

Mesurer les inégalités de genre en ligne avec le genre grammatical : Une étude du subreddit r/france

Marie Flesch¹ Heather Burnett¹

(1) Laboratoire de Linguistique Formelle, Université Paris Cité, Paris
marie.flesch@u-paris.fr, heather.burnett@cnrs.fr

RÉSUMÉ

Cet article présente un système de détection du genre basé sur le genre grammatical, conçu pour le français, créé afin de mesurer les inégalités de genre dans les espaces francophones en ligne. Il décrit tout d'abord la création et le test du système, qui extrait le genre grammatical dans les expressions de type *je suis* depuis un lexique, sur un corpus étiqueté. Ensuite, il propose une étude de cas en deux parties, avec l'application du système sur un corpus de 11.8 millions de commentaires publiés sur r/france, le plus grand forum francophone de Reddit, suivie d'une étude des dynamiques de participation des femmes et des hommes dans cet espace. Cette recherche montre qu'un système de détection du genre simple, basé sur du pattern-matching, atteint une haute performance (précision de 96% dans le corpus test), et permet de dévoiler d'importantes inégalités de participation sur un forum francophone de premier plan.

ABSTRACT

Measuring the online gender gap with grammatical gender : A study of the r/france subreddit

This article presents a gender inference system based on grammatical gender, designed for French, created to measure gender inequalities in Francophone online spaces. It first describes the creation and testing of the system, which extracts the grammatical gender in *je suis* expressions from a lexicon, on a tagged corpus. Then, it proposes a two-part case study, with the application of the system on a corpus of 11.8 million comments posted on r/france, the largest Francophone subreddit, followed by a study of the dynamics of women's and men's participation in this space. This research shows that a simple gender inference system, based on pattern-matching, achieves high performance (96% precision in the test corpus), and reveals substantial inequalities in participation.

MOTS-CLÉS : genre grammatical ; détection automatique du genre ; inégalités de genre en ligne ; Reddit.

KEYWORDS: grammatical gender ; gender inference ; online gender gap ; Reddit.

ARTICLE : **Accepté à TALN.**

1 Introduction

La détection du genre, ou inférence du genre, est une tâche de classification populaire en TAL depuis une vingtaine d'années, avec deux grandes familles d'études : celles qui reposent sur des méthodes d'apprentissage automatique et celles qui font appel à l'apprentissage profond. Les travaux

basés sur l'apprentissage automatique reposent sur des méthodes variées, comme la classification naïve bayésienne (Mukherjee & Bala, 2017; Yan & Yan, 2006), la régression logistique (Argamon *et al.*, 2007) ou encore les machines à vecteurs de support (Ciot *et al.*, 2013; De Vel *et al.*, 2002; Ljubešić *et al.*, 2017; Rao *et al.*, 2010; van der Goot *et al.*, 2018). Les systèmes utilisent plusieurs types de représentations du texte : par exemple, les mots fréquents (Bamman *et al.*, 2014), les n-grammes (Liu & Cocea, 2018), les mots grammaticaux (Koppel *et al.*, 2004), les « caractéristiques sociolinguistiques » comme les émojis, les émoticônes les abréviations (Rao *et al.*, 2010; Yan & Yan, 2006) ou encore des données non-linguistiques comme les « likes » sur Facebook (Bi *et al.*, 2013). La performance des systèmes est généralement comprise entre 70 et 80% (O'Connor *et al.*, 2024) et semble plus élevée pour les langues possédant le genre grammatical (par exemple, 92% pour le slovène chez Verhoeven *et al.*, 2017 ou 87.62% pour l'espagnol chez Verhoeven *et al.*, 2016). Depuis 2017, les approches basées sur l'apprentissage profond sont de plus en plus utilisées, ce qui s'est accompagné d'une amélioration de la performance (O'Connor *et al.*, 2024). Pour les deux types d'approches, apprentissage automatique et deep learning, l'inférence de genre est très majoritairement une tâche binaire, qui classe les personnes en deux groupes : femmes et hommes (Cesare *et al.*, 2017; O'Connor *et al.*, 2024). Cette binarité est problématique, d'une part parce que les identités de genre des femmes et des hommes sont complexes et qu'il est sans doute tout simplement impossible de classer certaines personnes dans une catégorie à partir de leurs productions textuelles (Liu & Cocea, 2018), et de l'autre parce qu'elle invisibilise les personnes non-binaires (Lauscher *et al.*, 2022).

Même si la majorité des travaux sur la détection du genre porte sur l'anglais, quelques travaux ont été consacrés au français (Wang *et al.*, 2019; Ljubešić *et al.*, 2017; Hashempour *et al.*, 2024; van der Goot *et al.*, 2018). Le travail de Ciot *et al.* (2013) est particulièrement intéressant, parce qu'il compare une approche « traditionnelle » de la détection du genre (une machine à vecteurs de support utilisant comme features des caractéristiques lexicales et des métadonnées) et une approche déterministe beaucoup plus simple, qui extrait le genre grammatical dans les expressions *je suis* depuis un lexique de participes passés et d'adjectifs. Même si elle ne permet pas de détecter le genre de tous les auteur-es de leur corpus de tweets, cette dernière atteint une performance bien plus élevée (accuracy de 90%) que l'algorithme de classification (76%).

Cette approche ouvre des perspectives intéressantes pour le domaine des sciences sociales, où la détection du genre n'est pas une fin en soi (comme elle est souvent présentée dans les travaux de TAL), mais une étape nécessaire pour obtenir des données sociodémographiques sur les internautes, afin d'ensuite pouvoir répondre à diverses questions de recherche. Elle semble notamment pertinente dans l'étude des inégalités de genre en ligne, qui est souvent difficile à mener à cause de l'anonymat qui est la norme dans de nombreux espaces, comme les forums.

Cet article s'inspire de cette approche et la met en application pour explorer les dynamiques de participation en ligne sur r/france, le plus grand subreddit (forum de Reddit) francophone. Il fait deux apports aux domaines du TAL et des sciences sociales computationnelles. Tout d'abord, il décrit et met en pratique une méthode d'identification automatique du genre pour le français, basée sur le genre grammatical, plus sophistiquée que celle de Ciot *et al.* (2013), et qui a pour avantage d'être à la fois simple (elle ne nécessite ni ressources computationnelles importantes, ni de connaissances poussées de l'apprentissage automatique) et robuste (elle mégenre le moins de personnes possible). Ensuite, il propose ce qui est sans doute la première étude de corpus quantitative des inégalités de genre sur une plateforme francophone de premier plan, qui compte aujourd'hui 2.3 millions de membres. Notre travail s'articule en trois volets :

1. La description d'un système d'identification automatique du genre grammatical dans les

expressions *je+être*, créé et testé sur un corpus étiqueté avec le genre des auteur-es, composé de commentaires mis en ligne sur divers subreddits (corpus d'entraînement de 6.3 millions de mots, corpus test de 2.2 millions de mots), et qui atteint une précision de 96%.

2. **L'application de ce système à un corpus de 584 millions de mots**, contenant les commentaires et posts mis en ligne par 216 558 internautes sur r/france. Il attribue un genre à 21.75% des auteur-es, qui ont écrit 93.57% des contributions contenues dans le corpus. L'examen manuel d'un échantillon permet d'estimer une marge d'erreur (inférieure à 2%).
3. **Des analyses quantitatives de la participation des femmes, des personnes non-binaires et des hommes** basées sur des données de comptage. Elles révèlent des inégalités de participation plus importantes que celles relevées sur des subreddits anglophones : il y a près de sept fois plus d'hommes que de femmes dans le corpus r/france, et les femmes ont écrit seulement 6.94% de l'ensemble des commentaires.

2 Travaux connexes : inégalités de genre en ligne

En France, la fracture numérique semble très réduite : 94.4% des foyers ont accès à internet (INSEE, 2025). Quelle soit leur identité de genre, la plupart des Français-es sont aujourd'hui considérés-es comme « numériquement favorisé-es », par contraste avec des sociétés où les inégalités d'accès à internet entre femmes et hommes restent importantes, voire s'accroissent (Acilar & Sæbø, 2023). Toutefois, même dans les sociétés numériquement favorisées, il existe un important fossé entre femmes et hommes en termes de numératie (littéracie numérique) et de participation dans les espaces numériques : c'est ce que l'on appelle le « second niveau » de la fracture numérique (Van Deursen & Helsper, 2015). Ce fossé, décrit comme une véritable « ségrégation digitale » (Friesen *et al.*, 2021) est notamment visible dans les forums de discussion et sections de commentaires de sites de média et de réseaux sociaux, où plus d'hommes participent, et plus fréquemment, que de femmes (Bode, 2017; Brown *et al.*, 2018; Küchler *et al.*, 2023; Meyer & Carey, 2015; Pierson, 2015; Sun *et al.*, 2020).

Les inégalités de genre en ligne ont des conséquences dans la « vraie vie ». Les communautés en ligne sont des places publiques qui participent à la construction d'identités (Lutz *et al.*, 2014), mais aussi d'arguments et d'opinions (Babeau, 2014) qui peuvent ensuite être réinvestis dans des actions concrètes dans les sphères privées et publiques (par le vote, notamment). D'autres conséquences, ou « inégalités de troisième niveau » (Van Deursen & Helsper, 2015), sont les gains liés à la participation à des communautés en ligne, qui peuvent être d'ordre économique ou professionnel (par l'élargissement du réseau social ou l'acquisition d'expertise). Mesurer les dynamiques de participation sur diverses plateformes est donc crucial pour comprendre comment la participation des femmes est découragée et sanctionnée, mais aussi pour sensibiliser le public à l'existence de ces inégalités de genre en ligne.

Le problème, c'est que, dans les espaces où les internautes sont anonymes ou pseudonymes, et n'indiquent pas leur identité de genre sur leur profil, il est difficile de savoir qui est une femme ou un homme. C'est la raison pour laquelle de nombreuses études s'intéressant aux inégalités de genre en ligne s'appuient sur des données issues de questionnaires ou de sondages (Brown *et al.*, 2018; Meyer & Carey, 2015; Peacock & Duyn, 2021; Van Duyn *et al.*, 2021; Vochocová *et al.*, 2016; Lilleker *et al.*, 2023). D'autres études ayant pour but de décrire la composition sociodémographique de communautés en ligne s'appuient sur les prénoms parfois présents dans les noms d'utilisateur-ices pour inférer l'identité de genre des internautes (Cannon *et al.*, 2021; Jagfeld *et al.*, 2021; Küchler *et al.*, 2023; Pierson, 2015). Sur Reddit, cette approche a été employée mais elle n'est pas très productive, la

majorité des Redditeur·ices utilisant des pseudonymes sans prénom (Thelwall & Stuart, 2018).

Enfin, d’autres études utilisent des expressions régulières pour extraire des informations sur le genre, en recherchant des expressions du type *je suis une femme*. Fabian *et al.* (2015) ont ainsi détecté le genre de 25 465 personnes sur divers subreddits anglophones, dévoilant une proportion de 78.5% d’hommes et de 21.5% de femmes. Dans cet article, nous proposons une approche alternative, pour la langue française (adaptable à d’autres langues possédant le genre grammatical), et l’utilisons pour quantifier les inégalités de genre sur un grand subreddit francophone.

3 Le système de détection du genre : création et test sur un corpus étiqueté

3.1 Description

Notre système de détection du genre¹ extrait le genre grammatical grâce à un lexique dans deux types d’expressions *je + être*. Il a deux composantes. La première, que nous appelons **détection +1**, extrait le genre grammatical du mot (adjectif, nom, participe passé) qui vient directement après une expression *je+être* : *je suis fatigué, je suis enseignante, je suis partie*. La seconde, appelée **détection SN**, détecte le genre dans des syntagmes nominaux : *je suis un grand garçon, je suis une femme*. Le lexique utilisé pour extraire le genre grammatical a été créé sur mesure en combinant, pour les adjectifs et participes passés, le Lefff (Sagot, 2010) et le GLÀFF (Sajous *et al.*, 2013) et, pour les noms, un sous-ensemble de Flexique (Bonami *et al.*, 2013) disposant d’une annotation d’animéité « humaine » (Chlebowsky & Bonami, 2015), plusieurs listes issues du Wiktionnaire (membres de la famille, synonymes de *femme* et *homme*), et deux listes de noms et de fonctions (Cerquilligni *et al.*, 1999; Otto-Bruc, 2022). Les noms épicènes n’ont pas été inclus ; les noms ayant une forme unique au masculin et au féminin ont été annotés comme étant « common gender » (Corbett, 1991).

3.2 Le corpus d’entraînement et de test

3.2.1 Constitution du corpus

Le corpus étiqueté a été créé à partir d’un corpus de 1 milliard de mots composé de posts et commentaires mis en ligne sur 17 subreddits francophones (AskFrance, AskMec, AskMeuf, besoindeparler, conseiljuridique, Elles, france, FranceDetendue, FranceLibre, jeuxvideo, lgbtfrence, Lyon, ParentingFR, paris, questionsante, SexualiteFR et vosfinances). Plusieurs opérations de normalisation et de nettoyage du corpus ont été effectuées : normalisation des apostrophes et des guillemets, suppression des doublons, des URLs, et des éléments de syntaxe HTML. Les commentaires cités à l’intérieur d’autres commentaires (identifiés à l’aide de la syntaxe Markdown) ont également été supprimés.

1. dont nous fournissons le code et le lexique sur [OSF \(lien\)](#).

3.2.2 Étiquetage du corpus

L'étiquetage a été réalisé par l'extraction de « gender tags » (De Candia *et al.*, 2022), des sigles parfois utilisés par les Redditeur·ices pour fournir du contexte à leurs témoignages ou commentaires : *je suis (F23)* (femme de 23 ans) ; *ici H35* (homme de 25 ans), par exemple. Ces labels peuvent être considérés comme une source fiable d'informations sur l'identité de genre d'une personne - ce type d'autodescription a en effet été décrit comme étant le « gold standard for ascribing gender categories » en TAL (Larson, 2017) (p. 36). Une expression régulière qui prend en compte les différentes formes de ces sigles (initiale *H* ou *M* utilisée pour un homme, entre parenthèses, entre crochets ou non, âge précisé avant le genre ou inversement, etc.) a été utilisée pour les extraire du corpus. Les résultats ont ensuite été examinés manuellement par la première autrice de cet article pour supprimer les faux positifs. Cette technique a permis d'identifier le genre de 1329 Redditeur·ices, dont 498 femmes, 827 hommes et 4 personnes non-binaires.

3.2.3 Description du corpus étiqueté

Nous avons ensuite créé un sous-ensemble du corpus initial, contenant uniquement les contributions de ces 1329 personnes : il comprend 146 916 contributions (144 942 commentaires et 1964 posts). 77.87% de ces contributions ont été écrites par des hommes ($n=114\,396$) et 22.12% par des femmes ($n=32\,500$). Seuls 10 textes ont été écrits par des personnes non-binaires ; à cause de la petite taille de ce sous-corpus, nous avons décidé de ne pas les utiliser dans le système de détection du genre. Dans l'étude de cas (section 4), nous avons ajouté une composante non-binaire qui permet de ne pas invisibiliser les personnes non-binaires. Enfin, le corpus a été divisé en trois parties : un corpus d'entraînement (70% des auteur·es, utilisé pour créer le système), un corpus de développement (15% des auteur·es, utilisé pour affiner le système) et un corpus de test (15% des auteur·es, utilisé pour mesurer les performances du système).

3.3 La chaîne de traitement

Voici les différentes étapes réalisées pour détecter l'identité de genre des auteur·es du corpus :

- **Extraction des expressions *je+être* du corpus** (avec variantes orthographiques et à divers temps verbaux ; avec négations). L'examen manuel d'un échantillon a révélé que les négations ne posent pas de problème dans la détection +1, mais peuvent poser des problèmes dans la détection SN, ce dont nous tenons compte par la suite.
- **Normalisation orthographique d'une fenêtre de six mots** dans le contexte droit de ces expressions, réalisée semi-automatiquement avec le package Hunspell (Ooms, 2024).
- **Suppression des adverbes** suivant les expressions *je+être* ; pour la détection SN, nous n'enlevons pas les adverbes de négation.
- **Ajout automatique d'une annotation de genre issue du lexique** pour chaque participe passé et adjectif suivant les expressions *je+être* ; pour les noms, annotation uniquement pour les noms d'humain non épïcènes. Pour la détection SN, la règle suivante est appliquée : si un déterminant genré (*un, la*, etc.) est suivi par un nom d'humain non épïcène (éventuellement précédé d'un adjectif), alors le genre grammatical est celui du déterminant ; cela permet de détecter un genre pour *je suis une fan* et pas pour *je suis une belle personne*.
- **Inférence du genre de chaque auteur·e** à partir des résultats fournis par les deux composantes

	Femmes	Hommes	Précision
Femmes	56	3	0.95
Hommes	3	93	0.97
Genre inconnu	16	28	NA
Recall	0.75	0.75	

TABLE 1 – Matrice de confusion (il n’y a pas d’étiquette « genre inconnu » dans le corpus étiqueté mais le système classe des auteur-es dans cette catégorie

du système. Si le nombre d’expressions au féminin est supérieur à celui d’expression au masculin, l’auteure est considérée comme étant une femme, et vice versa. Si les deux totaux sont égaux, l’auteur-e est considéré-e comme étant de genre inconnu. Le genre des personnes n’ayant pas utilisé d’expression *je+être* genrée est aussi considéré comme inconnu.

3.4 Ajustements et résultats

Après vérification manuelle de plusieurs échantillons, le système a fait l’objet de plusieurs modifications, principalement des ajouts au lexique, ou la neutralisation d’unités phraséologiques produisant des faux positifs. Le test réalisé sur la portion du corpus non-explorée permet de détecter correctement le genre de 149 auteur-es (ou 74.87% de tous les auteur-es dans le corpus test), avec une précision de 0.95 pour les femmes et de 0.97 pour les hommes (Table 1).

4 Étude de cas : mise en pratique du système de détection du genre dans un grand corpus du forum r/france

4.1 Le corpus

Le jeu de données contient les posts et commentaires mis en ligne sur r/france de mars 2008 (date de création du subreddit) à décembre 2022. En plus des opérations de normalisation et nettoyage décrites plus haut, nous avons supprimé les commentaires mis en ligne par des bots, comme le bot « AutoModerator », qui effectue des actions de modération automatique, ou les bots qui mettent en ligne automatiquement du contenu extrait d’autres plateformes (Wikipédia, Twitter, etc.). Pour les détecter, nous avons adapté la méthodologie décrite dans (Atcheson *et al.*, 2024), basée sur des recherches de mots clés (*transcriber*, *automoderator*, *savevideo*, etc.) contenus dans les noms d’utilisateur-ices, ainsi que par des requêtes effectuées dans les commentaires (*je suis un bot*, *ce commentaire a été écrit par un bot*, etc.). Cette méthode a permis d’identifier 245 bots, et de supprimer du corpus les 40 787 commentaires et 4008 posts dont ils étaient les auteurs. Après ces opérations, le corpus contient 11 804 428 posts et commentaires (583 855 440 tokens) écrits par 216 558 auteur-es uniques.

4.2 Méthodologie

Nous appliquons à ce corpus les étapes décrites plus haut, en faisant des ajustements supplémentaires au lexique et aux règles utilisées. Afin d'éviter de mégenrer et d'invisibiliser les personnes non-binaires, nous ajoutons une composante non-binaire au système, basée sur un principe simple. Avant la mise en œuvre du système, nous recherchons des déclarations d'identité de genre non-binaire en utilisant une liste d'expressions (*je suis non-binaire*, *je suis genderqueer*, *je suis agenre*, etc.) établie à partir de plusieurs sources web (Wilfried, 2021; Wikipédia, 2024; Espineira, sd; Klutz & Wallis, sd; Rézo, sd; lgbtqia.fandom, sd). Après vérification manuelle de ces expressions dans leur contexte, les auteur-es s'identifiant comme non-binaires sont classé-es en tant que telles.

4.3 Résultats de la détection automatique du genre

Le système a détecté un genre grammatical féminin ou masculin dans 467 520 expressions je+être écrites par 47 567 auteur-es. 48 784 expressions ont un genre grammatical féminin, et 418 736 un genre grammatical masculin. Après addition des marqueurs pour chaque personne, le système identifie 5992 femmes, et 41 107 hommes, et ne détecte pas de genre pour 468 auteur-es (qui ont utilisé autant de marqueurs féminins que de marqueurs masculins). Les requêtes par mots clés effectuées dans le corpus pour détecter les personnes non-binaires fournissent 15 résultats, validés par inspection manuelle (pour éliminer les faux positifs liés au discours rapporté ou à l'ironie, par exemple). Après avoir combiné les résultats, nous avons pu détecter un genre pour 47 102 Redditeur-ices, ou 21.75% des 216 558 auteur-es du corpus. Au niveau du corpus, le système a détecté un genre pour les auteur-es de 11 045 971 contributions (posts et commentaires), soit 93.57% de toutes les contributions contenues dans le corpus. Cela représente un total de 555 228 964 tokens ou 95.01% des tokens du corpus initial.

L'asymétrie entre le nombre d'auteur-es (21.75%) dont le genre a été détecté et le nombre de contributions dont le genre de l'auteur-e a été détecté (93.57%) est liée à deux facteurs. Tout d'abord, le système de détection du genre fonctionne mieux pour les personnes qui ont écrit un nombre relativement important de posts/commentaires : le nombre médian de contributions par personne est de 25 ($EI = 98$) pour les Redditeur-ices pour lequel-les un genre a été détecté, et de 2 pour les personnes dont le genre n'a pas été détecté ($EI = 3$). Ensuite, la distribution de la participation dans les communautés en ligne est fortement asymétrique, avec un nombre réduit de personnes qui contribuent beaucoup, une caractéristique que nous explorons plus en détail ci-dessous (section 5.1).

4.4 Performance du système

La performance du système a été évaluée par l'examen manuel d'un échantillon aléatoire de 320 profils de Redditeur-ices, par la première autrice de cet article et un étudiant stagiaire. Plusieurs indices ont été pris en compte pour annoter cet échantillon : utilisation du genre grammatical dans les commentaires, présence éventuelle de prénoms dans les pseudonymes, texte descriptif sur la page de profil. Dans 84 cas, la vérification n'a pas pu être réalisée (profils supprimés ou vides, indices insuffisants). Pour les 236 cas restants, le système a généré 233 annotations correctes et trois erreurs, ou 98.73% d'annotations correctes. Ces trois erreurs d'annotation correspondent, selon les observations des annotateur-ices, à un homme (utilisant entre autres pour son nom pseudonyme un prénom masculin) qui a utilisé un marqueur féminin (*je suis fière*), et à deux personnes non-binaires

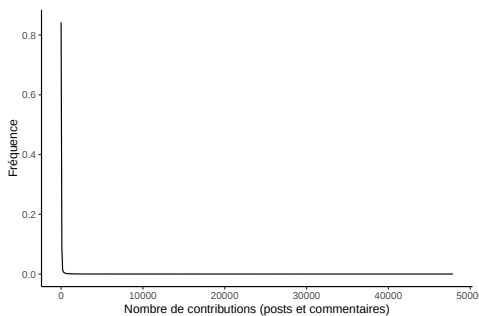


FIGURE 1 – Distribution de la participation

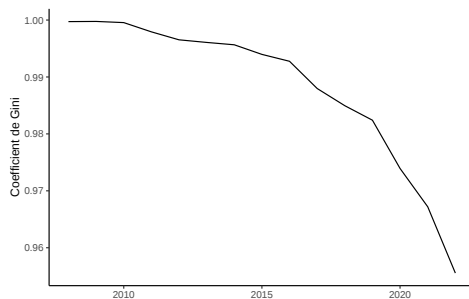


FIGURE 2 – Évolution du coefficient de Gini au fil du temps

annotées comme « hommes » par le système. La première avait utilisé exclusivement des marqueurs masculins, la second cinq marqueurs masculins et un féminin.²

5 Analyse des inégalités de participation sur r/france

5.1 Une participation fondamentalement asymétrique

Avant d’explorer les inégalités de genre, nous nous intéressons à la distribution de la participation sur r/france, en prenant en compte l’intégralité du corpus (y compris les personnes pour lesquelles le système n’a pas renvoyé d’identité de genre). La Figure 1 montre la distribution du nombre de contributions par personne. Celle-ci est fortement asymétrique, avec une longue queue à droite : une minorité de personnes a écrit la majorité des posts et commentaires du corpus. Le nombre médian de contributions par personne est de 2. Le neuvième décile (21 742 auteur-es, ou 10% des 216 558 auteur-es) a écrit 39 posts/commentaires ou plus ; l’ensemble de leurs contributions représente 92.4% du corpus (10 907 678 posts et commentaires). Les dynamiques de participation sur r/france sont donc très inégalitaires, avec un petit groupe de personnes très actives, ce qui a été observé dans d’autres communautés en ligne. Pour savoir si cette inégalité de participation est corrélée avec la croissance de la communauté (98 utilisateur-ices uniques en 2008, 79 718 en 2022), nous avons calculé le coefficient de Gini pour chaque année du corpus (Figure 2). Cette mesure statistique de la dispersion est généralement utilisée pour représenter les inégalités économiques ; un coefficient de 0 indique une égalité parfaite (dans notre cas, il représenterait le fait que toutes les personnes contribuent autant à la communauté), et un coefficient de 1 l’inégalité maximale (une seule personne a écrit tous les commentaires). Le coefficient baisse légèrement chaque année, passant de 0.99 en 2008 à 0.90 en 2022 ; il semble donc que, contrairement à ce que [Machado et al., 2025](#) ont observé sur Reddit, les inégalités de participation diminuent à mesure que le nombre de membres de r/france augmente.

2. Ce travail de vérification manuelle est très chronophage, car il nécessite une inspection détaillée des profils des internautes. Par conséquent, les deux annotateur-ices ont travaillé chacun-e sur une partie du jeu de données, et nous ne pouvons pas fournir un score d’accord inter-annotateur.

5.2 Proportion de femmes, d’hommes et de personnes non-binaires

La proportion d’hommes dans le corpus d’utilisateur-ices étiqueté est bien supérieur à celle des femmes : 87.25% (n=41 098) contre 12.71% (n=5989). Les personnes non-binaires sont très minoritaires 0.03% (n=15) (Table 2). La proportion de commentaires écrits par les hommes est de 93.89% ; les femmes ont écrit 6.07% des commentaires et les personnes non-binaires 0.07% des commentaires. Pour les posts, la proportion écrite par les femmes est plus faible : 4.87% contre 95.11% pour les hommes. Cela signifie que non seulement les femmes sont moins nombreuses à écrire des posts et commentaires, mais que lorsqu’elles participent, elles participent moins fréquemment que les hommes.

	femmes	hommes	personnes non-binaires	total
auteur-es	5989 (12.71%)	41 098 (87.25%)	15 (0.03%)	47 102 (100%)
posts	12 845 (4.87%)	250 725 (95.11%)	59 (0.02%)	263 629 (100%)
commentaires	654 252 (6.07%)	10 120 427 (93.86%)	7663 (0.07%)	10 782 342 (100%)
toutes contributions	667 097 (6.04%)	10 371 152 (93.89%)	7722 (0.07%)	1 1045 971 (100%)
tokens	38 251 874 (6.89%)	516 465 009 (93.02%)	512 081 (0.09%)	555 228 964 (100%)

TABLE 2 – Description du corpus étiqueté par le système de détection du genre

5.3 Analyses diachroniques

5.3.1 Proportion de la participation des femmes

La Figure 3 montre la proportion (en pourcentages) de femmes qui ont écrit des posts/commentaires, par an, ainsi que le nombre de posts/commentaires qu’elles ont écrits. La proportion de femmes varie avec le temps, avec un pic en 2010, où elles représentaient 12.59% des utilisateur-ices uniques sur le forum, une chute de 2014 à 2016, pour arriver en dessous des 8% en 2016, puis une remontée à partir de 2017 (11.80%). La proportion de commentaires écrits par les femmes a une évolution parallèle et atteint 7.02% en 2022. Le graphique met en lumière le fait que non seulement il y a beaucoup moins de femmes que d’hommes qui participent sur le forum, mais également que les femmes participent moins fréquemment.

5.3.2 Cycle de vie des Redditeur-ices

Par cycle de vie des Redditeur-ices, nous entendons la durée pendant laquelle une personne écrit des posts ou commentaires sur le site. Pour le calculer, nous avons identifié la date du premier post/commentaire mis en ligne par une personne, puis la date du dernier. Nous considérons qu’une personne cesse de participer activement si elle n’a pas écrit de commentaire pendant un an. Pour

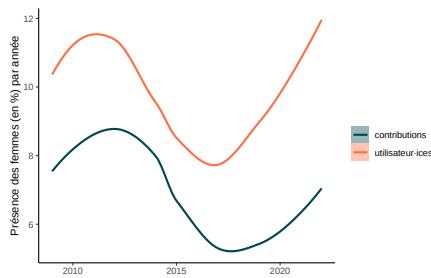


FIGURE 3 – Proportion de la participation des femmes, par an

chaque année du corpus (sauf la dernière, pour laquelle nous ne pouvons pas savoir si le dernier commentaire correspond à un arrêt de participation ou non), les Redditeur·ices sont placé·es dans quatre catégories (suivant [Danescu-Niculescu-Mizil et al. \(2013\)](#)) : 1. personne qui commence à participer (l'année où est posté son premier post/commentaire) ; 2. personne qui participe (écrit posts/commentaires) ; 3. personne qui arrête de participer (année du dernier post/commentaire) ; 4. personne qui commence et arrête de participer au cours de la même année.

La Figure 4 montre les évolutions de la participation sur le subreddit pour les hommes à droite, et pour les femmes à gauche. Elle montre le dynamisme du subreddit, dont le nombre de participant·es croît chaque année, et de façon assez parallèle pour les femmes et les hommes. Elle révèle également que la proportion de personnes qui commencent et arrêtent de participer au cours d'une même année est plus importante chez les femmes que chez les hommes. La durée médiane de participation active est effectivement plus longue chez les hommes : elle est de 202 jours ($EI=746$), contre 110 jours ($EI=495$) pour les femmes. Cette différence est significative (test de Wilcoxon-Mann-Whitney, $W=18850862$, $p<0.001$).

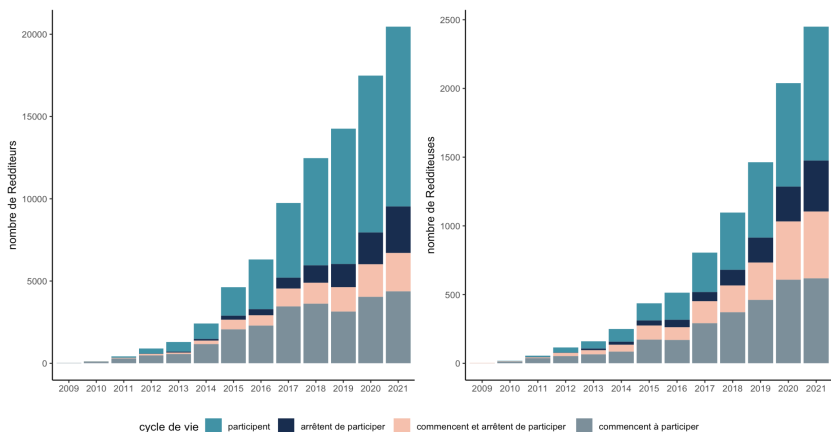


FIGURE 4 – Évolution de la communauté r/france (hommes à gauche, femmes à droite)

6 Conclusions

Cet article a montré qu'un système de détection du genre simple, basé sur du pattern-matching, est capable d'atteindre une précision élevée dans des corpus de commentaires publiés sur des forums. Bien qu'il ne permette pas d'identifier le genre de tous les auteur-es d'un corpus, il offre une couverture élevée au niveau du corpus. Cela est dû à la nature très asymétrique de la participation en ligne, un nombre limité de personnes étant auteures de la majorité des commentaires. La mise en pratique de ce système dans un corpus « écologique » (par opposition au corpus étiqueté utilisé dans la phase de création et de test) confirme l'intérêt de cette méthode, à la fois pour les domaines du TAL et des sciences sociales.

Notons, toutefois, que les résultats de l'annotation automatique du genre dans le corpus de *r/france* comportent une marge d'erreur qui, contrairement à celle du corpus étiqueté, ne peut être pleinement mesurée ; elle est liée à la fois à la variation orthographique, à la nature anonyme (ou pseudonyme) des interactions sur Reddit, et aux identités de genre en elles-mêmes. Il est notamment possible que certaines personnes « mentent » sur leur identité de genre, se faisant passer pour une femme dans un fil de discussion et pour un homme dans un autre, pour diverses raisons. De plus, certaines identités de genre sont fluides, et la méthode adoptée ne permet pas de prendre en compte cette fluidité (à l'exception des personnes déclarant explicitement être « genderfluid »).

Les importantes inégalités de participation entre femmes et hommes révélées par nos analyses ne sont pas surprenantes : Reddit a été décrit comme un site volontiers misogyne, dont le fonctionnement favorise l'émergence de « technocultures toxiques » (Massanari, 2017). Toutefois, l'étendue du *gender gap* sur *r/france*, avec seulement 12.71% de femmes dans notre corpus, semble particulièrement importante, comparée aux résultats de sondages (33% de femmes dans la population de Redditeur·ices, Barthel *et al.*, 2016) et d'autres études sur des subreddits anglophones (25% de femmes sur les 100 subreddits les plus populaires chez Thelwall & Stuart, 2018, 21.5% de femmes chez Fabian *et al.*, 2015). Cela est sans doute dû au fait que le subreddit est dominé par des sujets en lien avec la politique, un thème marqué par de fortes inégalités de participation, en ligne comme hors ligne (Lutz *et al.*, 2014). La façon dont le subreddit est modéré a peut-être également un effet, difficile à mesurer dans le cadre de cette étude, sur la participation des femmes ; c'est toutefois ce que suggère la politique de modération d'un subreddit plus jeune, *r/FranceDigeste*, qui se veut un « Espace inclusif et modéré » (semble-t-il par contraste avec *r/france*), qui fait « particulièrement attention à modérer les contenus oppressifs qui pourraient dissuader certaines personnes de participer » (FranceDigeste, 2025).

Notons également que notre étude ne prend en compte que la dimension la plus visible de la participation (l'écriture de posts et de commentaires). Une étude par questionnaire a montré que si les femmes écrivent moins que les hommes sur Reddit, elles fréquentent les forums autant que les hommes (Brown *et al.*, 2018) et « votent » (l'équivalent du « like » sur Reddit) également autant qu'eux, une participation qui, même si elle est silencieuse et impossible à mesurer, contribue à façonner les communautés. Enfin, les femmes ne sont sans doute pas la seule catégorie de la population sous-représentée sur *r/france* ; le forum semble fréquenté principalement par des personnes jeunes (Benzitoun & Flesch, 2024).

Remerciements

Les autrices souhaitent remercier Marie Candito et Guillaume Wisniewski pour leurs précieux conseils, Irvine Descout, et Tiffany Nguyen pour son travail sur les premières étapes de ce projet. Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet « ANR-24-AERC-0011-01 », et par l'ERC dans le cadre du programme de financement de la recherche et de l'innovation Horizon 2020 de l'Union européenne (accord de subvention n°850539).

Références

- ACILAR A. & SÆBØ O. (2023). Towards understanding the gender digital divide : A systematic literature review. *Global knowledge, memory and communication*, **72**(3), 233–249. DOI : <https://doi.org/10.1108/GKMC-09-2021-0147>.
- ARGAMON S., KOPPEL M., PENNEBAKER J. W. & SCHLER J. (2007). Mining the blogosphere : Age, gender and the varieties of self-expression. *First Monday*. DOI : <https://doi.org/10.5210/fm.v12i9.2003>.
- ATCHESON A., KOSHY V. & KARAHALIOS K. (2024). Not what it used to be : Characterizing content and user-base evolution in newly created online communities. In *CHI '24 : Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. DOI : <https://doi.org/10.1145/3613904.3642769>.
- BABEAU F. (2014). La participation politique des citoyens « ordinaires » sur l'internet. la plateforme youtube comme lieu d'observation. *Politiques de communication*, **3**(2), 125–150. DOI : <https://doi.org/10.3917/pdc.003.0125>.
- BAMMAN D., EISENSTEIN J. & SCHNOEBELEN T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, **18**(2), 135–160. DOI : <https://doi.org/10.1111/josl.12080>.
- BARTHEL M., STOCKING G., HOLCOMB J. & MITCHELL A. (2016). Reddit news users more likely to be male, young and digital in their news preferences.
- BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éd.s. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.
- BENZITOUN C. & FLESCHE M. (2024). La disparition qu'on a mis au jour : étude sur corpus de l'accord du participe passé avec avoir chez des adultes francophones. In *SHS Web of Conferences*, volume 191 : EDP Sciences. DOI : <https://doi.org/10.1051/shsconf/202419114004>.
- BI B., SHOKOUHI M., KOSINSKI M. & GRAEPEL T. (2013). Inferring the demographics of search users : Social data meets search queries. In *Proceedings of the 22nd international conference on world wide web*, p. 131–140. DOI : <https://doi.org/10.1145/2488388.2488401>.
- BODE L. (2017). Closing the gap : Gender parity in political engagement on social media. **20**(4), 587–603. DOI : [10.1080/1369118X.2016.1202302](https://doi.org/10.1080/1369118X.2016.1202302).
- BONAMI O., CARON G. & PLANCQ C. (2013). Flexique : An inflectional lexicon for spoken french. In *Actes du quatrième Congrès Mondial de Linguistique Française*, p. 2583-2596.
- BROWN D. K., NG Y. M. M., RIEDL M. J. & LACASA-MAS I. (2018). Reddit's veil of anonymity : Predictors of engagement and participation in media environments with hostile reputations. **4**(4), 2056305118810216. DOI : [10.1177/2056305118810216](https://doi.org/10.1177/2056305118810216).
- CANNON E., CROUSE B., GHOSH S., RIHN N. & CHUA K. (2021). “ don't downvote a\$\$\$\$\$s!! ” : An exploration of reddit's advice communities. In *Proceedings of the 55th Ha-*

waii International Conference on System Sciences : arXiv preprint arXiv :2109.09044. DOI : <https://doi.org/10.48550/arXiv.2109.09044>.

CERQUILIGNI, BECQUER A.-M., CHOLEWKA N., COUTIER M., MATHIEU M.-J. & FRÉCHER J. (1999). *Femme, j'écris ton nom... Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. Rapport interne, CNRS - Institut National de la Langue Française.

CESARE N., GRANT C. & NSOESIE E. O. (2017). Detection of user demographics on social media : A review of methods and recommendations for best practices. *arXiv preprint arXiv :1702.01807*, p. 1–25.

CHLEBOWSKI A. & BONAMI O. (2015). *Annotation sémantique des noms de Flexique*. [rapport de stage de master, Laboratoire de Linguistique Formelle.

CIOT M., SONDEREGGER M. & RUTHS D. (2013). Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 1136–1145.

CORBETT G. G. (1991). *Gender*. Cambridge University Press.

DANESCU-NICULESCU-MIZIL C., WEST R., JURAFSKY D., LESKOVEC J. & POTTS C. (2013). No country for old members : User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, p. 307–318 : ACM.

DE CANDIA S., DE FRANCISCI MORALES G., MONTI C. & BONCHI F. (2022). Social norms on reddit : A demographic analysis. In *Proceedings of the 14th ACM web science conference 2022*, p. 139–147. DOI : <https://doi.org/10.1145/3501247.3531549>.

DE VEL O., CORNEY M., ANDERSON A. & MOHAY G. (2002). Language and gender author cohort analysis of e-mail for computer forensics. In *Proceedings of digital forensics research workshop*, p. 1–16.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

ESPINEIRA K. S. (s.d.). Les genres non-binaires sur internet et facebook.

FABIAN B., BAUMANN A. & KEIL M. (2015). Privacy on reddit? towards large-scale user classification. In *Proceedings of ECIS'15*. DOI : <https://doi.org/10.18151/7217310>.

FRANCEDIGESTE (2025). Espace inclusif et modéré.

FRIESEN A., BURGE R. & BRITZMAN K. (2021). Digital segregation : gender, occupation, and access to politics. *Social Science Computer Review*, **39**(1), 38–55. DOI : [10.1177/0894439319851164](https://doi.org/10.1177/0894439319851164).

HASHEMPOUR R., PLANK B., VILLAVICENCIO A. & DE AMORIM R. C. (2024). A deep learning approach to language-independent gender prediction on twitter. *arXiv preprint arXiv :2411.19733*. DOI : <https://doi.org/10.48550/arXiv.2411.19733>.

INSEE (2025). Accès et utilisation de l'internet dans l'union européenne | insee.

JAGFELD G., LOBBAN F., RAYSON P. & JONES S. H. (2021). Understanding who uses reddit : Profiling individuals with a self-reported bipolar disorder diagnosis. In *The Seventh Workshop on Computational Linguistics and Clinical Psychology : Improving Access @NAACL 2021*. DOI : <https://doi.org/10.48550/arXiv.2104.11612>.

KLUTZ B. & WALLIS E. (s.d.). Lexique de la communauté lgbt.

KOPPEL M., ARGAMON S. & SHIMONI A. R. (2004). Automatically categorizing written texts by author gender. *Computing Reviews*, **45**(1), 43. DOI : <https://doi.org/10.1093/lilc/17.4.401>.

KÜCHLER C., STOLL A., ZIEGELE M. & NAAB T. K. (2023). Gender-related differences in online comment sections : findings from a large-scale content analysis of commenting behavior. **41**(3), 728–747. DOI : [10.1177/08944393211052042](https://doi.org/10.1177/08944393211052042).

- LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éd.s., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.
- LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.
- LARSON B. (2017). Gender as a variable in natural-language processing : ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, p. 1–11, Valencia, Spain : Association for Computational Linguistics. DOI : <https://doi.org/10.18653/v1/W17-1601>.
- LAUSCHER A., CROWLEY A. & HOVY D. (2022). Welcome to the modern world of pronouns : Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv :2202.11923*. DOI : <https://doi.org/10.48550/arXiv.2202.11923>.
- LGBTQIA.FANDOM (s.d.). Wiki lgbtqia fr.
- LILLEKER D., KOC-MICHALSKA K. & BIMBER B. (2023). *Women learn while men talk ? : Revisiting gender differences in political engagement in online environments*, In A. SCHIFFRIN, K. KOC-MICHALSKA & M. FERRIER, Éd.s., *Women in the digital world*, p. 46–62. Routledge.
- LIU H. & COCEA M. (2018). Fuzzy rule based systems for gender classification from blog data. In *2018 tenth international conference on advanced computational intelligence (ICACI)*, p. 79–84 : IEEE. DOI : <https://doi.org/10.1109/ICACI.2018.8377585>.
- LJUBEŠIĆ N., FIŠER D. & ERJAVEC T. (2017). Language-independent gender prediction on twitter. In D. HOVY, S. VOLKOVA, D. BAMMAN, D. JURGENS, B. O'CONNOR, O. TSUR & A. S. DOĞRUÖZ, Éd.s., *Proceedings of the second workshop on NLP and computational social science*, p. 1–6, Vancouver, Canada : Association for Computational Linguistics. DOI : <https://doi.org/10.18653/v1/W17-2901>.
- LUTZ C., HOFFMANN C. P. & MECKEL M. (2014). Beyond just politics : A systematic literature review of online participation. *First Monday*, **19**(7), 1–36. DOI : <https://doi.org/10.5210/fm.v19i7.5260>.
- MACHADO G., PACHECO D., MENEZES R. & BAXTER G. (2025). Super-linear growth and rising inequality in online social communities : Insights from reddit. *arXiv preprint arXiv :2503.02661*. DOI : <https://doi.org/10.48550/arXiv.2503.02661>.
- MASSANARI A. (2017). #Gamergate and The Fappening : How Reddit's algorithm, governance, and culture support toxic technocultures. **19**(3), 329–346. DOI : [10.1177/14614448156088](https://doi.org/10.1177/14614448156088).
- MEYER H. K. & CAREY M. C. (2015). Men more likely to post online newspaper comments. **36**(4), 469–481. DOI : [10.1177/0739532915618417](https://doi.org/10.1177/0739532915618417).
- MUKHERJEE S. & BALA P. K. (2017). Gender classification of microblog text based on authorial style. *Information Systems and e-Business Management*, **15**(1), 117–138. DOI : <https://doi.org/10.1007/s10257-016-0312-0>.
- OOMS J. (2024). *hunspell : High-performance stemmer, tokenizer, and spell checker*.
- OTTO-BRUC C. (2022). Les noms de métiers, titres honorifiques et grades au féminin et au masculin.
- O'CONNOR K., GOLDER S., WEISSENBACHER D., KLEIN A. Z., MAGGE A. & GONZALEZ-HERNANDEZ G. (2024). Methods and annotated data sets used to predict the gender and age of twitter users : Scoping review. *Journal of Medical Internet Research*, **26**, e47923. DOI : <https://doi.org/10.2196/47923>.
- PEACOCK C. & DUYN E. V. (2021). Monitoring and correcting : why women read and men comment online. *Information, Communication & Society*, p. 1–16. DOI : <https://doi.org/10.1080/1369118X.2021.1993957>.

- PIERSON E. (2015). Outnumbered but well-spoken : Female commenters in the New York Times. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*, p. 1201–1213. DOI : [10.1145/2675133.2675134Rosenberg](https://doi.org/10.1145/2675133.2675134Rosenberg).
- RAO D., YAROWSKY D., SHREEVATS A. & GUPTA M. (2010). Classifying latent user attributes in twitter. In *2nd International Workshop on Search and Mining UserGenerated Content. ACM*. DOI : <https://doi.org/10.1145/1871985.1871993>.
- RÉZO (s.d.). Lexique lgbtq+.
- SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on language resources and evaluation (LREC 2010)*.
- SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GlÀff, un gros lexique À tout faire du français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.
- SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.
- SUN B., MAO H. & YIN C. (2020). Male and female users' differences in online technology community based on text mining. **11**, 806. DOI : [10.3389/fpsyg.2020.00806](https://doi.org/10.3389/fpsyg.2020.00806).
- THELWALL M. & STUART E. (2018). She's reddit : A source of statistically significant gendered interest information ? *Information processing & management*, **56**(4), 1543–1558. arXiv : 1810.08091, DOI : <https://doi.org/10.1016/j.ipm.2018.10.007>.
- VAN DER GOOT R., LJUBEŠIĆ N., MATROOS I., NISSIM M. & PLANK B. (2018). Bleaching text : Abstract features for cross-lingual gender prediction. In I. GUREVYCH & Y. MIYAO, Éd., *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2 : Short papers)*, p. 383–389, Melbourne, Australia : Association for Computational Linguistics. DOI : <https://doi.org/10.18653/v1/P18-2061>.
- VAN DEURSEN A. J. & HELSPER E. J. (2015). The third-level digital divide : Who benefits most from being online ? *Communication and information technologies annual*, **10**, 29–52. DOI : <https://doi.org/10.1108/S2050-206020150000010002>.
- VAN DUYN E., PEACOCK C. & STROUD N. J. (2021). The gender gap in on-line news comment sections. *Social Science Computer Review*, **39**(2), 181–196. DOI : <https://doi.org/10.1177/0894439319864876>.
- VERHOEVEN B., DAELEMANS W. & PLANK B. (2016). Twisty : a multilingual twitter stylometry corpus for gender and personality profiling. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éd., *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, p. 1632–1637, Portorož, Slovenia : European Language Resources Association (ELRA).
- VERHOEVEN B., ŠKRJANEC I. & POLLAK S. (2017). Gender profiling for slovene twitter communication : The influence of gender marking, content and style. In *Proceedings of the 6th workshop on balto-slavic natural language processing*, p. 119–125. DOI : <https://doi.org/10.18653/v1/W17-1418>.
- VOCHOCOVÁ L., ŠTĚTKA V. & MAZÁK J. (2016). Good girls don't comment on politics ? gendered character of online political participation in the czech republic. *Information, Communication & Society*, **19**(10), 1321–1339. DOI : <https://doi.org/10.1080/1369118X.2015.1088881>.
- WANG Z., HALE S., ADELANI D. I., GRABOWICZ P., HARTMAN T., FLÖCK F. & JURGENS D. (2019). Demographic inference and representative population estimates from mul-

tilingual social media data. In *The world wide web conference*, p. 2056–2067. DOI : <https://doi.org/10.1145/3308558.3313684>.

WIKIPÉDIA (2024). Non-binarité. Page Version ID : 219993262.

WILFRIED C. (2021). Qu'est-ce qu'une personne non binaire, et la non binarité ?

YAN X. & YAN L. (2006). Gender classification of weblog authors. In *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, p. 228–230.