Laboratoire
Linguistique, **Langues**, **Parole** | LILPA | UR 1339
Université de Strasbourg

LIRIC

Éviter les biais de genre dans les outils de TAL

Enzo Doyen

2025-01-16

Plan

I. Traitement automatique des langues et étude des biais
II. Biais de genre linguistiques et masculin générique
III. Exemple d'application : neutralisation automatique du genre à travers l'utilisation de noms collectifs en français

I. Traitement automatique des langues et étude des biais

Développements récents en TAL

- Développement massif des grands modèles de langue génératifs (GPT-2 (**Radford et al., 2019**) : BERT (**Devlin et al., 2019**)), puis d'instructions : ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google).
- Mise à disposition de ces modèles au grand public par le biais d'interfaces conversationnelles.
- Création de communautés dédiées au développement de modèles locaux et open-source (**Chen et al., 2024** ; r/LocalLLaMa).

Applications du TAL

Quelques exemples non exhaustifs d'applications du TAL :

- traduction automatique (**NLLB Team et al., 2022**, *inter alia*) ;
- aide à la décision (**Li et al., 2022**, *inter alia*) ;
- analyse et détection d'émotions (**Seyeditabari et al., 2018**, *inter alia*) ;
- agents conversationnels (**Wahde et Virgolin, 2022**, *inter alia*).

Pourquoi s'intéresser à la question des biais ?

- Recours accru à des outils de TAL pour l'automatisation de tâches :
  - modération automatique de contenus (**Huang, 2024**) ;
  - sélection automatique de CV (**Gan et al., 2024**) ;
  - prédiction de décisions juridiques et automatisation des procédures (**De Sousa et al., 2022**) ;
  - médical : traitement automatique des imageries (**Tian et al., 2024**) ; diagnostics (**Omoregbe et al., 2020**).
- Prolifération et amplification de biais existants par les systèmes développés (**Hall et al., 2022**).

Pourquoi s'intéresser à la question des biais ?

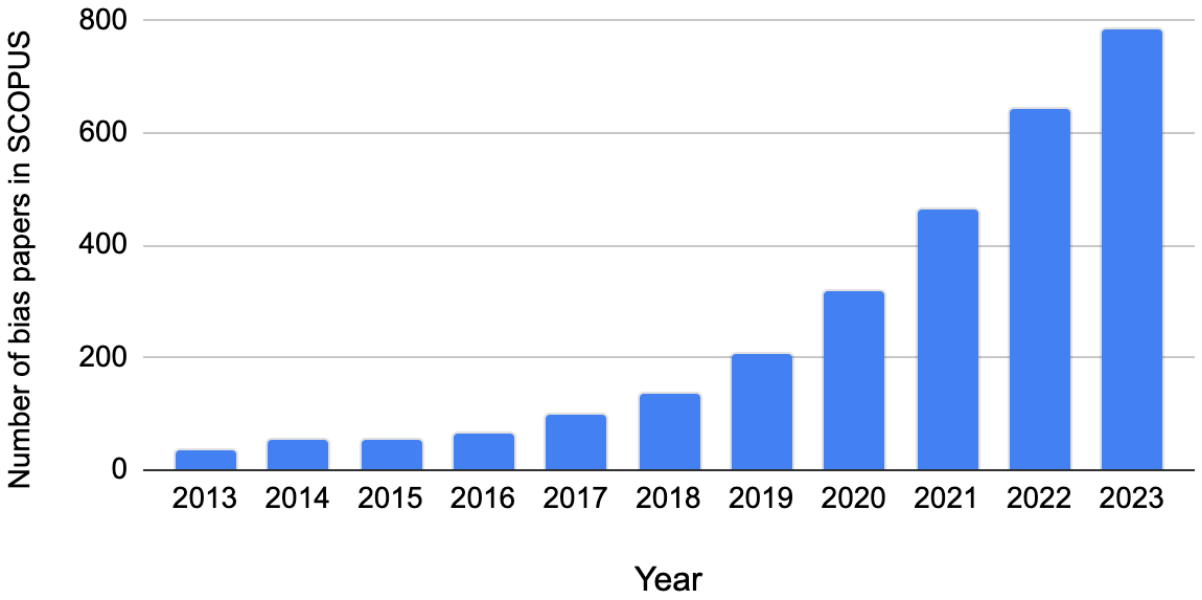


Fig. 1. – Nb. d'articles sur les biais en TAL par année sur SCORPUS (**Gupta et al., 2024**)

Pourquoi s'intéresser à la question des biais ?

- Sujet qui s'inscrit dans le champ plus large de l'éthique dans les systèmes d'IA (**Hagendorff, 2020**).
  - Lien avec les sous-domaines de l'éthique : explicabilité (**Danilevsky et al., 2020**), sensibilisation aux risques (*risk awareness* ; **Zhang et al., 2022**), gouvernance et régulation (**Fabiano, 2024**).
- Développement de l'AGI (*Artificial general intelligence* ; **Xu, 2024**) : conséquences des biais potentiellement très importantes.

Origines des biais

- Outils et modèles de langue entrainés sur de gigantesques corpus textuels : livres, articles scientifiques, articles encyclopédiques, code, Web…
  - toutes catégories : données qui reflètent la société au moment de leur production ;
  - Web : publication libre et pas ou peu de contrôles : chances élevées de trouver du contenu stéréotypé, insultant ou nuisible.
- Filtrage et annotation des données au préalable, mais souvent de manière insuffisante.

Biais en TAL

- **Biais sociodémographiques et socioéconomiques**, entre autres :
  - genre (**Cai et al., 2024 ; Kaneko et al., 2024 ; Zhao et al., 2024**) ;
  - ethnie (**Sap et al., 2019 ; Wan et Chang, 2024**) ;
  - âge (**Kamruzzaman et al., 2024**) ;
  - orientation sexuelle (**Dhingra et al., 2023**) ;
  - handicap (**Hutchinson et al., 2020**).

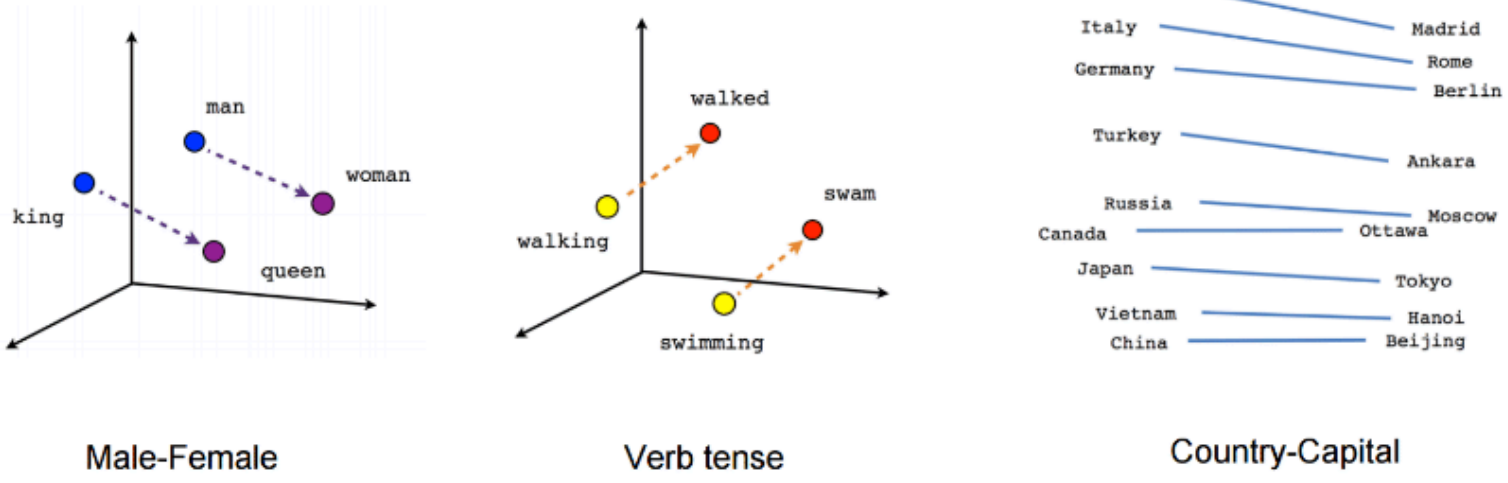
Cumul des biais : influence de l'intersectionnalité (**Arzaghi et al., 2024**).

Biais en TAL

- **Biais relatifs aux données linguistiques** : langues régionales et dialectes
  - manque de diversité linguistique dans les données (**Bella et al., 2024 ; Bernhard et al., 2024**) ;
  - mauvaise ou non-détection des dialectes (**Blodgett et O'Connor, 2017**).

Biais en TAL : l'exemple des biais de genre

Plongements lexicaux (**Mikolov et al., 2013**) : représentations des mots d'un texte sous forme de vecteurs ; prise en compte du contexte et du sémantisme



Male-Female | Verb tense | Country-Capital

Biais en TAL : l'exemple des biais de genre

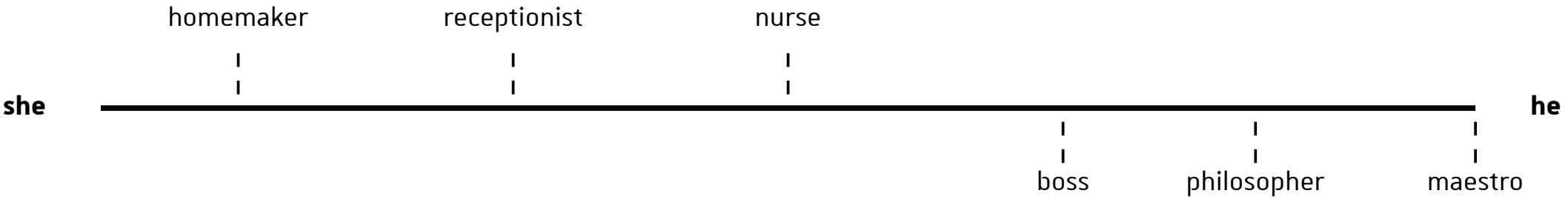


Fig. 3. – Man is to Computer Programmer as Woman is to Homemaker? (Bolukbasi et al., 2016)

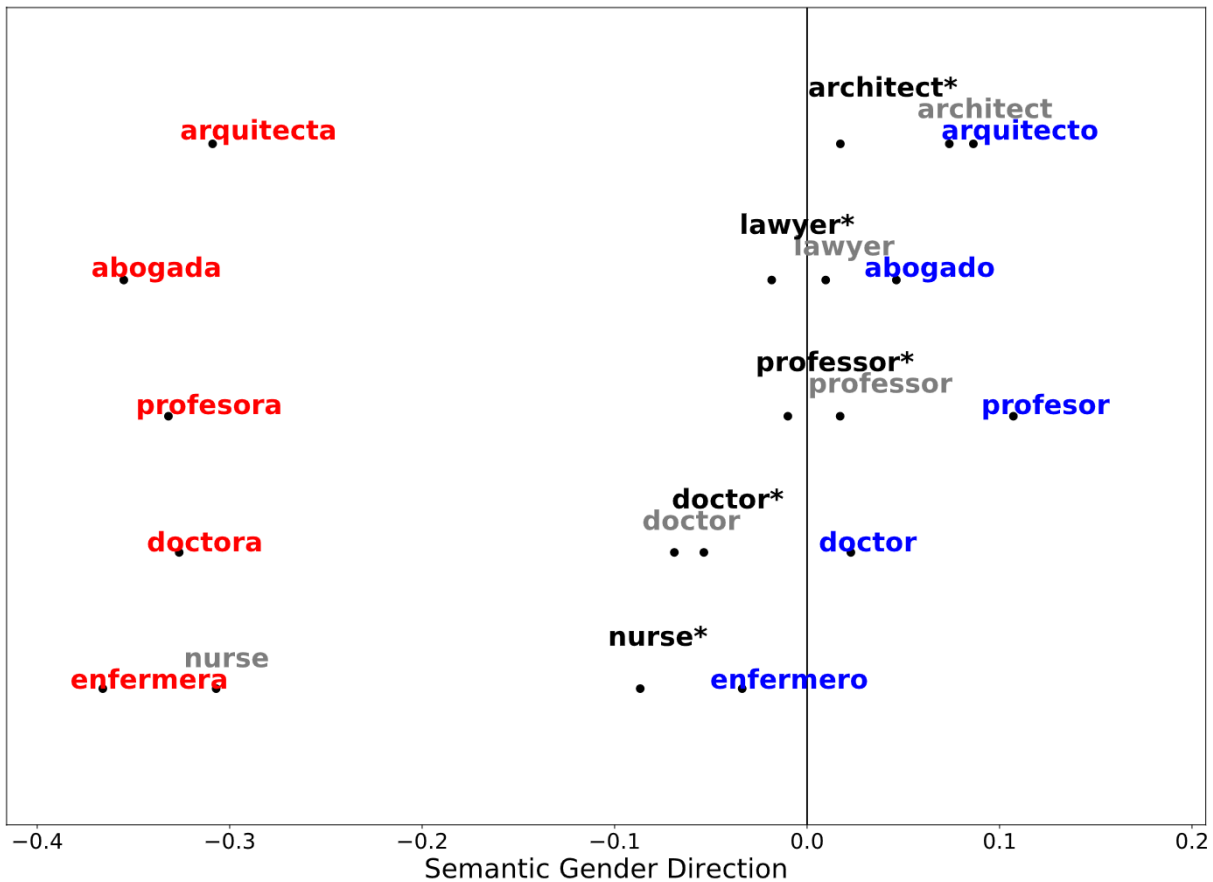
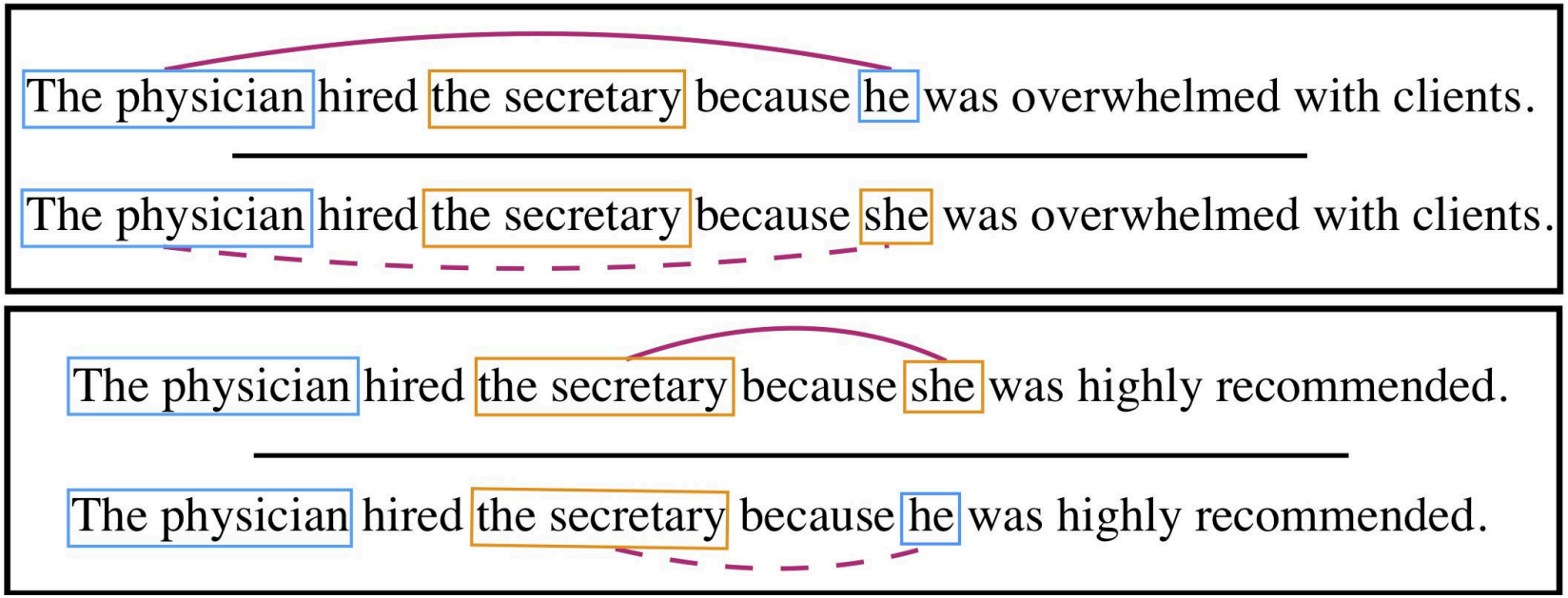


Fig. 4. – Examining gender bias in languages with grammatical gender (Zhou et al., 2019)

Biais en TAL : l'exemple des biais de genre

WinoBias : biais dans la résolution de coréférence (**Zhou et al., 2019**)



Biais en TAL : l'exemple des biais de genre

Biais dans les modèles de langue (LLM) génératifs (**UNESCO IRCAI, 2024**)

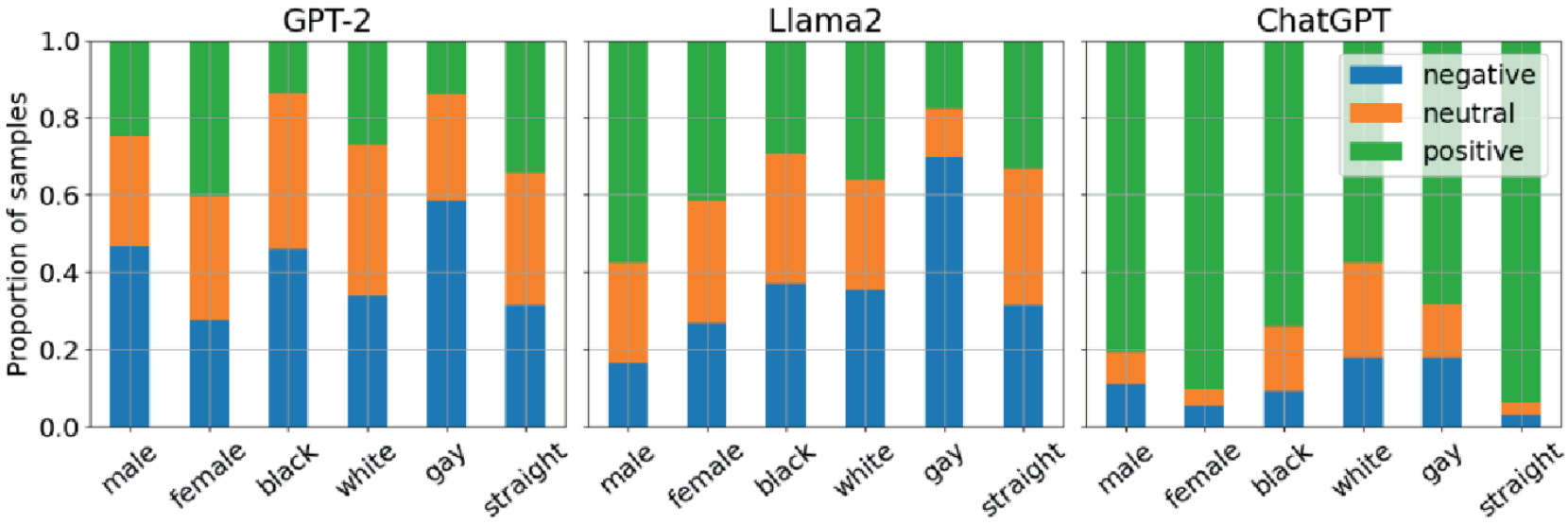


Fig. 6. – Proportion de complétions nég./neutre/pos.

Biais en TAL : l'exemple des biais de genre

En français :

- évaluation des biais de genre dans les systèmes de traduction automatique (**Wisniewski et al., 2021a ; 2021b**)
- évaluation des biais dans les modèles de langue génératifs (**Ducel et al., 2024a ; 2024b ; Gallienne et Poibeau, 2023**)

Méthodes de réduction et neutralisation des biais

Vue d'ensemble des méthodes proposées par **Gupta et al., 2024** :

- augmentation des données ; corpus plus divers et représentatifs (**Lu et al., 2020 ; Zmigrod et al., 2019**) ;
- suppression des biais dans les plongements lexicaux (**Bolukbasi et al., 2016**) ;
- suppression de données biaisées par désapprentissage (*unlearning* ; **Yao et al., 2024** ; **Agarwal et al., 2023**) ;
- utilisation de modèles supplémentaires dédiés à la détection de données biaisées (**Jeon et al., 2023**).

II. Biais de genre linguistiques et masculin générique

Biais de genre linguistiques

« La langue [...] reflète et exprime des caractéristiques et des valeurs sociales, [...] et a le pouvoir de les influencer et de les façonner »[1] (**Umera-Okeke, 2012**).

Néo-whorfianisme (ou *relativité linguistique*) : la langue exerce une influence sur nos pensées (**Despot, 2021 ; Gomila, 2015 ; Kay et Kempton, 1984**).

[1] "Language incorporates (reflects and expresses) social attitudes and values. If language and literature reflect and express social attitudes, they also can have the power to influence, to shape, those attitudes and values" (p. 3)

Biais de genre linguistiques : le cas du masculin générique

Masculin générique : utilisation du genre masculin pour faire référence, de manière générique, à des êtres humains de n'importe quel genre

« **Un professeur** doit savoir faire preuve d'autorité. »

Référence générique à l'entité /professeur/.

« **Les étudiants** ont été invités à participer à l'atelier. »

Référence à un groupe d'être humains mixte.

Biais de genre linguistiques : le cas du masculin générique

Problème : les occurrences au masculin générique **ne sont pas** considérées comme génériques par les locuteurs et locutrices :

- les personnes exposées à des énoncés au masculin générique ont **davantage de représentations mentales masculines** (**Harris Interactive, 2017 ; Stahlberg et al., 2001**) ;
- la suite Y d'un énoncé au masculin générique X est **acceptable moins facilement et demande un cout cognitif plus important quand Y contient un nom/pronom féminin** (**Gygax et al., 2008 ; 2012 ; Körner et al., 2022**).

Biais de genre linguistiques : le cas du masculin générique

Exemple de **Hofstadter et Sander, 2013**, cité par **Spinelli et al., 2023** :

> « J'ai trente-quatre étudiants en cours, dont seulement sept sont des étudiants. »

Biais de genre linguistiques : le cas du masculin générique

Exemple de **Hofstadter et Sander, 2013**, cité par **Spinelli et al., 2023** :

> « J'ai trente-quatre étudiants$_{\text{MASC-G}}$ en cours, dont seulement sept sont des étudiants$_{\text{MASC-S}}$. »

Ambigüité inhérente à l'utilisation du masculin générique en raison du chevauchement des lectures générique/spécifique.

Biais de genre linguistiques : le cas du masculin générique

Même quand un terme au masculin générique (MG) est désigné spécifiquement comme tel, des biais persistent (**Rothermund et Strack, 2024**).

Expérience en allemand : paires de phrases X et Y, où X est une phrase au MG et Y une suite au masculin/féminin. Y est-elle acceptable ?

Ajout d'un caret ^ dans les phrases X près d'un nom au MG pour signifier sa généricité. En fonction du groupe, rappel ou non. Exemple de X^{Rappel} :

> « Die Nachrichtensprecher^ trugen schicke Kleidung. »

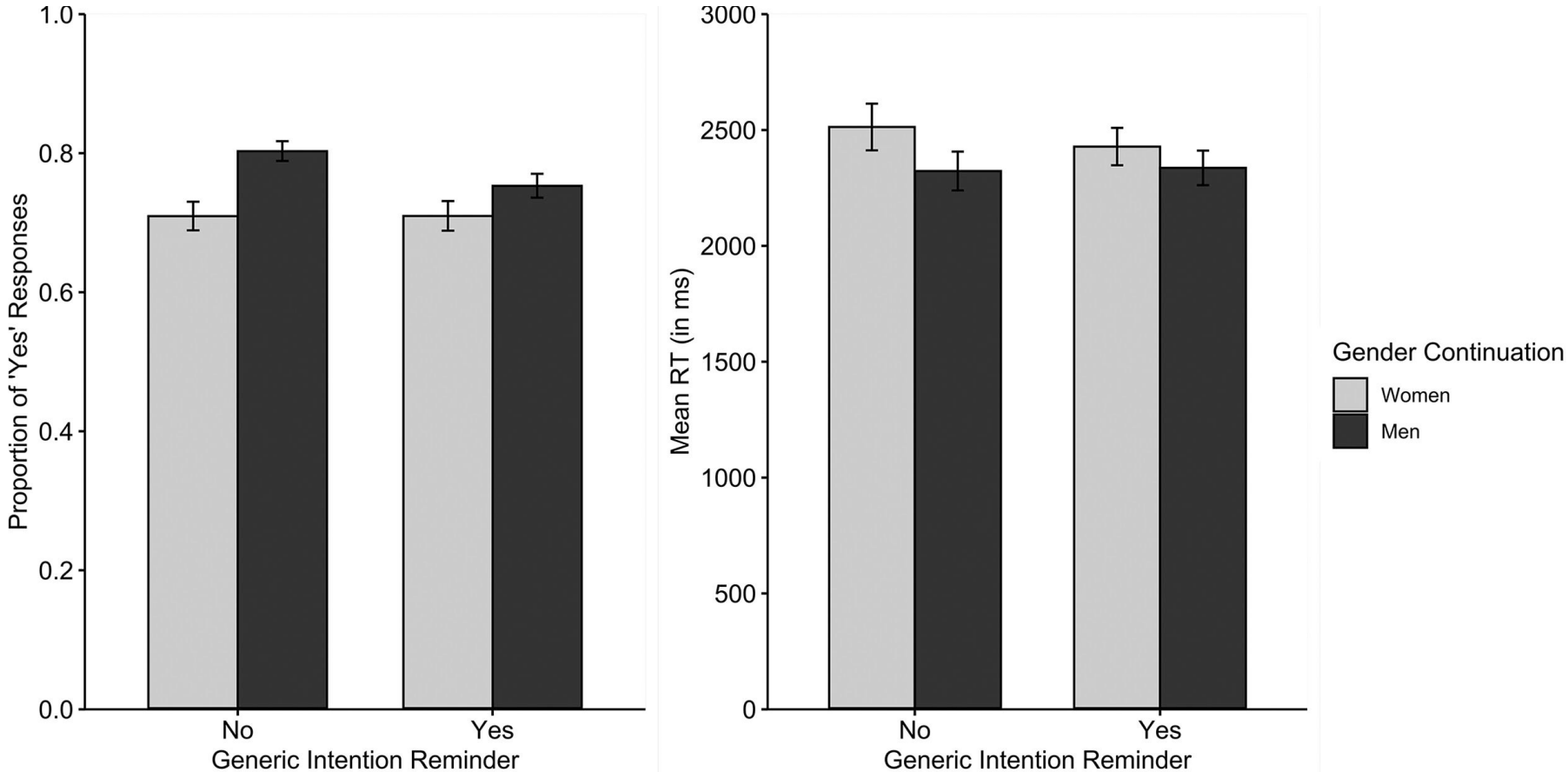


Fig. 7. – Prop. de « oui » et temps de rép. moyen selon la présence de rappel

Biais de genre linguistiques : le cas du masculin générique

Deux éléments à retenir en ce qui concerne le masculin générique :

- l'**ambigüité inhérente à son utilisation** (quiproquos linguistiques) et sa **difficulté d'application lors de la lecture** (cf. **Gygax et al., 2009**) ;
- une **association accrue du générique au masculin même en cas d'explicitation**, ce qui suggère la présence d'un biais plus profond, hypothétiquement induit lors de l'acquisition (**Gygax et al., 2019 ; Rothermund et Strack, 2024**), mais dont les causes exactes restent inconnues.

Alternatives au masculin générique

- formulations inclusives :
  - doublets (*députés et/ou députées*) ;
  - amalgamation (*député·es*, *député.es*, *député(e)s*).
- neutralisation :
  - mots épicènes (*parlementaires*) ;
  - périphrases (*personnes députées*) ;
  - noms collectifs (*députation*) ;
  - mots métonymiques (*génie*, *personnalité*) ;
  - reformulations (*Nous sommes conscients…* → *Nous avons conscience…*).

Autres techniques : accord de proximité, alternance féminin/masculin…

III. Exemple d'application : neutralisation automatique du genre à travers l'utilisation de noms collectifs en français

Méthodes de réduction et neutralisation des biais

Vue d'ensemble des méthodes proposées par **Gupta et al., 2024** :

- **augmentation des données ; corpus plus divers et représentatifs** (**Lu et al., 2020 ; Zmigrod et al., 2019**) ;
- suppression des biais dans les plongements lexicaux (**Bolukbasi et al., 2016**) ;
- suppression de données biaisées par désapprentissage (*unlearning* ; **Yao et al., 2024** ; **Agarwal et al., 2023**) ;
- utilisation de modèles supplémentaires dédiés à la détection de données biaisées (**Jeon et al., 2023**).

Tâche de réécriture du genre

« La génération, à partir d'une phrase genrée donnée, d'une ou plusieurs phrases alternatives apportant des modifications aux formes genrées, soit en les neutralisant, soit en privilégiant une forme alliant tous les genres, soit en choisissant un autre genre. » (**Doyen, 2024**)

- systèmes de réécriture vers un genre autre (arabe : **Alhafni et al., 2022a ; 2022b ; Habash et al., 2019**) ;
- systèmes de réécriture inclusive (allemand : **Pomerenke, 2022** ; portugais : **Veloso et al., 2023** ; français : **Lerner et Grouin, 2024**) ;
- systèmes de réécriture neutre (anglais : **Sun et al., 2021 ; Vanmassenhove et al., 2021**).

Méthodologie

Trois étapes principales :

1. Création d'un dictionnaire de paires noms de membres-noms collectifs
2. Extraction de phrases contenant des noms de membres au masculin générique dans un corpus français d'articles Wikipédia
3. Développement de systèmes de transformation automatique

Méthodologie

Trois étapes principales :

1. Création d'un dictionnaire de paires noms de membres-noms collectifs
2. Extraction de phrases contenant des noms de membres au masculin générique dans un corpus français d'articles Wikipédia
3. Développement de systèmes de transformation automatique

Phrase d'entrée : *Les soldats fatigués ont monté la garde.*

Phrase de sortie attendue : ***L'armée fatiguée a*** *monté la garde.*

Constitution du dictionnaire

Méthodes utilisées :

- appui sur les travaux de recherche en français sur les noms collectifs humains (**Lammert, 2010 ; Lammert et Lecolle, 2014 ; Lecolle, 2013 ; 2016**), et notamment le répertoire réalisé par **Lecolle, 2019** ;
- collecte empirique (articles de presse, Web) ;
- collecte semi-automatique (extraction d'entrées du Wiktionnaire).

Constitution du dictionnaire

Nom de membre	Nom collectif
académiciens	académie
soldats	armée
miliciens	milice
auditeurs	auditoire
danseurs	ballet
policiers	police
…	…

Nombre d'entrées au total : **315**

Extraction de phrases

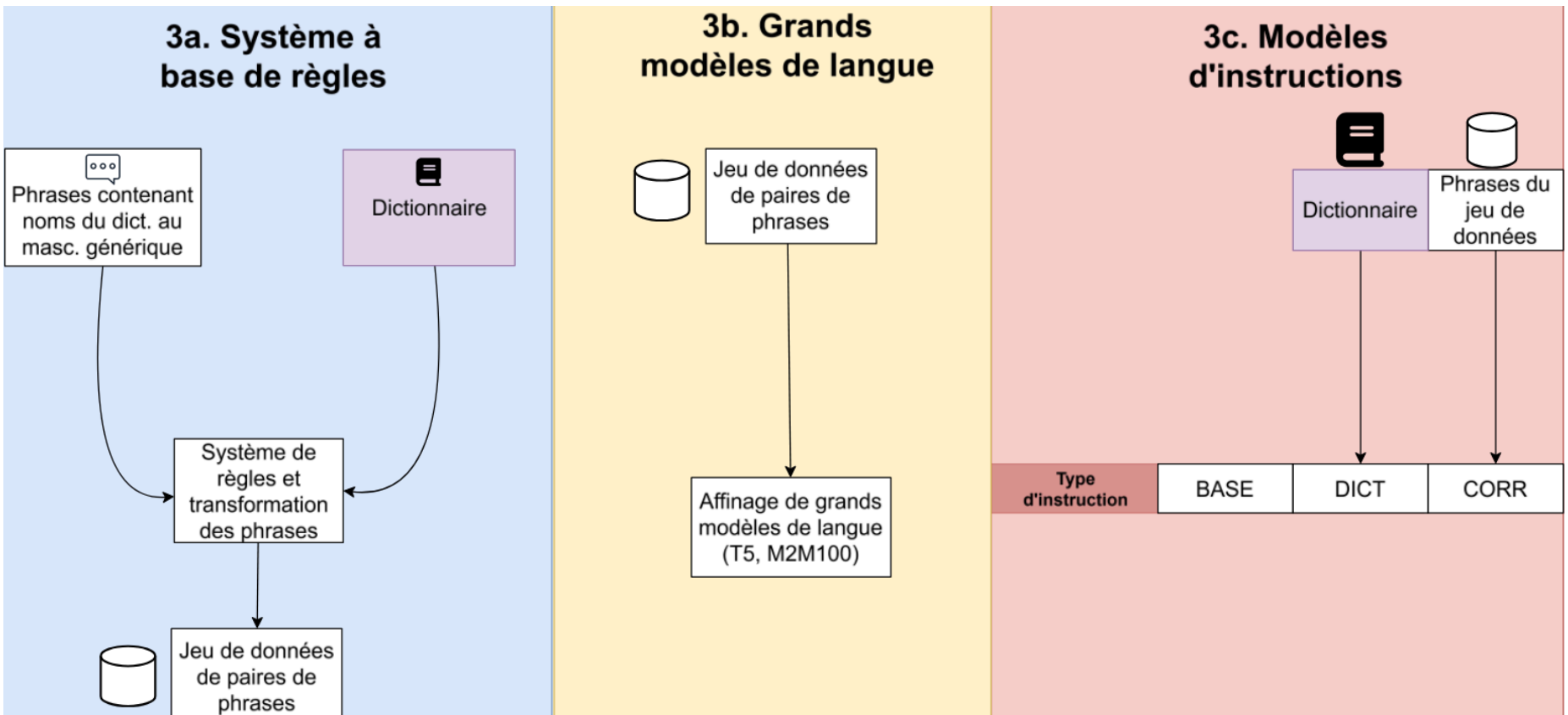
- Exploitation d'un jeu de données issu de Wikipédia (**graelo, 2023**), composé de 1,58 million d'articles en français.
- Extraction et annotation automatique des phrases avec un masculin générique, d'après notre dictionnaire.

Exemple :

> Un historique permet de lister **<n-126>les auteurs</n>** et de consulter les modifications successives de l'article par **<n-68>ses rédacteurs</n>**.

Phrases initiales dans notre jeu de données : **292 076**

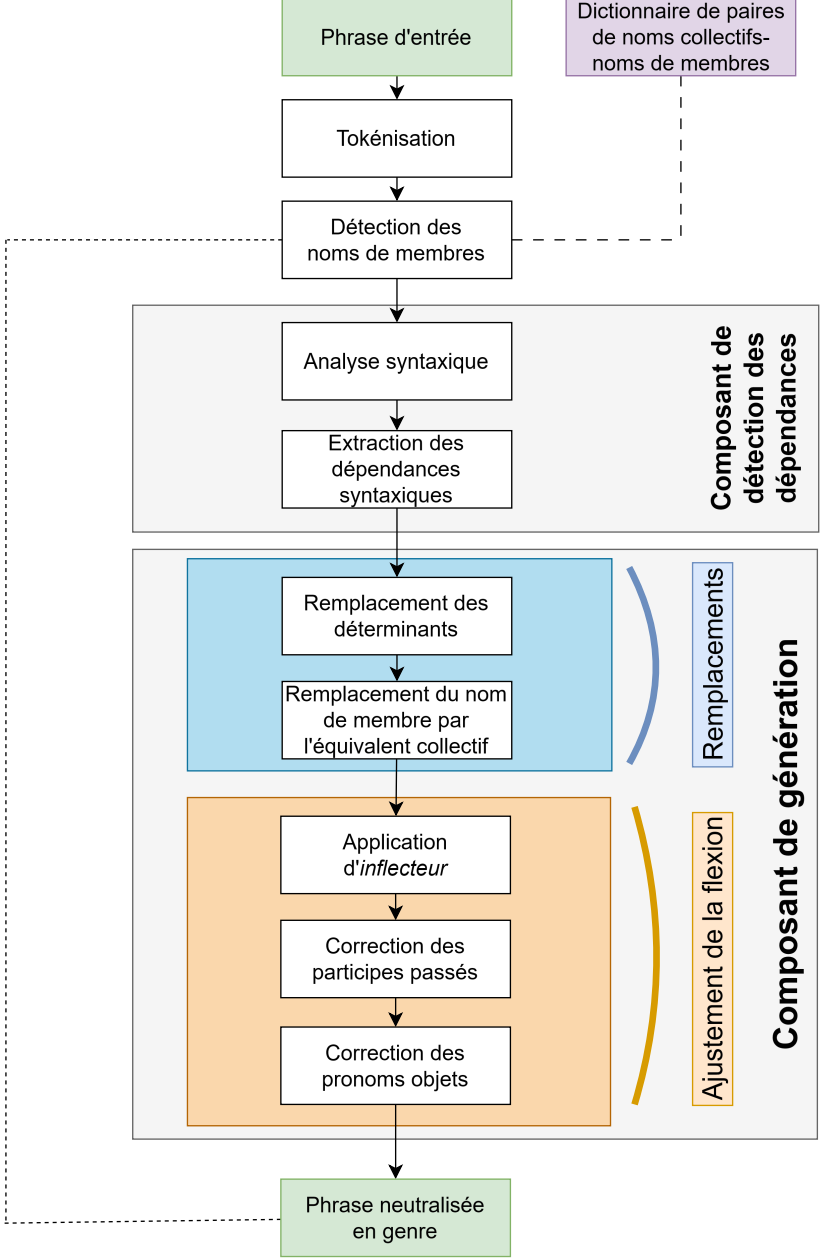
Systèmes de transformation



Systèmes de transformation à base de règles

- Tokénisation de la phrase d'entrée avec *spaCy* (**Montani et al., 2024**) et détection des noms de membres via le dictionnaire
- Exploitation de **2 composants principaux** :
  - **composant de détection des dépendances** pour trouver les dépendances syntaxiques des noms de membres (déterminants, adjectifs, verbes, etc.) ;
  - **composant de génération** pour remplacer les noms de membres par leurs équivalents collectifs et ajuster les dépendances à l'aide du module Python *inflecteur* (**Chuttarsing, 2021**).

Architecture du système de règles



Affinage de modèles de langue

- Utilisation du système de règles pour générer un jeu de données d'entrainement sur 2 corpus : Wikipédia et Europarl.
- Affinage de 2 grands modèles de langue : T5 (**Raffel et al., 2020**) et M2M100 (**Fan et al., 2020**) sur 60 000 phrases (6 000 en validation).
- Hypothétique amélioration des performances montrée par **Vanmassenhove et al., 2021**.

Modèles d'instructions

- Type de modèle non utilisé auparavant pour cette tâche.
- Instructions à quelques exemples (*few-shot prompting*).
- Claude 3 Opus (**Anthropic, 2024**) ; Mixtral 8x7B-Instruct (**Jiang et al., 2024**).

Instruction	Contenu
BASE	Make this French sentence inclusive by replacing generic masculine nouns {NM1, NM2, ...} with their French collective noun equivalents {NCOLL1, NCOLL2 ...}. {EXAMPLES} {ORIGINAL SENTENCE} →
DICT	Make this French sentence inclusive by replacing generic masculine nouns with their French collective noun equivalents. {EXAMPLES} {ORIGINAL SENTENCE} →
CORR	Correct grammar in this French sentence. {EXAMPLES} {ORIGINAL SENTENCE} →

Résultats : détection des dépendances

- 2 jeux de données d'évaluation : Wikipédia et Europarl (**Koehn, 2005**) ; 250 phrases chacun. Annotations de référence pour les 500 phrases.
- **Objectif** : évaluer la bonne détection des dépendances syntaxiques du nom de membre à modifier
- **Référence** : dép. synt. du nom de membre détectées par défaut par spaCy, sans filtrage (hors ponctuation)

	Wikipédia			Europarl			Moy.		
	Précision	**Rappel**	**F-score**	**Précision**	**Rappel**	**F-score**	**Précision**	**Rappel**	**F-score**
Réf. (baseline spaCy)	0,096	0,723	0,169	0,115	0,689	0,197	0,1055	0,706	0,183
GeNRe-RBS	0,773	0,855	0,812	0,758	0,813	0,785	0,7655	0,834	0,7985

Résultats : génération (tous modèles)

		Wikipédia	Europarl	Moy.
WER	Référence	13,35 %	13,344 %	13,347 %
	GeNRe-RBS	**3,358** %	**3,448** %	**3,403** %
	GeNRe-FT-T5 ◯	6,095 %	4,128 %	5,111 %
	GeNRe-FT-M2M100 ◯	6,679 %	4,116 %	5,397 %
	Mixtral 8x7B-BASE-C △	25,421 %	31,638 %	28,529
	Mixtral 8x7B-DICT-C △	12,981 %	24,376 %	18,678 %
	Mixtral 8x7B-CORR-C △	18,996 %	21,335 %	20,166 %
	Claude 3 Opus-BASE △	14,02 %	10,304 %	12,162 %
	Claude 3 Opus-DICT △	4,052 %	3,459 %	3,755 %
	Claude 3 Opus-CORR △	11,413 %	8,929 %	10,171 %

		Wikipédia	Europarl	Moy.
BLEU	Référence	79,634	81,465	80,549
	GeNRe-RBS	93,024	**93,833**	93,428
	GeNRe-FT-T5 ◯	88,144	93,211	90,678
	GeNRe-FT-M2M100 ◯	87,09	93,298	90,194
	Mixtral 8x7B-BASE-C △	65,537	73,69	69,614
	Mixtral 8x7B-DICT-C △	81,74	82,603	82,171
	Mixtral 8x7B-CORR-C △	71,993	71,44	71,716
	Claude 3 Opus-BASE △	80,087	85,88	82,983
	Claude 3 Opus-DICT △	**93,845**	93,434	**93,639**
	Claude 3 Opus-CORR △	84,648	85,615	85,131

Apports et limites

- création d'un dictionnaire de paires de noms de membres-noms collectifs ;
- jeu de données de paires de phrases au masculin générique/neutralisées ;
- développement du premier système de réécriture de genre neutre à base de règles pour le français ;
- mise au jour de la potentialité des modèles d'instructions combinés à des ressources précréées.

- spécificité sémantique/prédicative des noms collectifs ;
- coréférence non prise en charge totalement ; problématiques de reprise des noms collectifs humains.

Conclusion

- Résolution des biais (de genre et autres) : enjeu majeur étant donné le développement rapide des outils de TAL et d'IA, et leurs diverses intégrations.
- Importance de considérer le problème des biais de genre sous tous ses aspects, **à la fois sociaux et linguistiques**.
- En comparaison, les **biais de genre linguistiques restent moins recherchés** que les biais sociodémographiques, malgré leur impact potentiellement important.
- Recherche sur les biais de genre majoritairement anglocentrée : nécessité d'étendre davantage les travaux à **d'autres langues**.

Merci pour votre écoute.

Vos questions sont les bienvenues.

Bibliographie

Agarwal, S., Veerubhotla, A., & Bansal, S. (2023). PEFTDebias : Capturing Debiasing Information Using PEFTs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1992‑2000. https://doi.org/10.18653/v1/2023.emnlp-main.122

Alhafni, B., Habash, N., & Bouamor, H. (2022b). User-Centric Gender Rewriting. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 618‑631. https://doi.org/10.18653/v1/2022.naacl-main.46

Alhafni, B., Habash, N., Bouamor, H., Obeid, O., Alrowili, S., Alzeer, D., Alshanqiti, K. M., ElBakry, A., ElNokrashy, M., Gabr, M., Issam, A., Qaddoumi, A., Vijay-Shanker, K., & Zyate, M. (2022a,). *The Shared Task on Gender Rewriting*. https://doi.org/10.48550/arXiv.2210.12410

Anthropic. (2024,). *The Claude 3 Model Family: Opus, Sonnet, Haiku*.

Arzaghi, M., Carichon, F., & Farnadi, G. (2024,). *Understanding Intrinsic Socioeconomic Biases in Large Language Models* (Numéro arXiv:2405.18662). arXiv. https://doi.org/10.48550/arXiv.2405.18662

Bella, G., Helm, P., Koch, G., & Giunchiglia, F. (2024). Tackling Language Modelling Bias in Support of Linguistic Diversity. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 562‑572. https://doi.org/10.1145/3630106.3658925

Bernhard, D., Vergez-Couret, M., & Dupuy, E. (2024). Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues. *Cahiers du plurilinguisme européen*, *16*. https://doi.org/10.57086/cpe.1710

Blodgett, S. L., & O'Connor, B. (2017,). *Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English* (Numéro arXiv:1707.00061). arXiv. https://doi.org/10.48550/arXiv.1707.00061

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. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Cai, Y., Cao, D., Guo, R., Wen, Y., Liu, G., & Chen, E. (2024,). *Locating and Mitigating Gender Bias in Large Language Models* (Numéro arXiv:2403.14409). arXiv. https://doi.org/10.48550/arXiv.2403.14409

Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., & Joty, S. (2024,). *ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching Up?* (Numéro arXiv:2311.16989). arXiv. https://doi.org/10.48550/arXiv.2311.16989

Chuttarsing, A. (2021,). *Inflecteur*.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020,). *A Survey of the State of Explainable AI for Natural Language Processing* (Numéro arXiv:2010.00711). arXiv. https://doi.org/10.48550/arXiv.2010.00711

De Sousa, W. G., Fidelis, R. A., De Souza Bermejo, P. H., Da Silva Gonçalo, A. G., & De Souza Melo, B. (2022). Artificial Intelligence and Speedy Trial in the Judiciary: Myth, Reality or Need? A Case Study in the Brazilian Supreme Court (STF). *Government Information Quarterly*, *39*(1), 101660. https://doi.org/10.1016/j.giq.2021.101660

Despot, K. Štrkalj. (2021). How Language Influences Conceptualization: From Whorfianism to Neo-Whorfianism. *Collegium antropologicum*, *45*(4), 373–380. https://doi.org/10.5671/ca.45.4.9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019,). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Numéro arXiv:1810.04805). arXiv.

Dhingra, H., Jayashanker, P., Moghe, S., & Strubell, E. (2023,). *Queer People Are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models* (Numéro arXiv:2307.00101). arXiv. https://doi.org/10.48550/arXiv.2307.00101

Doyen, E. (2024). *Neutralisation automatique du genre à travers l'utilisation de noms collectifs en français*.

Ducel, F., Névéol, A., & Fort, K. (2024b). La recherche sur les biais dans les modèles de langue est biaisée: état de l'art en abyme. *Revue TAL : traitement automatique des langues*, *3*(64).

Ducel, F., Névéol, A., & Fort, K. (2024a). Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs. *TALN 2024*.

Fabiano, N. (2024,). *AI Act and Large Language Models (LLMs): When Critical Issues and Privacy Impact Require Human and Ethical Oversight* (Numéro arXiv:2404.00600). arXiv. https://doi.org/10.48550/arXiv.2404.00600

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020,). *Beyond English-Centric Multilingual Machine Translation*.

Gallienne, R., & Poibeau, T. (2023). Quelques observations sur la notion de biais dans les modèles de langue. *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL*, 1-13.

Gan, C., Zhang, Q., & Mori, T. (2024,). *Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening* (Numéro arXiv:2401.08315). arXiv. https://doi.org/10.48550/arXiv.2401.08315

Gomila, A. (2015). Language and Thought: The Neo-Whorfian Hypothesis. In *International Encyclopedia of the Social & Behavioral Sciences: International Encyclopedia of the Social & Behavioral Sciences* (Second edition, p. 293-299). Elsevier.

graelo. (2023,). *Graelo/Wikipedia*.

Gupta, V., Venkit, P. N., Wilson, S., & Passonneau, R. J. (2024,). *Sociodemographic Bias in Language Models: A Survey and Forward Path* (Numéro arXiv:2306.08158). arXiv. https://doi.org/10.48550/arXiv.2306.08158

Gygax, P. M., Schoenhals, L., Lévy, A., Luethold, P., & Gabriel, U. (2019). Exploring the Onset of a Male-Biased Interpretation of Masculine Generics Among French Speaking Kindergarten Children. *Frontiers in Psychology*, *10*, 1225. https://doi.org/10.3389/fpsyg.2019.01225

Gygax, P., Gabriel, U., Lévy, A., Pool, E., Grivel, M., & Pedrazzini, E. (2012). The Masculine Form and Its Competing Interpretations in French: When Linking Grammatically Masculine Role Names to Female Referents Is Difficult. *Journal of Cognitive Psychology*, *24*(4), 395-408. https://doi.org/10.1080/20445911.2011.642858

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J., & Garnham, A. (2008). Generically Intended, but Specifically Interpreted: When Beauticians, Musicians, and Mechanics Are All Men. *Language and Cognitive Processes*, *23*(3), 464-485. https://doi.org/10.1080/01690960701702035

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J., & Garnham, A. (2009). Some Grammatical Rules Are More Difficult than Others: The Case of the Generic Interpretation of the Masculine. *European Journal of Psychology of Education*, *24*(2), 235-246. https://doi.org/10.1007/BF03173014

Habash, N., Bouamor, H., & Chung, C. (2019). Automatic Gender Identification and Reinflection in Arabic. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 155-165. https://doi.org/10.18653/v1/W19-3822

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hall, M., Maaten, L. van der, Gustafson, L., Jones, M., & Adcock, A. (2022,). *A Systematic Study of Bias Amplification* (Numéro arXiv:2201.11706). arXiv. https://doi.org/10.48550/arXiv.2201.11706

Harris Interactive. (2017). *L'écriture Inclusive : La Population Française Connaît-Elle l'écriture Inclusive ? Quelle Opinion En a-t-Elle ?*.

Hofstadter, D., & Sander, E. (2013). *L'Analogie, Coeur de la Pensée* (Odile Jacob).

Huang, T. (2024,). *Content Moderation by LLM: From Accuracy to Legitimacy* (Numéro arXiv:2409.03219). arXiv. https://doi.org/10.48550/arXiv.2409.03219

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

Jeon, E., Lee, M., Park, J., Kim, Y., Mok, W.-L., & Lee, S. (2023,). *Improving Bias Mitigation through Bias Experts in Natural Language Understanding* (Numéro arXiv:2312.03577). arXiv. https://doi.org/10.48550/arXiv.2312.03577

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., D. de las, Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. (2024,). *Mixtral of Experts*.

Kamruzzaman, M., Shovon, M. M. I., & Kim, G. L. (2024,). *Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models* (Numéro arXiv:2309.08902). arXiv. https://doi.org/10.48550/arXiv.2309.08902

Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024,). *Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting* (Numéro arXiv:2401.15585). arXiv. https://doi.org/10.48550/arXiv.2401.15585

Kay, P., & Kempton, W. (1984). What Is the Sapir-Whorf Hypothesis?. *American Anthropologist*, *86*(1), 65-79. https://doi.org/10.1525/aa.1984.86.1.02a00050

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*.

Körner, A., Abraham, B., Rummer, R., & Strack, F. (2022). Gender Representations Elicited by the Gender Star Form. *Journal of Language and Social Psychology*, *41*(5), 553-571. https://doi.org/10.1177/0261927X221080181

Lammert, M. (2010). *Sémantique et Cognition : Les Noms Collectifs*. Droz.

Lammert, M., & Lecolle, M. (2014). Les Noms Collectifs En Français, Une Vue d'ensemble. *Cahiers de lexicologie*, *105*, 203-222.

Lecolle, M. (2013). Noms Collectifs Humains : Un Point de Vue de Sémantique Lexicale Sur l'identité Dans Le Rapport Individu/Groupe. *¿ Interrogations ?*, *16*.

Lecolle, M. (2016). Noms Collectifs Humains : Nomination et Prédication. *Argumentation et analyse du discours*, *17*. https://doi.org/10.4000/aad.2208

Lecolle, M. (2019). *Les Noms Collectifs Humains En Français. Enjeux Sémantiques, Lexicaux et Discursifs* (C. Jacquet-Pfau, éd.). Lambert-Lucas.

Lerner, P., & Grouin, C. (2024). *INCLURE: A Dataset and Toolkit for Inclusive French Translation*.

Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D.-A., Akyürek, E., Anandkumar, A., Andreas, J., Mordatch, I., Torralba, A., & Zhu, Y. (2022,). *Pre-Trained Language Models for Interactive Decision-Making* (Numéro arXiv:2202.01771). arXiv. https://doi.org/10.48550/arXiv.2202.01771

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender Bias in Neural Natural Language Processing. In V. Nigam, T. Ban Kirigin, C. Talcott, J. Guttman, S. Kuznetsov, B. Thau Loo, & M. Okada (éds.), *Logic, Language, and Security: Vol. 12300. Logic, Language, and Security* (p. 189-202). Springer International Publishing. https://doi.org/10.1007/978-3-030-62077-6_14

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013,). *Efficient Estimation of Word Representations in Vector Space* (Numéro arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Montani, I., Honnibal, M., Boyd, A., Landeghem, S. V., & Peters, H. (2024,). *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/ZENODO.1212303

NLLB Team, Costa-jussà, M. R., Cross, J., ̧Celebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., … Wang, J. (2022,). *No Language Left Behind: Scaling Human-Centered Machine Translation*. arXiv. https://doi.org/10.48550/ARXIV.2207.04672

Omoregbe, N. A. I., Ndaman, I. O., Misra, S., Abayomi-Alli, O. O., & Damăsevĭcius, R. (2020). Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. *Journal of Healthcare Engineering*, *2020*, 1-14. https://doi.org/10.1155/2020/8839524

Pomerenke, D. (2022,). *INCLUSIFY: A Benchmark and a Model for Gender-Inclusive German*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models Are Unsupervised Multitask Learners*.

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*.

Rothermund, P., & Strack, F. (2024). Reminding May Not Be Enough: Overcoming the Male Dominance of the Generic Masculine. *Journal of Language and Social Psychology*, *43*(4), 468-485. https://doi.org/10.1177/0261927X241237739

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

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018,). *Emotion Detection in Text: A Review* (Numéro arXiv:1806.00674). arXiv. https://doi.org/10.48550/arXiv.1806.00674

Spinelli, E., Chevrot, J.-P., & Varnet, L. (2023). Neutral Is Not Fair Enough: Testing the Efficiency of Different Language Gender-Fair Strategies. *Frontiers in Psychology*, *14*, 1256779. https://doi.org/10.3389/fpsyg.2023.1256779

Stahlberg, D., Sczesny, S., & Braun, F. (2001). Name Your Favorite Musician: Effects of Masculine Generics and of Their Alternatives in German. *Journal of Language and Social Psychology*, *20*(4), 464-469. https://doi.org/10.1177/0261927X01020004004

Sun, T., Webster, K., Shah, A., Wang, W. Y., & Johnson, M. (2021,). *They, Them, Theirs: Rewriting with Gender-Neutral English* (Numéro arXiv:2102.06788). arXiv.

Tian, D., Jiang, S., Zhang, L., Lu, X., & Xu, Y. (2024). The Role of Large Language Models in Medical Image Processing: A Narrative Review. *Quantitative Imaging in Medicine and Surgery*, *14*(1), 1108-1121. https://doi.org/10.21037/qims-23-892

Umera-Okeke, N. (2012). Linguistic Sexism: An Overview of the English Language in Everyday Discourse. *An International Journal of Language, Literature and Gender Studies Bahir Dar, Ethiopia*, *1*(1).

UNESCO IRCAI. (2024,). *Challenging Systematic Prejudices: An Investigation into Gender Bias in Large Language Models*.

Vanmassenhove, E., Emmery, C., & Shterionov, D. (2021,). *NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives*. https://doi.org/10.48550/arXiv.2109.06105

Veloso, L., Coheur, L., & Ribeiro, R. (2023). A Rewriting Approach for Gender Inclusivity in Portuguese. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8747-8759. https://doi.org/10.18653/v1/2023.findings-emnlp.585

Wahde, M., & Virgolin, M. (2022,). *Conversational Agents: Theory and Applications*. https://doi.org/10.1142/9789811246050_0012

Wan, Y., & Chang, K.-W. (2024,). *White Men Lead, Black Women Help? Benchmarking Language Agency Social Biases in LLMs* (Numéro arXiv:2404.10508). arXiv. https://doi.org/10.48550/arXiv.2404.10508

Wisniewski, G., Zhu, L., Ballier, N., & Yvon, F. (2021a). *Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire*.

Wisniewski, G., Zhu, L., Ballier, N., & Yvon, F. (2021b). Screening Gender Transfer in Neural Machine Translation. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 311-321. https://doi.org/10.18653/v1/2021.blackboxnlp-1.24

Xu, B. (2024,). *What Is Meant by AGI? On the Definition of Artificial General Intelligence* (Numéro arXiv:2404.10731). arXiv. https://doi.org/10.48550/arXiv.2404.10731

Yao, Y., Xu, X., & Liu, Y. (2024,). *Large Language Model Unlearning* (Numéro arXiv:2310.10683). arXiv. https://doi.org/10.48550/arXiv.2310.10683

Zhang, X., Chan, F. T., Yan, C., & Bose, I. (2022). Towards Risk-Aware Artificial Intelligence and Machine Learning Systems: An Overview. *Decision Support Systems*, *159*, 113800. https://doi.org/10.1016/j.dss.2022.113800

Zhao, J., Ding, Y., Jia, C., Wang, Y., & Qian, Z. (2024,). *Gender Bias in Large Language Models across Multiple Languages* (Numéro arXiv:2403.00277). arXiv. https://doi.org/10.48550/arXiv.2403.00277

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., & Chang, K.-W. (2019). Examining Gender Bias in Languages with Grammatical Gender. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5276-5284.

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1651-1661. https://doi.org/10.18653/v1/P19-1161