Exploration ortientée entités

étude du genre dans le *Mercure de France*

Yoann Dupont, Marguerite Bordry

27 juin 2022

ObTIC, Sorbonne Université

Plan

Introduction

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

Analyse sur corpus

Introduction

Introduction

Le genre dans la critique littéraire est un sujet déjà étudié (Clément 2016; Triaire, Planté, and Vaillant 2003) :

- Montrent l'aspect asymétrique de la critique en fonction du genre.

[1] https://github.com/YoannDupont/minerva

Introduction

Le genre dans la critique littéraire est un sujet déjà étudié (Clément 2016; Triaire, Planté, and Vaillant 2003) :

- Montrent l'aspect asymétrique de la critique en fonction du genre.

Nous proposons une étude sur un corpus de critique littéraire :

- Annotation en entités nommées
- Liage à la base Wikidata
- Analyse de sentiments
- Traitements et corpus seront accessibles sur un dépôt en ligne[1].

[1] https://github.com/YoannDupont/minerva

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

Le *Mercure de France*

Un corpus de critique littéraire :

- Le *Mercure de France* : importante revue parisienne de la fin du XIXe (1890-1918)
- Articles cibles : critiques sur la littérature italienne contemporaine
- Intégralité du corpus : https://obvil.sorbonne-universite.fr/corpus/mdf-italie/
- Fichiers TEI : https://obvil.huma-num.fr/ariane/mdf17032022/search.

Les articles du *Mercure de France*

227 articles, se répartissant en trois catégories:

- Articles tirés de la chronique de littérature italienne (« Lettres italiennes »), 1891–1918 par 4 chroniqueurs. Actualité du paysage éditorial en Italie et s'adressent à un public italophile.
- Articles traitant des traductions françaises d'oeuvres italiennes ou des spectacles italiens joués dans les théâtres français.
- Articles isolés, consacrés à un auteur ou à un mouvement littéraire.
- Licence `CC-BY-NC-ND` des textes au format XML-TEI[2].

Le corpus compte 296 879 mots (comptage effectué par `SpaCy`).

[2]Licence par défaut donnée par un outil de traitement d'ObTIC, peut être généré à nouveau avec une licence plus adaptée

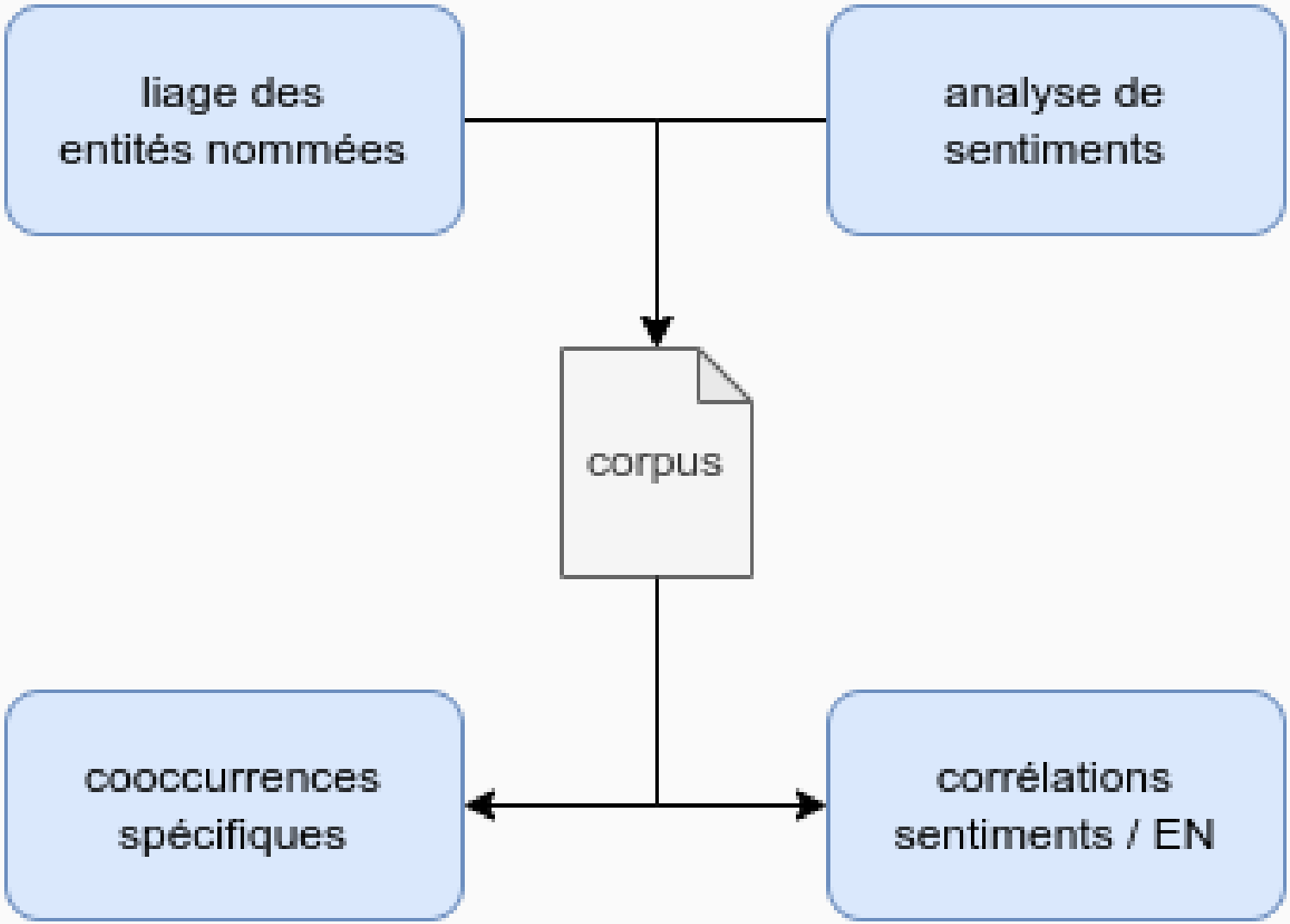
La littérature italienne dans le *Mercure de France*

Intérêt de reconstituer le canon italien du *Mercure de France*

- Le *Mercure de France* se distingue par son ouverture aux littératures étrangères et accorde une place importante à l'Italie et à sa littérature.
- Grande disparité selon les auteurs cités : certains cités plus de 100 fois entre 1890 et 1918, des centaines ne le sont qu'une seule fois.
- Connaître avec précision tous les auteurs cités au moins une fois dans la revue française permet de reconstituer l'histoire des circulations littéraires entre l'Italie et la France.
- Étape suivante : déterminer les auteurs qui sont appréciés par les critiques de la revue et ceux qui, au contraire, sont jugés plus sévèrement.

Analyse sur corpus

Étude du corpus : processus général



Constitution de la liste des auteurs

Une liste constituée semi-automatiquement

- Première liste établie en comparant les noms présents dans les articles aux auteurs italiens répertoriés dans la base de données `Data BnF` pour la période allant du Moyen Âge aux années 1960-1970
- Filtrage manuel pour éliminer doublons, homonymes et erreurs.

Constitution de la liste des auteurs

Une liste constituée semi-automatiquement

- Première liste établie en comparant les noms présents dans les articles aux auteurs italiens répertoriés dans la base de données `Data BnF` pour la période allant du Moyen Âge aux années 1960-1970
- Filtrage manuel pour éliminer doublons, homonymes et erreurs.

Projection sur le corpus

- 598 auteurs sont cités au moins une fois
  - Total de 4435 mentions répertoriées dans le corpus
  - 312 auteurs ne sont cités qu'une fois (rubrique *Memento* en fin d'article)
  - Des auteurs sont très visibles: Gabriele D'Annunzio est cité 589 fois, Giosuè Carducci 223 fois et Giovanni Pascoli 176 fois.

Liage à Wikidata (Vrandečić and Krötzsch 2014)

- Intérêts double : identification et requête dans la base de connaissance
  - Propriété P21 de Wikidata : « sexe ou genre »
  - Pseudonymes souvent connus au moment des articles
  - Utilisation du genre "véritable", pas celui évoqué par le pseudonyme

[3] https://github.com/UB-Mannheim/spacyopentapioca

Liage à Wikidata (Vrandečić and Krötzsch 2014)

- Intérêts double : identification et requête dans la base de connaissance
  - Propriété P21 de Wikidata : « sexe ou genre »
  - Pseudonymes souvent connus au moment des articles
  - Utilisation du genre "véritable", pas celui évoqué par le pseudonyme
- Algorithme adapté
  - désambiguisation : surcouche[3] `SpaCy` (Honnibal and Montani 2017) de la bibliothèque `Opentapioca` (Delpeuch 2019)
  - Requête Wikidata pour les entités non trouvées
  - Filtrage des personnes avec une activité "écrivain", "poète", "romancier" ou "essayiste"
  - Établir des liens pour les entités ambiguës
    - nom de famille lié au nom complet
    - Utilisation des pseudonymes de `Data BNF`

[3] https://github.com/UB-Mannheim/spacyopentapioca

Résultats du liage

- 21% des 598 mentions uniques non liées
- 7% des 4435 mentions au total non liées
- Manque de couverture sur Wikidata
- Auteurs/autrices à renseigner : export automatique des fiches de data BNF vers Wikidata[4].

[4] https://dicare.toolforge.org/bnf/

Annotation en sentiments

- Annotation automatique avec Ariane[5] (Alrahabi 2021)
  - Système à base de règles : attribue des sentiments à l'échelle de la phrase
  - 68 classes

[5] Accessible à l'adresse : https://obvil.huma-num.fr/ariane/

Annotation en sentiments

- Annotation automatique avec Ariane[5] (Alrahabi 2021)
  - Système à base de règles : attribue des sentiments à l'échelle de la phrase
  - 68 classes

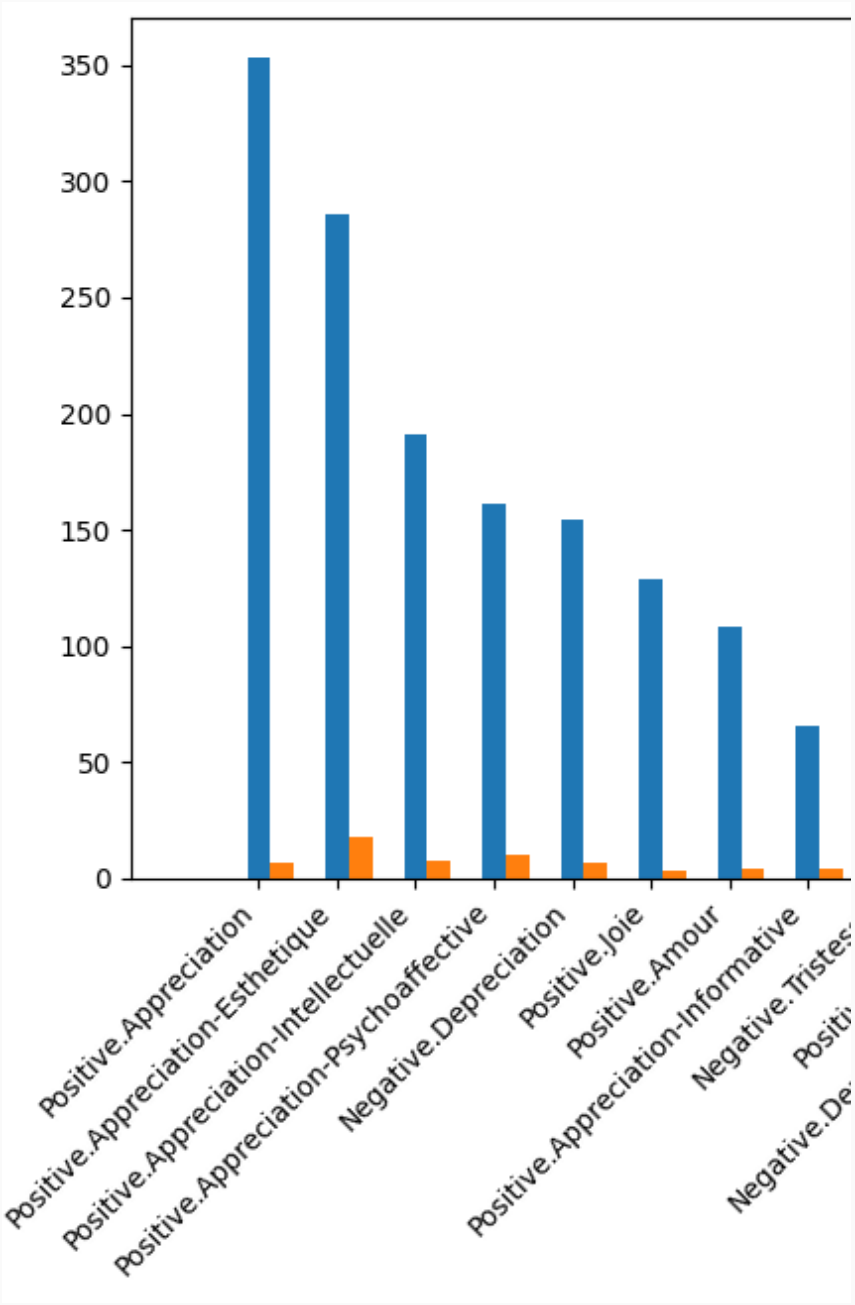
polarité	classe
positif	Accord, Amour, Appréciation, Appréciation esthétique, Appréciation informative, Appréciation intellectuelle, Appréciation psychoaffective, Appréciation éthique, Assurance, Comique, Confiance, Correct, Courage, Fierté, Force, Guérison, Joie, Paisible, Pardonner, Politesse, Précision,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

[5] Accessible à l'adresse : https://obvil.huma-num.fr/ariane/

Exemples d'annotations

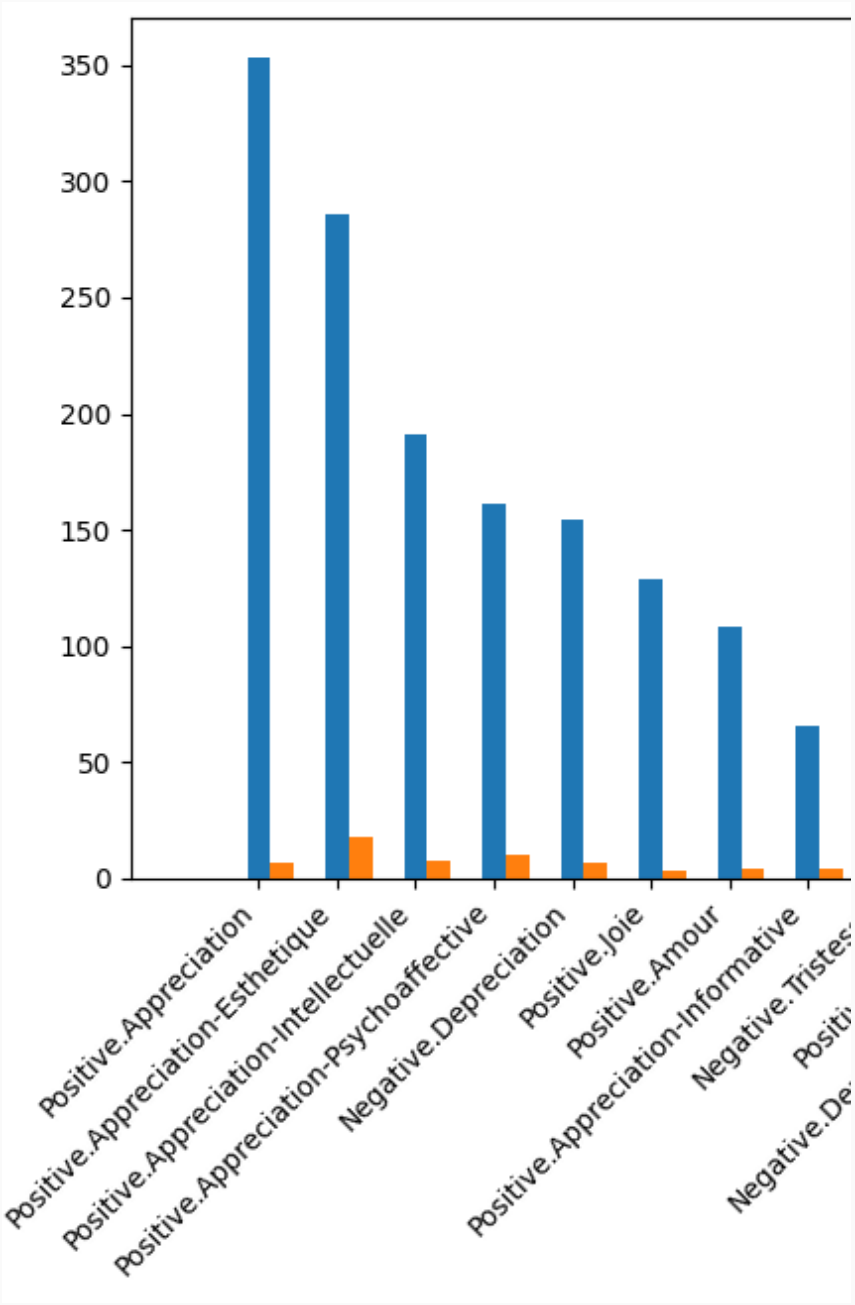
classe	phrase
Appréciation	De l'œuvre de M. Beltramelli je parlerai d'ailleurs un jour plus longuement: parmi les jeunes écrivains italiens, il est aujourd'hui celui qui est le plus puissant évocateur de la beauté et de la force de sa terre.
Appréciation esthétique	c'est un bon roman, audacieux et honnête, de cette honnêteté littéraire qui est le résultat d'un travail réfléchi et personnel.
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.
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.

Correspondances sentiments $\leftrightarrow$ personnes selon leur genre



Sentiments attribués à l'échelle de la phrase. Compte de la co-présence d'un sentiment et d'un auteur.

Correspondances sentiments $\leftrightarrow$ personnes selon leur genre



Sentiments attribués à l'échelle de la phrase. Compte de la co-présence d'un sentiment et d'un auteur.

Test exact de Fisher (1934) pour l'homogénéité $\rightarrow$ *p-value* $= 0,4244$.
Non concluant, mais sans doute dû aux faibles effectifs.

Cooccurrences spécifiques

Étant donnés une classification des pôles (ici, H/F), et un *score* de cooccurrence, on estime une spécificité de cooccurrence pour une classe :

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

Test de deux mesures, le dice (Dice 1945; Sorensen 1948) adouci :

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

La seconde testée : l'indice de cooccurrence de Lafon (1980), via logiciel TXM Heiden (2010).

Indice brut vs spécifique – féminin

	indice		rang	
token	brut	spécifique	brut	spécifique
Mme	63,87	63,87	1	1
.	21,99	0,17	2	170
féminin	20,24	4,17	3	35
:	19,49	0,11	4	179
,	19,40	0,14	5	173
—	19,17	0,22	6	165
Mathilde	12,65	12,65	7	2
Mlle	11,55	11,55	8	3
masculin	10,54	0,20	9	166
Milan	9,57	0,13	10	175

Indice brut vs spécifique – masculin

	indice		rang	
token	brut	spécifique	brut	spécifique
:	172,49	8,85	1	34
,	143,12	7,38	2	56
M	127,0	127,0	3	1
.	127,0	5,77	4	69
—	88,75	4,63	5	96
(	83,13	19,28	6	9
)	74,25	19,08	7	10
Milan	72,71	7,6	8	53
masculin	53,26	5,06	9	83
di	49,35	49,35	10	2

Dice et Lafon : comparaison

Sélection de mots spécifiques aux autrices (gauche) et aux auteurs (droite) :

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

Lexique autour des autrices inscrit dans la notion d'« écriture féminine » : divorce, romancière, farouche, sentimentalité, chefs-d'œuvre[6]

[6] Les « chefs-d'œuvre » parlent d'une femme mais écrits par des hommes

Conclusion et perspectives

Existence de stéréotypes de genre à travers le lexique

- Auteurs et autrices semblent avoir autant d'annotations positives/négatives
- Encore à démontrer sur une analyse de sentiments plus fine

Extension à des corpus plus larges pour des tests statistiques plus fiables

- Besoin d'automatiser, notamment au niveau du liage
- Étendre sur un corpus plus volumineux : corpus critique[7]
- Alimenter des données `Data BNF` la base Wikidata[8].

[7] https://obvil.sorbonne-universite.fr/corpus/critique/
[8] https://dicare.toolforge.org/bnf/

Bibliographie i

[Alr21] Motasem Alrahabi. "Ariane: dispositif de fouille et de lecture synthétique de textes". In: *DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis (Atelier Dahlia)*. 2021.

[Clé16] Michèle Clément. "Asymétrie critique. La littérature du XVI^{e} siècle face au genre". In: *Littératures classiques* 2 (2016), pp. 23–34.

[Del19] Antonin Delpeuch. "Opentapioca: Lightweight entity linking for wikidata". In: *arXiv preprint arXiv:1904.09131* (2019).

[Dic45] Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

[Fis34] Ronald Aylmer Fisher. *Statistical methods for research workers*. 5th. Edinburgh: Oliver & Boyd, 1934.

Bibliographie ii

[Hei10] Serge Heiden. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme". In: *24th Pacific Asia conference on language, information and computation*. Vol. 2. 3. Institute for Digital Enhancement of Cognitive Development, Waseda University. 2010, pp. 389–398.

[HM17] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7.1 (2017), pp. 411–420.

[Laf80] Pierre Lafon. "Sur la variabilité de la fréquence des formes dans un corpus". In: *Mots. Les langages du politique* 1.1 (1980), pp. 127–165.

Bibliographie iii

[Sor48] Th A Sorensen. "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". In: *Biol. Skar.* 5 (1948), pp. 1–34.

[TPV03] Sylvie Triaire, Christine Planté, and Alain Vaillant. *Féminin/Masculin: écritures et représentations. Corpus collectifs*. Montpellier: Presses universitaires de la Méditerranée, 2003, pp. 7–18.

[VK14] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

Analyse des résultats - Le cas Domenico Gnoli

Le poète Domenico Gnoli (1839-1915)

- différents pseudonymes : Giulio Orsini et Gina d'Arco

La visualisation des sentiments associés à Domenico Gnoli étant donné ses différents pseudonymes :