d'identifier ce qui est encodé dans les représentations avant de mettre en œuvre des sanctions. Une mauvaise interprétation ou une gestion inappropriée des biais pourrait non seulement compromettre la performance du modèle, mais aussi amplifier les biais existants et nuire davantage aux groupes marginalisés.

Ainsi, cette recherche encourage à adopter une approche plus réfléchie et nuancée dans l'utilisation des modèles de langue pour la modération des discours en ligne. Une meilleure compréhension des dynamiques encodées et une prudence dans l'application de méthode d'atténuation de biais sont essentielles pour garantir une modération juste, efficace et respectueuse des diversités dans les environnements de jeux vidéo multijoueurs en ligne.

6.4 Limites de la recherche

La recherche présentée présente plusieurs limites qu'il est important de souligner. En premier lieu, les biais identitaires à l'étude sont des biais qui se retrouvent dans une communauté linguistique nord-américaine, plus précisément canadienne. La diversité des participantes et participants reste inconnue, puisque la collecte et l'analyse d'informations sensibles à leur sujet n'étaient pas l'objet de cette recherche. Par contre, connaître plus de détails sur leur appartenance à différentes communautés aurait pu apporter un point de vue intéressant sur les résultats. Surtout, les substituts proposés auraient pu différer de manière significative si cette recherche avait été menée dans d'autres contextes culturels.

En deuxième lieu, certains participants ou participantes m'ont fait part de leur sentiment d'être « peu originaux » parce qu'ils proposaient fréquemment les mêmes termes, bien que les instructions les y autorisaient. Selon leurs dires, il semble qu'ils aient parfois évité de trop se répéter par peur de fausser les résultats, et ont évité d'entrer certains termes pour toutes les phrases. Ce comportement pourrait expliquer pourquoi certains termes étant des synonymes, comme *homosexual*, n'apparaissent pas partout dans les résultats. Malgré tout, comme indiqué dans le chapitre précédent, ce terme est davantage proposé dans les contextes non toxiques, suggérant que les participant-es se sentaient plus à l'aise de l'utiliser dans des situations perçues comme plus inclusives. Une autre limite tient au fait que tous-tes les participant-es n'ont pas eu accès à toutes les interactions cibles, ce qui peut affecter la représentativité des termes recueillis. Par exemple, un substitut peut apparaître de manière récurrente dans certaines phrases, mais être absent dans d'autres simplement parce que le-la participant-e n'a pas vu ces phrases. Cela rend difficile de déterminer si l'absence d'un terme est due à son inadéquation contextuelle ou simplement au fait que le-la participant-e avait tendance à l'utiliser systématiquement lorsqu'il en avait l'occasion.

On retrouve aussi une limite quant au contexte offert. Bien que chaque ligne était accompagnée de plusieurs lignes de texte avant et après, le contexte de ce qui se déroule dans le jeu n'est pas inclus. Il est fort possible que certaines lignes semblent déconnectées des autres, puisqu'elles offrent un commentaire sur une action qui se passe en jeu. Ces événements sont essentiels à la dynamique de l'interaction, mais ce genre d'information n'est pas encore inclus dans les jeux de données de clavardage. Cela a pu brouiller la compréhension de certain-es participant-es, affectant ainsi leur capacité à proposer des substituts qui fonctionnent en contexte.

Finalement, une autre limite est survenue lors de la réduction des dimensions des plongements, réalisée avant l'application de l'algorithme de regroupement. Une réduction est nécessaire avant le regroupement, mais ne garder que trois dimensions est drastique : il n'y a pas assez de dimensions pour effectuer des regroupements efficaces, et beaucoup d'information se perd. Le choix de trois dimensions a été fait pour être en mesure de visualiser directement les regroupements en trois dimensions. L'idéal serait d'appliquer une réduction avec un plus grand nombre de dimensions en premier lieu pour faire l'analyse de regroupement, puis utiliser les deux ou trois premières dimensions pour créer les visualisations. Tester plusieurs réductions et vérifier la plus efficace aurait aussi été crucial, toutefois le manque de temps n'a pas permis d'explorer cette hypothèse plus en profondeur.

ANNEXE A
APPROBATION ÉTHIQUE

UQÀM | Comités d'éthique de la recherche
avec des êtres humains

No. de certificat : 2024-6720
Date : 2024-03-28

CERTIFICAT D'APPROBATION ÉTHIQUE

Le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPE FSH) a examiné le projet de recherche suivant et le juge conforme aux pratiques habituelles ainsi qu'aux normes établies par la *Politique No 54 sur l'éthique de la recherche avec des êtres humains* (avril 2020) de l'UQAM.

Titre du projet : DÉTECTION NON BIAISÉE DE TOXICITÉ DANS LES CLAVARDAGES DE JEUX VIDÉO : ÉVALUATION THÉORIQUE

Nom de l'étudiant : Josiane Van Dorpe

Programme d'études : Maîtrise en linguistique

Direction(s) de recherche : Grégoire Winterstein

Modalités d'application

Toute modification au protocole de recherche en cours de même que tout événement ou renseignement pouvant affecter l'intégrité de la recherche doivent être communiqués rapidement au comité.

La suspension ou la cessation du protocole, temporaire ou définitive, doit être communiquée au comité dans les meilleurs délais.

Le présent certificat est valide pour une durée d'un an à partir de la date d'émission. Au terme de ce délai, un rapport d'avancement de projet doit être soumis au comité, en guise de rapport final si le projet est réalisé en moins d'un an, et en guise de rapport annuel pour le projet se poursuivant sur plus d'une année au plus tard un mois avant la date d'échéance (**2025-03-28**) de votre certificat. Dans ce dernier cas, le rapport annuel permettra au comité de se prononcer sur le renouvellement du certificat d'approbation éthique.



Sylvie Lévesque
Professeure, Département de sexologie
Présidente du CERPÉ FSH

APPENDICE A
FORMULAIRE DE CONSENTEMENT ÉTHIQUE

CONSENT FORM

Research project title

Détection non biaisée de toxicité dans les clavardages de jeux vidéo : évaluation théorique
(Unbiased toxicity detection in written video game chat : theoretical evaluation)

Student-researcher

Josiane Van Dorpe. Université du Québec à Montréal (UQAM)

Master's degree in Linguistics

514 546-3998

Van_dorpe.josiane@courrier.uqam.ca

Research supervisor

Winterstein, Grégoire. Université du Québec à Montréal (UQAM)

Department of Linguistics

(514) 987-3000 ext. 7032

winterstein.gregoire@uqam.ca

Preamble

You are invited to participate in a research project evaluating linguistics components that are at work behind biases in toxicity detection within video games written chat. Before accepting to participate in this project, please take the time to understand and carefully consider the information that follows.

This consent form explains the purpose of the study, the procedures, the benefits, the risks and disadvantages as well as the people to contact if necessary.

The present form might include words that you may not understand. Please do not hesitate to ask the researcher any questions you may have.

Description of project and its objectives

The objective of this experiment is to better understand the link between toxicity and identity biases

found in a toxicity detection system. For this, we need to collect data from native English speakers. We are looking for terms that can substitute others in a sentence with context. More specifically, we are looking at terms in a chatline from online multiplayer competitive games, which is why experience with this type of game is necessary for this task.

The link to complete the task will be available for two weeks, starting from May 4th 2024. We will then use large language models to extract the embeddings of terms and conduct a qualitative analysis on the clusters surrounding them in the vector space.

Nature and duration of your participation

This is a one-time task that should take around 35 minutes. The results will be sent when you answer the last prompt. Your task will be to provide terms that can substitute other terms in chatlines found in online multiplayer matches. The terms to substitute will be clearly highlighted.

TRIGGER WARNING: Some chatlines may contain vulgar or offensive language, namely hate speech of homophobic or racist nature. Other types of hate speech might be present in the data as well.

Benefits associated with participating in the present study

You will not personally benefit from participating in this study. However, you will have contributed to the advancement of science.

Risks associated with participating in the present study

Due to the nature of the data you will see, you will be exposed to offensive, vulgar, and toxic content that can be difficult to read. There might be harsh insults targeting specific groups of people, such as homophobic or racist chatlines and slurs. This implies a psychological risk if you identify or if you are sensitive to the targeted group.

If you feel like the text you read or write in this task affects you negatively, stop answering and close the window. Your data and answers will not be collected, and you are not required to contact the researchers if you decide to withdraw from the experiment. In Canada, for mental health support, please consult the section relevant to your province or territory on this webpage:

<https://www.canada.ca/en/public-health/services/mental-health-services/mental-health-get-help.html>, or contact 9-8-8 for suicide help and prevention.

Confidentiality

The results are confidential. You can provide your name and contact information on a voluntary basis if

you would like to be contacted with the study's results. Your personal information will only be known to researchers and will not be revealed when the results are disseminated. The results form will be numbered, and only the researchers will have the list of participants and the number assigned to them. If provided, the personal information collected will be destroyed 1 year after the last scientific communication or on request.

Secondary use of data

The research data collected in this project will not be used for other research projects and will be destroyed.

Voluntary participation and right to withdraw

Your participation in this project is entirely voluntary. You may refuse to participate or you may withdraw from the study at any time without the need to justify your decision. If you decide to withdraw from the study, you only need to verbally inform Josiane Van Dorpe; in this case, all data concerning you will be destroyed.

Compensation

No compensatory allowance is provided.

Questions concerning the research project?

If you have any further questions concerning your participation or the study itself, you may contact the people responsible for the project:

Student researcher:

Josiane Van Dorpe

514 546-3998

van_dorpe.josiane@courrier.uqam.ca

Supervisor:

Winterstein, Grégoire

(514) 987-3000 ext. 7032

winterstein.gregoire@uqam.ca

Any questions concerning your rights? The research ethics review committee involving human subjects (CERPE) has approved this research project in which you are involved. If you have any ethical concerns or complaints about your participation in this study, and want to speak to someone who is not on the research team, please contact the coordinator of CERPE (*CERPE plurifacultaire* (cerpe-pluri@ugam.ca) or *CERPE FSH* (cerpe.fsh@ugam.ca).

Acknowledgements

Your collaboration is essential to the realization of our project and the research team wishes to thank you.

Consent

I acknowledge having read about and understood the present research project, including the nature and extent of my participation as well as the potential risks and disadvantages to which I will be exposed, as indicated in this consent form. I have had the opportunity to ask questions concerning the various aspects of the study and to receive answers to my satisfaction.

I, the undersigned, voluntarily consent to participate in this study. I understand that I can withdraw at any time without prejudice of any kind. I certify that I have been given the time needed to make my decision.

First name, Surname

I hereby confirm that I currently reside in Canada.

Date (DD/MM/YYYY)

Contact

If you want to obtain a copy of this form or be notified when the results are published, enter your email address here and check the box(es) with your choice(s):

Send me a copy of the consent form.

Notify me when the research is published.

APPENDICE B

INSTRUCTIONS POUR LA TÂCHE DE SUBSTITUTION

INSTRUCTIONS

--The data collected in this research is anonymized. Your name will not be associated to your answers.--

What you will see: In the following pages, you will see chat conversations between players during matches in a multiplayer online competitive videogame. Information about the game has been redacted and will be specified in curly brackets {}.

Your task: First, read the whole exchange, with the context before and after a *Target interaction*.

Focus on the *Target interaction* and the **intentions** conveyed in this line.

A term will be highlighted. Your task is to find as many terms that can replace the highlighted term as possible. Think of all the substitutes you can and write them in the box found below the conversation, separated by a comma.

Important points about the substitutes :

- It is imperative that the substitutes work with the meaning and intentions conveyed by the *Target interaction*.
- As much as possible, write substitutes constituted of only a **single word**, as you will see in the example below.
- First, find substitutes that convey the same **intentions** rather than direct synonyms. Resort to synonyms only when necessary.
- You can reuse the same terms for different interactions.

TRIGGER WARNING: Some chatlines may contain vulgar or offensive language, namely hate speech of homophobic or racist nature. Other types of hate speech might be present in the data as well.

APPENDICE C

JEU DE DONNÉES ET SUBSTITUTS OBTENUS

Tableau des lignes cibles, les substituts obtenus et leur score de similarité, la fréquence d'apparition de la ligne pendant la tâche de substitution, les fréquences des substituts (si le substitut n'est pas indiqué dans cette colonne, c'est qu'il n'a été proposé qu'une seule fois), le contexte précédant la ligne et le contexte suivant. Les lignes 1 à 10 ont une annotation **toxique**, les lignes 11 à 20 ont une annotation **non toxique**.

Tableau 6.5 Tableau complet du jeu de données, avec les substituts obtenus.

#	Ligne cible	Substituts (score de similarité)	Fréq.	Fréq. des substituts	Contexte précédent	Contexte suivant
1	gay for not letting us win	lame (0.902), weak (0.901), dumb (0.901), stupid (0.895), rude (0.893), bad (0.889), cheap (0.881), rubbish (0.880), loser (0.877), trashy (0.872), null (0.866), zero (0.855), idiotic (0.854), noob (0.815)	7	bad (2)	[Player_1: i am crying], [Player_2: swallow?], [Player_1: pls], [Player_3: bitch your pc is so fucking slow], [Player_1: YES], [Player_4: www.creedthoughts.gov /creedthoughts.net/gov], [Player_3: that was sad], [Player_2: men]	[Player_1: sorry i dont make the rules], [Player_2: are you a homo], [Player_3: are you?], [Player_2: cuz we migh be friends], [Player_1: tyes i love fat cock], [Player_3: {character} is 35hp], [Player_2: nuttttt], [Player_4: can you just leave]
2	gayest thing ive ever seen	weakest (0.897), lamest (0.896), noobest (0.868), most stupid (0.858), best (0.857), most unfair (0.855), dumb (0.847), most broken (0.846), worst (0.840), most disappointing (0.838), stupid (0.831), cheap (0.823), most unbalanced (0.822), terrible (0.819), bad (0.818), most displeasing (0.753)	7		No lines before	[Player_2: bit mad?], [Player_3: peak], [Player_1: mad at what ?], [Player_3: cant lie], [Player_4: who], [Player_1: im happy bro XD], [Player_2: seem to be a bit upset Player_1], [Player_2: cant lie]
3	ur kinda gay for that	dumb (0.912), rude (0.910), lame (0.907), stupid (0.902), weak (0.898), bad (0.883), idiot (0.878), thief (0.866), idiotic (0.865)	6	stupid (3), dumb (3), lame (2)	[Player_2: Player_1 is mad already for some reason lol], [Player_1: MY INJURE], [Player_3: nice Player_9], [Player_1: LMAO], [Player_4: Player_10 afk], [Player_1: IMAO], [Player_1: hugp], [Player_1: FUCK YUOU Player_9]	[Player_1: I HOPE U DIE], [Player_2: he thinks i stole the kill on {character}], [Player_1: ONG U DID], [Player_2: damn], [Player_1: THAT WAS MY INJURE], [Player_1: I SWEAR ONG], [Player_2: boohoo], [Player_3: play time]

4	who puts exclamation marks at the end of their sentences thats gay	lame (0.910), stupid (0.904), dumb (0.901), irrelevant (0.892), useless (0.889), retarded (0.794)	6	stupid (2), dumb (2)	[Player_2: SHUT THE FUCK YOU BITCH], [Player_3: youre in {rank}], [Player_3: please shut up], [Player_4: ima watch and ima be armed], [Player_3: you suck], [Player_2: Player_9 is coming {map location} for knife fight], [Player_4: no cheating!!!], [Player_5: you gotta be given free wins two game sin a row]	[Player_3: okay!], [Player_5: stfu {rank} trash can], [Player_2: im {breacher character} no shooting will be done trust me], [Player_2:Player_5 is a fucking TRASH ASS PUSSY], [Player_3: im p1 main], [Player_1: nah ggs to them free win], [Player_1: .79 kd], [Player_3: much love guys!]
5	and you are gay now	homosexual (0.971), queer (0.960), homo (0.944), nice (0.887), weak (0.880), stupid (0.878), caring (0.873), bad (0.870), traitor (0.867), a loser (0.857), noob (0.851), pacifist (0.836), fag (0.829), backstabber (0.825), whinning (0.782), faggot (0.782)	8		[Player_1: do you want to die ?], [Player_2: what], [Player_2: you can actually try ?], [Player_2: i thought you already were], [Player_2: virgin], [Player_2: na na na na], [Player_2: don't], [Player_3: this is a friendly match]	[Player_3: I have no reason to], [Player_2: get bitch slapped], [Player_2: bitch], [Player_3: wait], [Player_3: so you guys are actually trying rn?], [Player_3: thinking about buying {character}], [Player_3: wait], [Player_3: what is the limit?]
6	why r u so gay	queer (0.970), homo (0.945), dumb (0.923), stupid (0.916), weak (0.911), pussy (0.910), terrible (0.899), idiot (0.898), bad (0.896), mean (0.894), overly sensitive (0.872), idiotic (0.868), fag (0.833)	8	stupid (3), fag (2), dumb (2), bad (2)	[Player_2: reported kid], [Player_1: so gay], [Player_1: ^)jigger stfu], [Player_1: ^)jigger stfu], [Player_1: hope u get beheaded by isis], [Player_2: wow youre so upset], [Player_1: gay], [Player_2: for the random on their team can u tk him hes a child]	[Player_1: i will lynch ur dad], [Player_1: if u even know him], [Player_1: did he go buy milk?], [Player_2: yeah i do hes rich], [Player_1: u ^i)jiggerrr], [Player_1: yeah rich gay man], [Player_2: vote to kick], [Player_1: im doing better ;_]
7	yall gay or somethin?	homo (0.927), homosexual (0.916), dumb (0.903), pussy (0.895), stupid (0.888), annoying (0.881), immature (0.881), sensitive (0.880), bad (0.878), unfair (0.876), queers (0.871), childish (0.870), weak (0.864), new (0.862), terrible (0.841), affectionate (0.836), noob (0.829), misunderstanding (0.828), fag (0.811), fags (0.809), ganging up (0.784), traitors (0.774), faggots (0.774), disloyal (0.772), cheaters (0.763), colliders (0.721)	9	bad (3), homosexual (2), homo (2)	[Player_1: Player_5 that shouldnt have worked but that was amazing], [Player_2: sdfihusdfy8ushduifhjsuidnmfjionsiudonfinsdf], [Player_3: so funny], [Player_1: had to take out the bottom 2], [Player_4: :3], [Player_5: ez mid], [Player_1: Yo wtf], [Player_3: ez]	[Player_1: stfu f(AG)], [Player_1: stfu f(AG)], [Player_1: Hey Player_5, i complimented you. Avenge me], [Player_5: try harder], [Player_5: kids], [Player_5: get gud]
8	if u ban ur gay	homosexual (0.946), queer (0.939), homo (0.925), lame (0.899), bad (0.893), stupid (0.892), a pussy (0.892), annoying (0.890), salty	9	stupid (2), bad (2), homosexual (2), homo (2)	[Player_2: stick your hand in your ass and itll warm up], [Player_3: nothing better to do], [Player_3: ill just use your ass like your dad], [Player_3: we got {map name}], [Player_2: No	[Player_4: pre gn], [Player_2: no bad or your dad touches your asshole], [Player_4: u guys suck], [Player_1: gge], [Player_1: pre gn], [Player_2: DONT LISTEN TO SIMP HIS MENTAL

		(0.885), overly sensitive (0.884), baby (0.882), un-fun (0.880), unfair (0.873), weak (0.873), a loser (0.861), fag (0.855), whiny (0.853), noob (0.847), cowardly (0.844), compensating (0.792)			ban for the win], [Player_3: not a bad map], [Player_3: im just not good at it], [Player_4: pre ggs]	IS CHALKED], [Player_3: call me elton John im banning], [Player_4: its over for yall]
9	that's pretty gay of him	homosexual (0.965), queer (0.923), weak (0.907), dumb (0.898), rude (0.897), mean (0.889), stupid (0.881), unfair (0.880), homo (0.866), bad balancing (0.863), cowardly (0.861), not fun (0.851), feminin (0.824), not like us (0.822), party-poooper (0.771), disrespectful (0.702)	6	Stupid (2), homosexual (2)	[Player_1: I made a mistake choosing {battle tech}], [Player_2: remove {battle tech2}], [Player_2: ty], [Player_1: never], [Player_2: pls], [Player_1: wtf], [Player_3: gay], [Player_1: lad just negated knockback]	[Player_1: broke his nose but didn't budge the bastard], [Player_1: can you quit just spamming from the side please], [Player_2: Also why does {character's ability} doesn't break {armor} but { other character} does], [Player_1: yeah ikr], [Player_1: it's bs], [Player_1: can't believe that worked for as long as it did xD], [Player_1: wheeze], [Player_3: fix ping Player_9]
10	that was a gay move, running away like that	dumb (0.875), lame (0.873), cheap (0.866), coward (0.862), nasty (0.859), boring (0.855), mean (0.854), loser (0.851), low (0.850), weak (0.848), bad (0.845), annoying (0.844), ungamely (0.792), cowardly (0.789)	6	Cheap (3), weak (2), lame (2), bad (2)	No lines before	[Player_2: ur whole team runs], [Player_3: rm?]
11	maybe he's gay?	homosexual (0.973), queer (0.960), weak (0.889), not good (0.846)	7	Homosexual (3)	[Player_2: fart in my mouth?], [Player_3: sure!!!], [Player_1: oh nooo], [Player_3: {female character} and {male character} should put a ring on eachother], [Player_4: they did irl], [Player_1: bro how do you know {male character} is a heterosexual?], [Player_4: but its not disability month anymore so they broke up]	[Player_4: no it was for pride month], [Player_4: hey gu7s==], [Player_4: wlcom to my epsode of fornte], [Player_2: {character} blow urself up], [Player_3: /Vigger], [Player_1: he's cheating], [Player_1: you should blow urself upo buddy], [Player_1: omg]
12	why so mad we got gay banners	homosexual (0.975), queer (0.963), ugly (0.929), dumb (0.924), useless (0.904), colorful (0.887), dumbass (0.854), unfitting (0.840), flashy (0.833), sucky (0.818)	7		[Player_1: lmao], [Player_2: im sorry my teammates exist bro], [Player_2: ggs to the other team], [Player_3: cry about it], [Player_4: shitters], [Player_4: no cap], [Player_2: bro no one cares that you like dick], [Player_2: lmao]	[Player_4: your the one with gay flag], [Player_3: your just mad you dad beats youi]
13	and my parents know im gay	homosexual (0.948), lesbian (0.938), queer (0.932)	6	Homosexual (2)	[Player_2: @Player_1], [Player_3: ill be your mommy], [Player_2: pl], [Player_4: low ammo], [Player_1: now who is the hacker], [Player_5: ASSIST KING!], [Player_2: Player_1	[Player_2: he has reporetd], [Player_2: did you guys know that ash ketchums name from pokemon actually means catch them], [Player_3: you like dik?], [Player_2: like],

					says star trek sucks], [Player_2: and hes hacking]	[Player_1: i have been disowned =(], [Player_2: catch em all], [Player_4: like], [Player_3: i like dick]
14	Nothing wrong with being gay	homosexual (0.958), queer (0.930)	6	Homosexual (3)	[Player_2: samurai squad], [Player_1: weeb gang], [Player_1: I would've felt out being the only chinese man on the team], [Player_3: lmao this gon be fun], [Player_1: Not as fun as when my tongue hit your dad's prostate], [Player_3: fag]	[Player_2: lol], [Player_1: sorry about that], [Player_1: my brain blanked], [Player_2: it happens lol], [Player_3: such a pussy], [Player_3: {character} is so ass], [Player_2: gg], [Player_2: close]
15	gay people are normal tooo	homosexual (0.969), lgbt (0.962), queer (0.959), homo (0.944), lgbtq (0.902), lgbtq+ (0.840), lgbtq+ (0.754)	8	Homosexual (6), lgbtq+ (2), queer (2)	[Player_2: ok], [Player_2: {other team} got {map location}], [Player_3: ^^], [Player_1: anyone apart of the lgbtq+], [Player_1: pls lmk], [Player_3: yo mama], [Player_4: ggwp], [Player_1: not cool]	[Player_2: cap], [Player_1: stand up for the gays], [Player_3: no one said otherwise?], [Player_1: we riot til equal], [Player_3: why are you turning this into a thing], [Player_1: {character} main lol], [Player_1: Player_3 this penis lol], [Player_4: ggwp]
16	yeah and I'm part of the gays	homosexuals (0.957), homos (0.933), lgbt community (0.849), dick-lovers (0.844), lgbtq (0.838), homosexual (0.837), lgbtq+ community (0.789)	8	Homosexual (2)	[Player_2: your*], [Player_3: your**], [Player_2: yall ever shove a razorblade up ur butt], [Player_2: gotchu bbg], [Player_2: damn:/], [Player_4: pls add my snap im 5'1 200 pounds black single gay hmu], [Player_2: not me yall stay safe just wondering], [Player_5: HAHHAHAHhahahahh HAH HA H LOLOLOLOO LMFAO LFMAO AHHAHAHAHAHAha]	[Player_5: {character} i hope u burn in hell], [Player_5: gg], [Player_4: shit balls], [Player_2: {character} on main], [Player_4: god im sooo horny rn], [Player_4: need penis asap], [Player_2: cap], [Player_3: i hope u have a good rest of ur day]
17	ill accept being gay :)	homosexual (0.978), lgbt (0.966), queer (0.964), homo (0.956), dumb (0.907), bad (0.896), right (0.894), lgbtq+ (0.893), good (0.891), different (0.891), odd (0.874), noob (0.864), correct (0.855), decisive (0.843), the winner (0.746)	9	Homosexual (4), queer (2), homo (2), the winner (2)	[Player_2: defuse if gay], [Player_3: cock], [Player_4: oof i panicked], [Player_2: your gay], [Player_5: wow um..... i have no words], [Player_6: u know it daddy :*], [Player_1: thx], [Player_2: no fucking way]	[Player_4: t lesbian], [Player_2: almost clutched], [Player_1: ur lesbian], [Player_2: sorry but no one gives a fuck], [Player_2: sandwich], [Player_4: it was cuz he called me gay lol], [Player_3: sorry but nobody asked for your opinoon fucker], [Player_3: onion]
18	im gay too :)	homosexual (0.963), lgbt (0.925), queer (0.925), homo (0.902), friendly (0.850), lgbtq+ (0.842), tram-player (0.800), homosexual (0.790)	9	Homosexual (6), homo (2)	[Player_2: ass or tits], [Player_3: you went 6-6 last game sto talking lmao], [Player_4: dick], [Player_4: im gay], [Player_2: i mean r u tryna say some here], [Player_4: i eat dick for a living], [Player_2: bet], [Player_2: perfect]	[Player_2: when and where?], [Player_4: never], [Player_2: holy shit bro i come back and my names been mentioned 3 times do you get off to it? virgins smh], [Player_2: also guys can have fat asses and tits too], [Player_4: trannies], [Player_3: I wasnt paying attention], [Player_2: where did the talk go], [Player_2: {character} do not advise you to stand in the open like that]

19	Indeed, I am gay	homosexual (0.962), lesbian (0.947), homo (0.930), witty (0.890), wicked (0.862), simple (0.856), noob (0.840), owning (0.836), trolling (0.820), unskilled (0.778), (doesn't gaf) (0.747)	6	Homosexual (3)	[Player_2: they will regret facing us today], [Player_3: they already lost], [Player_2: as my corpse will drop a stinky ass shit], [Player_4: after u hit them with stick], [Player_2: steeeeeeckkk], [Player_4: if u play {character} u gaiii], [Player_4: get hit with beeeeh stick], [Player_1: w]	[Player_4: i know], [Player_2: POIIIRAAAAAAAATEE], [Player_1: yo], [Player_2: YOU RUN FROM A FAT MAN], [Player_2: I CANNOT REACH YOU AS I HAVE AN APPOINTMENT WITH DIABETES], [Player_2: PLEASE DONT DO THAT], [Player_4: POIIRRATE], [Player_1: srry]
20	i do, im gay	homosexual (0.951), homo (0.929), witty (0.887), proud (0.864), a woman (0.851), a troll (0.843), good (0.833), an ally to gays (0.794)	6	Homosexual (5)	[Player_2: loose asshole], [Player_1: f, died again like in the bible huh], [Player_2: why you so mad lol], [Player_2: cant take a joke], [Player_1: istn it normal to complain about bitch bois], [Player_2: bet you only take dick in the ass], [Player_2: EZ], [Player_3: ez]	[Player_2: bet you are], [Player_1: are you homophobic you sad asshole], [Player_2: your parents dont love you]
21	i smell black	cheap (0.887), ghetto (0.882), bad (0.872), potato (0.870), poor (0.868), picky (0.852), stingy (0.850), poverty (0.842), oldness (0.841), low quality (0.841), broke (0.839), poorness (0.837), nigger (0.818), a poor person (0.800), lower class (0.799), afro-american (0.799), african american (0.797), helplessness (0.780), nigga (0.768), uneducated (0.766), penniless (0.764), lower-class (0.747), someone of socially low hierarchy (0.721), underprivileged (0.644)	18	Poor (9), cheap (4), underprivileged (2)	[Player_2: I can drop a follow], [Player_3: go for it], [Player_1: hey Player3 what kind of pc do you have], [Player_3: clearly a cheap ass one with that lag], [Player_3: like 49.99\$]	[Player_2: say less], [Player_3: lose or win im plugging my twitch], [Player_4: shameless], [Player_2: {twitch link}], [Player_2: {twitch link}]
22	{character}'s helmet so black it applied for welfare	white (0.963), red (0.961), blue (0.959), yellow (0.950), pink (0.947), dark (0.940), green (0.936), poor (0.889), rural (0.887), ghetto (0.872), null (0.871), cheap (0.867), ridiculous (0.867), bad (0.856), broke (0.834), unemployed (0.828), lazy (0.826), african american (0.825), color-coded (0.812), underpower (0.809), low-budget (0.803), of african descent (0.760), of low social standard (0.728)	18	Poor (10), african american (2), cheap (2), bad (2), dark (2)	No lines before	[Player_2: lol], [Player_3: ong XDDD], [Player_3: sssshhhhhiiiiiiit], [Player_2: k come to {map location}], [Player_3: nice], [Player_4: CMON], [Player_3: ohhh nice gg], [Player_4: no time]

23	bruh im black	black (1.000), dark (0.934), african (0.907), actually black (0.899), minority (0.896), good (0.891), man (0.878), dark skinned (0.877), dark skin colored (0.860), poc (0.828), a gangster (0.815), afro-american (0.806), not arab (0.805), a criminal (0.792), african american (0.783), not muslim (0.782), a terrorist (0.777), a person of color (0.775), a minority (0.766), african-descent (0.765), african-american (0.755)	18	African american (2), african-american (2), African (2), black (2), afro-american (2), dark skinned (2)	[Player_1: no], [Player_2: pig poo!!], [Player_1: grow up], [Player_2: stfu], [Player_2: bomb urself]	[Player_2: gr], [Player_3: Player1 is indeed dark], [Player_2: you sound like an afghan terrorist], [Player_4: ninjer]
24	youre black arent you	dark (0.952), gay (0.901), rural (0.899), poor (0.897), new (0.897), bad (0.893), chinese (0.893), needy (0.881), korean (0.880), bad race (0.870), japanese (0.869), african american (0.819), nigger (0.811), like big butts (0.809), unskilled (0.800), (missing context)? (0.790), unimportant (0.771), using slang (0.762), uneducated (0.746), slutty (0.746)	18	Poor (3), gay (2), african american (2), dark (2), bad (2)	[Player_2: stop it], [Player_2: you], [Player_3: im boutta eat ur ass bruh watch], [Player_3: wait just a bit], [Player_3: ur booty is mine]	[Player_2: better not use that lame ass {character} lol], [Player_3: nah ill use much better], [Player_3: LMAO haha], [Player_2: worst {character} ive seen tonigt], [Player_1: jj hahah lmao HAHHAH]

APPENDICE D
REGROUPEMENTS OBTENUS

Tableau 6.6 Tous les regroupements obtenus avec les plongements (BERT) des substituts.

#	Ligne cible	k	ID	Substituts	Plus représentatif	Inertia	Silhouette
1	gay for not letting us win	4	0	['noob']	noob	61.344	0.437
			1	['stupid', 'lame', 'weak', 'dumb', 'cheap', 'bad']	lame		
			2	['zero', 'null']	null		
			3	['idiotic', 'rubbish', 'loser', 'trashy', 'rude']	loser		
2	gayest thing ive ever seen	7	0	['lamest', 'weakest']	lamest	31.628	0.545
			1	['best', 'worst']	best		
			2	['bad', 'terrible']	terrible		
			3	['most unbalanced', 'most displeasing']	most unbalanced		
			4	['most unfair', 'most stupid', 'most broken', 'most disappointing']	most broken		
			5	['stupid', 'dumb', 'cheap']	stupid		
			6	['noobest']	noobest		
3	ur kinda gay for that	4	0	['thief']	thief	24.124	0.456
			1	['stupid', 'weak', 'lame', 'dumb', 'rude']	rude		
			2	['idiot', 'idiotic']	idiot		
			3	['bad']	bad		
4	who puts exclamation marks at the end of their sentences thats gay	4	0	['retarded']	retarded	5.393	0.477
			1	['stupid', 'dumb']	stupid		
			2	['irrelevant', 'useless']	useless		
			3	['lame']	lame		
5	and you are gay now	4	0	['backstabber', 'noob', 'whinning', 'pacifist', 'faggot', 'fag']	noob	128.154	0.474
			1	['nice', 'caring']	nice		
			2	['homosexual', 'queer', 'homo']	queer		

			3	['stupid', 'traitor', 'weak', 'a loser', 'bad']	weak		
6	why r u so gay	5	0	['idiotic', 'overly sensitive']	idiotic	23.447	0.543
			1	['pussy', 'queer', 'homo']	homo		
			2	['mean', 'bad', 'terrible']	mean		
			3	['fag']	fag		
			4	['stupid', 'idiot', 'weak', 'dumb']	weak		
7	yall gay or somethin?	2	0	['homosexual', 'stupid', 'annoying', 'terrible', 'misunderstanding', 'pussy', 'immature', 'bad', 'unfair', 'childish', 'new', 'sensitive', 'homo', 'weak', 'affectionate', 'dumb']	unfair	451.342	0.481
			1	['ganging up', 'cheaters', 'fag', 'disloyal', 'noob', 'traitors', 'colliders', 'fags', 'queers', 'faggots']	queers		
8	if u ban ur gay	4	0	['homosexual', 'homo', 'queer', 'baby']	baby	92.865	0.572
			1	['lame', 'stupid', 'annoying', 'salty', 'bad', 'unfair', 'weak', 'overly sensitive']	sensitive		
			2	['cowardly', 'whiny', 'fag', 'noob', 'un-fun', 'compensating']	un-fun		
			3	['a loser', 'a pussy']	pussy		
9	that's pretty gay of him	5	0	['not like us', 'not fun', 'bad balancing']	not fun	156.432	0.378
			1	['stupid', 'unfair', 'weak', 'cowardly', 'queer', 'dumb', 'rude']	cowardly		
			2	['disrespectful']	disrespectful		
			3	['homosexual', 'mean', 'homo']	homosexual		
			4	['feminin', 'party-poopers']	feminin		
10	that was a gay move, running away like that	2	0	['lame', 'weak', 'annoying', 'loser', 'boring', 'low', 'coward', 'dumb', 'mean', 'cheap', 'bad', 'nasty']	dumb	229.267	0.464
			1	['cowardly', 'ungamely']	ungamely		
11	maybe he's gay?	2	0	['homosexual', 'queer']	queer	89.985	0.249
			1	['not good', 'weak']	weak		
12	why so mad we got gay banners	4	0	['colorful']	colorful	71.861	0.388951
			1	['homosexual', 'dumb', 'queer', 'useless', 'ugly']	useless		
			2	['dumbass', 'flashy', 'sucky']	sucky		
			3	['unfitting']	unfitting		
15	gay people are normal tooo	3	0	['lgbtq+']	lgbtq+	52.011	0.46654
			1	['homosexual', 'lgbt', 'queer', 'homo']	lgbt		

			2	['lgbtq', 'lgbtq+']	lgbtq+		
16	yeah and I'm part of the gays	4	0	['homos', 'homosexuals']	homosexuals	30.512	0.452
			1	['lgbtq', 'lgbtq+ community', 'lgbt community']	lgbtq		
			2	['dick-lovers']	dick-lovers		
			3	['homosexual']	homosexual		
17	ill accept being gay :)	4	0	['lgbtq+', 'noob']	noob	63.303	0.556
			1	['homosexual', 'lgbt', 'queer', 'homo']	homo		
			2	['correct', 'right', 'good', 'odd', 'dumb', 'bad', 'decisive', 'different']	dumb		
			3	['the winner']	winner		
18	im gay too :)	4	0	['homosexual', 'lgbt', 'queer', 'homo']	lgbt	36.185	0.439
			1	['homsexual', 'tram-player']	homsexual		
			2	['friendly']	friendly		
			3	['lgbtq+']	lgbtq+		
19	Indeed, I am gay	4	0	['homosexual', 'lesbian', 'homo']	homo	60.391	0.615
			1	['noob', 'unskilled', '(doesnâ€™t gaf)']	noob		
			2	['simple', 'witty', 'wicked']	wicked		
			3	['owning', 'trolling']	trolling		
20	i do, im gay	4	0	['an ally to gays']	an ally to gays	29.359	0.590
			1	['proud', 'good', 'witty']	witty		
			2	['a troll', 'a woman']	a troll		
			3	['homosexual', 'homo']	homo		
21	i smell black	3	0	['african american', 'penniless', 'afro-american', 'nigger', 'nigga', 'underprivileged', 'uneducated']	nigger	369.164	0.440
			1	['poverty', 'ghetto', 'cheap', 'broke', 'oldness', 'bad', 'poor', 'a poor person', 'picky', 'stingy', 'potato', 'poorness', 'helplessness', 'low quality']	a poor person		
			2	['lower-class', 'lower class', 'someone of socially low hierarchy']	lower class		
22	{character}'s helmet so black it applied for welfare	3	0	['rural', 'ghetto', 'cheap', 'broke', 'unemployed', 'ridiculous', 'bad', 'poor', 'lazy', 'null', 'underpower']	ghetto	239.337	0.540
			1	['white', 'pink', 'dark', 'green', 'red', 'yellow', 'blue']	dark		
			2	['african american', 'of african descent', 'color-coded', 'of low social standard', 'low-budget']	color-coded		

23	bruh im black	4	0	['african american', 'afro-american', 'african-descent', 'african-american']	afro-american	130.698	0.619
			1	['not arab', 'dark skin colored', 'not muslim', 'dark skinned', 'actually black']	dark skinned		
			2	['a criminal', 'a gangster', 'a person of color', 'a terrorist', 'a minority']	a person of color		
			3	['good', 'dark', 'black', 'man', 'minority', 'african', 'poc']	poc		
24	youre black arent you	4	0	['unimportant', 'unskilled', 'nigger', 'uneducated', 'slutty']	nigger	134.340	0.579

Tableau 6.7 Tous les regroupements obtenus avec les plongements (RoBERTa) des substituts.

#	Ligne cible	k	ID	Substituts	Plus représentatif	Inertia	Silhouette
1	gay for not letting us win	2	0	['stupid', 'idiotic', 'lame', 'weak', 'rubbish', 'zero', 'null', 'trashy', 'cheap', 'rude']	stupid	180.072	0.531
			1	['loser', 'noob', 'dumb', 'bad']	loser		
2	gayest thing ive ever seen	5	0	['most unbalanced', 'most unfair', 'most displeasing', 'most disappointing']	most unfair	36.161	0.501
			1	['best', 'noobest', 'lamest', 'worst', 'weakest']	worst		
			2	['dumb', 'bad']	bad		
			3	['most broken', 'cheap', 'terrible']	cheap		
			4	['stupid', 'most stupid']	most stupid		
3	ur kinda gay for that	4	0	['thief']	thief	66.785	0.240
			1	['idiot', 'idiotic', 'dumb']	dumb		
			2	['stupid', 'lame', 'rude']	lame		
			3	['weak', 'bad']	weak		
4	who puts exclamation marks at the end of their sentences thats gay	4	0	['dumb', 'useless']	dumb	16.948	0.328
			1	['stupid', 'retarded']	stupid		
			2	['irrelevant']	irrelevant		
			3	['lame']	lame		
5	and you are gay now	6	0	['traitor', 'faggot']	faggot	45.894	0.414
			1	['nice', 'noob', 'bad']	noob		
			2	['homosexual', 'homo', 'pacifist']	pacifist		

			3	['weak', 'backstabber', 'queer', 'caring', 'whinning']	queer		
			4	['fag']	fag		
			5	['stupid', 'a loser']	loser		
6	why r u so gay	5	0	['weak', 'queer', 'homo', 'terrible', 'overly sensitive']	queer	55.823	0.401
			1	['idiot', 'dumb']	dumb		
			2	['pussy', 'fag']	pussy		
			3	['stupid', 'idiotic']	stupid		
			4	['mean', 'bad']	bad		
7	yall gay or somethin?	2	0	['homosexual', 'stupid', 'terrible', 'misunderstanding', 'ganging up', 'cheaters', 'immature', 'disloyal', 'unfair', 'childish', 'sensitive', 'traitors', 'colliders', 'homo', 'fags', 'queers', 'affectionate', 'faggots']	unfair	276.341	0.465
			1	['annoying', 'pussy', 'bad', 'fag', 'new', 'noob', 'weak', 'dumb']	weak		
8	if u ban ur gay	2	0	['homosexual', 'lame', 'stupid', 'cowardly', 'whiny', 'salty', 'unfair', 'homo', 'weak', 'un-fun', 'a loser', 'queer', 'baby', 'overly sensitive', 'compensating', 'a pussy']	queer	241.131	0.404
			1	['annoying', 'bad', 'fag', 'noob']	noob		
9	that's pretty gay of him	3	0	['homosexual', 'stupid', 'feminin', 'unfair', 'cowardly', 'party-pooper', 'disrespectful', 'homo', 'bad balancing', 'rude']	unfair	134.721	0.424
			1	['not like us', 'weak', 'not fun', 'queer']	queer		
			2	['mean', 'dumb']	dumb		
10	that was a gay move, running away like that	3	0	['mean']	mean	103.097	0.436
			1	['lame', 'weak', 'cowardly', 'boring', 'low', 'coward', 'cheap', 'nasty', 'ungamely']	low		
			2	['annoying', 'loser', 'dumb', 'bad']	dumb		
11	maybe he's gay?	2	0	['queer', 'not good', 'weak']	not good	107.929	0.063
			1	['homosexual']	homosexual		
12	why so mad we got gay banners	2	0	['dumb', 'useless']	dumb	156.742	0.489
			1	['homosexual', 'colorful', 'dumbass', 'flashy', 'queer', 'sucky', 'ugly', 'unfitting']	dumbass		
15	gay people are normal tooo	3	0	['lgbt', 'lgbtq', 'lgbtq+', 'lgbtq+']	lgbtq+	39.404	0.494
			1	['homosexual', 'homo']	homo		
			2	['queer']	queer		

17	ill accept being gay :)	3	0	['homosexual', 'correct', 'odd', 'queer', 'decisive', 'homo', 'different', 'the winner']	winner	137.527	0.518
			1	['right', 'good', 'noob', 'dumb', 'bad']	dumb		
			2	['lgbt', 'lgbtq+']	lgbtq+		
18	im gay too :)	3	0	['homosexual', 'homsexual', 'homo']	homsexual	66.925	0.551
			1	['lgbt', 'lgbtq+']	lgbtq+		
			2	['friendly', 'tram-player', 'queer']	tram-player		
19	Indeed, I am gay	4	0	['owning', 'trolling', '(doesn't gaf)']	trolling	40.345	0.380
			1	['noob']	noob		
			2	['homosexual', 'lesbian', 'homo']	lesbian		
			3	['simple', 'unskilled', 'witty', 'wicked']	simple		
20	i do, im gay	2	0	['homosexual', 'a troll', 'an ally to gays', 'proud', 'homo', 'witty', 'a woman']	a troll	140.287	0.406
			1	['good']	good		
21	i smell black	3	0	['broke', 'bad', 'poor', 'poorness']	broke	152.072	0.557
			1	['african american', 'penniless', 'poverty', 'ghetto', 'lower-class', 'lower class', 'afro-american', 'cheap', 'underprivileged', 'oldness', 'uneducated', 'someone of socially low hierarchy', 'a poor person', 'picky', 'stingy', 'potato', 'helplessness', 'low quality']	cheap		
			2	['nigger', 'nigga']	nigger		
22	{character}'s helmet so black it applied for welfare	4	0	['african american', 'rural', 'ghetto', 'unemployed', 'ridiculous', 'of african descent', 'color-coded', 'of low social standard', 'underpower', 'low-budget']	rural	83.401	0.415
			1	['white', 'cheap', 'pink', 'broke', 'dark', 'green', 'poor', 'red', 'lazy', 'yellow', 'blue']	pink		
			2	['bad']	bad		
			3	['null']	null		
23	bruh im black	7	0	['a criminal', 'a gangster', 'a person of color', 'not muslim', 'minority', 'a terrorist', 'a minority']	muslim	20.487	0.532
			1	['good']	good		
			2	['black', 'actually black']	black		
			3	['african american', 'afro-american', 'african-descent', 'african-american', 'african']	african-descent		

			4	['dark skin colored', 'dark', 'dark skinned']	dark skin colored		
			5	['not arab', 'poc']	not arab		
			6	['man']	man		
24	youre black arent you	4	0	['using slang', 'rural', 'needy', 'dark', 'nigger', 'like big butts', 'bad race', 'slutty']	bad race	70.722	0.530
			1	['bad', 'new', 'gay']	gay		
			2	['african american', 'korean', 'japanese', 'chinese']	korean		
			3	['unimportant', 'unskilled', 'uneducated', 'poor', '(missing context)?']	poor		

RÉFÉRENCES

- ADL. (2020, décembre). *Disruption and harms in online gaming framework*. Anti-Defamation League. <https://www.adl.org/fpa-adl-games-framework>
- ADL. (2022). *Hate is no game : Hate and harassment in online games 2022*. <https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022>
- Anderson, L., & Barnes, M. (2022). Hate speech. Dans E. N. Zalta (Éd.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/entries/hate-speech/>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, mai 23). *Machine bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=TiqCeZlj4uLbXI91e3wM2PnmnWbCVOvS>
- Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing subjective tasks : The case study of understanding toxicity in online discussions. *Companion Proceedings of The 2019 World Wide Web Conference*, 1100-1105. <https://doi.org/10.1145/3308560.3317083>
- Arsht, A., & Etcovich, D. (2018, mars 2). *The human cost of online content moderation*. Harvard Journal of Law & Technology. <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation>
- Ashwell, L. (2016). Gendered slurs. *Social Theory and Practice*, 42(2), 228-239. <https://doi.org/10.5840/soctheorpract201642213>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots : Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. <https://doi.org/10.1145/3442188.3445922>
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017, mars 27). *Fairness in criminal justice risk assessments : The state of the art*. arXiv.Org. <https://arxiv.org/abs/1703.09207v2>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power : A critical survey of “bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings* (No. arXiv:1607.06520). arXiv. <https://doi.org/10.48550/arXiv.1607.06520>
- Bordia, S., & Bowman, S. R. (2019). *Identifying and reducing gender bias in word-level language models* (No. arXiv:1904.03035). arXiv. <https://doi.org/10.48550/arXiv.1904.03035>
- Brown, J. R. (2011). No homo. *Journal of Homosexuality*, 58(3), 299-314. <https://doi.org/10.1080/00918369.2011.546721>

- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. <https://doi.org/10.1126/science.aal4230>
- Canada, E. et C. climatique. (2017, mai 5). *Définition de la toxicité* [Lignes directrices]. <https://www.canada.ca/fr/environnement-changement-climatique/services/registre-environnemental-loi-canadienne-protection/listes-substances/definition-toxicite.html>
- Carillo, K., & Marsan, J. (2016, décembre 1). *“the dose makes the poison”—Exploring the toxicity phenomenon in online communities*.
- Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it’s like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. Dans R. Fernández & T. Linzen (Éds.), *Proceedings of the 24th Conference on Computational Natural Language Learning* (p. 227-244). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.conll-1.17>
- Ciftci, T., Gashi, L., Hoffmann, R., Bahr, D., Ilhan, A., & Fietkiewicz, K. (2017, janvier). *Hate speech on Facebook*.
- Cruz, A. M. F. da. (2020). *Fairness-aware hyperparameter optimization*. <https://repositorio-aberto.up.pt/handle/10216/128959>
- Dastin, J. (2018, octobre 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11, 300-319. https://doi.org/10.1162/tacl_a_00550
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). *Racial bias in hate speech and abusive language detection datasets* (No. arXiv:1905.12516). arXiv. <http://arxiv.org/abs/1905.12516>
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language* (No. arXiv:1703.04009). arXiv. <http://arxiv.org/abs/1703.04009>
- Davis, C., & McCready, E. (2020). The instability of slurs. *Grazer Philosophische Studien*, 97(1), 63-85. <https://doi.org/10.1163/18756735-09701005>
- De Smedt, T., Jaki, S., Kotzé, E., Saoud, L., Gwózdź, M., De Pauw, G., & Daelemans, W. (2018). Multilingual cross-domain perspectives on online hate speech. *arXiv:1809.03944 [cs]*. <http://arxiv.org/abs/1809.03944>
- DeepAI. (2019, mai 17). *Hidden layer*. DeepAI. <https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [Cs]*. <http://arxiv.org/abs/1810.04805>

- Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67-73. <https://doi.org/10.1145/3278721.3278729>
- Dolphin, R. (2023, février 14). *The power of embeddings in machine learning*. Medium. <https://towardsdatascience.com/exploring-the-power-of-embeddings-in-machine-learning-18a601238d6b>
- Doucerein, M. M. (2020). Stéréotypes et préjugés. Dans É. Gagnon-St-Pierre, C. Gratton, & E. Muszynski (Éds.), *Raccourcis : Guide pratique des biais cognitifs Vol. 2*. www.shortcogs.com
- Du, W., & Wu, X. (2021). Fair and robust classification under sample selection bias. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2999-3003. <https://doi.org/10.1145/3459637.3482104>
- Erk, K., & Chronis, G. (2022). Word embeddings are word story embeddings (and that's fine). Dans S. Lappin & J.-P. Bernardy, *Algebraic Structures in Natural Language* (1^{re} éd., p. 189-218). CRC Press. <https://doi.org/10.1201/9781003205388-9>
- Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in reddit. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 87-96. <https://doi.org/10.1145/3292522.3326045>
- Fehn Unsvåg, E., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. Dans D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Éds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (p. 75-85). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-5110>
- Fellbaum, C. (1998). *WordNet : An electronic lexical database*. The MIT Press. <https://doi.org/10.7551/mitpress/7287.001.0001>
- Fillmore, C. J. (1982). Frame semantics. *Linguistics in the Morning Calm, The Linguistic Society of Korea*, 111-137.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. Dans *Studies in Linguistic Analysis* (p. 1-32). Philological Society.
- Garg, T., Masud, S., Suresh, T., & Chakraborty, T. (2023). Handling bias in toxic speech detection : A survey. *ACM Computing Surveys*, 55(13s), 264:1-264:32. <https://doi.org/10.1145/3580494>
- Gastaldi, J. L. (2021). Why can computers understand natural language? *Philosophy & Technology*, 34(1), 149-214. <https://doi.org/10.1007/s13347-020-00393-9>
- Geiger, A., Wu, Z., D'Oosterlinck, K., Kreiss, E., Goodman, N. D., Icard, T., & Potts, C. (2022, octobre 31). *Faithful, interpretable model explanations via causal abstraction*. <http://ai.stanford.edu/blog/causal-abstraction/>
- Gevers, I., Markov, I., & Daelemans, W. (2022). Linguistic analysis of toxic language on social media. *32nd meeting of computational linguistics in the netherlands, CLIN 2022*, 12, 33-48.

- Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases : The psychology of intuitive judgment*. Cambridge University Press.
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig : Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv:1903.03862 [cs]*. <http://arxiv.org/abs/1903.03862>
- Government of Canada. (2022, août 28). *2SLGBTQI+ terminology* [Glossary and common acronyms]. Gouvernement of Canada. <https://www.canada.ca/en/women-gender-equality/free-to-be-me/2slgbtqi-plus-glossary.html>
- Gratton, C., & Gagnon-St-Pierre, É. (2020). *Heuristiques et biais cognitifs*. Raccourcis : Guide pratique des biais cognitifs. <https://www.shortcogs.com>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition : The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464-1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Haas, C. (2019). The price of fairness—A framework to explore trade-offs in algorithmic fairness. *ICIS 2019 Proceedings*. https://aisel.aisnet.org/icis2019/data_science/data_science/19
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489-1501. <https://doi.org/10.18653/v1/P16-1141>
- Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. Dans J. Burstein, C. Doran, & T. Solorio (Éds.), *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)* (p. 4129-4138). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1419>
- Hoover, B., Strobelt, H., & Gehrmann, S. (2019). exBERT : A visual analysis tool to explore learned representations in Transformers models. *arXiv:1910.05276 [cs]*. <http://arxiv.org/abs/1910.05276>
- Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers : A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2), 11:1-11:52. <https://doi.org/10.1145/3631326>
- Ingersoll, C. (2019, juillet). *Free to play? Hate, harassment, and positive social experiences in online games*. Anti-Defamation League. <https://www.adl.org/free-to-play>
- Jaki, S., De Smedt, T., Gwóźdz, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum : Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2), 240-268. <https://doi.org/10.1075/jlac.00026.jak>
- Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651-3657. <https://doi.org/10.18653/v1/P19-1356>

- Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). *Are natural language inference models IMPPRESSive? Learning IMPLicature and PRESupposition*. <https://arxiv.org/abs/2004.03066v2>
- Jhaveri, M., Ramaiya, D., & Chadha, H. S. (2022). *Toxicity detection for indic multilingual social media content* (No. arXiv:2201.00598). arXiv. <https://doi.org/10.48550/arXiv.2201.00598>
- Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*, 924-929. 2012 IEEE 12th International Conference on Data Mining (ICDM). <https://doi.org/10.1109/ICDM.2012.45>
- Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2017). Exploiting reject option in classification for social discrimination control. *Information Sciences*, 425. <https://doi.org/10.1016/j.ins.2017.09.064>
- Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv:1805.04508 [cs]*. <http://arxiv.org/abs/1805.04508>
- Kowert, R. (2020). Dark participation in games. *Frontiers in Psychology*, 11, 598947. <https://doi.org/10.3389/fpsyg.2020.598947>
- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture : Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5), 905-949. <https://doi.org/10.1177/0003122419877135>
- Kremer, G., Erk, K., Padó, S., & Thater, S. (2014). What substitutes tell us—Analysis of an “all-words” lexical substitution corpus. Dans S. Wintner, S. Goldwater, & S. Riezler (Éds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (p. 540-549). Association for Computational Linguistics. <https://doi.org/10.3115/v1/E14-1057>
- Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). *Measuring bias in contextualized word representations* (No. arXiv:1906.07337). arXiv. <http://arxiv.org/abs/1906.07337>
- Kwak, H., & Blackburn, J. (2015). Linguistic analysis of toxic behavior in an online video game. Dans L. M. Aiello & D. McFarland (Éds.), *Social Informatics* (Vol. 8852, p. 209-217). Springer International Publishing. https://doi.org/10.1007/978-3-319-15168-7_26
- Kwak, H., Blackburn, J., & Han, S. (2015). *Exploring cyberbullying and other toxic behavior in team competition online games*. 22. <https://doi.org/10.1145/2702123.2702529>
- Le Robert. (s. d.). Restaurant. Dans *Le Robert*. <https://dictionnaire.lerobert.com/google-dictionnaire-fr?param=restaurant>
- Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). *A new generation of perspective API : Efficient multilingual character-level transformers*. <https://doi.org/10.48550/ARXIV.2202.11176>
- Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). *Toxic language detection in social media for brazilian portuguese : New dataset and multilingual analysis* (No. arXiv:2010.04543). arXiv. <https://doi.org/10.48550/arXiv.2010.04543>

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, juillet 26). *RoBERTa : A robustly optimized BERT pretraining approach*. arXiv.Org. <https://arxiv.org/abs/1907.11692v1>
- Loi sur les préjudices en ligne, Projet de loi C-63 § 44 (2021). <https://www.parl.ca/LegisInfo/fr/projet-de-loi/44-1/c-63>
- Lu, J., Ren, X., Ren, Y., Liu, A., & Xu, Z. (2020). Improving contextual language models for response retrieval in multi-turn conversation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1805-1808. <https://doi.org/10.1145/3397271.3401255>
- MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, 49(1), 259-287. <https://doi.org/10.1146/annurev.psych.49.1.259>
- Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What happens to BERT embeddings during fine-tuning? *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 33-44. <https://doi.org/10.18653/v1/2020.blackboxnlp-1.4>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (No. arXiv:1301.3781). arXiv. <https://doi.org/10.48550/arXiv.1301.3781>
- Miller, N. (2019). *Dispelling common player behavior myths (presented by fair play alliance)*. GDC19. <https://gdcvault.com/play/1025805/Dispelling-Common-Player-Behavior-Myths>
- Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on arabic social media. Dans Z. Waseem, W. H. K. Chung, D. Hovy, & J. Tetreault (Éds.), *Proceedings of the First Workshop on Abusive Language Online* (p. 52-56). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-3008>
- Padó, S., & Lapata, M. (2003). Constructing semantic space models from parsed corpora. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, 128-135. <https://doi.org/10.3115/1075096.1075113>
- Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799-2804. <https://doi.org/10.18653/v1/D18-1302>
- Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). *Toxicity detection : Does context really matter?* (No. arXiv:2006.00998). arXiv. <https://doi.org/10.48550/arXiv.2006.00998>
- Pennington, J., Socher, R., & Manning, C. (2014). Glove : Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. <https://doi.org/10.3115/v1/D14-1162>
- Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021). *Fair bayesian optimization* (No. arXiv:2006.05109). arXiv. <https://doi.org/10.48550/arXiv.2006.05109>

- Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019). Perturbation sensitivity analysis to detect unintended model biases. *arXiv:1910.04210 [cs]*. <http://arxiv.org/abs/1910.04210>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67.
- Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., & Scheffczyk, J. (2010). *FrameNet II : Extended theory and practice*.
- Sahlgren, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. <https://www.semanticscholar.org/paper/The-Word-Space-Model%3A-using-distributional-analysis-Sahlgren/4996835ae239edc4d213940bfefa6d2c7e0ffbcd>
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668-1678. <https://doi.org/10.18653/v1/P19-1163>
- Seal, M. (2021). No homo : Reactions to the interruption of heteronormativity on a youth and community work course. *Journal of LGBT Youth*, 18(4), 459-489. <https://doi.org/10.1080/19361653.2020.1727394>
- Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models : A conceptual framework and overview. Dans D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Éds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5248-5264). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.468>
- Single market for digital services (Digital Services Act), 2022/2065 102 (2022). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065>
- Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2023). *Smiling women pitching down : Auditing representational and presentational gender biases in image generative AI* (No. arXiv:2305.10566). arXiv. <http://arxiv.org/abs/2305.10566>
- Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1-9. <https://doi.org/10.1145/3465416.3483305>
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillcrap, T., Lazaridou, A., ... Vinyals, O. (2023, décembre 19). *Gemini : A family of highly capable multimodal models*. arXiv.Org. <https://arxiv.org/abs/2312.11805v3>

- Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P. G., Kumar, D., McCoy, D., Meiklejohn, S., Ristenpart, T., & Stringhini, G. (2021). *SoK : Hate, harassment, and the changing landscape of online abuse*. 247-267. <https://doi.org/10.1109/SP40001.2021.00028>
- Tizpaz-Niari, S., Kumar, A., Tan, G., & Trivedi, A. (2022). Fairness-aware configuration of machine learning libraries. *Proceedings of the 44th International Conference on Software Engineering*, 909-920. <https://doi.org/10.1145/3510003.3510202>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2 : Open foundation and fine-tuned chat models* (No. arXiv:2307.09288). arXiv. <https://doi.org/10.48550/arXiv.2307.09288>
- Unity. (2021). *Toxicity in multiplayer games report*. Harris On Demand - The Harris Poll. https://create.unity.com/toxicity-in-multiplayer-games-report?_ga=2.241060039.576245343.1642088471-1978310244.1639494506
- Usito. (s. d.). *Toxicité*. Usito. Consulté 1 janvier 2024, à l'adresse <https://usito.usherbrooke.ca/définitions/toxicité>
- Van Dorpe, J., & Sénéchal, S. (2022). Analyse automatique des biais revendiqués par la communauté d'incels.is. *Actes du Colloque des Étudiant.e.s de Linguistique*, 72-101.
- Van Dorpe, J., Yang, Z., Grenon-Godbout, N., & Winterstein, G. (2023). Unveiling identity biases in toxicity detection : A game-focused dataset and reactivity analysis approach. Dans M. Wang & I. Zitouni (Éds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing : Industry Track* (p. 263-274). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-industry.26>
- van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification : An in-depth error analysis. Dans D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Éds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (p. 33-42). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-5105>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, juin 12). *Attention is all you need*. arXiv.Org. <https://arxiv.org/abs/1706.03762v7>
- Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. (2014). Cursing in English on Twitter. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 415-424. <https://doi.org/10.1145/2531602.2531734>
- Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter : A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6, 13825-13835. <https://doi.org/10.1109/ACCESS.2018.2806394>
- Webster, C. S., Taylor, S., Thomas, C., & Weller, J. M. (2022). Social bias, discrimination and inequity in healthcare : Mechanisms, implications and recommendations. *BJA Education*, 22(4), 131-137. <https://doi.org/10.1016/j.bjae.2021.11.011>

- Weld, H., Huang, G., Lee, J., Zhang, T., Wang, K., Guo, X., Long, S., Poon, J., & Han, C. (2021). CONDA : A CONTEXTUAL Dual-Annotated dataset for in-game toxicity understanding and detection. Dans C. Zong, F. Xia, W. Li, & R. Navigli (Éds.), *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021* (p. 2406-2416). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.213>
- Wittgenstein, L. (2014). *Recherches philosophiques* (É. Rigal, Trad.). Gallimard.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). *HuggingFace's Transformers : State-of-the-art natural language processing* (No. arXiv:1910.03771). arXiv. <http://arxiv.org/abs/1910.03771>
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016, septembre 26). *Google's neural machine translation system : Bridging the gap between human and machine translation*. arXiv.Org. <https://arxiv.org/abs/1609.08144v2>
- Wulczyn, E., Thain, N., & Dixon, L. (2017). *Ex machina : Personal attacks seen at scale* (No. arXiv:1610.08914). arXiv. <https://doi.org/10.48550/arXiv.1610.08914>
- Yang, Z., Grenon-Godbout, N., & Rabbany, R. (2023). Towards detecting contextual real-time toxicity for in-game chat. Dans H. Bouamor, J. Pino, & K. Bali (Éds.), *Findings of the Association for Computational Linguistics : EMNLP 2023* (p. 9894-9906). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.663>
- Yang, Z., Grenon-Godbout, N., & Rabbany, R. (2024). Game on, hate off : A study of toxicity in online multiplayer environments. *ACM Games*, 2(2), 14:1-14:13. <https://doi.org/10.1145/3675805>
- Yu, Z., Chakraborty, J., & Menzies, T. (2023). *FairBalance : How to achieve equalized odds with data pre-processing* (No. arXiv:2107.08310). arXiv. <https://doi.org/10.48550/arXiv.2107.08310>
- Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)* [Software]. <https://doi.org/10.17605/OSF.IO/MD832>
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). *Men also like shopping : Reducing gender bias amplification using corpus-level constraints* (No. arXiv:1707.09457). arXiv. <https://doi.org/10.48550/arXiv.1707.09457>
- Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning gender-neutral word embeddings. Dans E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Éds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (p. 4847-4853). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1521>
- Zhou, X., Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). Challenges in automated debiasing for toxic language detection. *arXiv:2102.00086 [cs]*. <http://arxiv.org/abs/2102.00086>

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237-291.
https://doi.org/10.1162/coli_a_00502