Caractérisation du genre des auteurs dans l'écriture de romans français.

Adrian Tanasescu

Caractérisation du genre des auteurs dans l'écriture de romans français.

Adrian Tanasescu
Institut des Sciences de l'Homme de Lyon
69007, Lyon, France
adrian.tanasescu@ish-lyon.cnrs.fr

RESUME
L'utilisation des nouvelles technologies de l'information et particulièrement des techniques d'analyse de données textuelle dans les sciences sociales est un domaine actif de la recherche relative aux Humanités Numériques. Dans ce contexte, les travaux présentés dans cet article s'intéressent à l'analyse des romans francophones afin d'étudier les éventuelles caractéristiques liées au genre de l'auteur (Homme ou Femme) dont l'intérêt est connu dans la communauté de études sur le genre (gender studies). Plusieurs analyses sont ainsi proposées afin d'étudier cette la relation entre le style d'écriture et le genre de l'auteur.

Mots Clés
Analyse d'écriture ; analyse syntaxique ; analyse de données ; romans francophones.

INTRODUCTION

Quand on parle d'écriture de romans, beaucoup prétendent que l'écriture serait asexuée. Toutefois, nous sommes tous conscients que chaque auteur possède sa propre manière d'écrire ses textes, de les rendre uniques [1]. C'est cette diversité narrative qui permet, par exemple, aux lecteurs d'avoir des préférences pour certains auteurs plutôt que d'autres.

Partant de l'idée qu'il ne peut y avoir un effacement total de la part de féminité ou masculinité de l'auteur d'un roman, nous avons souhaité étudier l'hypothèse de l'existence d'un genre dans l'écriture, à travers une analyse de romans français. L'idée est de voir si cette hypothèse peut déjà être soutenue dans le cas du roman français avant d'investiguer plus loin.

La problématique du genre de l'écriture est un sujet qui a déjà suscité l'intérêt des investigateurs dans un cadre précis d'analyse d'emails [2] mais aussi des chercheurs [3] mais surtout en langue anglaise. Le systèmes de recommandation et de publicité sur Internet s'intéresse aussi à savoir qui est derrière un clavier (homme ou femme) afin d'envoyer les publicités ad hoc.

ETUDE DU GENRE DANS L'ECRITURE

Pour analyser l'hypothèse de l'existence d'un genre dans l'écriture de romans, nous avons réuni un ensemble de 100 romans écrits en français, la moitié écrits par des auteurs hommes et l'autre moitié par des auteurs femmes. Nous avons souhaité observer, à travers l'analyse syntaxique des textes, s'il existe des similarités entre les écrits d'auteurs de même genre et surtout si il existe des disparités détectables et notables entre les écrits d'auteurs hommes et femmes. A la différence des quelques études similaires déjà effectuées [2,3], nous nous sommes intentionnellement strictement limité aux indicateurs syntaxiques ayant l'intuition que ces derniers étaient capables de restituer la part du style d'écriture lié au genre.

Pour notre étude nous avons analysé les 100 romans à travers 32 indicateurs syntaxiques (ex : noms, verbes au présent, verbes au futur, adjectifs, énumérations, etc.). Tous ces indicateurs ont été générés pour l'ensemble des romans, puis normalisés selon le nombre total de mots dans l'œuvre afin de gommer l'effet d'échelle. Chaque indicateur était ainsi représenté comme étant la part du nombre d'occurrences correspondante dans l'ensemble des mots du romans.

Dans un premier temps nous avons cherché à savoir si tous les indicateurs étaient pertinents au regard du genre de l'auteur. Pour cela nous avons utilisé des méthodes de sélection d'attributs (Fisher filtering, forward logit, arbres de décision) afin de réduire la dimensionnalité compte tenu du faible nombre d'observations. Nous avons ainsi retenu un nombre limité (8) de ses indicateurs selon leur capacité à discriminer entre les genres des auteurs.

Indicateurs syntaxiques
Verbes à l'infinitif
Déterminants possessifs
Enumérations
Noms communs
Verbes au futur
Pronoms relatifs
Verbes au présent
Verbes à l'imparfait du subjonctif

Tableau 1. Indicateurs retenus après sélection d'attributs

Plusieurs méthodes de fouille de données ont été utilisées, parmi lesquelles la *classification ascendante hiérarchique (CAH)*, les *arbres de décision* ou encore la *régression logistique*. Les résultats son très encourageants puisque nous avons obtenu des résultats de détection du genre des auteurs plutôt justes. La validation de ces méthodes sur échantillons indépendants présente des résultats ayant des taux

d'erreur compris entre 19 et 25%.

Comparativement aux résultats des études citées [2,3], même si les contextes (ainsi que les langues de rédaction des textes) sont bien différents, la précision du modèle présenté est plus qu'encourageante du fait du faible nombre de variables indicatives considérées dans l'apprentissage.

	Arbres de décision		Régression logistique	
Genre	**F**	**H**	**F**	**M**
	Apprentissage			
Rappel	**0.900**	**0.980**	**0.860**	**0.776**
Précision	**0.978**	**0.906**	**0.796**	**0.884**
Accuracy*	**0.939**		**0.818**	
	Cross-validation			
Rappel	**0.790**	**0.740**	**0.780**	**0.710**
Précision	**0.752**	**0.779**	**0.729**	**0.763**
Accuracy*	**0.765**		**0.745**	

Tableau 2. Précision des modèles étudiés

L'écart entre les taux d'erreurs en apprentissage et en cross-validation, indicateur de la robustesse des modèles, nous indique qu'il existe encore une variabilité élevée des indicateurs parmi les romans analysés.

Nous pensons que l'élargissement des échantillons des romans analysés pourrait réduire cette variabilité et de minimiser le taux d'erreur de détection du genre de l'auteur. Ces travaux sont encore en cours de réalisation.

Bibliographie

1. Françoise Wuilmart. Traduire un homme, traduire une femme… est-ce la même chose ? Dans « Traduire le genre : femmes en traduction », pp. 23-39, 2009.
2. Vel OD, Corney M, Anderson A, Mohay G. Language and gender author cohort analysis of e-mail for computer forensics. In Proc. digital forensic research workshop, 2002
3. Na Cheng, R. CHandramouli, K.P. Soubbalakshmi. Author gender identification from text. In Digital Investigation, Vol.8. No.1, pp. 78-88, 2011