DOI: http://dx.doi.org/10.4995/XXVColloqueAFUE.2016.3054

Analyse textométrique et lexicométrique de l'eau dans *Manon des Sources* de Marcel Pagnol

De Oliveira, Ana Paula

Universidad de Salamanca, AnaDeOliveira@usal.es

Resumen

Mediante las nuevas inteligencias artificiales, intentaremos medir el impacto de las palabras que derivan del campo léxico del agua en el conjunto de la obra de Manon des Sources *de Marcel Pagnol. Se tratará de un estudio lexicométrico y textométrico generado por un programa de análisis lingüístico dónde el léxico y su contexto serán puestos en adecuación para dar lugar a unas interpretaciones precisas. Nuestro interés se basará concretamente en un estudio comparativo de las diferentes potencialidades de la maquina frente al razonamiento humano. Intentaremos, a lo largo de este análisis, descubrir con qué nivel de frecuencia y hasta qué punto nuestro artificio es capaz de analizar concretamente y de forma coherente un extracto de la obra. Nuestro objetivo final será de contrastar los diferentes puntos de divergencia como los de convergencia, con el fin de demostrar en qué medida nuestra herramienta puede ser útil y eficaz en el marco de la enseñanza de la lengua y de la literatura francesa.*

Palabras clave : *Lexicometría ; textometría ; TXM ; agua ; fuente.*

Résumé

Grâce aux nouvelles intelligences artificielles, nous tenterons de mesurer l'impact des mots découlant du champ lexical de l'eau dans l'ensemble de l'ouvrage de Manon des Sources *de Marcel Pagnol. Il s'agira d'une étude lexicométrique et textométrique générée par un programme d'analyse linguistique où le lexique et son contexte seront mis en adéquation pour donner lieu à des interprétations précises. Notre intérêt se portera concrètement sur une étude comparative des différentes potentialités de la machine face au raisonnement humain. Nous tâcherons, au cours de cette analyse, de découvrir avec quel niveau de fréquence et jusqu'à quel point notre artifice peut être capable d'analyser correctement et de façon cohérente un extrait de cet ouvrage. Notre objectif final sera celui de constater les différents points de divergence tout comme ceux de convergence afin de démontrer dans quelle mesure notre outil peut être utile et efficace dans le cadre de l'enseignement de la langue et de la littérature française.*

Mots-clés : *Textométrie ; lexicométrie ; TXM ; eau ; source.*

Abstract

Thanks to the artificial intelligence, we will try to measure the impact of words from the lexical field of the water in Manon des Sources *of Marcel Pagnol. It will be about a lexicometric and textometric study generated by a software of linguistic analysis in which words and its context will be study in order to achieve the most precise interpretation. Our aim is to carry out a comparative study of the potential of the computer and the human reason. During this study, we will try to discover the frequency and the effectiveness of the machine in analysing correctly and with coherence extracts of the book* Manon des Sources*. Our major goal is to show the points of convergence and divergence in order to verify how useful and effective the tool could be in the area education, and more precisely, in teaching French language and literature.*

Keywords : *Textometry ; lexicometry , TXM , water , spring.*

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Introduction

Cette étude s'inscrit dans le contexte d'une expérimentation lexicométrique et textométrique autour de l'eau dans l'œuvre de *Manon des Sources* de Marcel Pagnol. Elle vise principalement les différentes approches qui peuvent être réalisées dans le cadre d'un enseignement de la langue ou de la littérature françaises.

1. Hypothèses de recherche

1.1. Contextualisation

La textométrie est une analyse strictement appliquée au texte et la lexicométrie une analyse strictement appliquée au lexique. La textométrie extrait les caractéristiques significatives des données textuelles (terminologie, étiquetage morphosyntaxique), les attirances contextuelles des mots (concordances, cooccurrences), la linéarité et organisation interne du texte et les contrastes intertextuels (mesure statistique fiable d'un mot dans un texte et repérage des mots et des phrases caractéristiques d'un texte).

L'interprétation des calculs se fonde sur des indicateurs chiffrés mais aussi sur l'examen systématique des contextes, maintenant facilité par des liens hypertextes pertinents (Heiden *et al.*, 2008). La textométrie exige une vue globale des textes, mais aussi une consultation des textes locaux puisqu'il s'agit d'analyses quantitatives et qualitatives.

Elle démontre ainsi les nouvelles possibilités de lecture proposées par les corpus numériques. Ces intelligences artificielles sont indispensables pour réaliser des analyses complexes sur des données textuelles d'un large corpus.

1.2. À propos de notre analyseur textuel

Notre analyseur textuel est la plateforme TXM car elle combine des techniques puissantes et originales pour l'analyse de corpus de textes de grands volumes. Elle permet de réaliser des opérations très variées et complexes lors de l'analyse d'un corpus de textes numérisés. De plus, elle présente l'avantage d'être open-source.

Les différentes constructions auxquelles nous ferons appel au cours de notre étude seront les suivantes :

- Concordances à partir des propriétés des mots.
- Calcul du vocabulaire d'ensemble d'un corpus.
- Calcul de la liste des mots apparaissant de façon préférentielle dans les mêmes contextes qu'un motif lexical complexe (cooccurrents statistiques) ;
- Calcul des mots ou des propriétés de mots particulièrement présents dans une partie du corpus (spécificités statistiques) ;
- Application automatique d'outils de traitement automatique de la langue (TAL) sur les textes afin d'obtenir un étiquetage morphosyntaxiquement.
- Exportation de certains résultats au format CSV pour les listes et au format SVG pour les graphiques.

1.3. La textométrie, une approche originale dans notre contexte de travail

Il nous semble important de préciser que, dans un contexte de travail comme l'enseignement du Français Langue Étrangère, cette approche méthodologique des sciences humaines qui envisage les textes comme des données organisées qui peuvent être analysées à travers un traitement informatique, représente une certaine innovation car l'analyse complexe de données nous permet d'extraire des informations sur lesquelles nous pouvons nous interroger sans même avoir lu l'ouvrage, ce qui est une approche différente dans le contexte de l'enseignement.

Nous pouvons ainsi proposer d'autres parcours de lecture sur des corpus de texte par l'observation de la fréquence ou la disposition des mots et leur contextualisation. Cela nous permet d'observer des contrastes et des similitudes. L'ensemble du corpus nous apporte des connaissances.

Il s'agit d'une première prise de contact synthétique d'un corpus à travers lequel nous allons nous interroger sur certains phénomènes comme les répétitions, les mots-clefs, l'usage des temps verbaux, les pronoms personnels, etc. D'entrée, lorsque nous parcourons notre corpus, la textométrie nous permet de pointer sur des éléments curieux, singuliers qu'il faudra ensuite interpréter car il y a une multiplicité de résultats et d'interprétations possibles.

Cela ne nous dispensera en aucun cas de lire l'ouvrage, mais cela nous permettra une approche différente où il sera question de procéder à une analyse comparative entre le traitement automatique et le traitement manuel de l'information.

2. *Manon des Sources* de Marcel Pagnol comme objet d'étude

2.1. Description générale du corpus

Notre corpus est le texte de *Manon des Sources* dans son intégralité. Il s'agit du tome II de « L'eau des collines ». Cet ouvrage est, par excellence, un hymne à la Provence et à la nature où l'eau y est décrite comme l'élément principal de la vie. Que ce soit les personnages, le village ou encore les collines, tout converge autour de l'eau. L'eau y symbolise la fertilité, la sociabilité, la pureté, l'érotisme, la bénédiction, les croyances, le renouveau, mais aussi et inversement la trahison, le péché, le silence, la solitude ou encore la mort. L'eau y adopte tour à tour des visages différents et, en plus d'y être à la fois « eau », elle y est aussi « source ». Ce qui apporte encore des nuances concernant son origine et donc ses vertus, mais aussi sa symbologie.

La source est l'origine, mais aussi la principale source de vie car elle peut être bue. L'eau qui étanche la soif, que ce soit celle du corps ou de l'esprit comme la soif de vérité, de connaissance ou même de vengeance. L'eau de la source est claire, pure, fraîche et thérapeutique. Manon est très fortement reliée à l'élément de l'eau, elle en est presque la métaphore, contrairement aux autres éléments tels que le feu, l'air ou encore la terre qui semblent parfois lui échapper. Cet ouvrage est donc articulé sur deux oppositions : l'eau comme métaphore de la vie, mais aussi, par extension, comme métaphore de la mort.

Cette étude porte principalement sur l'analyse du lexique de l'eau et de la source dans *Manon des Sources*. Il s'agit d'une étude textométrique et lexicométrique car le lexique et son contexte vont être sollicités.

2.2. Structure générale du corpus

Nous allons nous interroger sur l'importance de l'eau dans cet ouvrage et donc sur la fréquence de ses entrées. Concernant les statistiques générales, *Manon des Sources* se compose de 100.295 mots, ce qui est un volume important qui requiert l'aide de la machine pour une quelconque analyse approfondie de cette envergure. Par la suite, nous interpréterons les résultats obtenus par la machine face à nos propres résultats en analysant les variations et hypothèses de contrastes.

Grâce à TXM, nous allons donc pouvoir observer le lexique et dégager les premières impressions concernant le corpus. Pour ce faire, nous lançons une requête de fréquences afin de connaître les mots qui ont le plus d'occurrences au sein de notre corpus. Nous balisons notre requête à un minimum de 10 entrées (inclus).

Dans un premier temps, nous remarquons que les lemmes (« frlemma ») les plus fréquents du corpus correspondent à la liste constituée par le linguiste et lexicologue Etienne Brunet des lemmes les plus fréquents de la langue française. Dans notre corpus, nous avons, en effet, dans l'ordre de fréquence, ces trois lemmes : « de » (2685), « la » (2120), « et » (2010)[1] ainsi que de très nombreuses prépositions ou articles. Ce qui paraît évident et révélateur concernant la syntaxe de la langue française.

Cependant, et étant donné que ces résultats de requête sont plutôt des éléments à caractère général, nous optons pour une recherche plus minutieuse qui pourra nous apporter davantage d'informations. Nous choisissons donc de nous intéresser aux noms car ce sont des lemmes qui portent un sens.

[1] Cf. Tableau 1.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

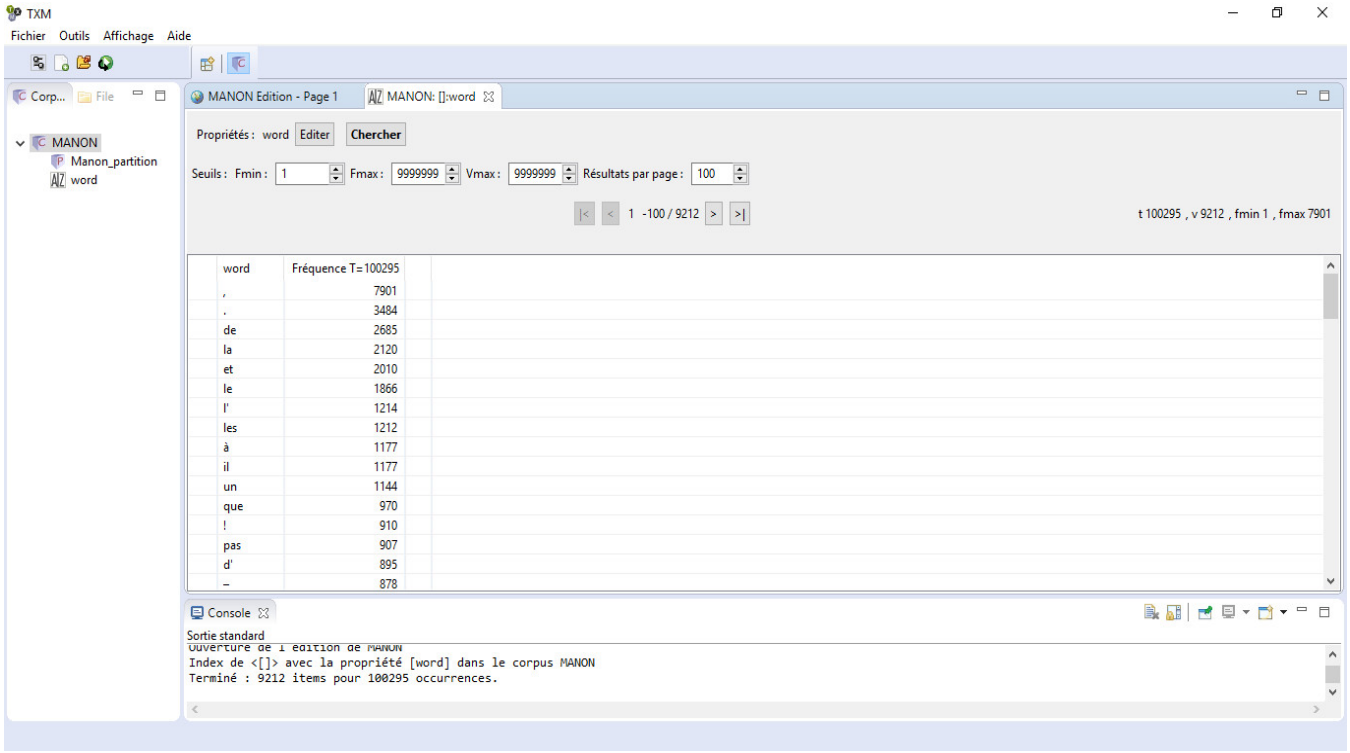


Fig. 1. Fréquences des lemmes dans l'ensemble du corpus

2.3. Analyse générale du corpus

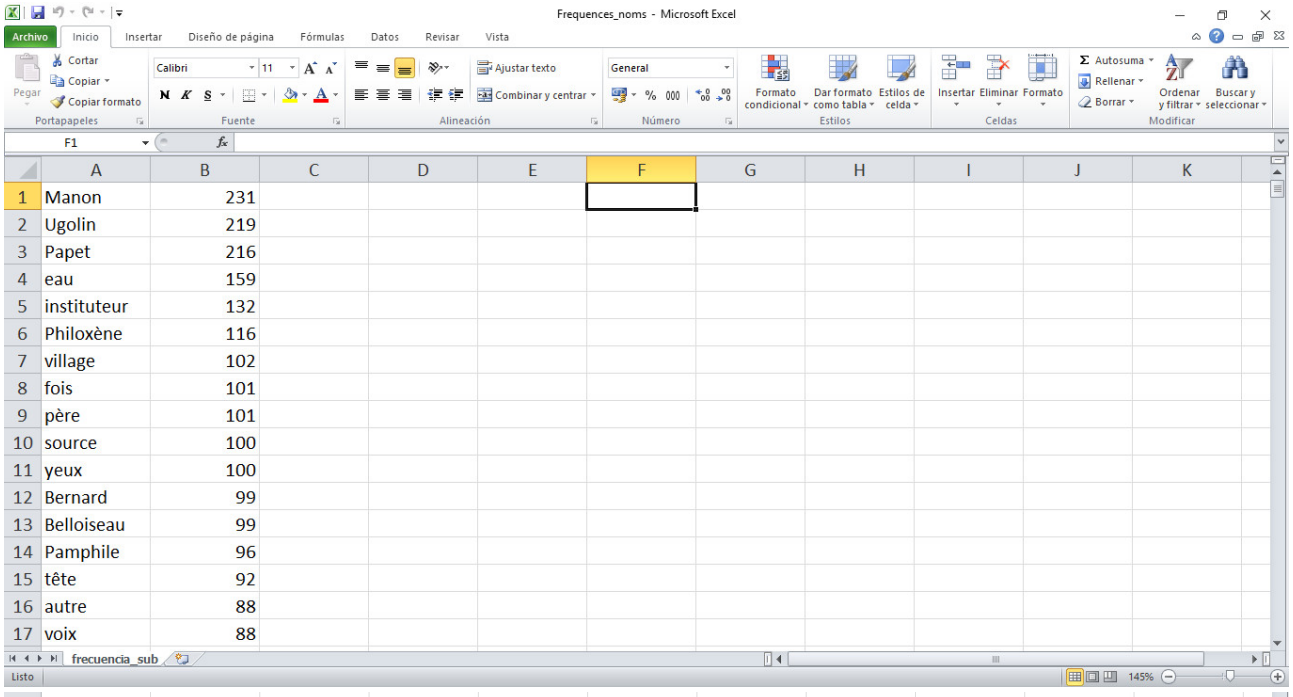
Par la suite, nous lançons donc une requête concernant les noms et nous nous apercevons rapidement que les premiers lemmes qui ont le plus de fréquences sont les noms propres. De fait, la plateforme en a extrait la « hiérarchie » suivante : Manon, Ugolin, Papet, instituteur[2]. Nous corroborons effectivement sur le fait que le personnage principal de l'œuvre est Manon, suivi des trois autres personnages.

Les autres substantifs, par ordre d'occurrence, sont les suivants[3] : village (102 entrées), fois (101), père (101), yeux (100), tête (92), autre (88), voix (88), temps (85), coup (79), jour (75) [+ jours (49) = (124)] , heures (68), matin (63), bras (62), fille (60), œillets (59), homme (54), pauvre (53), mère (53), chien (53), mains (53) [+ main (50) = (103)], belle (52), vieux (50), curé (49), soir (48), gens (48), maison (47), beau (46), nuit (46), place (46), or (46), chèvres (46), fontaine (46), chose (45), bassin (45), roche (44), jeune (44), été (42), mort (42), air (41), pierre (41), silence (41), moment (41), bout (39), table (39), barre (39), bonne (38), peine (38), soleil (38), Dieu (38), personne (37), vieillard (37), visage (37), derrière (36), cause (36), vue (36), rire (35), vallon (35), monde (35), maire (35), ingénieur (35), bruit (34), seul (33), cheveux (33), collines (33) [+ colline (33) = (66)], ville (33), bord (33), enfants (33), couteau (33), fleurs (32), mal (32), vieille (32), pieds (32), travers (32), pain (32), chapeau (32), épaules (31), heure (31), fond (31), femme (31), boulanger (31), tour (30), passage (30), pierres (30).

[2] Cf. Tableau 2.

[3] Dans un premier temps, nous nous intéressons uniquement aux noms ayant un résultat de fréquences allant jusqu'à 30 (inclus) afin d'en extraire les premières hypothèses.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA



	A	B	C	D	E	F	G	H	I	J	K
1	Manon	231									
2	Ugolin	219									
3	Papet	216									
4	eau	159									
5	instituteur	132									
6	Philoxène	116									
7	village	102									
8	fois	101									
9	père	101									
10	source	100									
11	yeux	100									
12	Bernard	99									
13	Belloiseau	99									
14	Pamphile	96									
15	tête	92									
16	autre	88									
17	voix	88									
18	temps	85									
19	coup	79									
20	jour	75									
21	heures	68									
22	matin	63									
23	bras	62									
24	fille	60									
25	oeillets	59									
26	homme	54									
27	pauvre	53									
28	mère	53									
29	chien	53									
30	mains	53									
31	belle	52									
32	Anglade	51									
33	main	50									
34	vieux	50									
35	Soubeyran	49									
36	jours	49									
37	curé	49									
38	soir	48									
39	gens	48									
40	maison	47									
41	beau	46									
42	nuit	46									
43	place	46									
44	or	46									
45	chèvres	46									
46	fontaine	46									
47	chose	45									
48	bassin	45									
49	roche	44									
50	jeune	44									
51	été	42									

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

52	Casimir	42
53	mort	42
54	air	41
55	pierre	41
56	silence	41
57	moment	41
58	Baptistine	40
59	bout	39
60	table	39
61	barre	39
62	bonne	38
63	peine	38
64	soleil	38
65	Magali	38
66	Dieu	38
67	personne	37
68	vieillard	37
69	Victor	37
70	visage	37
71	derrière	36
72	cause	36
73	vue	36
74	Romarins	35
75	rire	35
76	vallon	35
77	monde	35
78	maire	35
79	Ange	35
80	ingénieur	35
81	bruit	34
82	seul	33
83	Éliacin	33
84	cheveux	33
85	collines	33
86	ville	33
87	colline	33
88	bord	33
89	enfants	33
90	couteau	33
91	fleurs	32
92	mal	32
93	Aubagne	32
94	vieille	32
95	pieds	32
96	travers	32
97	pain	32
98	chapeau	32
99	épaules	31
100	heure	31
101	fond	31
102	femme	31
103	boulanger	31
104	tour	30
105	passage	30
106	pierres	30

Fig. 2. Fréquences des substantifs dans l'ensemble du corpus

Nous pouvons désormais procéder à une première interprétation de ces résultats qui découlent de notre analyse lexicométrique puisque nous pouvons en extraire différents champs lexicaux et les mettre en relation avec l'ensemble de l'œuvre.

Comme nous l'avons mentionné antérieurement, nous observons le champ lexical des personnages de l'œuvre : Manon, Ugolin, Papet, instituteur, Philoxène, père, Bernard, Belloiseau, Pamphile, fille, homme, pauvre, mère, belle, Anglade, vieux, Soubeyran, curé, gens, beau, jeune, Casimir, Baptistine, bonne, Magali, personne, vieillard, Victor, maire, Ange,

ingénieur, Éliacin, enfant(s), Aubagne, vieille, femme, boulanger, femmes, petits, petites, bergère, mécréants, Galinette, hommes, Bicou, Amélie, bossu, foule.

Puis, nous pouvons constater le champ lexical de la nature et des éléments : eau, source, œillets, roche, air, pierre(s), soleil, vallon, colline(s), fleurs, ciel, pluie.

Le champ lexical des lieux : village, maison, place, fontaine, bassin, Romarins, ville, Plantier, terrasse, ferme, Ombrées, café, église, cimetière.

Le champ lexical de la temporalité ou de la fréquence : fois, temps, jour(s), heures, matin, soir, nuit, été, moment, heure, mois, dimanche, habitude.

Le champ lexical de la morphologie de l'humain : yeux, tête, voix, bras, main(s), visage, vue, rire, cheveux, pied(s), épaules, larmes, regard, sourire, dos.

Tous ces différents éléments nous donnent déjà des pistes pour une possible interprétation et compréhension de l'œuvre.

3. L'eau dans *Manon des Sources* de Marcel Pagnol

3.1. Analyse lexicométrique

Néanmoins, et étant donné notre thématique d'étude, nous reviendrons sur le premier substantif qui apparait dans notre liste (si on omet les noms propres), c'est-à-dire, le substantif « eau » avec 159 fréquences[4]. Ce qui nous conforte sur la pertinence de notre corpus concernant notre contexte d'étude.

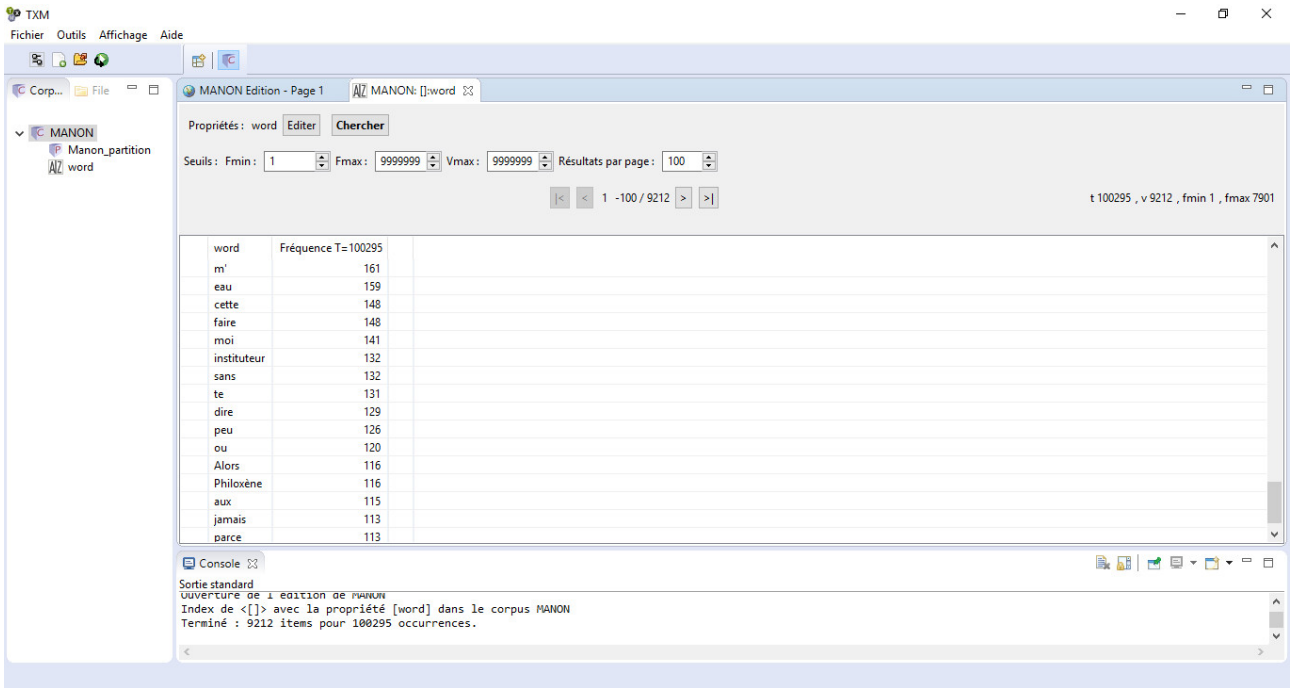


Fig. 3. Fréquence du substantif « eau » dans le corpus

Soulignons tout de même que le substantif « source » apparait à son tour à la septième position (si on omet les noms propres) avec 100 fréquences[5]. Ce qui nous permet de constater que le champ lexical de l'eau atteint un total de 259

[4] Cf. Tableau 3.

[5] Cf. Tableau 4.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

fréquences si nous additionnons ces deux résultats. La machine nous indique donc l'importance de cet élément dans l'ouvrage. Nous corroborons ce premier résultat.

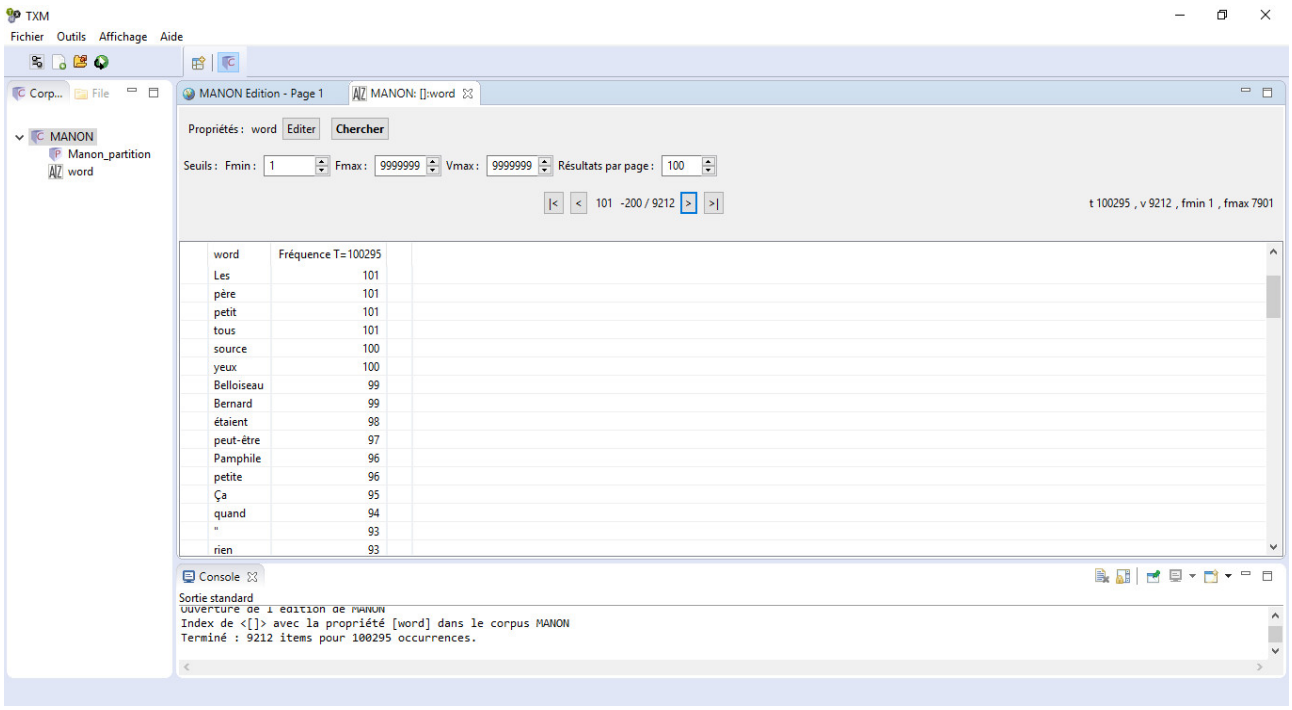


Fig. 4. Fréquence du substantif « source »

3.2. Analyse textométrique

Suite à cette première analyse lexicométrique qui nous permet d'observer les occurrences lexicales, il nous faut observer les résultats de l'analyse textométrique qui vont nous permettre d'observer le contexte dans lequel évolue ce substantif « eau » car c'est grâce au contexte que nous allons interpréter le sens et la fonction de l'eau dans le récit.

Dans un premier temps, nous pouvons diviser cette analyse textométrique en deux catégories. D'un côté « l'eau » dans un contexte positif (en bleu) et de l'autre « l'eau » dans un contexte négatif (en rouge).

Concernant son aspect positif, nous pouvons en extraire des concordances telles que « l'eau pure », « il baigna d'eau froide son visage», « l'eau fraîche », « l'eau brillante », « l'eau limpide et musicale », « l'eau claire », « l'eau glacée », « l'eau miroitante », « l'eau bénite », « l'eau en abondance », « l'eau si pure, si abondante, si constante », « l'eau délivrée », « l'eau parut », « l'eau arrive », « l'eau revenue ».

En référence à son aspect négatif, nous obtenons les concordances suivantes : « l'eau jaillissante qu'elle avait si cruellement refusée au meilleur des hommes », « l'eau des collines, celle qui aurait pu sauver son père », « sans l'eau, tout est foutu ! », « le village se trouve totalement privé d'eau », « si demain je n'ai pas mon eau, je viendrai foutre le feu à la baraque ! », « je ne veux pas prier pour l'eau des criminels qui ont volé celle de mon père », « comment ils ont fait pour voler l'eau de votre père ? », « son père a manqué d'eau toute sa vie, et c'est peut-être á qui l'a tué », « sans eau, cette ferme ne valait rien », « il est allé chercher l'eau au Plantier, et il est mort à la peine », « […] pour les punir ; seulement pour leur couper l'eau, il a été forcé de couper la nôtre ! », « un homme se tuait à transporter de l'eau avec sa femme et ses enfants ? », « le vieux Médéric annonçait son intention, si l'eau ne revenait pas, de se retirer en ville », « le criminel veut réparer sa faute, l'eau reviendra », « […] vont crever les œillets parce que l'eau ne reviendra jamais et ce village aussi, il va crever », « le village n'avait plus d'eau, son père était vengé, et la brûlante sécheresse […] », « l'eau n'est pas revenue, et au village, ça va mal », « […] sont en train de crever, et l'eau du camion, elle a un goût pas bien naturel », « et puis si l'eau ne revient pas, moi aussi j'irai là-bas ».

Nous pouvons, dans un second temps, réaliser le même type d'analyse grâce au synonyme du mot « eau » ; c'est-à-dire « source ».

D'un côté, nous relevons les concordances considérées comme étant positives (en bleu) : « l'arroser à profusion, grâce à l'intarissable source », « nous avons une source dans la cuisine, une source absolument pure, et glacée », « nous avons cherché cette source et nous avons eu de la chance ; nous l'avons trouvée ».

D'un autre côté, nous relevons les concordances ayant un caractère négatifs (en rouge) : « la source perfide », « le mauvais, c'est la source », « ceux qui avaient gardé le secret de la source, et qui voyaient mourir la leur », « la source ne coule plus ! », « ils ont bouché la source, voilà la vérité », « fallait pas se faire d'illusions : la source était morte, et le village allait forcément se dépeupler », « […] et se demandait si l'arrêt de la source pendant plus d'une semaine ne l'avait pas désamorcée à jamais ».

3.3. Vers une possible synthèse

Cette simple analyse nous permet également de procéder à des suppositions par rapport à l'œuvre dans son ensemble. Effectivement, les différents éléments lexicaux ayant été extraits par la machine et identifiés dans leur contexte, nous pouvons désormais procéder à des suppositions par rapport à l'œuvre. Ce type d'analyse nous permet de survoler l'œuvre et d'émettre de premières hypothèses. Par exemple, nous percevons que l'action se situe dans un village où l'eau est l'élément central puisqu'elle est source de vie. Nous comprenons que l'eau vient à manquer dans ce village et que cela provoque le désespoir des habitants qui tentent de trouver des solutions voire des coupables. Toute l'intrigue s'articule donc autour de l'eau et du triangle amoureux qui existe entre les personnages.

Cette analyse lexicométrique et textométrique, nous permet donc, par extension, d'obtenir les différents éléments nécessaires pour procéder à une synthèse de l'œuvre. Nous rappelons, néanmoins, qu'il s'agit là d'interprétations et que nous devrons, par la suite, contraster ces résultats avec une lecture complète de l'œuvre.

Conclusion

En guise de conclusion, nous pouvons donc affirmer que notre plateforme est capable de relever, avec un niveau de pertinence assez élevé, de nombreuses données qui vont nous permettre d'interpréter l'œuvre dans son ensemble car il existe de nombreux points de convergence entre la lecture automatique et la lecture manuelle.

Cependant, il peut exister quelques points de divergences, notamment dans les détails car la machine va extraire principalement les fréquences et il va nous falloir lire l'ouvrage dans son ensemble pour comprendre certains détails ou subtilités. Nous pouvons citer, par exemple, le dénouement de l'œuvre ou le personnage du Papet découvre qu'il est finalement le grand-père de Manon. Ce fait, qui est pourtant l'un des plus importants de l'ouvrage, ne va pas être perçu par la machine.

Néanmoins, dans un contexte d'enseignement de la langue ou de la littérature, cette première approche totalement différente peut se révéler être très stimulante pour les apprenants car ils vont devoir remettre en question les nombreux éléments obtenus par la machine et, il va s'en dire que, étant actuellement dans l'ère du numérique, remettre en question l'information, n'est-ce pas là l'une des principales perspectives d'enseignement ?

Références bibliographiques

HEIDEN, Serge (2010). « The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme ». Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto et Yasunari Harada (éds.). Dans : *24th Pacific Asia Conference on Language, Information and Computation–PACLIC24* (p. 389-398). Sendai : Institute for Digital Enhancement of Cognitive Development, Waseda University. <https://halshs.archives-ouvertes.fr/halshs-00549764> [Consulté le 26 novembre 2016].

HEIDEN, Serge ; MAGUE, Jean-Philippe et PINCEMIN, Bénédicte (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement ». Sergio Bolasco, Isabella Chiari et Luca Giuliano (éds.). Dans : *Proc. of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010)* (Vol. 2, p. 1021-1032). Rome : Edizioni Universitarie di Lettere Economia Diritto. <https://halshs.archives-ouvertes.fr/halshs-00549779/fr> [Consulté le 26 novembre 2016].

LABORATOIRE ICAR – ENS DE LYON. *Présentation projet Textométrie.* http://textometrie.ens-lyon.fr/spip.php?rubrique96. [Consulté le 26 novembre 2016].

PAGNOL, Marcel (1988). *L'eau des collines*. Tome II. *Manon des Sources*. Paris : Éditions de Fallois.

PAGNOL, Marcel (1993). *L'eau des collines*. Tome II. *Manon des Sources*. Paris : Éditions Relié.