Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Romanciers et romancières du XIX^{e} siècle : une étude automatique du genre sur le corpus GIRLS

Marco Naguib Marine Delaborde Blandine Andrault Anaïs Bekolo
Olga Seminck
Laboratoire Langues Textes Traitements Informatiques Cognition (Lattice), UMR 8094
olga.seminck@cnrs.fr

RÉSUMÉ

Cette étude porte sur les différences entre les romans français du XIX^{e} siècle écrits par des hommes et ceux écrits par des femmes en trois étapes. Premièrement, nous observons que ces textes peuvent être distingués par apprentissage supervisé selon ce critère. Un modèle simple a un score de 99% d'exactitude sur cette tâche si d'autres œuvres de la même personne figurent dans le jeu d'entraînement, et de 72% d'exactitude sinon. Cette différence s'explique par le fait que le langage de l'individu est plus distinctif qu'un éventuel style propre au genre. Deuxièmement, notre étude textométrique met au jour des stéréotypes de genre chez les hommes et les femmes. Troisièmement, nous présentons un modèle de coréférence entraîné sur des textes littéraires pour étudier le genre des personnages. Nous montrons ainsi que les personnages féminins sont plus nombreux chez les femmes, et prennent généralement une place plus proéminente que chez les hommes.

ABSTRACT

Male and female novelists : an automatic study of gender of authors and their characters

This study focuses on the differences between French novels of the 19th century written by men and those written by women in three steps. First, we observe that these texts can be distinguished by supervised learning according to this criterion. A simple model achieves 99% accuracy if other novels by the same person appear in the training set, and 72% accuracy otherwise. This gap is explained by the fact that the language of the individual is more distinctive than a possible gender-specific style. Second, our textometric study reveals gender stereotypes in the subcorpus of men and women. Third, we present a coreference model trained on literary texts to study the gender of characters. We show that female characters are more numerous among female authors, and generally take a more prominent place than among men authors.

MOTS-CLÉS : Genre, littérature, apprentissage automatique, modèle tranformer, coreference, textometrie, calcul de spécificités.

KEYWORDS: Gender, literature, machine learning, transformer models, coreference, textometry, specifities of the vocabulary.

1 Introduction

Il est possible d'identifier automatiquement le genre [1] de la personne qui écrit un texte. Par exemple, Koppel *et al.* (2002) obtiennent une exactitude moyenne de 74 % pour cette tâche sur le British National Corpus (Consortium *et al.*, 2007). Verhoeven *et al.* (2017) ont un score de 93 % d'exactitude sur un corpus de tweets slovènes. Avec un corpus d'emails en anglais, Safara *et al.* (2020) réalisent 98 % d'exactitude. Et pour le russe, Sboev *et al.* (2016) obtiennent 86% d'exactitude en obtenant des traits morphologiques et syntaxiques indépendants des thématiques des textes. Pour d'autres exemples de travaux sur ce sujet, voyez la table 1 de Safara *et al.* (2020) pour un état de l'art plus complet et détaillé.

Notre étude porte sur les romans du XIX^{e} siècle en français. Dans un premier temps, nous voulons vérifier s'il est possible d'apprendre automatiquement la différence entre des écrits des femmes et des hommes dans ces romans. Pour nos expériences dans cette matière, nous utilisons *the Corpus for Idiolectal Research* (CIDRE) (Seminck *et al.*, 2021), un corpus de romans de 37 millions de mots du XIX^{e} siècle écrits par 4 femmes et 7 hommes. De plus, nous constituons un nouveau corpus paritaire *the Gender Identification Resource for Literature and Style* (GIRLS) (Lattice, 2022). Ce dernier corpus ne comporte qu'un seul livre par personne, ce qui permet de voir l'influence de l'idiolecte (le langage de l'individu) sur la réussite des modèles appris. Nous démontrons en effet que celle-ci ne peut être ignorée bien qu'il s'avère que le genre (masculin ou féminin) soit également un facteur pertinent.

Après avoir établi que le genre peut être distingué automatiquement, nous nous intéressons dans un deuxième temps à la question des différences entre nos deux groupes. Nous présenterons une étude du vocabulaire spécifique à chaque partie, qui nous amène à nous interroger sur le genre des personnages évoqués par les hommes et les femmes dans leurs romans. Nous émettons l'hypothèse que le corpus GIRLS, de par son genre littéraire et son époque d'écriture, favorise l'émergence des stéréotypes de genre tels qu'évoqués par Argamon *et al.* (2009). Underwood *et al.* (2018) ont trouvé que les différences de vocabulaire lié aux hommes et aux femmes s'atténuent avec le temps entre le XIX^{e} siècle et le XX^{e}e siècle sur un corpus de littérature anglaise.

Pour étudier la représentation des hommes et des femmes dans GIRLS, nous développons et utilisons un modèle de coréférence basé sur une architecture *Transformer* et adapté aux œuvres littéraires. Ce modèle permet d'y identifier les chaînes de coréférence, et ainsi, les personnages distincts. Nous développons ensuite une méthode qui permet de déterminer le genre de chaque personnage ainsi que son nombre de mentions. Nos résultats montrent des phénomènes qui ont été observés pour la littérature anglophone (Underwood *et al.*, 2018; Nagaraj & Kejriwal, 2022) : il y a plus de personnages féminins dans les livres des romancières que dans ceux de romanciers.

	Nombre de livres	Nombre de de tokens			
		Min	Médiane	Max	Total
Femmes	32	12 279	50 588,5	274 268	1 779 214
Hommes	32	20 028	72 631	451 755	2 822 293

TABLE 1 – Comptage des tokens du corpus

1. Il faut remarquer que pour tous les travaux, les nôtres inclus, la représentation du genre est limitée du fait qu'on ne distingue que des hommes et des femmes, ce qui est discutable étant donné l'existence d'autres genres ou la fluidité du genre (Land, 2020). Cependant, il faut comprendre que c'est le résultat de manque de données sur d'autres genres (Verhoeven & Daelemans, 2019), surtout pour des données anciennes comme nos corpus du XIX^{e} siècle.

2 Corpus GIRLS

Notre corpus de travail est constitué de 64 romans français du XIX^{e} siècle, dont la moitié a été écrite par des femmes et l'autre par des hommes. Il y a un seul livre par personne. Ces textes sont dans le domaine public et ont été récupérés de manière automatique sur les sites du projet Gutenberg, Gallica, Wikisource, Google Books et ebooksgratuits en utilisant les scripts fournis avec le corpus CIDRE (Seminck *et al.*, 2021) qui servent à convertir les fichiers epub en TEI et ensuite en format texte brut tout en coupant le texte liminaire. Seul le titre de l'ouvrage apparaît au début du texte. Les métadonnées concernant le nombre de tokens sont reportées dans la table 1. Notre corpus est disponible dans le répertoire suivant : `https://www.ortolang.fr/market/corpora/girls` .

3 Détection du genre par apprentissage automatique

Notre première expérience consiste à contrôler si, comme dans les études citées dans l'introduction de cet article, un modèle simple arrive à distinguer les écrits des hommes et des femmes. Nous utilisons pour cela le corpus CIDRE (Seminck *et al.*, 2021) qui contient 421 romans écrits par 4 femmes et 7 hommes (respectivement 151 livres et 270 livres). Nous avons utilisé un classifieur de régression logistique de la bibliothèque `sklearn` (Pedregosa *et al.*, 2011) en python entraîné sur des n-grammes de tokens de taille 1 à 3 avec une occurrence minimale de 10. [2] Nous évaluons les résultats avec une validation croisée à 8 blocs et obtenons un score d'exactitude (*accuracy*) de 99 %.

Il nous semble que cette tâche est simple à réaliser sur ce corpus. Cependant, nous nous interrogeons sur l'influence de l'idiolecte, le langage de l'individu ou style personnel d'écriture, sur les résultats obtenus. En effet, l'idiolecte est souvent considéré comme étant aussi personnel qu'une empreinte digitale (Eder, 2011). Pour vérifier cela, nous entraînons un classifieur multiclasse de régression logistique [3] à reconnaître l'auteur (parmi les 11 romancières et romanciers présents dans le corpus) d'un texte donné. Nous obtenons une exactitude de 94 % sur cette tâche. Il semble donc que le style personnel est un signal très fort. Selon nous, la tâche est plus simple lorsque la même personne a écrit des textes qui se retrouvent à la fois dans le jeu d'entraînement et le jeu de test. Le modèle reconnaîtrait alors l'idiolecte plutôt que le genre.

Le corpus GIRLS ne contient qu'un seul texte par romancière ou romancier. Le modèle ne peut donc pas utiliser l'idiolecte pour apprendre la distinction du genre. Nous avons de nouveau entraîné un classifieur de régression logistique avec les mêmes techniques et paramètres mentionnés ci-dessus. Nous obtenons une exactitude de 72 %. L'écart avec les scores obtenus sur CIDRE s'explique en partie par la différence de taille : le corpus GIRLS est huit fois plus petit. Nous sommes cependant convaincus que ce résultat montre que le signal de l'idiolecte doit être neutralisé pour pouvoir étudier correctement celui du genre. De plus, ce score paraît correspondre à celui trouvé par Underwood *et al.* (2018) qui obtenaient 76 % d'exactitude quand ils cherchaient à prédire le genre de l'auteur à partir de la description des personnages dans un roman.

2. Tous les scripts utilisés pour les expériences décrites dans cette section se trouvent sur `https://github.com/oseminck/scripts_article_genre_TALN2022/tree/main/scripts_section3_apprentissage_auteurs_et_genre` .

3. Avec les mêmes paramètres et outils que dans l'expérience précédente.

4 Explorations textométriques : spécificités liées aux genres

Pour exploiter nos données et effectuer des explorations textométriques avec le logiciel Itrameur (Fleury, 2008), nous avons annoté notre corpus en lemmes et parties du discours à l'aide de la chaîne de traitements UDPipe (Straka *et al.*, 2016). L'indice de spécificité (SP) de Lafon (1984) permet de mettre au jour le vocabulaire spécifique à une partie du corpus. De nombreux noms propres, qui sont des noms de personnages, sont présents dans le vocabulaire spécifique à la partie. Il semble y avoir beaucoup de prénoms féminins et de noms de famille. On peut se demander si les personnages féminins sont plutôt appelés par leur prénom et les personnages masculins par leur nom.

La figure 2, en annexe, présente les lemmes surreprésentés dans chaque partie de notre corpus. Nous avons supprimé les noms propres de ces résultats pour plus de visibilité concernant le vocabulaire spécifique à chaque partie. Nous pouvons observer une différence de champ lexical entre les deux parties. Les mots spécifiques aux romans écrits par des hommes sont plutôt liés au champ lexical de la guerre (guerre, guerrier, soldat, cadavre) alors que les mots spécifiques aux romans écrits par des femmes sont plutôt liés au champ lexical des émotions et des sentiments (sentiment, bonheur, éprouver, aimer, affection, âme, douleur, heureux, cœur, malheureux), mais aussi de la famille (mère, cousin, enfant). Cela semble représenter les mêmes stéréotypes qui ont été détectés dans la littérature anglaise. Ainsi, Underwood *et al.* (2018) trouvaient que *felt* était surreprésenté chez les femmes. Ce résultat est également en ligne avec les stéréotypes des personnages du théâtre français du XVI^{e} au XIX^{e} siècle. Benamar *et al.* (2022) trouvaient que les personnages féminins sont entre autres fortement associés à des termes du champ lexical de la famille (*e.g.* belle-mère, fille, veuve, belle-sœur). On remarque aussi plus de titres de noblesse (baron, Monseigneur, grand-duc, prince) et de fonctions (gendarme, pape, roi, ministre, préfet, prêtre) chez les hommes, alors que les femmes privilégient les titres de civilité courants (Madame, Monsieur, Mademoiselle). Le lemme « il » est surreprésenté dans la partie « Femmes » (SP = 102) par rapport à la partie « Hommes ». Or, selon l'annotation obtenue par UDPipe, le lemme de la forme « elle » est « il ».

Étiquettes	FQ	Fq H	SP H	Fq F	SP F
G=Masc | N=Sing	446 265	280 316	**70**	165 949	-69
G=Fem | N=Sing	388 803	235 059	-54	153 744	**54**
G=Masc | N=Plur	156 011	103 625	**$9.0*10^{15}$**	52 386	$-9.0*10^{15}$
G=Fem | N=Plur	114 007	74 438	**151**	39 569	-151

TABLE 2 – Spécificités des étiquettes morphologiques dans chaque partie (FQ = Nombre d'occurrences total ; Fq = Nombre d'occurrences sur la partie « Hommes » ou « Femmes » ; SP = Indice de spécificité)

Pour effectuer une analyse linguistique plus fine, il est nécessaire de regarder les formes spécifiques à chaque partie. Dans la partie des romans écrits par des hommes, la forme « Il » est surreprésentée (SP = 282). À l'inverse, dans la partie des romans écrits par des femmes, la forme « elle » (SP = 233) est surreprésentée. Cela nous conduit à nous interroger sur le genre grammatical des mots utilisés dans chaque partie.

Le tableau 2 présente les spécificités des étiquettes de genre et de nombre les plus fréquentes des items du corpus. Les étiquettes des mots féminins singuliers sont spécifiques à la partie des romans écrits par des femmes. Les étiquettes des mots masculins (singuliers et pluriels) sont quant à elles spécifiques à la partie des romans écrits par des hommes, mais c'est aussi le cas des étiquettes des

mots féminins pluriels. Cela peut paraître surprenant, mais l'étude de ces mots en contexte nous rappelle qu'un mot dont le genre grammatical est féminin ne désigne pas nécessairement un référent féminin, par exemple « une personne », « une hyène » ou « la foule ».

Ces observations nous amènent à nous demander si les romancières de notre corpus ont plutôt tendance à mentionner des personnages féminins et si les romanciers ont plutôt tendance, quant à eux, à mentionner des personnages masculins. Pour vérifier cela, les chaînes de coréférence mentionnant des personnages humains peuvent être utiles.

5 Étude de genre des personnages par modèle de résolution de coréférence adapté à la littérature

Dans cette partie, nous cherchons à examiner si un lien apparent existe entre le genre d'un romancier ou d'une romancière et le genre de leurs personnages. Pour cela, nous procédons à la résolution de la coréférence dans GIRLS. Puis, nous sélectionnons des chaînes référant à des personnes. Enfin, nous étudions le nombre de chaînes de coréférence selon le genre du personnage grâce à une méthode basée sur des heuristiques.

5.1 Modèle de coréférence

Notre modèle est téléchargeable sur https://www.ortolang.fr/market/tools/modele-coref-fr-litb . Il a été entraîné sur le corpus Fr-LitBank.[4] Il s'agit des 10000 premiers mots de 15 romans des XIX^{e} et XX^{e} siècles, annotés en mentions et en chaînes de coréférence. Nous appelons mention une expression référentielle (groupe nominal, nom propre ou pronom) faisant référence à un personnage (PER), un lieu (LOC), une installation (FAC)... selon les définitions données par Bamman *et al.* (2019). Une chaîne de coréférence, annotée suivant la définition dans Democrat (Landragin, 2021), associe entre elles les mentions référant à la même entité, par exemple, au même personnage. Chaque mention est donc annotée par un identifiant indiquant l'entité à laquelle elle réfère.

5.1.1 Architecture

Nous utilisons un modèle CamemBERT pré-entraîné (Martin *et al.*, 2020) et procédons à un *fine-tuning* de ses paramètres pour reconnaître les mentions et résoudre la coréférence au sein d'un *chunk* (fenêtre glissante) de $n = 256$ tokens[5].

Pour pouvoir détecter des mentions imbriquées, nous utilisons un schéma d'étiquetage *BIOES* (**B**eginning, **I**nside, **O**utside, **E**nding, **S**ingle-word). L'étiquette attribuée à chaque mot dépend donc du type de la mention, mais également de la position du mot dans celle-ci. L'ensemble des étiquettes possibles est $M = \{$O, B-PER, I-PER, S-PER, E-PER, B-LOC, ...$\}$.

Concrètement, pour chaque token w_i, le modèle est entraîné à prédire deux étiquettes :

4. https://github.com/lattice-8094/fr-litbank

5. CamemBERT, comme la plupart des modèles basés sur une architecture *Transformer*, a une complexité quadratique en fonction de la longueur de séquence d'entrée et est donc plus adapté à des entrées de taille fixe relativement petite.

— $m_i \in M$ correspondant à la mention la plus courte contenant le mot, et à sa position dans celle-ci
— $r_i \in \{0, 1, ...n\}$ indiquant l'indice du premier mot qui coréfère avec w_i dans la fenêtre glissante, si un tel mot existe, et 0 sinon.

Pour cela, nous entraînons le modèle à attribuer à w_i trois représentations $a(w_i)$, $q(w_i)$ et $k(w_i)$ telles que :

— $a(w_i)$ représente le mot en tant que mention, on modélise ainsi la distribution de probabilité sur l'ensemble des étiquettes possibles comme

$$P\big(m_i = M_s | a(w_i)\big) = \sigma\left[f(a(w_i))\right]_s$$

où f est une projection linéaire apprise dans l'ensemble des étiquettes et σ la fonction *softmax*.

— $q(w_i)$ représente le mot en tant que référence et $k(w_i)$ en tant que référent, de façon à ce que plus w_i est susceptible de référer à w_j, plus $q(w_i)$ est proche de $k(w_j)$. On modélise ainsi la distribution de probabilité sur l'ensemble des étiquettes comme

$$P\big(r_i = t | \left(q(w_i), k(w_t)\right)\big) = \sigma\left[q(w_i) \cdot k(w_t)\right]_t$$

où $\cdot$ désigne le produit scalaire.

Pour pouvoir utiliser le modèle dans notre cas avec des textes longs, ceux-ci sont découpés en $chunks$ chevauchés chacun de taille $n = 256$ tokens. Un nouveau $chunk$ commence tous les $l = 16$ tokens. La plupart des mots du texte sont donc présents dans $\frac{n}{l} = 16$ *chunks*.

Une fois les prédictions du modèle calculées, on sélectionne parmi les 16 prédictions attribuées à chaque mot en fonction du nombre d'occurrences et la position du mot dans chaque *chunk*, puis un parcours par profondeur est utilisé pour récupérer les chaînes de coréférence au niveau global. Nous obtenons ainsi, comme montré dans l'exemple figure 1, pour une entrée de taille quelconque, une suite d'étiquettes identifiant les mentions, et une suite d'étiquettes identifiant les entités distinctes mentionnées.

Et	c'est	ainsi	que	je	fis	la	connaissance	du	petit	prince.	Voilà	le	meilleur	portrait	que,	plus	tard,	j'	ai	réussi	à	faire	de	lui.
0	0	0	0	S-PER	0	0	0	B-PER	I-PER	E-PER	0	0	0	0	0	B-TIME	E-TIME	S-PER	0	0	0	0	0	S-PER
				(1)				(2)	(2)	(2)						(3)	(3)	(1)						(2)

FIGURE 1 – Exemple d'entrée et de sortie du modèle

5.1.2 Entraînement et évaluation

Deux textes sont séparés comme jeu de validation, deux autres comme jeu de test (cf. tableau 3). Nous entraînons un modèle CamemBERT-Large pendant 10 *epochs* avec un *batch size* de 10 *chunks*, en démarrant l'apprentissage avec un *learning rate* de 5×10^{-4}. Le tableau 4 détaille les performances du modèle sur le jeu de test pour la détection de mentions et la résolution de coréférence selon différents scores (Vilain *et al.*, 1995; Bagga & Baldwin, 1998; Luo, 2005; Recasens & Hovy, 2011; Moosavi & Strube, 2016).

	train	*dev*	*test*	total
# tokens	15 9591	22 304	22 102	184 107
# chunks	9 974	1 394	1 381	11 506
# mentions PER	20 712	2 111	2 950	25 773
# entités PER	5 078	720	771	6 569

TABLE 3 – Statistiques de FR-LitBank

		précision	rappel	F_1
Mentions		90,65	90,08	90,37
Coréférence	MUC	85,06	85,10	85,08
	B^3	82,66	56,49	67,11
	$CEAF_e$	28,50	91,89	43,50
	$BLANC$	85,81	62,99	69,22
	LEA	64,73	62,47	63,58

TABLE 4 – Résultats sur le test de FR-LitBank

5.2 Prédiction du genre par méthode heuristique

On utilise le modèle de coréférence pour identifier les mentions et les chaînes dans le corpus GIRLS. Après sélection des chaînes de type PER, on détecte de façon heuristique le genre du personnage mentionné dans chaque chaîne, s'il est explicite. Pour ce faire, nous nous basons sur un vocabulaire prédéfini. [6] On compte le nombre d'occurrences de mots distinctifs du genre masculin (*Monsieur*, *père*, *homme*...) et du genre féminin (*Madame*, *mère*, *femme*...) dans chaque chaîne. On attribue ainsi à la chaîne le genre qui a reçu le comptage le plus élevé, ou l'étiquette "Mixte/Non reconnu" en cas d'absence de mots distinctifs ou d'égalité. Enfin, on peut estimer le nombre de personnages masculins et féminins, ainsi que le nombre de mentions des personnages masculins et féminins.

5.2.1 Évaluation de la performance

Nous avons évalué la méthode heuristique de la détection du genre de la façon suivante : dans un premier temps, nous avons tiré vingt chaînes de coréférence au hasard du corpus GIRLS. Puis, nous avons discuté en équipe de la façon dont nous les classerions manuellement dans les classes 'Hommes', 'Femmes' et 'Mixte/Non reconnu'. Une fois ce schéma d'annotation établi, nous avons tiré 500 autres chaînes de coréférence au hasard et deux annotatrices les ont annotées séparément. Leur accord inter-annotateur par kappa de Cohen (Cohen, 1960) était de 0,85, ce qui indique un accord presque parfait. Ensuite, les annotatrices ont procédé à une mise en accord afin de produire un ensemble de test 'gold' de 500 chaînes annotées selon le genre [7]. Sur ce jeu de test, nous avons évalué la performance de la méthode heuristique en déterminant les scores de précision et de rappel par classe donnés dans le tableau 5, ainsi qu'une exactitude (*accuracy*) globale de 64%.

Nous observons une très haute précision aussi bien pour les hommes et les femmes, mais néanmoins un rappel assez bas pour ces deux catégories, avec tout de même un score un peu plus haut pour la catégorie 'Femmes'. Après inspection manuelle, nous nous sommes rendu compte que cela est dû en grande partie à la non-reconnaissance de la méthode heuristique des prénoms. Une amélioration évidente serait l'inclusion d'une liste de prénoms masculins et féminins dans le vocabulaire prédéfini.

	Précision			**Rappel**		
Classe	**Masc.**	**Fém.**	**Mixte/N.R.**	**Masc.**	**Fém.**	**Mixte/N.R.**
Score	0,95	0,97	0,47	0,45	0,58	0,96

TABLE 5 – Évaluation de la détection du genre par méthode heuristique.

6. Le script pour déterminer le genre et qui inclut également ce vocabulaire se trouve dans le répertoire https://github.com/oseminck/scripts_article_genre_TALN2022/tree/main/script_section5_2_methode_heuristique .

7. Voir le répertoire de la note de bas de page n°7 pour plus de détails sur notre schéma d'annotation et le gold.

5.2.2 Résultats sur GIRLS

Dans le tableau 6, nous pouvons observer une symétrie en termes de mentions : en moyenne [8], les romanciers et romancières mentionnent plus souvent les personnages de leur propre genre. En termes de chaînes, et donc de personnages, nous constatons une tendance globale à présenter davantage de personnages masculins que féminins, ce qui a également été trouvé par Nagaraj & Kejriwal (2022). La partie « Femmes » présente tout de même plus de personnages féminins que la partie « Hommes ». Ce résultat est également le même que celui de Nagaraj & Kejriwal (2022) et Underwood *et al.* (2018) pour l'anglais.

Nous constatons dans la partie « Femmes » un écart entre les chaînes et les mentions : une majorité de chaînes y sont classées « Masc. », mais une majorité de mentions sont classée « Fém. ». Cela semble refléter la présence de personnages masculins plus nombreux, ayant des rôles plutôt secondaires. Le comportement du modèle sur les chaînes longues est également à mettre en question : si un personnage est mentionné à une distance supérieure de 256 tokens, la chaîne le concernant est brisée et compte pour deux personnages. Cela est plus susceptible d'arriver avec des personnages secondaires.

	Mentions			**Chaînes**		
Partie	**Masc.**	**Fém.**	**Mixte/N.R.**	**Masc.**	**Fém.**	**Mixte/N.R.**
Hommes	**35,73%**	24,24%	40,03%	**19,23%**	11,03%	69,74%
Femmes	27,75%	**36,83%**	35,42%	**18,84%**	16,58%	64,58%

TABLE 6 – Pourcentage de mentions et de chaînes selon le genre

6 Conclusion et perspectives

Les écrits des romanciers et romancières peuvent se distinguer automatiquement. Notre étude du vocabulaire spécifique aux romanciers et aux romancières a montré que des stéréotypes de genre sont présents dans notre corpus. De plus, notre méthode a aussi permis de montrer que les romans écrits par des hommes mentionnent plus de personnages masculins que de personnages féminins. Quant à elles, les romancières mentionnent autant de personnages féminins que de personnages masculins, cependant les personnages féminins ont davantage de poids dans ces écrits car ils totalisent un nombre de mentions supérieur.

Nos résultats nous laissent penser que la détection automatique du genre des référents pourrait représenter une aide dans la tâche de détection automatique du genre des romancières et romanciers. Nos résultats portent sur un corpus de 64 romans du XIX^{e} siècle. Il serait intéressant de voir si les mêmes conclusions peuvent être tirées à partir de données issues d'autres genres littéraires et d'autres époques. Nous pensons par exemple que les romans contemporains présentent moins de stéréotypes de genre.

Remerciements

Ce projet a été en partie financé par le CNRS à travers l'IRN (International research Network) Cyclades (Corpora and Computational Linguistics for Digital Humanities). Merci à Mathieu Dehouck pour la suggestion de 'GIRLS' comme titre de corpus.

8. Les détails par livre sont disponibles dans l'annexe section 7, figure 3.

Références

ARGAMON S., KOPPEL M., PENNEBAKER J. W. & SCHLER J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, **52**(2), 119–123.

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, p. 563–566 : Citeseer.

BAMMAN D., POPAT S. & SHEN S. (2019). An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2138–2144, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1220.

BENAMAR A., GROUIN C., BOTHUA M. & VILNAT A. (2022). Étude des stéréotypes genrés dans le théâtre français du xvi^{e} au xix^{e} siècle à travers des plongements lexicaux. In *Actes de la 29ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)* : Association pour le Traitement Automatique des Langues (ATALA).

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

CONSORTIUM B. *et al.* (2007). British national corpus. *Oxford Text Archive Core Collection*.

EDER M. (2011). Style-markers in authorship attribution : a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, **6**(1).

FLEURY S. (2008). Textométrie : Le trameur (itrameur) aka le métier lexicométrique. programme de génération puis de gestion de la trame et du cadre d'un texte.

KOPPEL M., ARGAMON S. & SHIMONI A. R. (2002). Automatically categorizing written texts by author gender. *Literary and linguistic computing*, **17**(4), 401–412.

LAFON P. (1984). *Dépouillements et statistiques en lexicométrie*, volume 24. Slatkine.

LAND K. (2020). Predicting author gender using machine learning algorithms : Looking beyond the binary. *Digital Studies/Le champ numérique*, **10**(1).

LANDRAGIN F. (2021). Le corpus democrat et son exploitation. présentation. *Langages*, **224**(4), 11–24.

LATTICE (2022). Girls. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 25–32.

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

MOOSAVI N. S. & STRUBE M. (2016). Which coreference evaluation metric do you trust ? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 632–642.

NAGARAJ A. & KEJRIWAL M. (2022). Robust quantification of gender disparity in pre-modern english literature using natural language processing. *arXiv preprint arXiv :2204.05872*.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RECASENS M. & HOVY E. (2011). Blanc : Implementing the rand index for coreference evaluation. *Natural language engineering*, **17**(4), 485–510.

SAFARA F., MOHAMMED A. S., POTRUS M. Y., ALI S., THO Q. T., SOURI A., JANENIA F. & HOSSEINZADEH M. (2020). An author gender detection method using whale optimization algorithm and artificial neural network. *IEEE Access*, **8**, 48428–48437.

SBOEV A., LITVINOVA T., GUDOVSKIKH D., RYBKA R. & MOLOSHNIKOV I. (2016). Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science*, **101**, 135–142.

SEMINCK O., GAMBETTE P., LEGALLOIS D. & POIBEAU T. (2021). The corpus for idiolectal research (cidre). *Journal of Open Humanities Data*, **7**, 15. DOI : https ://doi.org/10.5334/johd.42.

STRAKA M., HAJIC J. & STRAKOVÁ J. (2016). Udpipe : trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4290–4297.

UNDERWOOD T., BAMMAN D. & LEE S. (2018). The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, **3**(2), 11035.

VERHOEVEN B. & DAELEMANS W. (2019). Discourse lexicon induction for multiple languages and its use for gender profiling. *Digital Scholarship in the Humanities*, **34**(1), 208–220.

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.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

7 Annexe

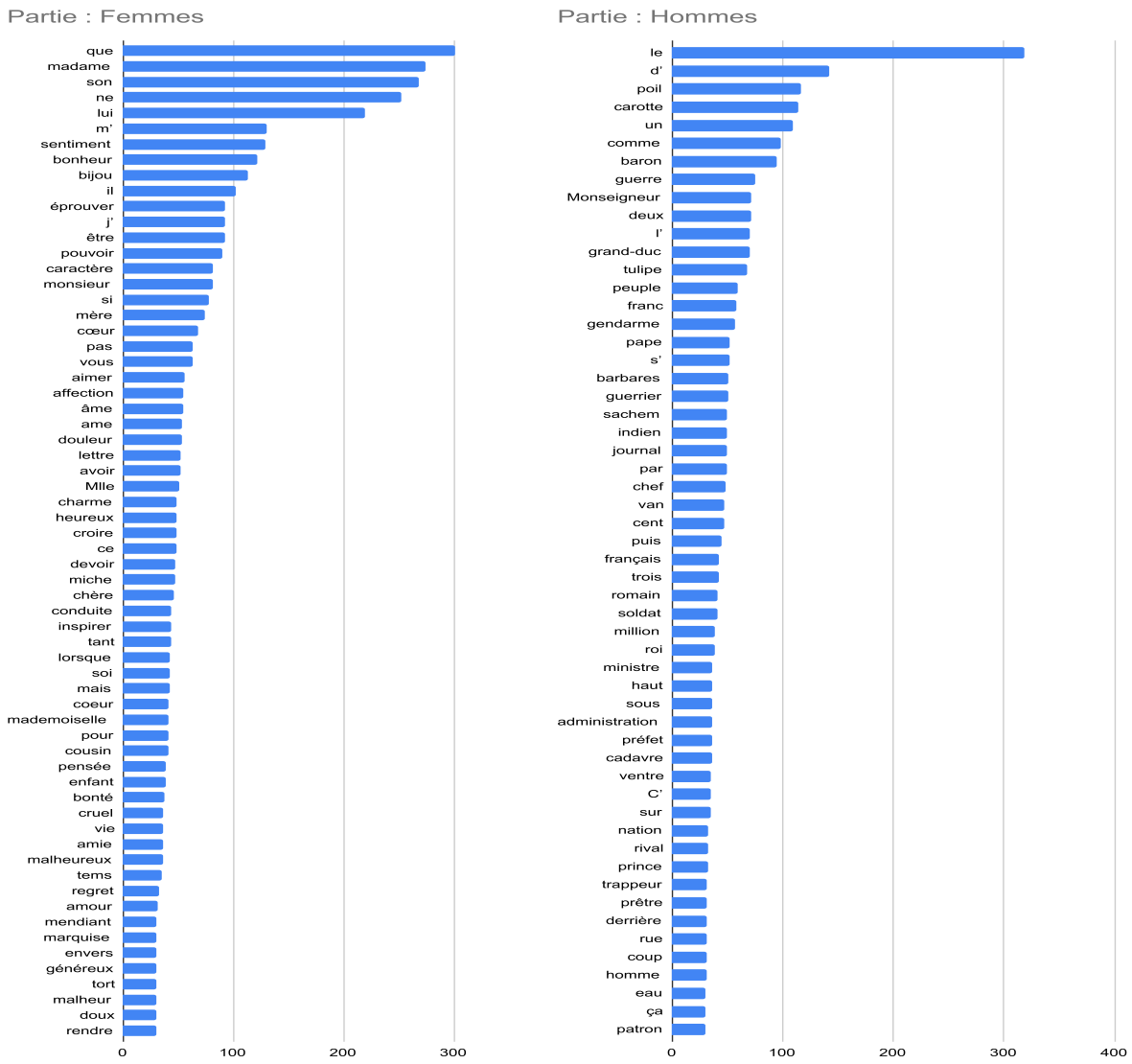
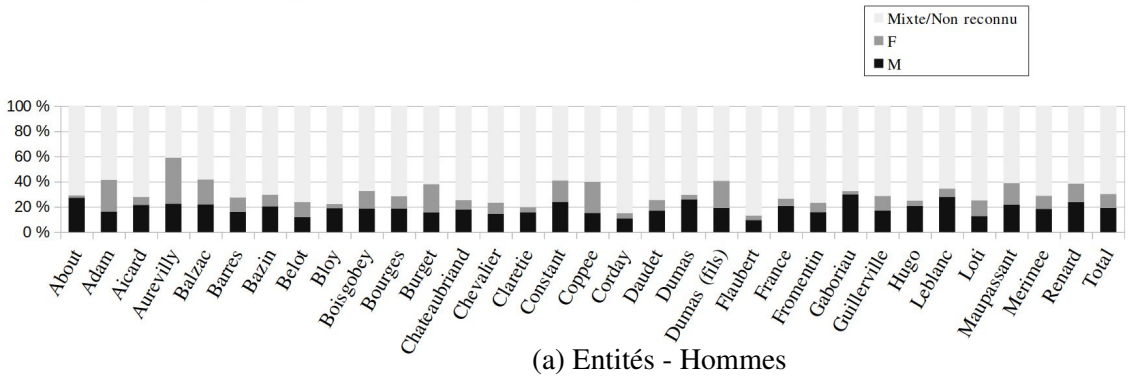
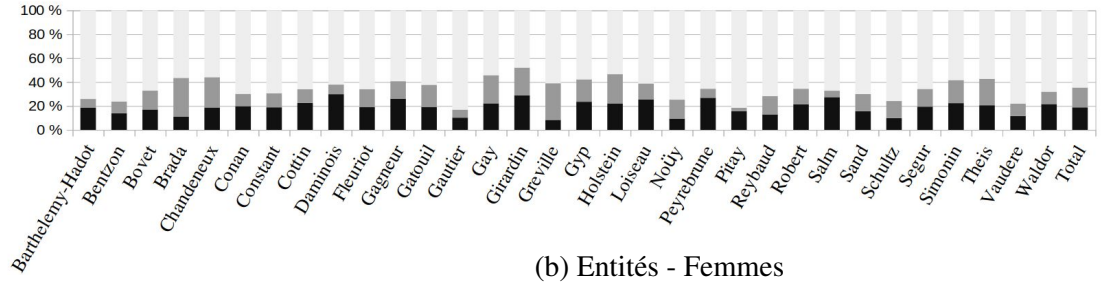


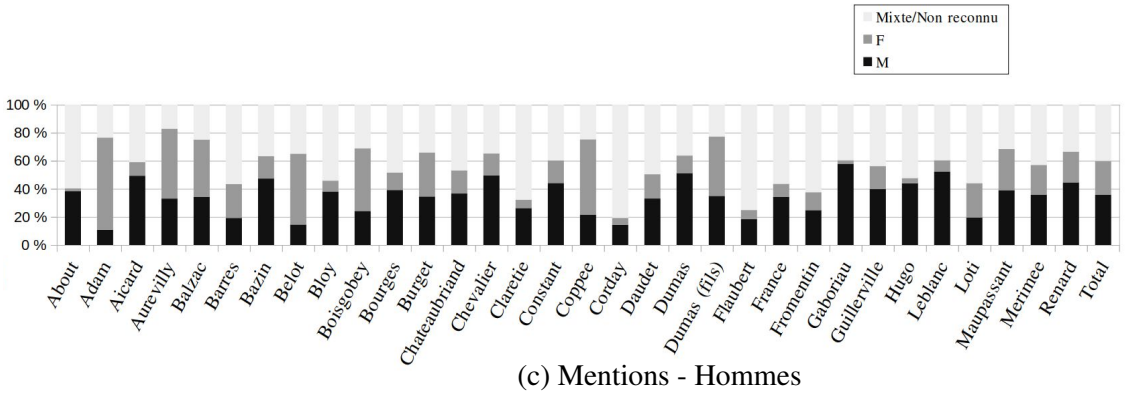
FIGURE 2 – Lemmes spécifiques (SP > 30) à la partie « Hommes » et « Femmes », sans les NP



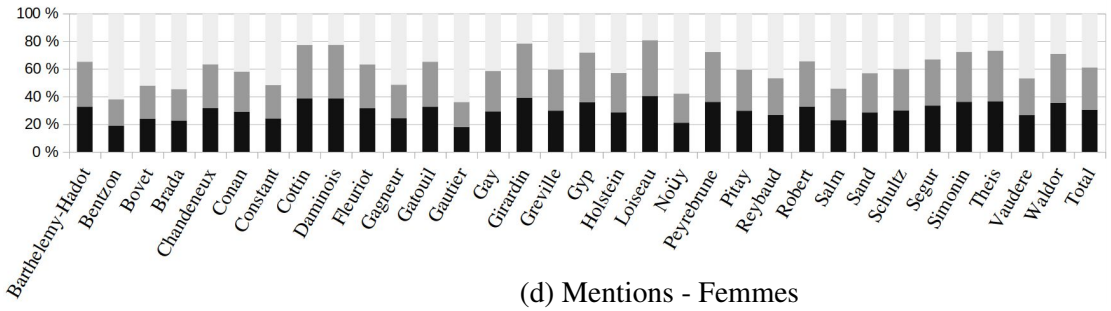
(a) Entités - Hommes



(b) Entités - Femmes



(c) Mentions - Hommes



(d) Mentions - Femmes

FIGURE 3 – Statistiques détaillées pour chaque romancière et romancier