Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Les représentations contextuelles stéréotypées dans les modèles de langue français : mieux les identifier pour ne pas les reproduire

Léandre Adam-Cuvillier[1] Pierre-Jean Larpin[1] Antoine Simoulin[2]
(1) Capgemini, 11 Rue de Tilsitt, 75017 Paris, France
(2) Université de Paris, Olympe de Gouges, 8 Rue Albert Einstein, 75013 Paris, France
leandre.adamcuvillier@gmail.com, pierre-jean.larpin@capgemini.com,
antoine.simoulin@etu.u-paris.fr

RÉSUMÉ

Nous présentons une étude pour mieux identifier comment les stéréotypes se reflètent dans les modèles de langue français. Nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) à la langue française et suivons le même protocole expérimental que celui utilisé pour l'anglais. Alors que les stéréotypes sont connus pour évoluer en fonction des contextes culturels et temporels, notre étude identifie des similitudes avec les résultats observés pour l'anglais, notamment en ce qui concerne la corrélation entre les capacités linguistiques des modèles et la présence de biais mesurables. Nous étendons notre étude en examinant des architectures de réseaux neuronaux pré-entraînées sur des corpus linguistiques différents. Nos résultats soulignent l'impact crucial des données de pré-entraînement sur les biais constatés dans les modèles français. De plus, nous observons que l'utilisation de corpus multilingues pour le pré-entraînement peut avoir un effet positif sur l'atténuation des biais.

ABSTRACT

Stereotyped contextual representations in French language models : better identifying them to avoid reproducing them

We present a study to identify better how stereotypes are reflected in French language models. We adapt the StereoSet dataset (Nadeem *et al.*, 2021) to the French language and follow the same experimental protocol used for English. While stereotypes are known to evolve based on cultural and temporal contexts, our study identifies similarities to the findings observed for English, particularly regarding the correlation between the models' language ability and the presence of measurable biases. Furthermore, we extend our investigation by examining neural network architectures pre-trained on different language corpora. Our results highlight the crucial impact of the pretraining data on the biases found in the French models. Moreover, we observe that leveraging multilingual corpora for pretraining can have a positive effect in mitigating biases.

MOTS-CLÉS : stéréotype, modèle de langue, pré-entraîné, français.

KEYWORDS: stereotype, language model, pre-trained, French.

1 Introduction

Un stéréotype est un biais inconscient qui nous amène à déformer la réalité en supposant qu'un groupe de personnes, présentant des caractéristiques similaires—physiques, morales, comportementales, réelles, ou supposées—partagent des attributs communs sans tenir compte des différences individuelles

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 130–143.

et en les réduisant à celles-ci (Katz & Braly, 1933; Allport *et al.*, 1954; Goffman, 1963). Les stéréotypes s'appuient sur des croyances simplificatrices qui s'apprennent et se renforcent par les interactions sociales et d'autres formes de communication. Les modèles de langue pré-entraînés (pLMs) peuvent contribuer à les véhiculer dans nos sociétés (Kirk *et al.*, 2021; Bender *et al.*, 2021; Bommasani *et al.*, 2021) car ils sont susceptibles d'apprendre un modèle probabiliste reflétant les biais statistiques associés aux stéréotypes présents dans les corpus de pré-entraînement (Zhao *et al.*, 2018; Sheng *et al.*, 2019; Jia *et al.*, 2020). Maîtriser les biais des pLMs constitue un enjeu majeur pour prévenir la propagation de stéréotypes susceptibles de perpétuer des préjugés et des comportements discriminatoires.

De nombreuses recherches ont été menées sur les pLMs en anglais. Notre objectif est de voir dans quelle mesure ces résultats peuvent être appliqués aux pLMs pré-entraînés en français. Pour ce faire, nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) en français et suivons le même protocole expérimental pour évaluer les biais statistiques dans divers modèles de langue pré-entraînés en français, tels que BERT, GPT, BART, XGLM et BLOOM. Nous étendons le périmètre des études existantes pour le français en examinant l'effet du corpus de pré-entraînement.

Notre article est organisé comme suit : la section 2 commence par détailler l'adaptation du corpus et rappeler le protocole expérimental correspondant utilisé par Nadeem *et al.* (2021). Nous détaillons ensuite les travaux connexes (§ 3). Nous présentons nos résultats expérimentaux (§ 4.1) que nous comparons avec ceux obtenus en anglais (§ 4.2) et avec les études menées sur la base d'autres corpus en français (§ 4.3). Finalement nous discutons de pistes permettant de combattre les biais identifiés (§ 4.4) et nous discutons des limites de notre approche (§ 5).

2 Méthode

Nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) en français (§ 2.1). Nos contributions ne concernent ni la création du jeu de données original StereoSet, ni le développement de la méthode d'évaluation qui lui est associée. Cette dernière est rappelée brièvement en section 2.2 afin de faciliter la compréhension de la méthode d'adaptation et la lecture des sections suivantes.

2.1 Adaptation de StereoSet au français : StereoSet-fr

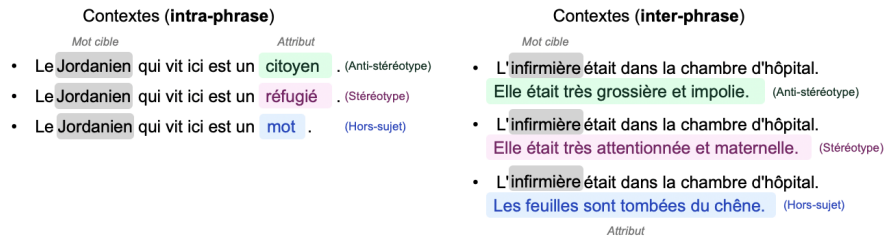

FIGURE 1 – Exemple de test d'association intra-phrastique **(gauche)** et inter-phrastique **(droite)**.

StereoSet (Nadeem *et al.*, 2021) est un test d'association qui mesure les biais stéréotypés dans les représentations de mots contextuelles. Étant donné un mot cible, par exemple "infirmière", le jeu de données propose trois instances de contexte en langage naturel contenant des classes d'attributs

correspondant à une association stéréotypée, anti-stéréotypée, ou hors sujet. Le jeu de donnée est constitué d'une collection de triplets de phrases complètes (intra-phrastique) et de paires de phrases (inter-phrastique). Nous illustrons un exemple d'instances en FIGURE 1.

Nous avons adapté StereoSet pour le français en trois étapes. Nous avons commencé par un audit du jeu de données original. Nous avons supprimé 11 exemples qui contenaient des erreurs de formes manifestes (par exemple plusieurs mots cibles) ou difficiles à adapter en français [1]. Par ailleurs, nous avons dédupliqué un petit nombre de tests originaux en remplaçant les attributs par des synonymes.

Nous avons ensuite traduit automatiquement les contextes de chacun des exemples en utilisant le logiciel DeepL Pro [2]. Nous avons ensuite revu manuellement chaque contexte pour s'assurer de la cohérence de la structure du jeu de données. En particulier, les tests intra-phrastiques doivent respecter une structure précise selon laquelle chaque contexte doit être identique, à l'exception du mot attribut. Comme nous avons traduit chaque contexte indépendamment, cette structure n'était pas respectée dans près de la moitié des traductions. Nous avons donc uniformisé les traductions des trois contextes de chacun de ces tests intra-phrastiques afin qu'ils ne diffèrent que par le mot attribut.

Nous avons également procédé à des adaptations plus larges pour s'assurer de la validité des tests. Nous avons passé environ 5 % des contextes au pluriel pour éviter des accords en genres différents des adjectifs possessifs [3]. Pour environ 1 % des contextes, nous avons aussi remplacé par des synonymes les mots pour lesquels il n'existe pas de traduction exacte de l'anglais vers le français [4]. Finalement, les exemples hors sujet s'orientaient vers une traduction qui privilégie la cohérence globale de la phrase. Nous avons repris environ 40 % de ces exemples pour qu'ils restent indiscutablement absurdes.

Nous résumons les statistiques descriptives du jeu de données StereoSet-fr en TABLE 1.

	Contextes intra-phrastiques		Contextes inter-phrastiques	
Domaine	**# Mots cibles**	**# Tests**	**# Mots cibles**	**# Tests**
Genre	11	256	10	242
Profession	32	806	30	827
Origine ethnique	43	958	36	976
Religion	3	76	3	78
Total	89	2 096	79	2 123

TABLE 1 – Statistiques descriptives du jeu de données StereoSet-fr.

2.2 Méthode d'évaluation

Nadeem *et al.* (2021) définissent le biais d'une représentation contextuelle—et par extension du modèle sous-jacent—comme la différence dans les probabilités estimées par le modèle des exemples stéréotypés et anti-stéréotypés. Nous avons utilisé la même méthode pour estimer ces probabilités et

1. Par exemple si le mot attribut est un adjectif possessif, qui s'accorde en genre avec ce qui est possédé en français : "son cahier", mais avec celui ou celle qui possède en anglais : "his/her notebook".

2. https://www.deepl.com/

3. Par exemple "Les Ukrainiens aiment boire des < vodkas/whiskies >." plutôt que "Les Ukrainiens aiment boire de la/du < vodka/whisky >." qui modifient la phrase au delà du mot attribut indiqué entre les chevrons.

4. Par exemple "nurse" en anglais qui se traduit à la fois par "infirmière" et "infirmier" en français. Nous avons utilisé les termes "soignants" et "sauveurs" dans ce cas.

les mêmes scores pour comparer différents modèles français, que nous rappelons brièvement ci-après.

Dans le cas du test intra-phrase, Nadeem *et al.* (2021) estiment la vraisemblance de chaque attribut en s'appuyant sur la tâche de modèle de langue (LM) utilisée pour le pré-entraînement des modèles. La log-probabilité de chaque token u du contexte est estimée selon l'équation 2 avec h le vecteur d'état caché de la dernière couche du modèle, C le contexte et W_{lm} la matrice de projection du vocabulaire apprise pendant la tâche LM. La log-probabilité de chaque attribut $P_{\text{intra}}\,(\text{attribut})$ est estimée comme une somme des log-probabilités des tokens u_i du mot attribut selon l'équation 2 [5].

$$P\,(u|C) = \text{softmax}(hW_{\text{lm}}^{\top}) \tag{1}$$

$$\log P_{\text{intra}}\,(\text{attribut}) = \sum_{i \in \text{attribut}} \log P\,(u_i|C) \tag{2}$$

Dans le test inter-phrase, Nadeem *et al.* (2021) estiment la vraisemblance de chaque attribut en s'appuyant sur la tâche de prédiction de la prochaine phrase (NSP) proposée dans Devlin *et al.* (2019). Pour chaque modèle, une tête de prédiction W_{nsp} est entraînée à estimer la log-probabilité que deux phrases soient consécutives dans le corpus de pré-entraînement [6]. La log-probabilité $P_{\text{inter}}\,(\text{attribut})$ de l'ensemble de la seconde phrase contenant l'attribut est estimée selon l'équation 3 [7].

$$P_{\text{inter}}\,(\text{attribut}) = \text{softmax}(hW_{\text{nsp}}^{\top}) \tag{3}$$

Finalement, Nadeem *et al.* (2021) classent un attribut parmi le triplet sur la base de ces probabilités (en choisissant celui qui a la probabilité la plus élevée). Afin de tester à la fois la conservation de la cohérence et la capacité à éviter les biais d'un modèle, Nadeem *et al.* (2021) proposent trois scores que nous utilisons directement tels qu'ils ont été définis. Nous rappelons simplement les définitions correspondantes ci-après pour simplifier la lecture des résultats expérimentaux en section 4 [8].

Le **LM score (*LMs*)** correspond au pourcentage de tests dans lesquels le modèle choisit une proposition qui ne soit pas hors sujet. Il évalue la capacité du modèle à conserver le sens. Le score idéal est 100. Plus le *LMs* est proche de 100, plus le modèle conserve une cohérence.

Le **Stereo score (*Ss*)** évalue la prédisposition d'un modèle à s'orienter vers un choix biaisé. Il correspond au pourcentage de tests pour lesquels le modèle choisit une proposition stéréotypique plutôt qu'anti stéréotypique. Le score idéal de 50 est atteint lorsque le modèle n'affiche pas de préférence particulière entre des choix biaisés et non biaisés.

Le **score d'association contextuelle (*ACs*)** est le score principal. Il combine à la fois le score *LMs* et le score *Ss*, afin de représenter la capacité d'un modèle de langue à se comporter de manière impartiale

5. Pour les architectures encodeurs, les tokens i de l'attribut sont remplacés par le token `[MASK]`, puis démasqués itérativement les sous-mots de gauche à droite. La probabilité finale est calculée comme la moyenne des probabilités pour chaque sous-mot de l'attribut. Pour les architectures décodeurs, les log-probabilités de l'ensemble des tokens du contexte sont sommées.

6. Certains modèles comme CamemBERT sont pré-entraînés en utilisant cette tâche mais pas tous. Nous avons donc ré-entraîné cette couche spécifique pour l'ensemble des modèles en reproduisant la procédure et les hyper-paramètres utilisés pour StereoSet sur un extrait de Wikipédia en français. Les modèles avec moins de 360 millions de paramètres ont été entraînés sur deux Nvidia T4 16Gb et ceux plus larges sur une carte Nvidia V100 32Gb.

7. Pour les architectures encodeurs, h correspond à l'état caché du premier token `[CLS]` de la dernière couche du modèle. Pour les architectures décodeur et encodeur-décodeur, h correspond à la moyenne de tous les états cachés de la dernière couche du modèle. Pour l'évaluation inter-phrastique, nous utilisons le modèle adapté incrémentalement sur la tâche NSP.

8. Ces scores permettent de comparer plusieurs modèles entre eux, à l'inverse de métriques comme la perplexité.

tout en excellant dans la modélisation du langage. Un modèle idéal possède un *ACs* score de 100, et un modèle totalement biaisé possède un *ACs* score de 0. Un modèle aléatoire, quant à lui, aura un *ACs* score de 50. Sa formule est la suivante : $ACs = LMs \times \min(Ss, 100 - Ss)/50$.

3 Travaux connexes

Les biais présents dans les plongements lexicaux sont traditionnellement analysés à travers des tests d'analogies et d'associations de mots. Les tests d'analogies transposent une relation syntaxique ou sémantique entre deux mots, par exemple (homme, chirurgien), pour compléter une nouvelle paire, telle que (femme, ·), en effectuant des opérations algébriques sur les plongements lexicaux (Mikolov *et al.*, 2013). Les tests d'association, comme WEAT (Islam *et al.*, 2016), mesurent la similarité relative de deux ensembles de mots cibles—par exemple des noms masculins ou féminins—à deux ensembles de mots attributs—comme des attributs sur la situation professionnelle.

May *et al.* (2019) étendent la méthode WEAT à des représentations contextuelles (SEAT). À partir d'un terme cible et son attribut, ils créent des phrases de manière semi-automatique selon la forme "Ceci est [cible]." et "Ils sont [attribut]." pour obtenir des plongements lexicaux contextualisés. StereoSet (Nadeem *et al.*, 2021) raffine cette approche en considérant des contextes rédigés en langue naturelle et pas selon une procédure semi-automatique. L'approche s'étend également à l'échelle du discours avec un test d'association considérant des ensembles de phrases plutôt que de mots.

Finalement l'étude la plus proche de la nôtre est celle de Nangia *et al.* (2020) introduisant CrowS-Pairs, et de Névéol *et al.* (2022a,b) qui adapte le jeu de données pour le français. CrowS-Pairs analyse les biais stéréotypés en utilisant des paires minimales. Néanmoins, CrowS-Pairs et French CrowS-Pairs n'étudient le biais qu'au sein d'une seule phrase (intra-phrastique). Aussi, StereoSet contient un jeu d'évaluation plus important et permet de mesurer le biais pour les architectures de modèles transformers encodeurs mais aussi décodeur et encodeur-décodeur, tandis que CrowS-Pairs ne mesure le biais que pour les architectures encodeurs.

4 Experiences

Cette section détaille les résultats d'évaluation sur StereoSet-fr. Nous présentons les résultats (§ 4.1) que nous comparons avec les métriques du papier original StereoSet (§ 4.2), puis avec des travaux similaires sur le français (§ 4.3), pour finalement analyser des pistes d'atténuations des biais (§ 4.4).

4.1 Résultats experimentaux et impact des paramètres du modèle

Nous commençons par présenter les résultats globaux sur les tâches inter-phrastique et intra-phrastique. La TABLE 2 reporte les scores d'évaluations sur StereoSet-fr de plusieurs pLMs français. Nous comparons des modèles entraînés sur des corpus majoritairement français ou multilingues. Nous considérons des modèles encodeurs : FlauBERT (Le *et al.*, 2020b,a), CamemBERT (Martin *et al.*, 2020) et m-BERT (Devlin *et al.*, 2019), décodeurs : GPT-fr (Simoulin & Crabbé, 2021), PAGnol (Launay *et al.*, 2022), XGLM (Lin *et al.*, 2021), et BLOOM (Scao *et al.*, 2022), et encodeur-décodeurs avec mBART (Tang *et al.*, 2020) et Barthez (Kamal Eddine *et al.*, 2021).

Modèles	*LMs*	*Ss*	*ACs*				
			Genre	Profession	Origine Ethnique	Religion	Global
FlauBERT	78,9	49,7	54,6	68,2	69,8	68,7	67,2
CamemBERT	87,4	58,9	76.6	70,3	72,2	66,4	71,9
m-BERT	73,5	54,3	71,0	62,5	66,3	76,4	66,7
GPT-fr	83,2	58,4	68,3	67,9	69,0	71,6	68,7
PAGnol	87,6	59,1	79,2	73,9	69,0	76,9	71,8
BLOOM	81,9	40,6	51,2	62,2	55,3	43,6	56,9
XGLM	87,0	56,3	78,4	73,3	76,5	89,8	76,0
Barthez	83,8	56,7	69,3	69,6	75,4	74,7	72,6
mBART	80,3	54,0	75,8	71,3	74,1	75,4	73,9
Moyenne	82,6	54,2	69.2	68.8	69.9	71.5	69.5

TABLE 2 – Nous reportons la moyenne des scores *LMs*, *Ss* et *ACs* sur les tâches inter-phrastique et intra-phrastique pour FlauBERT-large (Le *et al.*, 2020b,a), CamemBERT-large (Martin *et al.*, 2020), m-BERT-base (Devlin *et al.*, 2019), GPT-fr-base (Simoulin & Crabbé, 2021), PAGnol-large (Launay *et al.*, 2022), Barthez (Kamal Eddine *et al.*, 2021), mBART-50 (Tang *et al.*, 2020), BLOOM-560m (Scao *et al.*, 2022) et XGLM-564M (Lin *et al.*, 2021). Nous soulignons les meilleurs résultats pour chaque score de chaque sous ensemble.

Une première analyse des performances des modèles met en évidence le fait qu'un modèle ne peut pas être défini comme biaisé ou non-biaisé. Cette analyse doit être effectuée sur une typologie précise de biais. En effet, nous observons qu'à score *ACs* équivalent, les modèles CamemBERT et Barthez ont des résultats par type de biais totalement différents. Alors que CamemBERT présente un score *ACs* élevé sur le genre, Barthez présente un score élevé sur l'origine ethnique ou la religion. Ainsi, chaque modèle présente des biais variables, rendant l'analyse de ces derniers difficile à généraliser et plus spécifique à des cas d'utilisation particuliers.

Par ailleurs, nous n'observons pas de relation entre la nature du modèle (encodeur, décodeur) et ses biais. Par exemple, CamemBERT (Encodeur) et PAGnol (Décodeur) présentent des scores élevés pour le genre, alors que m-BERT (Encodeur) et XGLM (Décodeur) présentent un score élevé pour la religion et Barthez (Encodeur-Décodeur) pour l'origine ethnique.

4.2 Étude comparée avec Stereoset

À l'instar de Nadeem *et al.* (2021), nous observons que les modèles avec des scores *LMs* plus élevés—indiquant une meilleure capacité à modéliser le français—ont tendance à présenter des *Ss* plus éloignés du score idéal de 50—signalant ainsi une présence accrue de biais. Nous mesurons une corrélation de Spearman (ρ) de 0,63 entre les scores *LMs* et les écarts par rapport au scores *Ss* idéaux de 50. La corrélation est moindre que celle de 0,9 mesurée pour StereoSet, mais le test reste significatif avec une p-valeur associée de 0,04 (calculée avec un test de permutation).

Nadeem *et al.* (2021) observent également que plus un modèle a de paramètres, plus sa capacité de modélisation linguistique (*LMs*) augmente et par conséquent plus son score stéréotypé augmente également. Afin de vérifier si cette relation s'applique pour les modèles pré-entraînés en français, nous comparons en TABLE 3 les résultats sur la tâche intra-phrastique pour plusieurs modèles similaires

Modèles	# Param. ($\times 10^6$)	Tâche intra-phrastique *LMs*	*Ss*	*ACs*
CamemBERT-base	110	83.5	61.0	65.1
CamemBERT-large	335	**84.0**	**57.5**	**71.4**
FlauBERT-small	54	58.3	**52.1**	57.3
FlauBERT-base	138	**80.8**	52.8	**76.2**
FlauBERT-large	373	80.3	57.1	68.9
PAGnol-small	124	87.3	**60.3**	**69.4**
PAGnol-medium	355	**87.6**	60.9	68.5
PAGnol-large	773	87.3	61.5	67.3
GPT-fr-small	124	87.4	**61.5**	67.2
GPT-fr-base	1 000	**89.2**	62.2	**67.4**
BLOOM-560	560	86.9	**56.5**	**75.7**
BLOOM-1b1	1 100	**87.2**	60.6	68.7

TABLE 3 – Nous reportons les résultats de la tâche intra-phrastique pour des modèles entraînés avec différents nombres de paramètres. Nous reportons CamemBERT entraîné sur le corpus CCNet (Martin *et al.*, 2020), FlauBERT (Le *et al.*, 2020b,a), PAGnol (Launay *et al.*, 2022), GPT-fr (Simoulin & Crabbé, 2021) et BLOOM (Scao *et al.*, 2022). Nous soulignons les meilleurs résultats pour chaque score et mettons en **gras** les meilleurs résultats pour chaque architecture de modèle.

mais présentant un nombre différent de paramètres [9]. Cette fois encore, nos résultats sont cohérents avec ceux obtenus pour l'anglais à l'aide de StereoSet. Pour chaque modèle considéré séparément—à l'exception de CamemBERT—l'augmentation du nombre de paramètres du modèle est associée à une augmentation du score *Ss*. De manière générale, la capacité à modéliser le langage (scores *LMs*) augmente également avec le nombre de paramètres du modèle. Nous interprétons qu'avec plus de paramètres, le modèle peut mieux capturer la distribution du corpus d'entraînement mais aussi les biais inhérents. L'observation n'est valable que lorsque les modèles sont analysés séparément. Le coefficient de corrélation de rang de Spearman (ρ), pour l'ensemble des modèles de la TABLE 3, entre le nombre de paramètres et les écarts par rapport aux scores *Ss* optimaux de 50, est de 0,31 (p-valeur associée 0,16). Nous avons calculé des valeurs proches pour les résultats de StereoSet, avec un coefficient ρ de 0,32 (p-valeur associée 0,21).

4.3 Étude comparée avec French CrowS-Pairs

Névéol *et al.* (2022a,b) proposent également une analyse des biais stéréotypés pour les modèles pré-entraînés en français en comparant notamment CamemBERT, FlauBERT et m-BERT. Nous cherchons à vérifier la cohérence de notre analyse en comparant les mêmes modèles. Les résultats de StereoSet-fr ne sont pas directement comparables avec ceux de French CrowS-Pairs car la méthode d'évaluation et la définition des scores diffèrent. Nous comparons ainsi simplement des performances relatives des modèles sur les deux jeux de données sous la forme d'un test de contrôle.

9. Les différentes versions de chaque modèle peuvent avoir été pré-entraînés sur des corpus différents. Pour le modèle CamemBERT-base, nous avons sélectionné celui entraîné sur le corpus CCNet 135Gb afin de pouvoir le comparer avec CamemBERT-large qui est entraîné sur le même corpus.

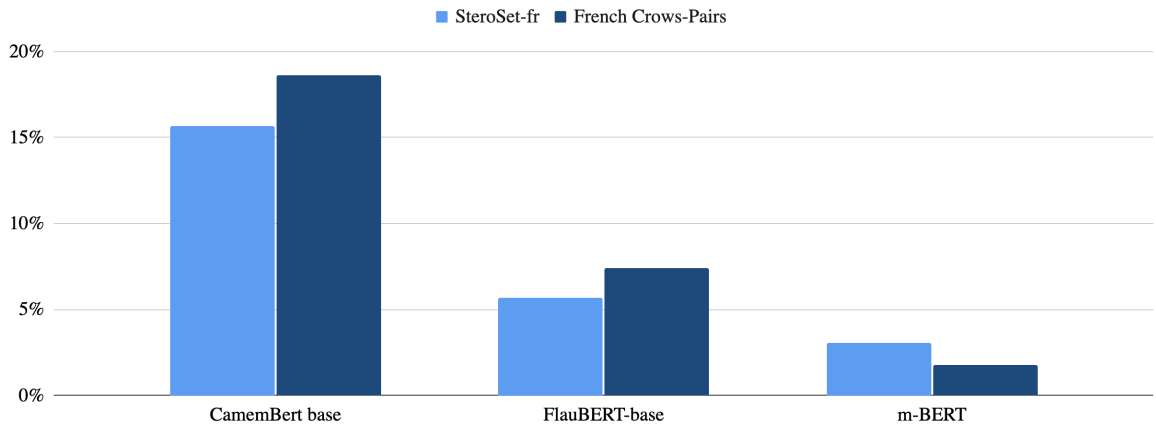

FIGURE 2 – Ecart relatif des scores *Ss* pour la tâche intra-phrastique de StereoSet avec un modèle non biaisé (score *Ss* de 50) avec l'écart relatif des scores de French CrowS-Pairs (Névéol *et al.*, 2022a,b). Il est important de noter que les scores ne sont pas directement comparables car CrowS-Pairs évalue les modèles en utilisant la pseudo-vraisemblance afin de ne pas pénaliser les termes moins frequents.

En FIGURE 2, nous comparons l'écart relatif des scores *Ss* pour la tâche intra-phrastique de StereoSet-fr avec un modèle théoriquement non biaisé (score *Ss* de 50) avec l'écart relatif des scores de French CrowS-Pairs. Plus l'écart relatif est important, plus le modèle présente des biais stéréotypés importants dans chacune des deux études. De manière générale, nous observons la même relation d'ordre entre les performances des modèles. Le modèle CamemBERT est celui qui présente les écarts les plus importants au score idéal. Les apparentes mauvaises performances de m-BERT sont à moduler par de mauvais *LMs* scores. Cette composante montre l'intérêt de la métrique *ACs* de StereoSet-fr qui permet de quantifier les biais en tenant compte de la capacité du modèle à modéliser le langage.

4.4 S'émanciper des biais

Dans cette section, nous explorons des stratégies clés pour réduire les biais stéréotypés dans les modèles de langue français. En particulier, nous examinons l'impact du corpus d'entraînement sur les biais manifestés par le modèle. Nous cherchons ainsi à comparer des modèles similaires qui ne diffèrent que par le corpus utilisé pour le pré-entraînement. À cette fin, la TABLE 4 présente les performances de modèles avec des architectures identiques mais entraînés sur des corpus différents.

Le choix du corpus de pré-entraînement semble jouer un rôle crucial dans la détermination des biais au sein des modèles de langue. Par exemple, les modèles CamemBERT, lorsqu'ils sont entraînés avec le corpus OSCAR, affichent systématiquement de meilleurs scores *ACs*, en particulier pour les catégories d'origine ethnique et de religion, par rapport à ceux entraînés avec CCNet ou Wikipédia. De plus, nous remarquons une sensibilité accrue aux biais religieux chez les modèles CamemBERT pré-entraînés spécifiquement avec le corpus CCNet.

Nous observons par ailleurs que la taille du corpus d'entraînement peut influencer le biais d'un modèle. Le modèle CamemBERT affiche de meilleurs scores *ACs* lorsqu'il est entraîné sur des versions plus petites du corpus OSCAR (4GB) et CCNet (4GB) par rapport à leurs versions plus volumineuses (135GB). Cette fois encore, cette observation n'est valable que lorsque l'on considère chaque corpus séparément. Elle n'est pas significative si on mesure le coefficient de corrélation de rang de Spearman,

Corpus d'entraînement	Genre	Profession	Origine ethnique	Religion	Global
		CamemBERT			
OSCAR (4 GB)	80,6	**70,1**	**81,5**	71,4	**76,8**
OSCAR (138 GB)	73,3	66,5	80,1	**71,5**	73,8
CCNet (4 GB)	**82,8**	68,6	75,9	63,1	73,6
CCNet (135 GB)	71,6	65,0	72,7	67,9	69,5
Wikipedia (4 GB)	72,8	65,4	70,4	68,9	68,8
		mBART			
CC25 (1 100 GB)	67,1	67,5	73,2	70,0	70,6
CC25 + 25 (1 160 GB)	**75,8**	**71,3**	**74,1**	**75,4**	**73,9**

TABLE 4 – Nous reportons la moyenne des scores *ACs* sur les tâches inter-phrastique et intra-phrastique pour CamemBERT (Martin *et al.*, 2020) entraîné sur différents corpus d'entraînements, et pour mBART (Liu *et al.*, 2020) et mBART-50 (Tang *et al.*, 2020), pour lesquel le pré-entraînement a été étendu sur des jeux de données supplémentaires. Nous avons estimé les tailles des corpus CC25 et CC25 + 25 à partir des statistiques descriptives des articles originaux. Nous soulignons les meilleurs résultats pour chaque score et mettons en **gras** les meilleurs résultats pour chaque modèle.

entre la taille du corpus d'entraînement et les écarts par rapport au *Ss* idéal de 50, sur l'ensemble des modèles de la TABLE 4. Nous mesurons un coefficient ρ de 0,29 avec une p-valeur associée de 0,35 (calculée avec un test de permutation). Nadeem *et al.* (2021) affirment ne mesurer aucune corrélation significative entre la taille du corpus et les performances du modèle en termes de scores *LMs* ou *Ss*. Il est important de souligner que cette analyse portait sur différents modèles et architectures, alors que nous nous concentrons sur le même modèle CamemBERT entraîné sur divers corpus, modifiant ainsi uniquement ce paramètre. Cela nous amène à conclure que la qualité du jeu d'entraînement prévaut sur la quantité. Cette philosophie guide par exemple le développement du modèle Phi-1, qui est entraîné sur une sélection remarquablement réduite de textes (Gunasekar *et al.*, 2023).

Finalement, pour le modèle mBART, nous constatons que le pré-entraînement sur des corpus multilingues pourrait contribuer à réduire les biais. Ainsi, Le modèle mBART-25, entraîné sur un corpus comprenant 25 langues, montre des scores *ACs* plus bas comparé au modèle mBART-50, enrichi avec 25 langues supplémentaires. Ceci suggère que les biais, souvent liés à une culture spécifique, sont potentiellement moins prononcés dans les modèles entraînés sur des données multilingues. Cela pourrait s'expliquer par le fait que les données d'entraînement, produites par les locuteurs d'une langue, reflètent les biais culturels associés. En intégrant plusieurs langues, on réduit le risque de renforcer des préjugés propres à une culture particulière. Bien que cette hypothèse demande à être validée par des recherches dédiées, elle ouvre une voie prometteuse pour explorer comment les influences culturelles et historiques façonnent les biais dans les modèles de langue.

5 Discussion et limites de notre analyse

Notre étude est basée sur l'adaptation du jeu de données StereoSet, pour lequel Nadeem *et al.* (2021) ont fait appel à des annnotateurs résidants aux Etats-Unis afin de capturer des variations locales des

stéréotypes. Ainsi, le corpus se concentre sur l'analyse de types de biais stéréotypés, qui peuvent être spécifiques à chaque contexte culturel, notamment les stéréotypes raciaux et religieux. Au contraire, les stéréotypes liés au genre et à l'âge ont tendance à présenter des similitudes dans différentes cultures (Fiske, 2017). Les scores *ACs* de la TABLE 2 sont ainsi généralement plus élevés pour les catégories liées à la religion et l'origine ethnique que pour ceux du genre et de la profession. Vraisemblablement, cet écart reflète une spécificité culturelle des stéréotypes liés à des aspects sociaux et non pas à une sensibilité moindre des pLMs français aux stéréotypes raciaux et religieux.

Par ailleurs, le jeu de données StereoSet et notre adaptation en français se restreignent à des exemples sur des biais de genre, origine ethnique, religion et profession qui ne sont pas nécessairement exhaustifs et représentatifs des biais actuels de notre société. La taille du jeu de données et la nature des exemples ne permettent en aucun cas d'affirmer qu'un modèle français exhibant des scores élevés est exempt de biais. Nous cherchons plutôt à analyser les facteurs influant sur la nature des biais dans les pLMs en français, la mesure dans laquelle ces facteurs se comparent à ceux mis en lumière pour les études en anglais et des pistes de développements d'atténuation des biais stéréotypés.

Finalement Blodgett *et al.* (2020) alertent sur la nécessité de définir clairement la notion de biais et des préjudices éventuels que ces derniers peuvent causer. Notre étude se limite à mieux identifier les biais stéréotypés capturés dans les représentations contextuelles des pLMs en français, indépendamment de leur utilisation finale et du contexte social dans lequel ils seront utilisés. Cette étude devra donc être précisée, et le jeu de données adapté, en fonction des cas d'usages et du contexte social dans lequel s'inscrivent les processus que l'on cherche à optimiser ou automatiser à l'aide de ces modèles.

6 Conclusion et travaux futurs

Dans une démarche visant à évaluer et à maîtriser les modèles de langue en français, nous avons réalisé une étude cherchant à évaluer les stéréotypes qu'ils capturent. Pour ce faire, nous avons adapté le jeu de données StereoSet (Nadeem *et al.*, 2021) et reproduit la démarche expérimentale proposée, cette fois pour des modèles français. Nous avons effectué une comparaison entre plusieurs pLMs français et noté des différences significatives entre eux. Chaque modèle présente des biais qui varient selon les types de stéréotypes examinés, ce qui rend l'analyse des biais difficile à généraliser et plus spécifique à des cas d'utilisation particuliers. Malgré ces variations, nous constatons des tendances similaires à celles observées pour l'anglais, notamment en ce qui concerne la corrélation entre les compétences linguistiques des modèles et la présence de biais mesurables. Dépassant le cadre des études existantes en français (Névéol *et al.*, 2022a,b), notre étude compare les mêmes modèles entraînés sur des corpus distincts. Nos résultats soulignent l'impact crucial des données de pré-entraînement sur les biais présents dans les modèles. Contrairement à ce qui est observé pour les modèles anglais, la taille du corpus d'entraînement peut influencer les biais d'un modèle en français. En général, les modèles français pré-entraînés sur des versions moins volumineuses des corpus sont moins sujets aux biais que ceux entraînés sur des versions plus larges. De plus, nous remarquons que l'utilisation de corpus multilingues pour l'entraînement initial peut contribuer à atténuer les biais. Nous espérons que cette étude permettra de développer des modèles français moins sensibles à ces comportements indésirables pour des applications académiques et industrielles.

Références

ALLPORT G. W., CLARK K. & PETTIGREW T. (1954). *The nature of prejudice*. Addison-wesley Reading, MA.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In M. C. ELISH, W. ISAAC & R. S. ZEMEL, Éds., *FAccT '21 : 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, p. 610–623 : ACM. DOI : 10.1145/3442188.3445922.

BLODGETT S. L., BAROCAS S., III H. D. & WALLACH H. M. (2020). Language (technology) is power : A critical survey of "bias" in NLP. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 5454–5476 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.485.

BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R. B., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E., BRYNJOLFSSON E., BUCH S., CARD D., CASTELLON R., CHATTERJI N. S., CHEN A. S., CREEL K., DAVIS J. Q., DEMSZKY D., DONAHUE C., DOUMBOUYA M., DURMUS E., ERMON S., ETCHEMENDY J., ETHAYARAJH K., FEI-FEI L., FINN C., GALE T., GILLESPIE L., GOEL K., GOODMAN N. D., GROSSMAN S., GUHA N., HASHIMOTO T., HENDERSON P., HEWITT J., HO D. E., HONG J., HSU K., HUANG J., ICARD T., JAIN S., JURAFSKY D., KALLURI P., KARAMCHETI S., KEELING G., KHANI F., KHATTAB O., KOH P. W., KRASS M. S., KRISHNA R., KUDITIPUDI R. & ET AL. (2021). On the opportunities and risks of foundation models. *CoRR*, **abs/2108.07258**.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186 : Association for Computational Linguistics. DOI : 10.18653/v1/n19-1423.

FISKE S. T. (2017). Prejudices in cultural contexts : Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science*, **12**(5), 791–799.

GOFFMAN E. (1963). *Stigma : Notes on the Management of Spoiled Identity*. Englewood Cliffs : Prentice-Hall.

GUNASEKAR S., ZHANG Y., ANEJA J., MENDES C. C. T., GIORNO A. D., GOPI S., JAVAHERIPI M., KAUFFMANN P., DE ROSA G., SAARIKIVI O., SALIM A., SHAH S., BEHL H. S., WANG X., BUBECK S., ELDAN R., KALAI A. T., LEE Y. T. & LI Y. (2023). Textbooks are all you need. *CoRR*, **abs/2306.11644**. DOI : 10.48550/ARXIV.2306.11644.

ISLAM A. C., BRYSON J. J. & NARAYANAN A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, **abs/1608.07187**.

JIA S., MENG T., ZHAO J. & CHANG K. (2020). Mitigating gender bias amplification in distribution by posterior regularization. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 2936–2942 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.264.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

KATZ D. & BRALY K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, **28**(3), 280.

KIRK H. R., JUN Y., VOLPIN F., IQBAL H., BENUSSI E., DREYER F. A., SHTEDRITSKI A. & ASANO Y. M. (2021). Bias out-of-the-box : An empirical analysis of intersectional occupational biases in popular generative language models. In M. RANZATO, A. BEYGELZIMER, Y. N. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, p. 2611–2624.

LAUNAY J., TOMMASONE E. L., PANNIER B., BONIFACE F., CHATELAIN A., CAPPELLI A., POLI I. & SEDDAH D. (2022). Pagnol : An extra-large french generative model. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, p. 4275–4284 : European Language Resources Association.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020a). Flaubert : des modèles de langue contextualisés préentraînés pour le français (flaubert : Unsupervised language model pre-training for french). In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelle, Nancy, France, June 8-19, 2020*, p. 268–278 : ATALA et AFCP.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020b). FlauBERT : Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LIN X. V., MIHAYLOV T., ARTETXE M., WANG T., CHEN S., SIMIG D., OTT M., GOYAL N., BHOSALE S., DU J., PASUNURU R., SHLEIFER S., KOURA P. S., CHAUDHARY V., O'HORO B., WANG J., ZETTLEMOYER L., KOZAREVA Z., DIAB M. T., STOYANOV V. & LI X. (2021). Few-shot learning with multilingual language models. *CoRR*, **abs/2112.10668**.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MAY C., WANG A., BORDIA S., BOWMAN S. R. & RUDINGER R. (2019). On measuring social biases in sentence encoders. In 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 622–628 : Association for Computational Linguistics. DOI : 10.18653/v1/n19-1063.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, p. 3111–3119.

NADEEM M., BETHKE A. & REDDY S. (2021). Stereoset : Measuring stereotypical bias in pretrained language models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1 : Long Papers), Virtual Event, August 1-6, 2021*, p. 5356–5371 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.416.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). Crows-pairs : A challenge dataset for measuring social biases in masked language models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, p. 1953–1967 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022a). French crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, p. 8521–8531 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.583.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022b). French crows-pairs : Extension à une langue autre que l'anglais d'un corpus de mesure des biais sociétaux dans les modèles de langue masqués (french crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. Z. BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, TALN-RECITAL 2022, Avignon, France, June 27 - July 1, 2022*, p. 355–364 : ATALA.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIC S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I. & ET AL. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, **abs/2211.05100**. DOI : 10.48550/arXiv.2211.05100.

SHENG E., CHANG K., NATARAJAN P. & PENG N. (2019). The woman worked as a babysitter : On biases in language generation. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3405–3410 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1339.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle transformer génératif pré-entrainé pour le_____ français. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 246–255, Lille, France : ATALA.

TANG Y., TRAN C., LI X., CHEN P., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, **abs/2008.00401**.

ZHAO J., WANG T., YATSKAR M., ORDONEZ V. & CHANG K. (2018). Gender bias in coreference resolution : Evaluation and debiasing methods. In M. A. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, p. 15–20 : Association for Computational Linguistics. DOI : 10.18653/v1/n18-2003.

Remerciements

Ce travail a été soutenu par le Quantlab de Quantmetry part of Capgemini invent. Nous tenons à remercier Nicolas Brunel, Florian Arthur et Gregoire Martinon pour leur relecture et leur précieux commentaires.