Exploration orientée entités : étude du genre dans le Mercure de France

Yoann Dupont, Marguerite Bordry

Exploration orientée entités : étude du genre dans le *Mercure de France*

Yoann Dupont[1] Marguerite Bordry[1]
(1) ObTIC, Sorbonne Université, 4 place Jussieu, 75006 Paris
yoann.dupont@sorbonne-universite.fr,
marguerite-marie.bordry@sorbonne-universite.fr

RÉSUMÉ

Dans cet article, nous étudions la façon dont le genre influence les critiques littéraires et plus précisément le *Mercure de France*, l'une des plus importantes revues parisiennes de la fin du XIX^{e} siècle. Nous nous intéressons aux auteurs et autrices italiennes. Nous avons utilisé Wikidata afin de lier les entités repérées à un identifiant unique de la base. Ainsi, nous avons pu récupérer le genre d'un auteur, quel que soit le pseudonyme sous lequel ce dernier écrivait, ce qui nous a permis d'obtenir des cooccurrents spécifiques pour chaque genre.

ABSTRACT

Entity oriented exploration : studying gender in the Mercure de France.

In this article, we study how gender influences literary critics. More precisely, we study the *Mercure de France*, one of the most important parisian journal at the end of the 19^{th} century. Our subject of interest are Italian authors. We used Wikidata to link entities to a unique ID in the database. We then could level the knowledge base to gather the gender of each author, regardless of their pseudonyms. We finally gathered specific cooccurrent terms for authors depending on their gender.

MOTS-CLÉS : Entités nommées, liage des entités nommées, analyse de sentiment.

KEYWORDS: Named entities, named entity linking, sentiment analysis.

1 Introduction

La question du genre dans la critique littéraire a fait l'objet de nombreuses études. Si Clément (2016) a mis en lumière l'« asymétrie critique » dont pâtissent les femmes du XVI^{e}, Christine Planté rappelle, elle, qu'il s'agit d'« un système à deux termes (hommes et femmes, masculin et féminin) qui ne sont ni égaux ni symétriques, car ils relèvent d'un rapport de hiérarchie et de domination » (Triaire *et al.*, 2003) Dès lors, la question que l'on peut se poser est celle de savoir si l'on peut trouver un moyen de visualiser ces rapports de dominations qui ne sont, comme le rappelle Christine Planté, « ni égaux, ni symétriques ». On se propose donc de mener une étude de genre dans un corpus de critique littéraire en alliant la reconnaissance automatique des entités nommées, reliées à la base Wikidata, et l'identification automatique des modalités du discours critique, sur un corpus préalablement annoté selon une ontologie qui comprend le jugement et les émotions. L'objectif est donc de visualiser l'impact des stéréotypes de genre à l'échelle d'un corpus critique dans son intégralité. Les différents traitements que nous avons effectués pour créer les données seront rendus accessibles sur un dépô

accessible publiquement [1].

2 Cadre d'étude : Les auteurs italiens critiqués dans le *Mercure de France*

2.1 Le *Mercure de France*

Nous travaillons sur un corpus de critique littéraire tiré du *Mercure de France*, l'une des plus importantes revues parisiennes de la fin du XIX^{e} siècle. Notre corpus couvre la période allant de 1890, date de la fondation de la revue par Alfred Vallette, à la fin de la Première Guerre mondiale, en 1918. La variété des sujets traités dans le *Mercure de France* est grande, mais nous nous intéresserons en particulier à la réception, dans la revue, de la culture et, plus particulièrement, de la littérature italiennes. Nous nous appuierons pour cela sur l'édition électronique « La culture italienne dans le *Mercure de France* [2, 3] », réalisée par le LabeX OBVIL, au sein de laquelle nous avons rassemblé un corpus de plusieurs centaines d'articles portant sur la littérature italienne contemporaine.

2.2 Les articles du *Mercure de France*

Les articles composant notre corpus peuvent être classés en trois catégories. Les articles tirés de la chronique de littérature italienne du *Mercure de France*, inaugurée dès 1891 par Remy de Gourmont sous le titre « Littérature italienne », bientôt devenue « Lettres italiennes », en constituent la partie la plus importante. Entre 1890 et 1918, quatre chroniqueurs se succèdent à sa tête : Remy de Gourmont jusqu'en 1897, en partie sous le pseudonyme d'A. Zanoni, puis Luciano Zuccoli (1897-1904), Ricciotto Canudo (1904-1913) et, enfin, Giovanni Papini (1913-1918). Leurs articles couvrant les publications paraissant en langue italienne en Italie, ils s'adressent donc en priorité à un public italophone, ou au moins italophile, qu'ils tiennent au courant des principales tendances de l'actualité littéraire italienne. Les articles écrits par d'autres contributeurs réguliers du *Mercure de France* et traitant d'œuvres italiennes traduites en français ou représentées sur les scènes françaises constituent le deuxième sous-ensemble de notre corpus. Enfin, on trouve une dernière catégories d'articles abordant la littérature italienne dans la revue. Il s'agit cette fois de publications occasionnelles, consacrées à un auteur ou à un mouvement littéraire particuliers. Notre corpus comporte 296 879 mots, comptage effectué par `SpaCy` (Honnibal & Montani, 2017).

2.3 Pourquoi s'intéresser aux auteurs italiens dans ce corpus ?

Il faut d'abord souligner la grande exhaustivité des différents chroniqueurs du *Mercure de France*. Certains auteurs sont cités beaucoup plus fréquemment que d'autres. Toutefois, parmi eux, une part importante voient leurs noms apparaître beaucoup plus rarement voire, pour certains, une fois seulement en trente ans. De telles variations, importantes, d'un auteur à l'autre, sont riches d'enseignements pour les chercheurs qui s'intéressent à la réception de la littérature italienne en

1. Disponible à l'adresse : `https://github.com/YoannDupont/minerva`
2. Textes accessibles à l'adresse : `https://obvil.sorbonne-universite.fr/corpus/mdf-italie/`
3. Fichiers XML-TEI disponibles à : `https://obvil.huma-num.fr/ariane/mdf17032022/search`

France à la fin du XIX^{e} siècle. Connaître tous les noms des auteurs cités au moins une fois permet en effet de reconstituer avec précision le canon italien du *Mercure de France*. L'étape suivante constitue à déterminer les auteurs appréciés par les chroniqueurs de la revue et ceux qui, au contraire, sont jugés sévèrement, voire, dans certains cas, relativement ignorés.

3 Constitution de la liste des auteurs

Notre liste d'auteurs a été obtenue de manière semi-automatique. Une première liste a été établie en comparant les noms présents dans les articles du corpus aux auteurs italiens répertoriés dans la base de données Data BnF. Cette opération d'extraction a été menée en comparant les noms présents dans les deux cent vingt-six articles du corpus aux auteurs italiens répertoriés dans la base de données Data BnF [4] pour la période allant du Moyen Âge aux années 1960-1970. Un filtrage manuel a ensuite été nécessaire, afin d'éliminer les doublons, les erreurs et les entités nommées non pertinentes, tels que les noms de lieux et les noms de maisons d'édition. La licence `CC-BY-NC-ND` des données d'origine ne nous permet pas de diffuser directement les documents prétraités pour nos analyses. Les scripts que nous avons utilisés dans le cadre de cet article seront diffusés sur le dépôt mentionné en note de bas de page 1.

3.1 Résultats de l'annotation

Au total, nous avons recensé 598 auteurs italiens cités au moins une fois dans les deux cent vingt-six articles formant notre corpus, pour un total de 4435 mentions répertoriées dans le corpus. Si ce nombre peut *a priori* sembler important, plus de la moitié des auteurs ne sont en réalité cités qu'à une seule reprise. Le plus souvent, il s'agit de brèves mentions à la fin des chroniques, dans la section *Memento* des « Lettres italiennes » : les chroniqueurs y dressent une liste de publications récentes, sans émettre à leur endroit le moindre jugement critique. À l'inverse, certains noms reviennent extrêmement fréquemment et bien au-delà des seules « Lettres italiennes » : Gabriele D'Annunzio est cité à cinq cent quatre-vingt-neuf reprises, par de très nombreux chroniqueurs du *Mercure*, tandis que les poètes Giosuè Carducci et Giovanni Pascoli le sont respectivement deux cent vingt-trois et cent soixante-seize fois.

3.2 Encodage du corpus

Chaque article formant notre corpus est encodé en XML-TEI, ce qui nous a permis de l'annoter automatiquement grâce à la plateforme Ariane [5] (Alrahabi, 2021), outil développé par l'équipe ObTIC [6] de Sorbonne Université. Construit sur une approche symbolique à base de règles et de patrons syntaxiques, l'outil identifie des marqueurs de surface et annote automatiquement le discours critique. Les annotations peuvent être positives (*Appréciation*, *Appréciation intellectuelle*, *Joie*, etc.) ou négatives (*Dépréciation*, *Dépréciation psychoaffective*, *Tristesse*, etc.). La liste complète des sentiment est données dans le tableau 1. Elles permettent une lecture du texte fondée sur le jugement

4. Site web : `https://data.bnf.fr/`
5. Accessible à l'adresse : `https://obvil.huma-num.fr/ariane/`
6. Site web de l'équipe : `https://obtic.sorbonne-universite.fr/`

et les émotions. Une sélection d'exemples de phrases annotées par la plateforme Ariane est disponible dans le tableau 2.

POLARITÉ	CLASSE
positif	Accord, Accord, Amour, Appréciation, Appréciation esthétique, Appréciation éthique, Appréciation informative, Appréciation intellectuelle, Appréciation psychoaffective, Assurance, Comique, Confiance, Correct, Courage, Fierté, Force, Guérison, Joie, Paisible, Pardonner, Politesse, Précision, s'Excuser, Sincérité, Soutien
négatif	Folie, Accusation, Agacement, Ambiguïté, Ambition, Avertissement, Colère, Critique, Déception, Découragement, Dégoût, Dénonciation, Dépréciation, Dépréciation esthétique, Dépréciation éthique, Dépréciation informative, Dépréciation intellectuelle, Dépréciation psychoaffective, Désaccord, Étrangeté, Faiblesse, Haine, Honte, Ignorance, Incorrect, Indignation, Insulte, Ironie, Malice, Mensonge, Mépris, Parodie, Peur, Plainte, Prétendre, Sévérité, Souffrance, Surprise, Suspicion, Tristesse, Vantardise, Violence, Voix

TABLE 1 – La liste des sentiments et opinions proposés par Ariane.

CLASSE	PHRASE	TOME, N°
Appréciation	De l'œuvre de M. Beltramelli je parlerai d'ailleurs un jour plus longuement : parmi les jeunes écrivains italiens, il est aujourd'hui celui qui est le plus puissant évocateur de la beauté et de la force de sa terre.	LVI, 195
Appréciation esthétique	c'est un bon roman, audacieux et honnête, de cette honnêteté littéraire qui est le résultat d'un travail réfléchi et personnel.	XXXVIII, 137
Amour	Quelques-unes, comme la comtesse Lara, dont la vie belle d'amoureuse fut brisée par un galant assassin, eurent des accents de liberté qui apportèrent quelques aperçus de vraie psychologie féminine.	LXXX, 292
Correct	Ce petit drame, pourtant, ne laisse pas que d'être assez habilement conduit, et il est joué avec émotion par MM. Bauer et Bourny.	XLVIII, 168

TABLE 2 – Quelques exemples de sorties de la plateforme Ariane sur des textes issus du *Mercure de France*.

3.3 Base de connaissance Wikidata

Nous avons utilisé la base de connaissance Wikidata [7] (Vrandečić & Krötzsch, 2014) afin de désambiguïser les entités d'intérêt de notre corpus. Un intérêt de Wikidata est de contenir des noms canoniques et des alias. Nous pouvons ainsi étudier les différents auteurs avec un certain degré de finesse. Nous pouvons comparer les sentiments associés à différents alias, par exemple entre un nom italien et un nom francisé, comme Gabriele D'Annunzio, souvent cité sous le nom de « Gabriel D'Annunzio »,

7. Site web : https://www.wikidata.org

ou Matilde Serao, dont le nom est souvent orthographié « Sérao ». Nous pouvons également étudier les sentiments et termes associés au auteurs en fonctions des différentes valeurs des catégories qui leurs sont associées, comme par exemple « sexe ou genre », la propriété Wikidata P21. Il est à noter qu'à l'époque du *Mercure de France*, les alias des auteurs que nous avons étudiés étaient, à quelques exceptions près, connus. Les critiques évoquent ainsi les auteurs selon leur genre de référence et non celui évoqué par le pseudonyme.

Nous avons utilisé une surcouche [8] `SpaCy` (Honnibal & Montani, 2017) de la bibliothèque `Opentapioca` (Delpeuch, 2019) pour attribuer un identifiant Wikidata aux mentions dans le texte. Lorsqu'une mention n'était pas reliée à une entrée dans Wikidata, nous avons récupéré 10 candidats à l'aide des bibliothèques `pywikibot` [9] et `wikidataintegrator` [10]. Ces candidats ont ensuite été filtrés pour conserver uniquement les éléments Wikidata correspondant à des êtres humains ayant au moins une des occupations suivantes : écrivain, poète, romancier ou essayiste.

Sur les 598 auteurs et autrices d'intérêt dans notre étude, 128 n'ont pas pu être reliés à Wikidata, soit 21%. Cela représente 308 mentions sur les 4435 que nous avons pu identifier dans le texte, soit 7%. Bien que la couverture à l'échelle des auteurs soit encore améliorable, cela ne représente qu'une faible partie du nombre total de mentions. Il conviendra de renseigner ces auteurs et autrices dans la base.

3.4 Cooccurrences spécifiques

Une fois les entités identifiées et catégorisées à l'aide de leur identifiant Wikidata, nous avons effectué une recherche de cooccurrents spécifiques selon chaque classe. Afin de mesurer la spécificité d'un score de cooccurrence en un pôle A et un cooccurrent B, nous avons simplement utilisé la formule suivante :

$$specificity(A, B) = \frac{score(A, B)}{score(\bar{A}, B)} \tag{1}$$

Où $score$ est une fonction mesurant l'indice de cooccurrence entre un pôle et un mot cooccurrent. La spécificité de la cooccurrence est donc la valeur du rapport entre les mot appartenant à une classe A par rapport à ceux en dehors de la classe A. Pour reprendre l'exemple de la catégorie "sexe ou genre", la spécificité est le rapport entre l'indice de cooccurrence entre "femme" et "non femme" ("homme" dans le cas du Mercure de France). Ce principe peut être étendu bien évidemment à d'autres cas de figure, comme "les auteurs écrivant en italien" et "les auteurs n'écrivant pas en italien".

Afin de valider cette approche, nous avons utilisé deux métriques dans notre étude. La première est le dice (Dice, 1945; Sorensen, 1948) adouci :

$$smoothed\_dice(A, B) = \frac{2|A \cup B| + 1}{|A| + |B| + 1} \tag{2}$$

Nous avons préféré cette métrique à la *Pointwise Mutual Information* (Church & Hanks, 1990), fréquemment utilisée car dice donne toujours une valeur positive. La seconde est l'indice de cooccur-

8. Disponible à l'adresse : https://github.com/UB-Mannheim/spacyopentapioca
9. Disponible à l'adresse : https://github.com/wikimedia/pywikibot
10. Disponible à l'adresse : https://github.com/SuLab/WikidataIntegrator

rence de Lafon (1980), comme utilisé notamment dans le logiciel TXM (Heiden, 2010), que nous avons utilisé pour calculer ces indices. Pour cela, nous avons remplacé les mentions des auteurs et autrices par l'identifiant de leur genre tel que renseigné dans Wikidata, à savoir `Q6581072` pour "féminin" et `Q6581097` pour "masculin". Un comparatif des spécificités obtenues selon le score est disponible dans le tableau 3. De nos observations, les deux mesures semblent fournir des ensembles de mots spécifiques similaires, classés différemment. Nous n'avons ici gardé que les cooccurrents ayant un indice de cooccurrence supérieur ou égal à 1.

MOT	SPÉCIFICITÉ		RANG	
	Dice	Lafon	Dice	Lafon
autrices				
Mme	764	63,8	38	1
autrice	4370	6,3	7	10
divorce	3146	3,72	16	50
féminine	6867	2,15	2	85

MOT	SPÉCIFICITÉ		RANG	
	Dice	Lafon	Dice	Lafon
auteurs				
M	642	127	1	2
poète	329	8,07	4	42
œuvre	145	7,36	238	46

TABLE 3 – Comparaison entre les scores de spécificité obtenus en utilisant dice ou l'indice de Lafon. À gauche, le tableau pour les spécificités pour les autrices. À droite, le tableau pour les spécificités des auteurs.

4 Analyse des résultats

Notre approche permet d'effectuer plusieurs types d'analyses liées au genre dans la critique littéraire. En premier lieu, il est possible de s'intéresser au lexique lié aux genres, en examinant les cooccurrences reliées à la propriété Wikidata P21, celle qui correspond au genre de l'autrice ou de l'auteur concerné. On obtient ainsi « critique », « poète », « Pascoli » ou « Carducci » parmi les résultats associés aux auteurs de sexe masculin : il s'agit là des auteurs les plus célèbres de l'époque, ainsi que de genres littéraires (essai critique, poésie), prisés par les chroniqueurs de la revue. En ce qui concerne les autrices de sexe féminin, on compte parmi les résultats « divorce » et « Mlle », qui situent les autrices en question vis-à-vis du mariage, considéré, à l'époque, comme le seul destin possible pour une femme, ainsi que l'adjectif « féminine ». Ce résultat s'explique par la tendance, typique de l'époque, à regrouper les autrices sous la bannière unique de l'écriture dite « féminine », qui diffère par essence de l'écriture dite masculine et qui a pour corollaire une série de qualités typiquement associées à la féminité, comme les sentiments, la finesse, le charme ou la psychologie. On observe en second lieu des résultats intéressants du point de vue du genre et des pseudonymes. Le poète Domenico Gnoli (1839-1915) se distingue par ses différents pseudonymes, dont Giulio Orsini et un pseudonyme féminin, Gina d'Arco. Le panorama de toutes les annotations, positives et négatives, associées à Gnoli, quelle que soit son appellation, comme illustré dans la figure 1, montre des résultats différents selon le genre supposé du pseudonyme. Pour « Gina d'Arco », on retrouve en effet « Appréciation psychoaffective » et « Appréciation esthétique », deux annotations qui dominent dans les articles consacrés aux autrices, alors que, pour « Gnoli » ou « Domenico Gnoli », on retrouve par exemple « Force », le retour au texte permettant de comparer les phrases reliées à chaque annotation. Ce cas ouvre la possibilité d'étudier les différences d'annotation entre pseudonymes masculins et féminins, afin de voir si l'on peut repérer une réception genrée en fonction du pseudonyme choisi par l'autrice ou l'auteur. Une telle analyse serait d'autant plus intéressante que les autrices avaient au XIXe siècle

souvent recours aux pseudonymes masculins pour dissimuler leur véritable identité.

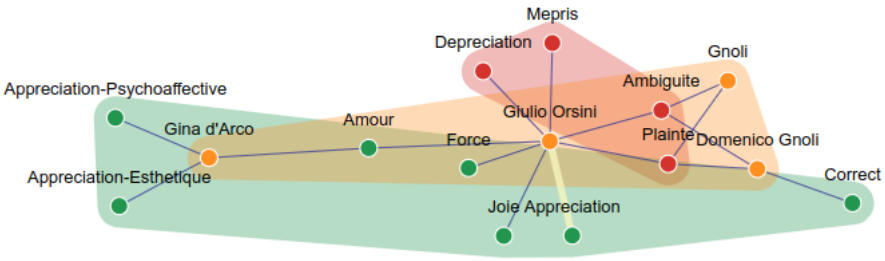

FIGURE 1 – La visualisation des sentiments associés à Domenico Gnoli étant donné ses différents pseudonymes. Il est possible d'effectuer un retour au texte en cliquant sur les nœuds.

Nous avons également analysé les sentiments associés aux auteurs et autrices. La plateforme Ariane (Alrahabi, 2021) a pu fournir des annotations de sentiments au niveau de la phrase. Nous avons relié les sentiments aux personnes en vérifiant leur co-présence. Au total, nous avons observé 66 sentiments reliés à au moins un auteur ou autrice. Le graphe montrant la répartition des 30 sentiments les plus fréquents selon le genre est donné dans la figure 2.

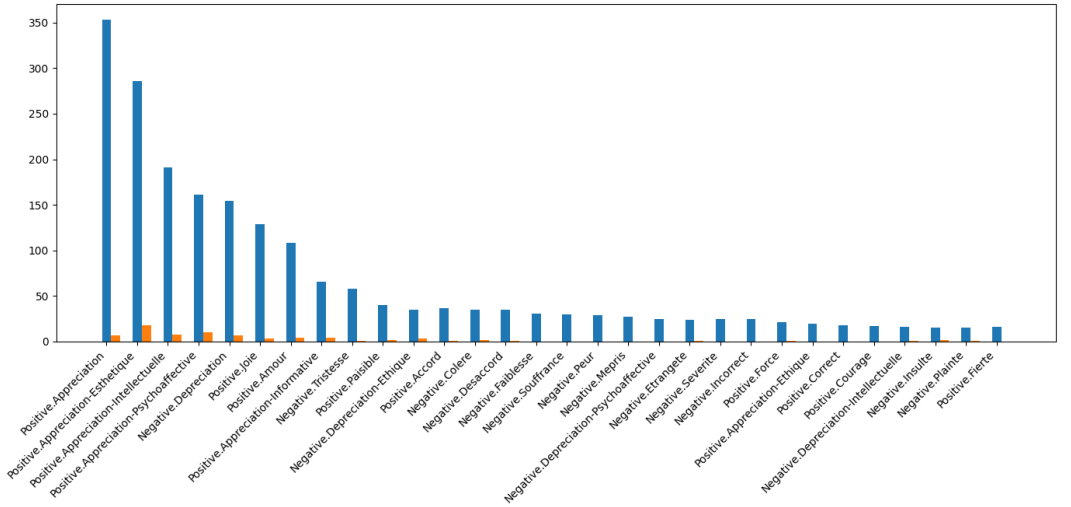

FIGURE 2 – Les comptes des sentiments les plus fréquents selon le genre dans le *Mercure de France*. Les comptes pour les hommes sont en bleu, ceux pour les femmes en orange. Graphe généré par la bibliothèque `matplotlib` (Hunter, 2007)

.

Nous avons testé l'homogénéïté de la répartition des sentiments chez les auteurs et les autrices. Afin de vérifier si les sentiments associés aux autrices suivaient une même loi ou non que les sentiments associés aux auteurs, nous avons utilisé un test exact de Fisher (Fisher, 1934). Le choix de ce test a été motivé par les faibles effectifs sur de nombreuses classes, qui n'étaient soit représentées que chez les auteurs, soit n'étaient présents qu'une fois chez les autrices. Ces faibles effectifs associés à

un relativement grand degré de liberté rendent inadapté l'utilisation d'un test du χ^2 (Pearson, 1900) d'homogénéité. Nous avons utilisé le langage de programmation R (Team *et al.*, 2013) pour réaliser le test exact de Fisher, en utilisant une méthode de Monte-Carlo (Metropolis & Ulam, 1949) avec un échantillonnage aléatoire répété un million de fois. Nous avons ainsi obtenu une *p-value* de $0,4244$. Nous ne pouvons donc pas rejeter l'hypothèse nulle sur notre corpus.

5 Conclusion

Nos résultats d'annotation confirment l'existence de stéréotypes de genre dans notre corpus, ce qui était classique pour l'époque. Néanmoins, l'intérêt principal de la visualisation à l'échelle du corpus tout entier est que nous avons pu en saisir les dynamiques. Nous avons ainsi pu saisir au premier coup d'œil le fait qu'auteurs et autrices sont substantiellement traités de la même façon par les critiques, sans que la proportion d'annotations positives ou négatives varie considérablement entre eux. Notre visualisation nous permet cependant de pousser l'analyse plus avant : la possibilité de visualiser toutes les annotations de la plateforme Ariane en les reliant à une entité nommée précise permet immédiatement de voir que ce sont les annotations elles-mêmes qui varient, avec l'émotion pour les femmes, et la raison pour les hommes. Cette possibilité d'offrir une vue d'ensemble à l'échelle d'un corpus tout entier et ce pour tous les auteurs cités au moins une fois dans ce corpus est un atout inestimable pour l'analyse littéraire.

Il serait sans doute intéressant de pousser plus avant cette étude. Pour ce travail, nous sommes en effet partis d'une liste d'entités nommées préétablie, portant sur un corpus de plusieurs centaines d'articles. Il serait pertinent de faire le même type de travail sur un corpus critique plus large, tel que le corpus critique du LabeX Obvil de Sorbonne Université [11], qui regroupe un ensemble conséquent de textes critiques d'auteurs variés, parmi lesquels on compte des adversaires acharnés de l'écriture dite « féminine », tel que Jules Barbey d'Aurevilly, auteur des *Bas-bleus* (1878). Cela pose toutefois la question de repenser notre outil : étant donné la taille du corpus critique en question, obtenir une liste préétablie d'entités nommées s'avèrerait sans doute beaucoup plus difficile, ce qui imposerait alors de ne plus travailler dans un cadre clos, mais dans un cadre ouvert.

Références

ALRAHABI M. (2021). Ariane : dispositif de fouille et de lecture synthétique de textes. In *DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis (Atelier Dahlia)*.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, **16**(1), 22–29.

CLÉMENT M. (2016). Asymétrie critique. La littérature du XVI^{e} siècle face au genre. *Littératures classiques*, (2), 23–34.

DELPEUCH A. (2019). Opentapioca : Lightweight entity linking for wikidata. *arXiv preprint arXiv :1904.09131*.

DICE L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**(3), 297–302.

11. Ce corpus est accessible ici : https://obvil.sorbonne-universite.fr/corpus/critique/

FISHER R. A. (1934). *Statistical methods for research workers*. Edinburgh : Oliver & Boyd, 5^{th} édition.

HEIDEN S. (2010). The TXM platform : Building open-source textual analysis software compatible with the TEI encoding scheme. In *24^{th} Pacific Asia conference on language, information and computation*, volume 2, p. 389–398 : Institute for Digital Enhancement of Cognitive Development, Waseda University.

HONNIBAL M. & MONTANI I. (2017). spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, **7**(1), 411–420.

HUNTER J. D. (2007). Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, **9**(3), 90–95. DOI : 10.1109/MCSE.2007.55.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, **1**(1), 127–165.

METROPOLIS N. & ULAM S. (1949). The Monte Carlo method. *Journal of the American statistical association*, **44**(247), 335–341.

PEARSON K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**(302), 157–175.

SORENSEN T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, **5**, 1–34.

TEAM R. C. *et al.* (2013). R : A language and environment for statistical computing.

TRIAIRE S., PLANTÉ C. & VAILLANT A. (2003). *Féminin/Masculin : écritures et représentations. Corpus collectifs*. Montpellier : Presses universitaires de la Méditerranée.

VRANDEČIĆ D. & KRÖTZSCH M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, **57**(10), 78–85.