Le deep learning comme défi pour identifier le style d'un écrivain : l'exemple de Jean Giono

Véronique Magri

Le deep learning comme défi pour identifier le style d'un écrivain : l'exemple de Jean Giono

Véronique Magri[1]

[1]Université Côte d'Azur, CNRS, BCL – Veronique.MAGRI@univ-cotedazur.fr

Abstract 1

What if artificial intelligence could identify a writer's style ? What if, automatically, the machine could identify the characteristics of a piece of writing, in other words the formal elements recognizable from one work to another, as well as the differences between a corpus of study and a reference corpus ? If finally a writing could be deciphered by an algorithm ? This is precisely the challenge of deep learning applied to literature.
This is exactly the experimentation that is being attempted on Giono, from an unpublished digital text base, a very large corpus that brings together Giono's novels. The necessary differential measurement guides the constitution of the corpus ; two bases were thus constituted by É. Brunet : one brings together Giono's works, processed by the Hyperbase software. The other is a vast reference corpus whose generic and temporal homogeneity is guaranteed since they are 50 novels from the twentieth to the twenty-first centuries. The corpus was constituted by É. Brunet and includes two texts written by the same author, i.e. 50 texts for 25 authors.
From prediction to *deconvolution*, an interpretative path is built, tending towards the horizon of the discovery of an author's style.

Keywords : deep learning, literature, Giono, Hyperbase, stylistics.

Abstract 2

Et si l'intelligence artificielle parvenait à identifier le style d'un écrivain ? Et si, de manière automatique, la machine parvenait à identifier les caractéristiques d'une écriture, autrement dit les éléments formels reconnaissables d'une œuvre à l'autre, de même que les différences d'un corpus d'étude par rapport à un corpus de référence ? Si finalement une écriture pouvait être décryptée par un algorithme ? C'est précisément le défi que lance le deep learning appliqué à la littérature.
C'est exactement l'expérimentation qui est tentée sur Giono, à partir d'une base de textes numériques inédite, un très grand corpus qui rassemble les œuvres romanesques de Giono. La mesure différentielle nécessaire guide la constitution des corpus ; deux bases ont ainsi été constituées par É. Brunet : l'une réunit les œuvres de Giono, traitées par le logiciel Hyperbase. L'autre est un vaste corpus de référence dont l'homogénéité générique et temporelle est garantie puisqu'il s'agit de 50 romans du XX^{e} au XXI^{e} siècles. Le corpus a été constitué par É. Brunet et comporte deux textes du même auteur soit 50 textes pour 25 auteurs.
De la prédiction à la *déconvolution* se construit un parcours interprétatif tendu vers l'horizon de la découverte d'un style d'auteur.

Mots clés : deep learning, littérature, Giono, Hyperbase, stylistique.

1. Introduction

Cet article[1] entend mener une expérimentation sur une base de textes numériques inédite, un très grand corpus qui rassemble les œuvres romanesques de Giono. La mesure différentielle

[1] Nos remerciements chaleureux vont à Étienne Brunet pour avoir constitué la *base Giono* et pour avoir assuré plusieurs des graphiques ici présentés.

nécessaire guide la constitution des corpus ; deux bases ont ainsi été constituées par É. Brunet : l'une réunit les œuvres romanesques de Giono, où on distingue les deux manières d'écriture bien connues des spécialistes gioniens depuis *Colline* (1929) : d'un côté les romans d'avant-guerre, de l'autre les chroniques postérieures à 1946 - à partir du *Roi sans divertissement*. Les critiques tendent à voir même trois manières en distinguant dans les *Chroniques* le Cycle d'Angelo inauguré par *Le Hussard*[2] et regroupant les œuvres suivantes : *Le Hussard, Angelo, Bonheur fou, Récits de la demi-brigade*[3]. L'autre corpus qui sert de référence compte 50 textes pour 25 auteurs[4].

Depuis la prédiction jusqu'à l'étape de la *déconvolution*[5], se construit un parcours interprétatif tendu vers l'horizon de la découverte d'un style d'auteur.

1. L'étape de la prédiction

1.1. Les méthodes traditionnelles

Pour mémoire, on rappelle ici les résultats obtenus avec les calculs de la distance intertextuelle dans Hyperbase appliqués à la base Giono uniquement, ce qui permet de mettre en valeur les regroupements des œuvres selon leurs affinités lexicales ou syntaxiques. Les textes du corpus d'étude sont comparés entre eux et deux à deux avec pour norme interne l'ensemble du corpus. L'analyse factorielle suivante (Figure 1) propose le calcul sur les formes du corpus.

---

[2] Voir Robert Ricatte, 1970, Préface Giono, *Œuvres complètes*, volume 1, Paris, Gallimard, La Pléiade, 1971, p. XLVI.

[3] *Colline, Un de Baumugnes, Regain, Naissance de l'Odyssée, Le Grand Troupeau, Solitude de la pitié, Jean le Bleu, Le Chant du monde, Que ma joie demeure, Batailles dans la montagne, Pour saluer Melville, L'Eau vive, Un Roi sans divertissement, Noé, Fragments d'un paradis, Mort d'un personnage, Les âmes fortes, Les Grands chemins, Deux cavaliers de l'orage, Le Hussard sur le toit, Le Moulin de Pologne, Le Bonheur fou, Angelo, Hortense, Les Récits de la demi-brigade, Ennemonde et autres caractères, L'Iris de Suze*. Œuvres rangées par ordre chronologique.

[4] Gide : *La Symphonie pastorale* et *L'Immoraliste* ; Giraudoux : *Simon le Pathétique* et *Bella* ; Montherlant : *Les Célibataires* et *Les Bestiaires* ; Saint-Exupéry : *Courrier Sud* et *Terre des hommes* ; Colette : *Sido* et *La Vagabonde* ; Queneau : *Le Chiendent* et *Zazie dans le métro* ; Camus : *L'Etranger* et *La Chute* ; Duras : *Un Barrage contre le Pacifique* et *l'Amant* ; Yourcenar : *Mémoires d'Hadrien* et *L'œuvre au noir* ; Gary : *La Promesse de l'aube* et *Les Racines du ciel* ; Tournier : *Vendredi ou Les Limbes du Pacifique* et *Eléazar* ; Gary : *Clair de femme* et *Au-delà de cette limite* ; Ernaux : *La Honte* et *Les Années* ; Proust : *Du côté de chez Swann* et *Le Temps retrouvé* ; Mauriac : *Le Baiser au lépreux* et *Le Mystère Frontenac* ; Malraux : *L'Espoir* et *Les Conquérants* ; Breton : *Nadja* et *L'Amour fou* ; Giono : *Le Grand Troupeau* et *Le Hussard sur le toit* ; Aragon : *Les Beaux Quartiers* et *Blanche ou l'oubli* ; Vian : *L'Ecume des jours* et *L'Automne à Pékin* ; Gracq : *Le Rivage des Syrtes* et *Un Balcon en forêt* ; Mammeri : *La Colline oubliée* et *La Traversée* ; Perec *: W ou le souvenir* et *L'Homme qui dort* ; Ajar : *Gros-Câlin* et *La Vie devant soi* ; Le Clézio : *Hasard* et *Désert.*

[5] Sur ce terme, voir L. Vanni, M. Ducoffe, D. Mayaffre, F. Precioso, D. Longrée, et al.. Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis. 56th Annual Meeting of the Association for Computational Linguistics, Jul 2018, Melbourne, France. hal-01804310.

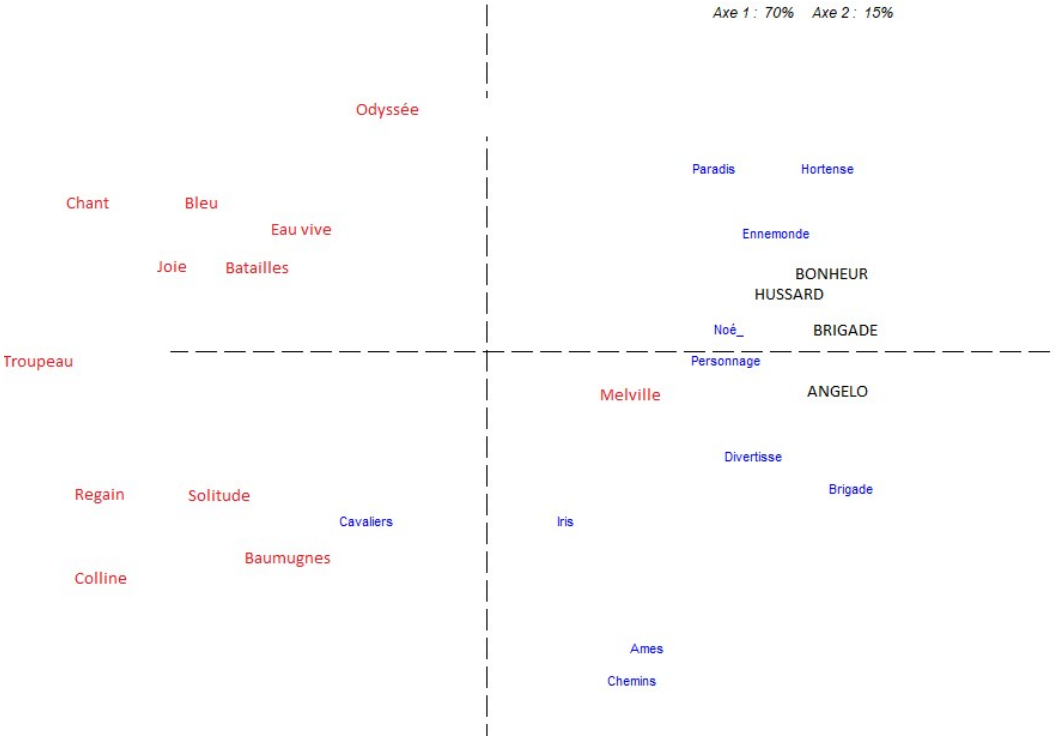

Figure 1- Analyse factorielle sur les formes – Base GIONO

On constate la répartition attendue entre les deux manières de Giono (la première dans les quadrants gauches du graphique, la deuxième dans les quadrants droits). On n'observe pas de défection du *Grand Troupeau*, même si cette œuvre correspond à une chronique de guerre et qu'elle a, à ce titre, un statut hybride – appartenant à la première manière sur le plan chronologique mais à la seconde, *a priori*, sur le plan thématique. *Pour saluer Melville* et *Deux cavaliers de l'orage* sont finalement les deux seules œuvres qui occupent une place non attendue sur l'analyse factorielle.

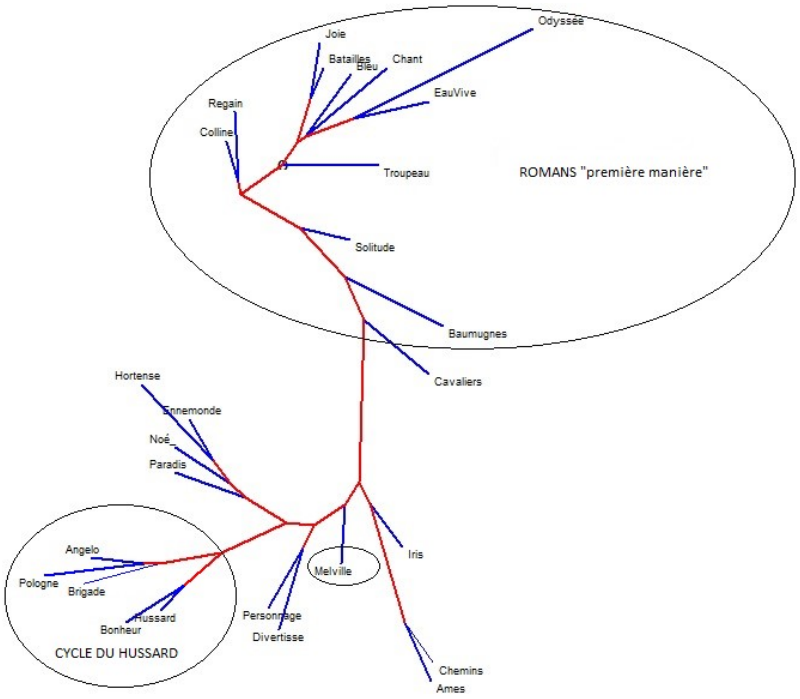

Figure 2- Analyse arborée sur les lemmes – Base Giono

La distribution est un peu plus perturbée quand elle porte sur les lemmes (Figure 2), qui font la part belle au lexique, et mettent donc l'accent sur une répartition thématique. La bipartition subsiste avec un regroupement en un bouquet du cycle du Hussard. Cependant, les œuvres paraissent affirmer leurs connexions selon les champs lexicaux qu'elles développent. *Regain* et *Colline* sont portées par une seule branche par exemple, tandis que *Le Grand Troupeau* prend quelque liberté par rapport aux romans de la première manière, comme *Pour saluer Melville* qui se retrouve isolé sur une seule branche.

1.2. Le deep learning : du modèle à la modélisation

Un point sur la méthode doit être fait en préambule[6] : le travail se fait à partir de la base de cinquante romans ; celle-ci est entraînée avec tous ces textes, autrement dit l'ingénieur apprend à la machine à reconnaître les œuvres et leurs auteurs à partir des deux textes donnés pour chaque écrivain ; pour ce qui concerne Giono, la machine a appris à identifier *Le Grand Troupeau* et *Le Hussard sur le toit* comme des œuvres écrites par Giono. Il s'agit ensuite de présenter les autres œuvres de Giono et d'observer le taux de reconnaissance effectué par la machine : c'est la première étape dite de **prédiction.** On envoie des textes non inclus dans la base et l'algorithme énonce un degré de reconnaissance évalué en termes de pourcentages qui pourrait être glosé par une formule du type « ce texte est plutôt du… ». L'algorithme procède par le biais d'une fenêtre de trois mots, ce qui veut dire que la séquentialité est le seul critère et que les appariements ne sont pas choisis en fonction de critères syntaxiques. Le deep learning travaille par couches et chaque mot est converti en valeurs numériques. La fréquence ne concerne pas seulement des formes mais des séquences, des suites linéaires de formes. Le deep learning est sensible aux

[6] Voir l'article de L. Vanni, ici même.

différents niveaux de granularité des textes – la valeur grammaticale est par exemple indépendante de la forme – et se distingue de la méthode traditionnelle par la combinaison du séquentiel et du fréquentiel, par le repérage de motifs complexes et multidimensionnels, par l'intertextualité à découvrir. Les résultats pour notre corpus d'étude présentés sous forme de pourcentages sont rassemblés dans le tableau ci-après – en rouge les plus forts taux de reconnaissance :

	Ajar	Aragon	Breton	Camus	LeClézio	Colette	Duras	Ernaux	Exupéry	Gary1	Gary2	Gide	Giono	Giraudoux	Gracq	Malraux	Mammeri	Mauriac	Montherlant	Pérec	Proust	Queneau	Tournier	Vian	Yourcenar	Total
Baumugnes	4.25	6.15	0	5.14	8.6	0.34	2.01	2.23	0.89	1.23	1.56	11.28	**28.94**	0.34	1.9	2.23	1.9	2.46	4.58	0.11	2.68	9.94	0.11	0.67	0.45	*100*
Regain	1.35	3.44	0	1.88	15.42	0.42	3.02	1.35	0.31	0.21	0.52	4.17	**50.94**	0.21	0.21	5.83	1.15	1.88	1.15	0.31	0.52	4.79	0.52	0.21	0.21	*100*
Odyssée	0.2	1.6	0.2	0.9	6.71	0.5	1.7	2.8	1.4	1	0.9	7.61	**8.71**	3.8	3.6	**12.21**	4.7	3.2	**12.71**	2.1	0.9	3.2	3	0.5	**15.82**	*100*
Solitude	1.78	9.63	0.12	3.92	8.2	0.48	1.31	1.9	0.95	1.78	2.73	8.8	**26.75**	0.95	1.9	5.11	2.62	2.14	4.28	0.83	0.83	11.18	0.59	0.59	0.59	*100*
Bleu	3.04	6.21	0.13	4.5	8.23	0.13	4.62	3.86	0.7	2.91	1.71	6.21	**22.17**	3.74	2.03	4.43	5.7	2.79	3.86	0.57	1.08	7.73	1.39	0.82	1.46	*100*
Chant	1.4	2.74	0	2.85	7.15	0	2.47	1.24	0.32	0.38	0.75	3.87	**39.76**	0.48	1.24	2.31	6.99	1.72	9.67	0.21	0.43	8.28	0.86	4.67	0.21	*100*
Joie	2.62	2.74	0.16	3.26	11.68	0	7.98	1.66	0.35	0.41	1.24	5.33	**31.38**	1.34	1.44	2.97	4.4	1.02	7.09	0.32	0.64	8.68	1.44	1.37	0.48	*100*
Batailles	2.32	3.77	0.03	2.71	11.44	0.08	3.8	3.49	0.28	0.95	1.03	6.08	**30.76**	1.17	2.79	3.54	4.72	1.67	4.83	0.33	0.87	8.68	1.4	2.18	1.09	*100*
Melville	2.41	4.56	1.21	4.02	7.37	0.94	4.16	5.09	0.94	3.08	1.47	9.12	**14.88**	2.01	1.74	4.16	3.89	3.08	3.75	2.01	2.82	12.06	1.74	0.8	2.68	*100*
EauVive	2.06	4.16	0.64	3.25	9.96	0.49	3.01	4.52	0.94	2.49	1.64	7.47	**20.77**	1.73	2.28	6.74	3.64	2.19	4.4	2.03	1.4	8.04	2.61	1.18	2.37	*100*
Divertisse	1.76	10.67	0.28	4.99	4.92	0.42	1.97	3.79	0.42	1.26	0.91	10.96	**16.92**	2.74	1.12	4.42	3.72	2.18	5.2	2.46	4.78	10.88	1.05	0.84	1.33	*100*
Noé_	1.17	5.22	1.7	3.24	8.05	0.65	1.9	8.62	1.29	2.79	0.93	8.33	**13.51**	2.95	1.54	7.48	3.11	2.75	4.25	4.49	1.54	8.13	1.74	0.65	3.96	*100*
Paradis	0.37	2.34	1.61	2.56	**12.52**	0.37	7.76	5.56	1.17	1.98	0.37	3.37	**8.78**	2.56	4.39	4.54	2.34	1.46	6.66	2.2	1.98	5.05	**15.89**	1.68	2.49	*100*
Personnage	4.44	**8.42**	0.58	4.56	3.39	1.17	3.39	6.43	0.58	3.86	1.29	**13.68**	**6.43**	1.64	1.99	1.4	1.05	1.75	2.92	**6.67**	5.85	**14.97**	1.75	0.7	1.05	*100*
Ames	3.75	**11.6**	0.16	6.72	7.42	0.39	3.4	5.35	0.55	1.56	1.68	**10.74**	**10.63**	3.32	0.66	2.54	10.12	0.78	7.46	0.78	1.91	7.46	0.23	0.51	0.27	*100*
Chemins	6.04	9.99	0.55	9	5.8	0.12	6.17	3.58	0.86	2.59	3.88	7.83	**15.6**	1.73	0.62	3.95	2.4	0.62	0.68	1.48	1.48	13.32	0.74	0.43	0.55	*100*
Cavaliers	1.93	8.13	0.06	3.45	8.07	0	2.57	2.17	0.35	0.76	1.46	8.25	**30.49**	2.11	2.4	4.27	4.1	1.29	4.1	1.29	0.76	8.37	1.7	1.17	0.76	*100*
Pologne	4.2	2.89	1.03	4.94	2.24	0.47	**13.62**	**6.25**	1.03	**7.18**	1.4	**9.7**	**5.5**	3.45	1.31	1.59	1.87	4.85	**8.77**	1.87	3.54	5.32	2.99	0.28	3.73	*100*
Bonheur	2.3	2.18	0.17	3.47	6.45	0.05	3.32	4.2	0.41	2.16	1.24	5.14	**33.31**	2.04	2.52	4.2	5	2.09	4.29	0.39	0.87	5.82	3.4	1.82	3.15	*100*
Angelo	2.13	1.98	0.91	2.13	2.51	0.15	3.19	2.96	0.38	3.12	1.14	6.69	**50.15**	1.9	1.37	0.91	1.22	1.06	3.88	0.99	2.28	3.65	2.05	0.23	3.04	*100*
Hortense	11.53	2.78	0	0.6	5.57	0	1.99	8.35	0.2	2.78	0.8	5.17	**24.65**	2.78	0.8	2.58	4.97	5.37	5.17	0.2	0.8	2.39	6.36	0.8	3.38	*100*
Brigade	2.87	5.35	0.48	5.64	4.11	0.29	2.87	4.11	0.96	3.63	3.82	15.01	**16.83**	4.11	2.39	3.44	3.73	1.05	3.25	0.86	1.63	8.32	1.24	0.38	3.63	*100*
Ennemonde	0.75	4.16	1.28	1.49	11.73	0.11	3.2	8.85	0.75	2.03	0.53	3.52	**13.75**	3.3	1.49	6.93	3.73	5.01	9.06	3.52	2.03	4.37	1.92	0.85	5.65	*100*
Iris	1.98	4.39	0.25	3.09	4.21	0	1.11	1.24	0.74	1.42	1.67	10.89	**24.75**	2.78	1.3	6.37	5.63	4.64	6	1.86	0.87	9.84	0.74	1.73	2.48	*100*
total	67.29	127.9	11.68	89.33	193.3	8.34	92.33	102.5	17.67	51.82	35.44	194.9	**581.3**	53.31	43.54	125.5	95.13	58.07	131.6	39.03	42.62	193.8	56.23	25.83	61.47	*2500*
moyenne	2.69	5.12	0.47	3.57	7.73	0.33	3.69	4.1	0.71	2.07	1.42	7.79	**23.3**	2.13	1.74	5.02	3.81	2.32	5.26	1.56	1.7	7.75	2.25	1.03	2.46	*100*

Figure 3- Tableau des reconnaissances

Regain et Angelo sont les textes qui présentent le plus fort taux de reconnaissance : 50 % d'extraits, composés de 100 mots, sont reconnus comme du Giono.
Quatre textes sont mal reconnus : *Naissance de l'Odyssée, Fragments d'un paradis, Mort d'un personnage, Le Moulin de Pologne*. Les noms propres peuvent perturber la reconnaissance quelquefois et brouiller l'exacte identification : par exemple, le nom propre « Antinoüs » semble expliquer l'attribution de près de 16 % des extraits de *Naissance de l'Odyssée* à M. Yourcenar (*Mémoires d'Hadrien*), contre 8 % à Giono ou encore le nom propre « Catherine » explique l'attribution majoritaire (13.68 %) de *Mort d'un personnage* à Queneau contre 6.43 % à Giono. Les noms propres, qui ont un rôle perturbateur, ont été retirés des bases dans les versions les plus récentes de l'algorithme.

Si on met à part ces quelques artefacts, la reconnaissance des textes de Giono est prioritaire pour toutes les autres œuvres : elle se situe entre 30 % et 40 % pour *Un de Baumugnes, Le Chant du monde, Batailles dans la montagne*, *Que ma joie demeure*, pour la première manière de Giono ; 30 % environ pour *Deux cavaliers de l'orage, Le Bonheur fou,* quant à la seconde manière.
Un nouvel apprentissage, pratiqué à titre expérimental, avec deux textes supplémentaires, *Le Moulin de Pologne* et *Un roi sans divertissement,* qui appartiennent à la deuxième manière, permet la reconnaissance de tous les textes cette fois. Ces deux romans concentrent les ingrédients de la production romanesque : le cadre campagnard, l'histoire familiale et la psychologie des gens ordinaires. Seule demeure une proximité de *Mort d'un personnage* avec Queneau, dont la reconnaissance devance légèrement Giono.

1		Ajar	Aragon	Breton	Camus	LaClézio	Colette	Duras	Ernaux	Exupéry	Garl	Garl2	Gide	Giono	Giraudoux	Gracq	Malraux	Mauriac	Mauriac	Montherlant	Pérec	Proust	Queneau	Tournier	Vian	Yourcenar
2	Colline	1.54	6.02	0.26	1.15	2.82	0.13	2.94	0.26	2.69	0.13	1.28	1.66	29.58	0.38	3.46	8.96	3.59	3.20	5.89	3.46	1.02	15.36	0.90	2.30	1.02
3	Baumugnes	4.02	2.46	0.11	1.45	0.78	0.89	4.92	0.22	0.67	1.23	3.24	3.24	30.39	0.89	1.12	2.91	0.34	1.79	8.04	1.68	6.37	20.89	0.45	0.78	1.12
4	Regain	1.67	2.50	0.10	1.04	2.92	0.10	6.98	0.00	0.63	0.31	1.77	0.94	43.02	0.26	1.88	6.77	0.83	1.77	6.77	1.46	1.46	13.75	0.94	1.67	0.52
5	Odyssée	0.20	0.90	0.60	2.00	0.80	0.30	0.50	0.50	2.10	1.70	0.80	0.33	21.12	0.98	6.61	3.90	6.61	8.71	16.72	1.20	0.90	2.90	4.80	6.11	8.71
6	Solitude	1.78	4.64	0.00	1.78	0.71	0.24	4.88	0.12	1.55	0.95	4.28	1.43	33.41	1.55	2.38	4.40	1.31	1.07	9.04	2.02	1.43	14.74	1.31	2.85	2.14
7	Bleu	1.58	3.04	0.06	2.98	1.33	0.19	7.92	0.51	2.60	2.98	2.98	0.44	27.11	2.98	3.04	1.08	2.15	1.90	9.94	0.82	2.22	11.08	2.85	5.00	3.23
8	Chant	1.56	1.07	0.00	2.47	0.97	0.00	7.74	0.27	0.54	1.24	2.79	0.81	38.53	0.27	1.88	1.18	2.63	0.70	14.83	0.54	0.54	9.08	2.96	6.07	1.34
9	Joie	1.82	1.09	0.03	2.65	1.79	0.10	7.85	0.35	0.89	3.29	2.90	0.67	31.50	0.45	1.31	0.93	2.94	0.96	13.50	0.45	0.73	9.35	3.22	9.19	2.04
10	Batailles	1.70	2.51	0.28	1.79	1.76	0.25	6.06	0.45	1.03	1.26	1.87	0.33	33.46	0.28	3.71	1.67	1.79	1.65	11.02	0.81	2.07	14.90	2.65	4.94	1.76
11	Melville	3.75	1.61	0.80	2.01	1.21	0.54	6.17	1.07	0.67	1.88	2.95	1.21	24.40	0.94	3.35	2.55	1.88	0.94	10.72	4.16	4.69	13.54	1.88	4.02	3.08
12	EauVive	1.70	1.91	0.43	2.64	2.03	0.49	5.34	0.97	1.64	1.94	3.07	0.97	31.79	1.24	3.95	2.70	3.67	1.43	9.90	1.85	2.09	7.83	3.04	3.37	4.01
13	Noé_	0.53	2.67	0.97	1.54	1.33	0.44	3.03	1.29	1.17	2.75	1.70	1.25	29.85	0.77	4.29	5.83	1.70	2.22	7.97	5.46	5.74	8.25	1.78	2.18	5.26
14	Paradis	2.20	1.54	0.95	3.29	3.88	0.00	5.34	2.20	2.93	2.93	1.17	0.44	17.20	0.37	6.66	1.61	2.64	1.46	13.76	1.83	4.03	2.86	11.79	1.46	7.47
15	Personnage	4.09	5.73	0.47	3.39	0.94	0.58	6.32	1.17	1.87	3.16	6.43	1.64	12.87	2.81	2.11	1.75	1.40	1.99	13.22	1.99	7.49	12.98	1.29	1.64	2.69
16	Ames	2.50	7.85	0.16	3.09	0.63	0.35	7.66	0.20	0.51	1.48	5.39	2.11	26.29	0.74	2.42	2.50	1.33	0.63	8.24	1.48	1.41	20.04	0.39	1.48	1.13
17	Chemins	4.07	9.31	0.49	4.75	0.25	0.25	10.30	0.18	2.34	1.97	6.10	1.36	23.61	0.18	2.16	2.10	0.99	0.18	1.66	1.42	2.22	20.10	0.49	1.23	2.28
18	Cavaliers	2.11	1.70	0.06	1.29	0.35	0.18	6.96	0.18	0.35	1.76	3.10	1.76	33.88	0.35	1.93	1.40	1.58	0.76	12.70	2.57	1.52	13.98	1.76	5.50	2.28
19	Bonheur	0.68	2.18	0.10	2.55	0.49	0.12	3.37	0.39	0.73	2.69	3.06	0.19	41.90	0.51	3.08	2.55	3.66	1.04	8.95	1.31	1.02	9.46	2.43	4.29	3.25
20	Angelo	1.29	0.91	0.15	2.43	0.38	0.15	3.34	0.38	0.84	2.89	1.98	0.23	59.73	0.76	2.36	1.60	0.99	0.76	4.86	1.29	1.52	5.47	1.52	0.61	3.57
21	Hortense	1.59	1.99	0.40	0.20	1.99	0.00	10.34	0.40	0.20	1.99	0.80	0.00	39.56	1.79	2.58	2.19	2.58	0.80	8.95	0.00	0.99	7.75	4.57	5.57	2.78
22	Brigade	1.15	4.21	0.00	2.68	0.10	0.19	2.77	0.57	1.72	3.15	4.21	1.24	30.31	1.91	4.02	2.87	2.10	1.05	6.31	2.68	3.73	11.66	1.24	2.49	7.65
23	Ennemonde	0.75	2.56	0.85	0.85	2.24	0.32	2.56	1.49	0.96	2.24	0.96	0.00	25.05	1.81	3.52	4.16	2.45	2.99	18.87	6.08	1.92	6.18	1.71	4.16	5.33
24	Iris	0.56	2.23	0.25	1.30	0.19	0.12	1.61	0.00	0.62	1.24	1.86	1.98	33.60	0.93	3.65	3.16	3.28	1.73	10.33	2.48	1.11	18.69	0.99	4.58	3.53
25	*total*	*42.84*	*70.63*	*7.52*	*49.32*	*29.89*	*5.93*	*124.90*	*13.17*	*29.25*	*45.16*	*64.69*	*24.23*	*718.16*	*23.15*	*71.47*	*68.77*	*52.44*	*39.73*	*232.19*	*47.04*	*56.22*	*270.84*	*54.96*	*81.49*	*76.19*
26	*moyenne*	*1.86*	*3.07*	*0.33*	*2.14*	*1.30*	*0.26*	*5.43*	*0.57*	*1.27*	*1.96*	*2.81*	*1.05*	*31.22*	*1.01*	*3.11*	*2.99*	*2.28*	*1.73*	*10.10*	*2.05*	*2.44*	*11.78*	*2.39*	*3.54*	*3.31*

Figure 4- Tableau des reconnaissances – 2

L'analyse factorielle suivante croise les écrivains et les textes de Giono. On y voit dans la partie supérieure les textes de Giono qui sont les plus typiques, ou du moins qui sont les plus proches des quatre textes témoins de Giono :

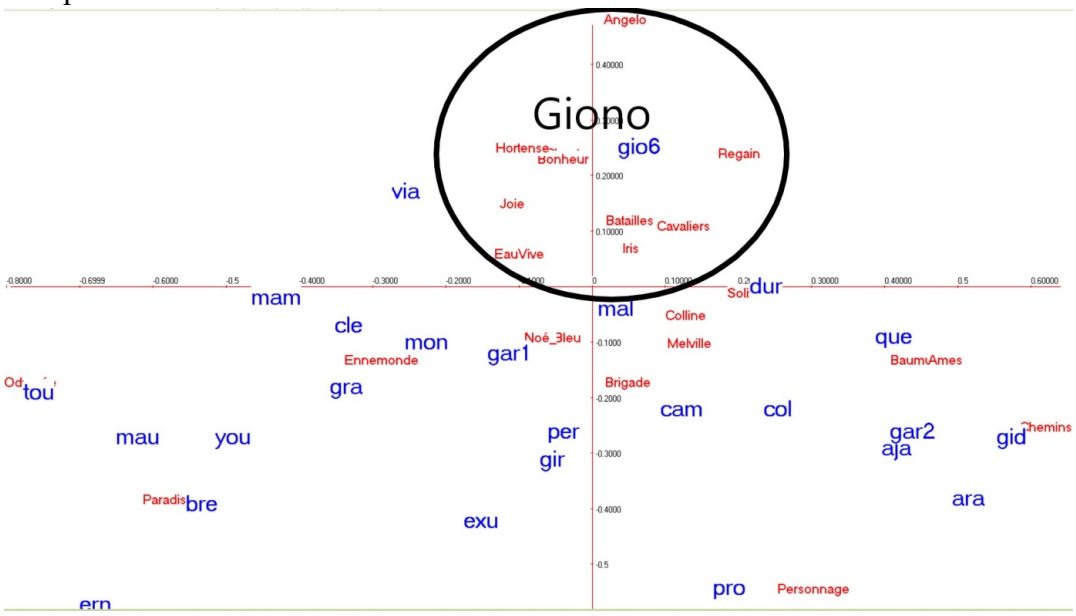

Figure 5- Les œuvres et les auteurs

Les œuvres qui affichent la plus forte proximité avec les œuvres d'entraînement sont *Regain, Que ma joie demeure, Batailles dans la montagne, Deux cavaliers de l'orage, Angelo, Le Bonheur fou, Hortense, L'Iris de Suze*, *L'Eau vive.*

2. L'étape de la *déconvolution*

2.1. À partir du deep learning

Le deep learning est un *terminus ad quem* et un *terminus a quo,* correspondant à deux étapes successives : à l'établissement préalable d'un taux de reconnaissance fait suite l'essai d'interprétation de ce taux de reconnaissance : quels sont les observables qui ont permis l'identification ? L'étape de la déconvolution conduit aux confins de la caractérisation d'une écriture. Des saillances sont repérées qui justifient la reconnaissance réussie.
L'étape de la *déconvolution* croise plusieurs critères : Prenons l'exemple de *Regain*, qui affiche le plus fort taux de reconnaissance.

L'extrait analysé est le suivant :

[…] un essieu au vieux rouleau. Il est bon d'avoir, sur la cheminée, une petite boîte, même si, sur la petite boîte il y a marqué poivre. Il est bon d'avoir cette boîte toute prête pour le cas où on aurait l'occasion d'un bon mulet. ça peut arriver. Il faudra voir. On ne peut pas toujours vivre d'emprunt. Dans le chemin qui descend, il y a Arsule et ses galoches ; on les entend toutes les deux. Arsule chante. […]

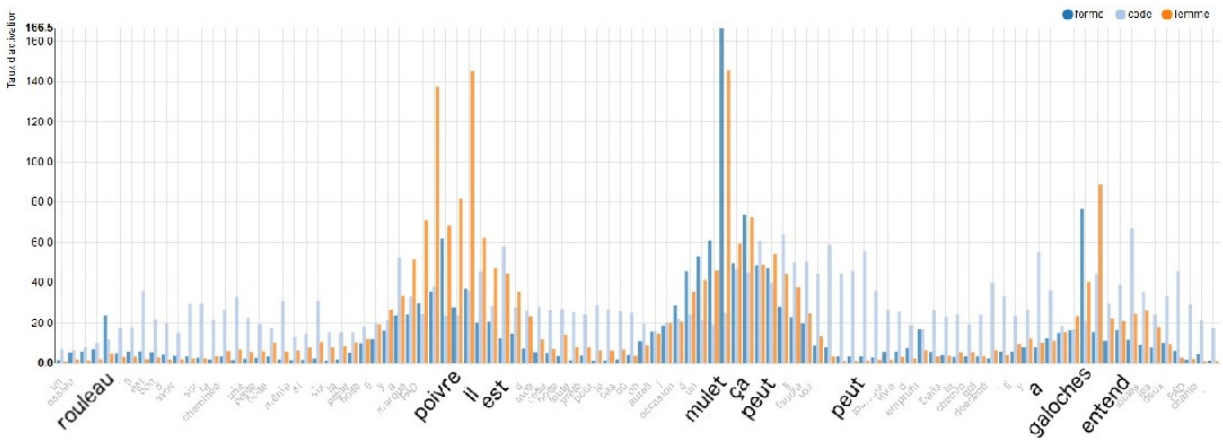

Figure 6- Histogramme – étape de la déconvolution

On obtient un histogramme à trois données - forme, code, lemme - et on peut observer tantôt l'une tantôt l'autre variable. Il s'avère que les saillances portent sur les formes « mulet », « ça », « galoches », « poivre », « rouleau » pour le premier extrait de 100 mots qui apparaît, « labours », « velours » pour le second ainsi que sur les codes grammaticaux ou catégories grammaticales « est », « a », « peut », « entend » pour le premier extrait – ce qui veut dire que ces formes ne sont retenues qu'en vertu de leur nature verbale.
Cela ressemble à de la spécificité pour les saillances de formes mais il ne faut pas s'y limiter car si on fait une recherche croisée de spécificités, au moyen des outils traditionnels implémentés dans Hyperbase, pour *Regain* dans la base Giono, seul le mot « ça » apparaît dans la liste avec un écart réduit de 29,7. Les autres mots ne s'y trouvent pas parce qu'ils n'atteignent pas le seuil de significativité – ils ne sont donc pas proposés comme spécificités lexicales ; cela signifie que ce n'est pas la seule forme qui constitue le pic de significativité mais la forme associée à d'autres paramètres. Quant aux codes, tous les éléments saillants sont des verbes de troisième personne du singulier, au présent (« est, a, peut, entend ») sans qu'on puisse décider si ces formes

émergent parce qu'il s'agit de formes verbales ou de formes de présent de l'indicatif en particulier. L'interprétation peut encore s'interroger sur la personne grammaticale récurrente. On peut supposer simplement que le deep learning conduit à de nouveaux observables, à des corrélations nouvelles entre les niveaux d'analyse, qui peuvent croiser, par exemple, un signe de ponctuation et un tiroir verbal.

Cette étape de l'élucidation des causes de la reconnaissance est ardue. Elle fait naître des interrogations ; on peut croiser d'autres fonctionnalités plus « classiques » pour tenter de comprendre les mécanismes de la reconnaissance.

2.2. Avec les outils hypertextuels

Les contrastes observés entre les textes de Giono et d'autres textes littéraires du XX^{e} siècle, par exemple, permettent de déceler des particularités de l'écriture gionienne. Il a été traité 2.3 millions de mots chez Giono et 32 millions pour le corpus de référence[7]. Ces contrastes portent sur les propriétés grammaticales d'une part, sur les champs lexicaux d'autre part, des textes en présence. La fonction « corpus extérieurs » du logiciel *Hyperbase* est activée : elle permet cette mise en contraste.

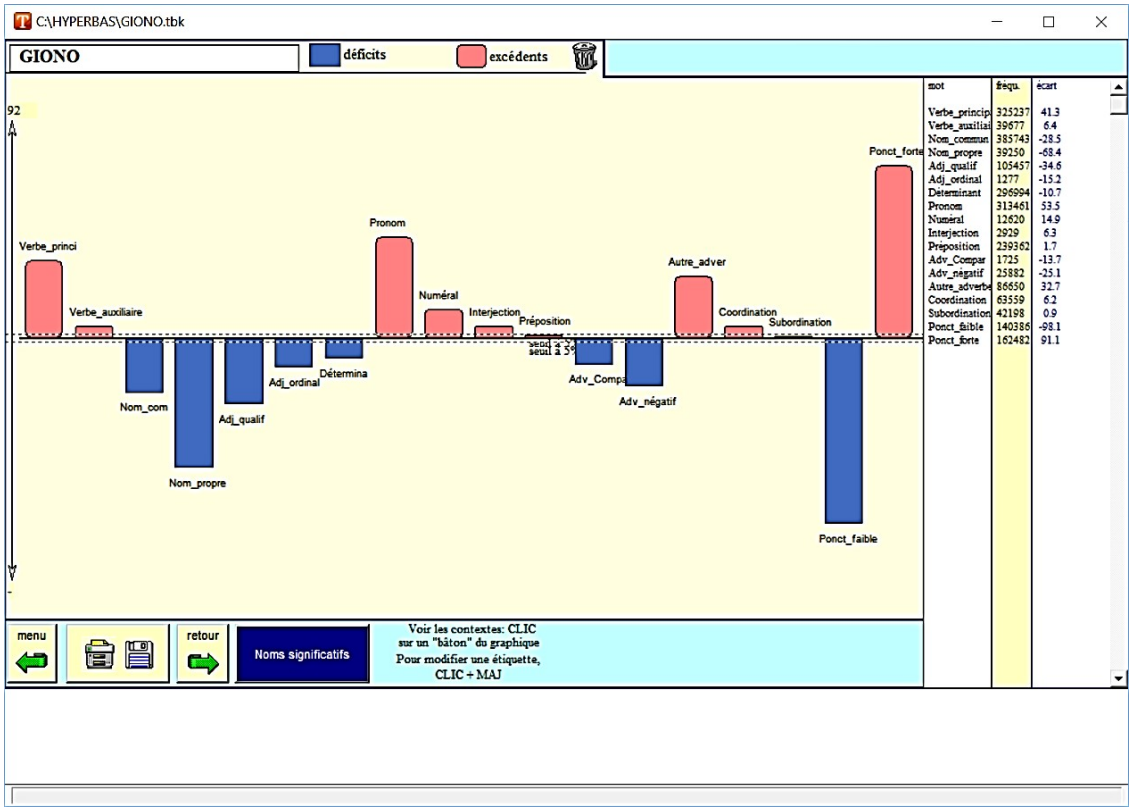

Figure 7- La distribution des catégories grammaticales chez Giono

Giono est du côté du verbe et de l'action. Sa phrase est courte, comme on peut le déduire des ponctuations fortes excédentaires.

[7] Voir note 4.

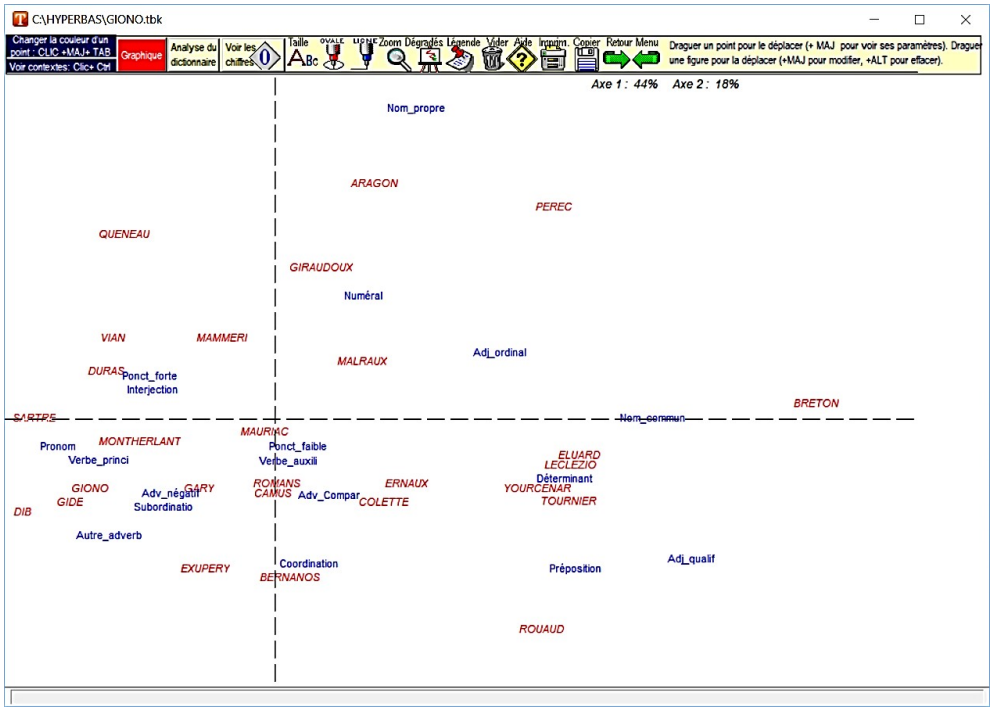

Figure 8- Analyse factorielle des catégories chez les auteurs du corpus

L'analyse factorielle qui projette les affinités entre les catégories grammaticales et les écrivains fait ressortir une bipartition assez claire. Dans les quadrants droits, du côté du substantif et de ses acolytes que sont l'adjectif, le déterminant et la préposition, se regroupent les poètes, les écrivains de la description et ceux de l'analyse. On peut craindre cependant une perturbation exercée par le genre littéraire : les écrivains qui ont une grande production au théâtre, comme Sartre, Gide ou Montherlant, se trouvent attirés du côté du verbe. Cette caractéristique pourrait alors, par conséquent, être davantage un trait générique qu'une propriété du style de l'auteur.

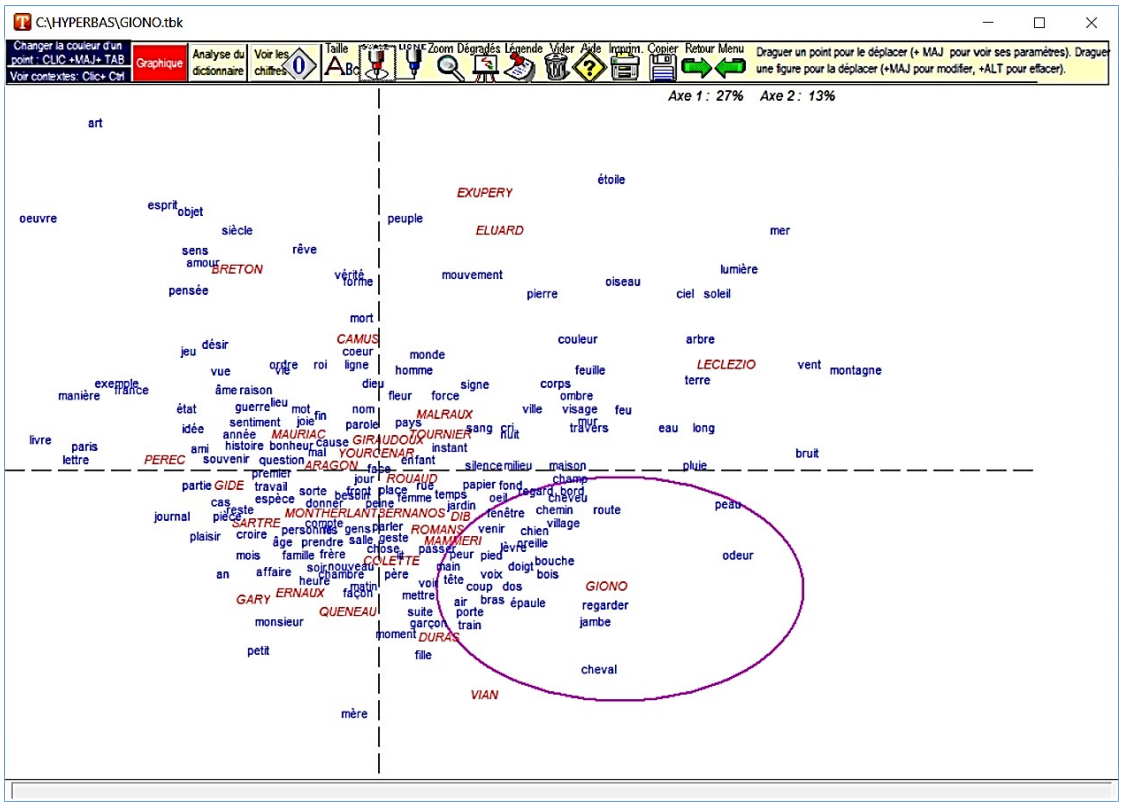

Figure 9- La carte thématique

La carte thématique, ici présentée, regroupe les 300 substantifs et verbes les plus fréquents du corpus. On ne considère que ceux qui sont représentés partout parmi les 300 premiers. Autour de Giono, se dégagent des verbes comme « regarder », « venir » et surtout des noms qui développent le champ lexical du corps : « bras », « dos », « épaule », « oreille », « pied », « tête », par exemple et le vocabulaire pastoral, « chemin », « village », « cheval », tandis qu'autour de Breton s'organise le réseau lexical de l'art et autour de Le Clézio celui de la nature.

Conclusion

Les calculs du deep learning se sont révélés efficaces pour la reconnaissance d'un auteur : les œuvres de Giono ont été toutes reconnues (sauf une à très peu près), après un apprentissage sur 4 romans de Giono et 2 romans de 25 autres écrivains contemporains. L'apport du deep learning se trouve aussi, et c'est sans doute là que réside la plus-value indéniable de cette méthode, dans les liens entre textes et intertextes, dans les liaisons établies entre les écrivains, le plus souvent insoupçonnées. L'étape de l'interprétation, autrement dit l'étape qui consiste à découvrir quels sont les éléments qui ont permis la reconnaissance et l'intertextualité est plus difficile ; un nouvel observable est à déchiffrer, qui réside dans la corrélation entre les niveaux d'analyse – un point-virgule associé à un code grammatical par exemple – puisqu'est privilégié le caractère séquentiel des unités, reliées par-delà les liens d'ordre syntaxique. Le recours aux outils plus traditionnels pour observer les propriétés lexicales et grammaticales de l'objet d'étude s'avère complémentaire du deep learning, en attendant les progrès à venir pour ce qui concerne l'étape de la « déconvolution », en marche au sein de l'UMR 7320.

References

Brunet É. et Vanni L. (2019). Deep learning et authentification des textes, Vol. XXIV(1) http://www.revue-texto.net/docannexe/file/4194/texto_brunetvanni_deep_final.pdf. Consulté le 15/01/2020.

Brunet É., Lebart L. et Vanni L. (2020). Littérature et intelligence artificielle. In Mayaffre D. et al., *L'intelligence artificielle des textes. Points de vue critique, points de vue pratique*, Paris, Champion.

Vanni L., Mayaffre D., Longrée D. ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. JADT 2018, Jun 2018, Rome, Italie. hal-01823560.

Ducoffe M., Precioso F., Arthur, A., Mayaffre D., Lavigne F., et Vanni L. Machine learningunder the light of phraseology expertise : use case of presidential speeches, De Gaulle - Hollande (1958-2016). JADT 2016 : 155-168.

Lebart L. (1997). Réseaux de neurones et analyse des correspondances. *Modulad* (18) : 21-37.

Rastier F. (2007). Passages. *Corpus* (6) : 25-54.

Roe G. (2014), L'étude littéraire à l'ère du numérique : du texte à l'intertexte dans les "digital humanities", *Philologie im Netz Beiheft* (7) : 85-111.

Vanni L. *et al*. (2018), Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis, *56th Annual Meeting of the Association for Computational Linguistics*, July 2018, Melbourne, [hal-01804310].

Vanni L., Corneli M., Longree D., Mayaffre D., Precioso F. (2020). Hyperdeep : deep learning descriptif pour l'analyse de données textuelles. JADT.