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

Marco Naguib, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, Olga Seminck

Objectifs et méthodologie

Questions de recherche sur un corpus de romans du XIX^{e} s.

- Les livres écrits par des hommes et par des femmes sont-ils différents ? Si oui, en quoi ?
- Quel impact de l'idiolecte dans la détection du genre de l'auteur·ice ?
- Les hommes et les femmes traitent-ils différemment leurs personnages ?

3 expériences

- Classification automatique des écrits entre H/F
- Comparaison du vocabulaire spécifique H/F
- Comparaison des personnages H/F selon le genre de l'auteur·ice

Corpus existant et construction d'une nouvelle ressource

Corpus CIDRE **(Corpus of Idiolectal Research)**	**Corpus GIRLS** **(Gender Identification Resource for Literature and Style)**
- 4 autrices et 7 auteurs du XIXe - 421 oeuvres - 37 millions de mots	- 32 autrices et 32 auteurs du XIXe - un livre par auteur·ice - 4,6 millions de mots - domaine public - sources : projet Gutenberg, Gallica, Wikisource et ebooksgratuits - lien : https://www.ortolang.fr/market/corpora/girls

Détection du genre par apprentissage automatique

- Modèle de distinction des écrits en fonction du genre de l'auteur·ice
  - ⇨ Classification multiclasse par régression logistique (bibliothèque sklearn en python) :
    - **paramètres** : n-grammes de tokens de taille 1 à 3 avec occurrence minimale de 10
    - **score d'exactitude** :
      - corpus CIDRE : 99 %
      - corpus GIRLS : 72 %

- Modèle de distinction des écrits en fonction de l'idiolecte de l'auteur·ice sur CIDRE
  - **score d'exactitude** : 94%

Lien : https://github.com/oseminck/scripts_article_genre_TALN2022/tree/main/scripts_section3_apprentissage_auteurs_et_genre

Explorations textométriques : spécificités liées aux genres

- **Un vocabulaire spécifique au genre de l'auteur·ice ?**
  - Outils : iTrameur (Fleury, 2008) + UDPipe (Straka *et al.*, 2016)
  - **Indice de spécificité** (Lafon, 1984) : vocabulaire surreprésenté dans une partie du corpus par rapport à l'autre
    - Partie « Hommes » :
      - Champ lexical de la guerre (lemmes)
        - Ex : guerre, guerrier, soldat, cadavre
      - Titres de noblesse, fonctions (lemmes)
        - Ex : baron, Monseigneur, grand-duc, prince
        - Ex : gendarme, pape, roi, ministre, préfet, prêtre
      - Pronom « Il » (formes) : indice de spécificité = 282
    - Partie « Femmes » :
      - Champ lexical des émotions / sentiments et de la famille (lemmes)
        - Ex : sentiment, bonheur, éprouver, aimer, affection, âme, douleur, heureux, coeur, malheureux
        - Ex : mère, cousin, enfant
      - Titres de civilité « courants » (lemmes)
        - Ex : madame, monsieur, mademoiselle
      - Pronom « elle » (formes) : indice de spécificité = 233

Explorations textométriques : spécificités liées aux genres

- **Un genre grammatical spécifique au genre de l'auteur·ice ?**

Étiquettes	FQ	Hommes / Fq	Hommes / SP	Femmes / Fq	Femmes / SP
Féminin Singulier	388803	235059	-54	153744	54
Masculin Singulier	446265	280316	70	165949	-69
Masculin Pluriel	156011	103625	9E+15	52386	-9E+15
Féminin Pluriel	114007	74438	151	39569	-151

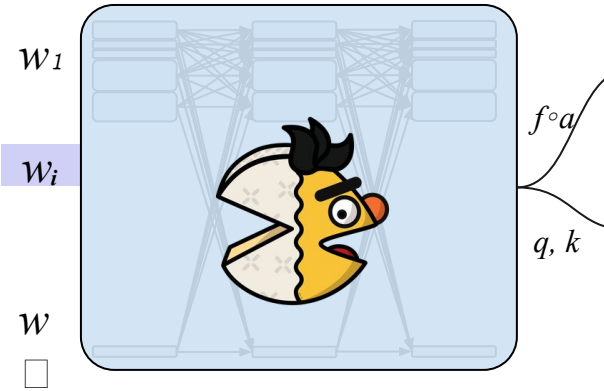
Tableau : spécificités des étiquettes morphologiques dans chaque partie
(FQ = Fréquence totale ; Fq= Fréquence sur la partie « Hommes » ou « Femmes » ; SP = Indice de spécificité)

- Mots féminins singuliers : spécifiques aux romans écrits par des femmes
- Mots masculins (singuliers et pluriels) : spécifiques aux romans écrits par des hommes
- Mots féminins pluriels : spécifiques aux romans écrits par des hommes

→ Genre grammatical féminin ⇏ référent féminin (ex : « les personnes »)

Un lien entre le genre de l'auteur·ice et le genre de ses personnages ?

Modèle de résolution de coréférence entraîné sur Democrat + Fr-LitBank



$m_i \in M = \{\text{O, B-PER, I-PER, S-PER, E-PER, B-LOC, ...}\}$

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

$r_i \in \{0, 1, ...n\}$

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

w_i	*Puis*	*il*	*me*	*dit*	*:*	*Alors*	*,*	*toi*	*aussi*	*tu*	*viens*	*du*	*ciel*	*!*
m_i	*O*	*S-PER*	*S-PER*	*O*	*O*	*O*	*O*	*S-PER*	*O*	*S-PER*	*O*	*B-LOC*	*E-LOC*	*O*
r_i après propagation	-	*<un petit bonhomme-7>*	*<Je-3>*	-	-	-	-	*<Je-3>*	-	*<Je-3>*	-	*<du ciel-9>*	-	-

Un lien entre le genre de l'auteur·ice et le genre de ses personnages ?

Méthode heuristique à base de vocabulaire : prédiction du genre des personnages

1. Identification automatique des mentions et des chaînes de coréférence + sélection des chaînes de type PER
2. Utilisation d'un vocabulaire prédéfini[1] pour l'attribution d'un genre à chaque chaîne
3. Estimation du nombre de personnages masculins / féminins + le nombre de leurs mentions

Evaluation de la méthode

Tirage de 500 chaînes → annotation manuelle par 2 annotatrices (Kappa de Cohen = 0,85) → établissement d'un GOLD

➢ Exactitude (*accuracy*) globale = 64%

pas de reconnaissance des prénoms

	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

[1] https://github.com/oseminck/scripts_article_genre_TALN2022/tree/main/script_section5_2_methode_heuristique

Un lien entre le genre de l'auteur·ice et le genre de ses personnages ?

Prédictions du modèle :

Mentions

Partie	Masc.	Fém.	Mixte/N.R.
Hommes	**35,73 %**	24,24 %	40,03 %
Femmes	27,75 %	**36,83 %**	35,42 %

Chaînes

Partie	Masc.	Fém.	Mixte/N.R.
Hommes	**19,23 %**	11,03 %	69,74 %
Femmes	**18,84 %**	16,58 %	64,58 %

⇨ Les romanciers et romancières mentionnent plus souvent des personnages de leur propre genre.

⇨ Les personnages féminins sont plus présents chez les femmes que chez les hommes.

⇨ Chez les femmes, les personnages féminins sont moins nombreux, mais plus mentionnés.

Conclusion

- Confirmation des stéréotypes de genre

Romans écrits par des hommes	Romans écrits par des femmes
+ personnages masculins - personnages féminins	autant de personnages masculins que féminins MAIS + de mentions de personnages féminins

⇨ Le genre du référent représente-t-il une aide dans la détection du genre du romancier ?

Perspectives

Les résultats seraient-ils différents avec un corpus :

- plus grand ?
- avec un thème commun (thème que notre étude aurait classifié comme plus spécifique à un genre) ?

Observe -t-on des différences entre les romanciers et romancières dans :

- le choix du narrateur (interne, omniscient, externe)
- le poids des dialogues
- l'usage de l'impératif selon le genre des personnages
- l'emploi du passif et de l'actif par le narrateur pour décrire les actions des personnages
- la proportion des rôles de personnages passifs VS actifs selon leur genre
- la longueur des phrases
- la construction syntaxique

Références bibliographiques

- 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 des genres dans le théâtre français du xvie au xixe 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., SU. REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE ..,
- 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. & STRAKOV. 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 modeltheoretic coreference scoring scheme. In Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.