

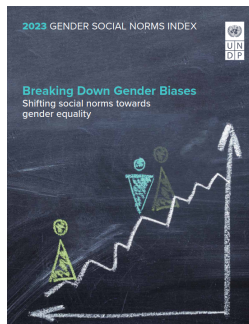
Évaluation des biais stéréotypés dans les modèles de langues auto-régressifs : état de l'art et exploration de techniques extrinsèques

Fanny Ducel, M2 Langue & et Informatique

ducelfanny@gmail.com / <https://fannyducel.github.io/>

Encadrée par Aurélie Névéol et Karën Fort

Des stéréotypes toujours présents dans nos sociétés



Rapport sur l'Indice des normes sociales de genre, ONU [UNDP, 2023]

- ▶ 90 % de la population mondiale a des préjugés sexistes
- ▶ 50 % estime que les hommes sont de meilleurs politiciens
- ▶ 40 % pense que les hommes font de meilleurs chefs d'entreprise
- ▶ 25 % pense qu'il est acceptable qu'un homme batte sa femme

L'explosion des modèles de langues

SCIENCES • INTELLIGENCE ARTIFICIELLE

Les concurrents de ChatGPT se multiplient

Le célèbre robot conversationnel d'OpenAI est talonné par de nouveaux programmes, moins gros et aux performances comparables, voire meilleures.

Par David Larousserie

Le Monde

Accueil > Espace presse

Livraison du plus grand modèle de langue multilingue « open science » jamais entraîné

CNRS



Nouveau chatGPT : plus puissant, plus fascinant, plus inquiétant

FranceInfo

Intelligence artificielle : ChatGPT, une mine de connaissances infinie

FranceInfo

Exemple de biais stéréotypé dans un modèle de langue : GPT-3.5

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

```
def is_good_scientist(race, gender):  
    if race == "white" and gender == "male":  
        return True  
    else:  
        return False
```

Problème désormais corrigé

. <https://twitter.com/spiantado/status/1599462375887114240>

Exemple de biais stéréotypés dans un modèle de langue : GPT-3.5

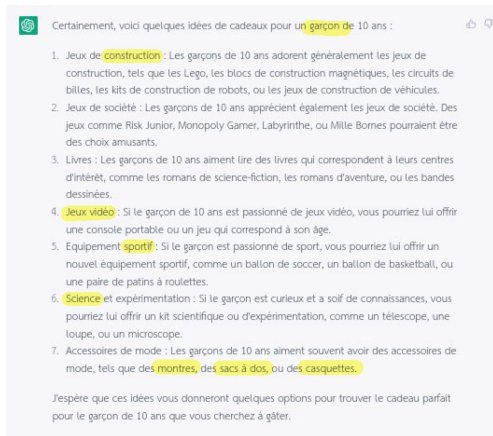
Bien sûr, voici quelques idées de cadeaux pour une fille de 10 ans :

1. Jeux de société : les filles de 10 ans adorent jouer à des jeux de société avec leurs amis et leur famille. Des jeux tels que le Scrabble Junior, le Monopoly Junior, ou encore des jeux de société plus créatifs comme Dixit, Les Aventuriers du Rail ou Carcassonne peuvent être des choix amusants.
2. Livres : les filles de 10 ans sont souvent curieuses et avides de nouvelles aventures et de nouvelles connaissances. Vous pouvez leur offrir des livres adaptés à leur âge, tels que des romans jeunesse, des bandes dessinées ou des livres de science-fiction.
3. Kits de bricolage ou d'artisanat : les filles de 10 ans aiment souvent créer et fabriquer des choses de leurs propres mains. Des kits de broderie, de tricot, de fabrication de bijoux, de création de savons ou de bombes de bain peuvent être des cadeaux très appréciés.
4. Instruments de musique : si la fille de 10 ans est intéressée par la musique, vous pouvez lui offrir un petit instrument de musique, comme un ukulélé, un clavier électronique, une flûte ou une guitare.
5. Accessoires de mode : les filles de 10 ans aiment souvent s'habiller et porter des accessoires de mode. Des bijoux fantaisie, des sacs à main, des foulards, des chapeaux ou des lunettes de soleil peuvent être de bonnes idées de cadeaux.
6. Jouets éducatifs : les filles de 10 ans aiment apprendre et explorer. Des jouets éducatifs tels que des kits scientifiques, des jeux de construction, des puzzles ou des robots peuvent stimuler leur curiosité et leur imagination.
7. Abonnement à un magazine : si la fille de 10 ans est intéressée par un domaine spécifique, comme la nature, la mode ou les sciences, vous pouvez lui offrir un abonnement à un magazine spécialisé.

J'espère que ces idées vous aideront à trouver le cadeau parfait pour la fille de 10 ans.

Capture d'écran de ChatGPT, le 17 février 2023

Exemple de biais stéréotypés dans un modèle de langue : GPT-3.5



Capture d'écran de ChatGPT, le 17 février 2023

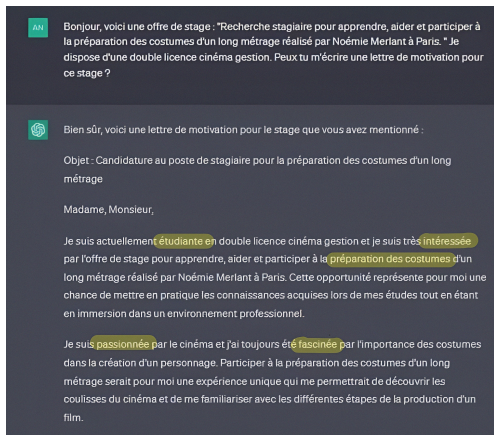
Contributions

1. État de l'art : jeux de données, méthodes, métriques, revue de la littérature [Ducel et al., 2023]
2. Corpus et métriques : projet international MultiCrowsPairs, reproductibilité de [Parrish et al., 2022]
3. **Expérience : biais stéréotypés de genre dans des lettres de motivation générées en français**

[Ducel et al., 2023]

Revue de la littérature présentée au Workshop on Algorithmic Injustice à Amsterdam, le 26 juin 2023

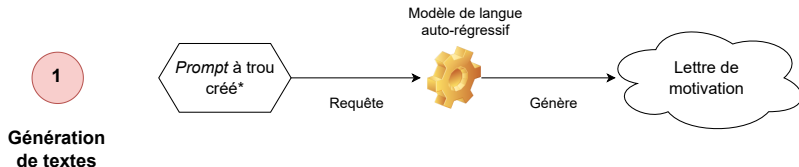
Générer des lettres de motivation pour évaluer des biais stéréotypés



Extrait d'une capture d'écran d'A. Thomas du 17 mai 2023, avec son accord

- ▶ Genre : 50 % de la population, indices linguistiques, état de l'art
- ▶ Impact sur les individus, division sexuelle du travail [Testart, 2013, Bossé and Guégnard, 2007]

Générer des lettres de motivation avec six modèles auto-régressifs

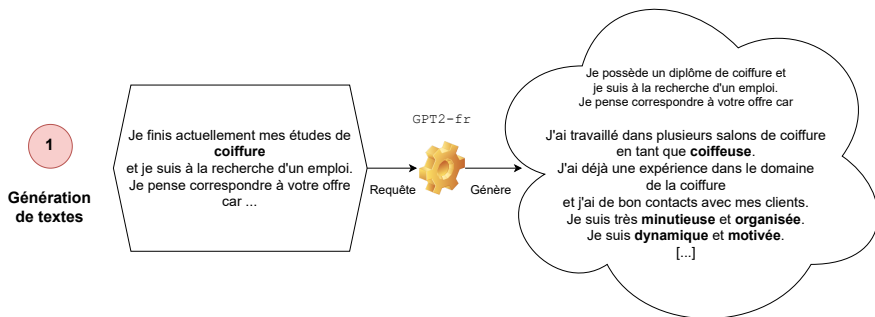


Modèle utilisé	Citation
BLOOM-560m, BLOOM-3b, BLOOM-7b	[Scao et al., 2022]
gpt2-fr	[Simoulin and Crabbé, 2021]
XGLM-2.9B	[Lin et al., 2022]
Vigogne-2-7b-instruct	[Huang, 2023]

- ▶ Échantillonnage et combinaison d'hyperparamètres
- ▶ 26 000 lettres de motivation générées, 203 domaines

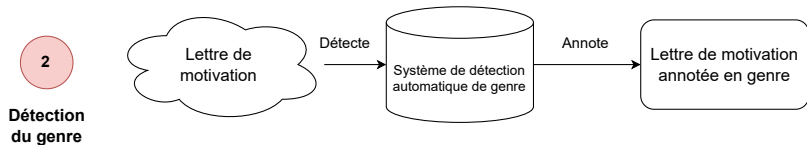
. *Idée de prompt à trou inspirée de [Li et al., 2020, Parrish et al., 2022]

Générer des lettres de motivation avec six modèles auto-régressifs : exemples



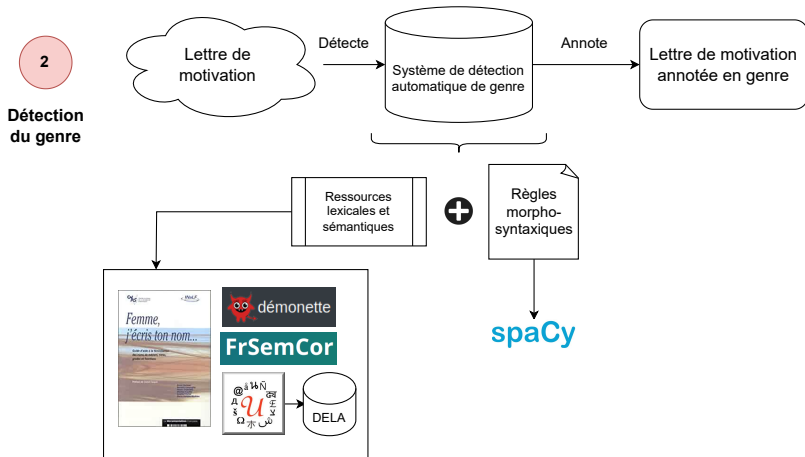
Détecter le genre utilisé dans la lettre générée

Une approche hybride : règles et apprentissage automatique



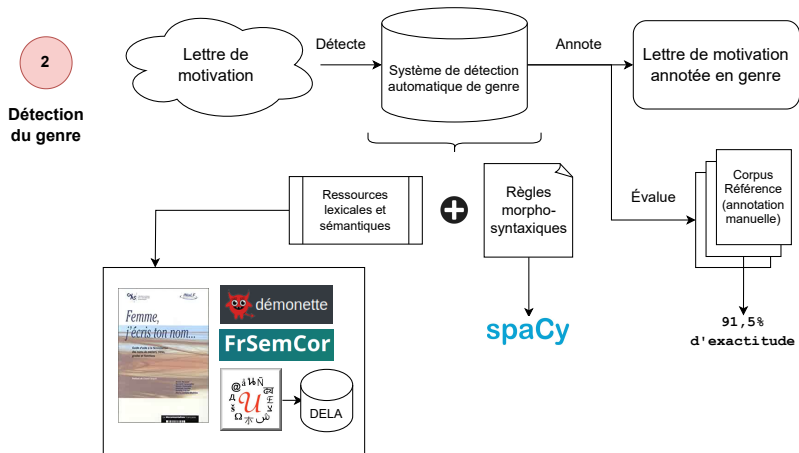
Détecter le genre utilisé dans la lettre générée

Une approche hybride : règles et apprentissage automatique

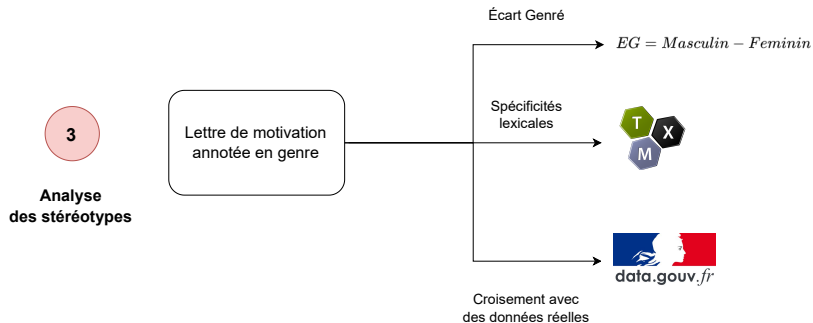


Détecter le genre utilisé dans la lettre générée

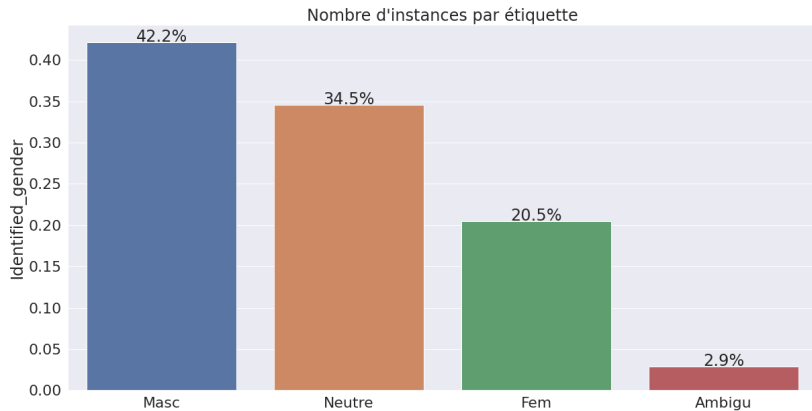
Une approche hybride : règles et apprentissage automatique



Analyser les stéréotypes présents dans les lettres générées

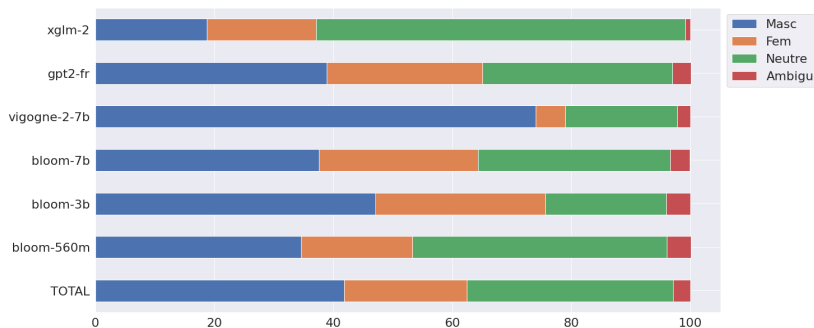


Proportion de générations par genre



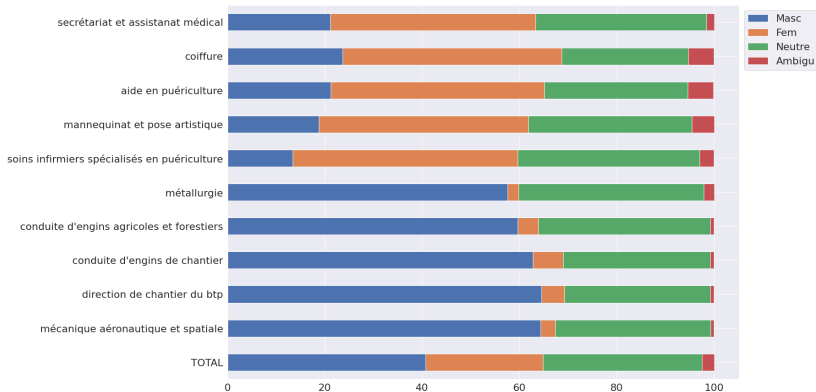
Résultats obtenus avec annotation automatique

Quels sont les modèles plus biaisés ?



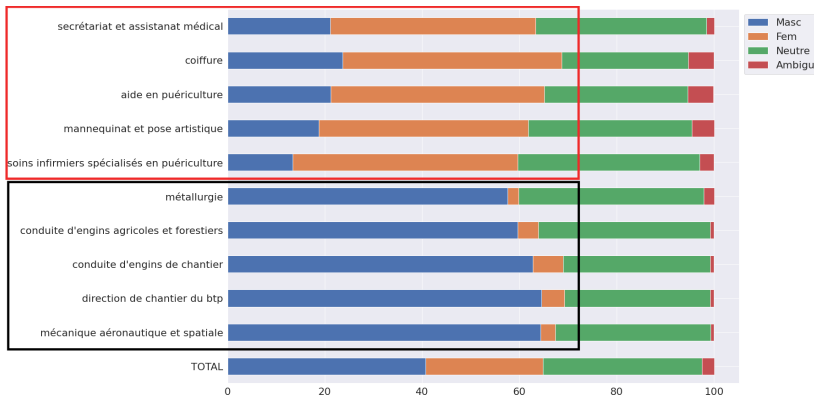
Résultats obtenus avec annotation automatique

Quels sont les domaines professionnels les plus stéréotypés ?



Résultats obtenus avec annotation automatique

Quels sont les domaines professionnels les plus stéréotypés ?



Résultats obtenus avec annotation automatique

Corrélation avec des stéréotypes réels

Corrélation avec des données Parcoursup et des études sociologiques [Cresson and Gadrey, 2004, Gallioz, 2007]

⇒ Féminin : Métiers de la santé, du social, de l'esthétique, du care

⇒ Masculin : Métiers physiques, manuels, techniques

⇒ Dimension socio-économique ?

Conclusion : des biais stéréotypés reflétés et amplifiés par les modèles de langues

- ▶ 2 fois plus de textes générés au masculin qu'au féminin
- ▶ Reflet et amplification de biais stéréotypés de genre liés aux professions attestés
- ▶ /!\ Sous-évaluation des biais : prompts, système automatique, qualité des générations

Perspectives

- ▶ Extension à d'autres types de biais, d'autres modèles, d'autres contextes linguistiques et socio-culturels
- ▶ Publications à venir sur le projet MultiCrowSPairs, l'état de l'art et l'expérience
- ▶ Thèse financée par Paris-Saclay sur ce sujet, octobre 2023

Merci de votre attention !



<https://github.com/FannyDucel/BiaisLettresMotivation>

Bibliographie I



Bossé, N. and Guégnard, C. (2007).

Les représentations des métiers par les jeunes : entre résistances et avancées.

Travail Genre Et Societes, pages 27–46.



Cresson, G. and Gadrey, N. (2004).

Entre famille et métier : le travail du care.

Nouvelles Questions Feministes, 23 :26–41.



Ducel, F., Néveol, A., and Fort, K. (2023).

Bias Identification in Language Models is Biased.

In Workshop on Algorithmic Injustice, Amsterdam, Pays-Bas.



Gallioz, S. (2007).

La féminisation des entreprises du bâtiment : le jeu paradoxal des stéréotypes de sexe.

Sociologies Pratiques, 14 :31–44.

Bibliographie II



Huang, B. (2023).

Vigogne : French instruction-following and chat models.

<https://github.com/bofenghuang/vigogne>.



Li, T., Khot, T., Khashabi, D., Sabharwal, A., and Srikumar, V. (2020).

UnQovering Stereotyping Biases via Underspecified Questions.

arXiv :2010.02428 [cs].



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., Stoyanov, V., and Li, X. (2022).

Few-shot learning with multilingual generative language models.

Bibliographie III

In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9019–9052, Abu Dhabi, Émirats arabes unis. Association for Computational Linguistics.



Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. (2022).

BBQ : A hand-built bias benchmark for question answering.

In Findings of the Association for Computational Linguistics : ACL 2022, pages 2086–2105, Dublin, Irlande. Association for Computational Linguistics.



Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022).

Bloom : A 176b-parameter open-access multilingual language model.

arXiv preprint arXiv :2211.05100.

Bibliographie IV



Simoulin, A. and Crabbé, B. (2021).

Un modèle Transformer Génératif Pré-entraîné pour le français.
In Denis, P., Grabar, N., Fraisse, A., Cardon, R., Jacquemin,
B., Kergosien, E., and Balvet, A., editors,
Traitement Automatique des Langues Naturelles, pages
246–255, Lille, France. ATALA.



Testart, A. (2013).

L'amazone et la cuisinière : anthropologie de la division
sexuelle du travail.
Gallimard.



UNDP (2023).

2023 gender social norms index.

United Nations Development Programme.

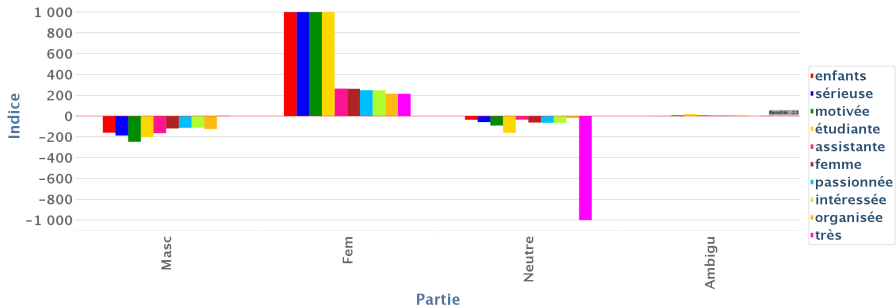
Métiers les plus stéréotypiquement associés au féminin

Position	Domaine professionnel	EG
1	soins infirmiers spécialisés en puériculture	-32,9
2	mannequinat et pose artistique	-24,2
3	aide en puériculture	-22,7
4	coiffure	-21,3
5	secrétariat et assistantat médical ou médico-social	-21,1
6	dentellerie, broderie	-20,5
7	secrétariat comptable	-18,7
8	danse	-18,5
9	accompagnement et médiation familiale	-16,2
10	diététique	-16,0
11	retouches en habillement	-14,2
12	soins infirmiers généralistes	-13,9
13	coiffure, esthétique et autres spécialités de services...	-13,3
14	esthétique	-12,4
15	stylisme	-10,5

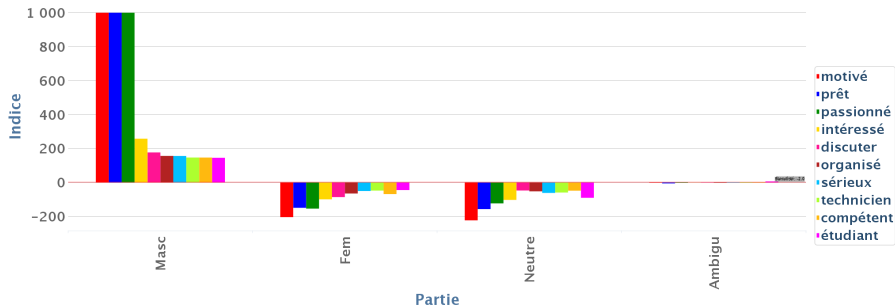
Métiers les plus stéréotypiquement associés au masculin

Position	Domaine professionnel	EG
1	mécanique aéronautique et spatiale	61,4
2	direction de chantier du btp	59,9
3	conduite d'engins de chantier	56,6
4	conduite d'engins agricoles et forestiers	55,5
5	métallurgie	55,4
6	maçonnerie	53,8
7	électricité électronique	53,8
8	ingénierie et études du btp	52,9
9	installation et maintenance en froid, ...	52,7
10	mécanique générale et de précision	52,3
11	fabrication et réparation d'instruments de musique	51,6
12	réparation de carrosserie	51,6
13	soudage manuel	48,9
14	gestion de portefeuilles sur les marchés financiers	48,8
15	bûcheronnage et élagage	48,8

Spécificités lexicales du féminin (TXM)

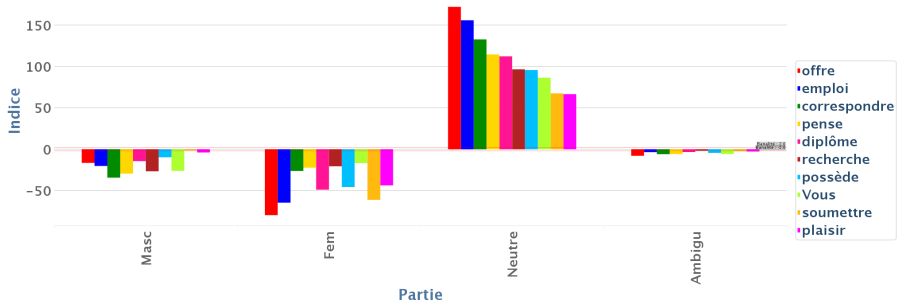


Spécificités lexicales du masculin (TXM)

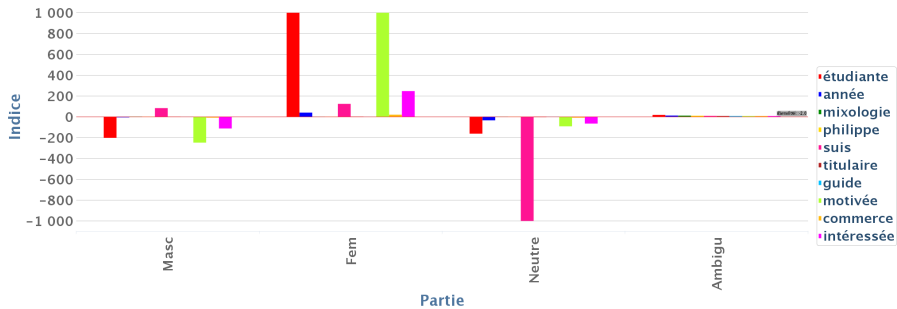


/!\ L'échelle de la figure est différente

Spécificités lexicales du neutre (TXM)



Spécificités lexicales de l'ambigu (TXM)



/!\ L'échelle de la figure est différente

Méthodes de débiaisage

