PANORAMA D'UN CORPUS MILLIONNAIRE. GÉNÉRALITÉS DU CORPUS JULES VERNE*

José Gregorio Parada**
Universidad de Los Andes, Mérida, Venezuela
Université Sophia Antipolis de Nice, France.

RESUMEN

Más de 6 millones de palabras componen el corpus Julio Verne que hemos establecido. Gracias a los dispositivos informáticos de los que disponemos en la actualidad, puede hacerse enfoques estadísticos para corpus extensos bastante fiables y objetivos y, en consecuencia, análisis imparciales y objetivos. Para esto recurrimos a la logometría como ciencia de la estadística lexical con el fin de hacer el inventario de la producción literaria del célebre escritor francés y analizarlo bajo la lupa de la estadística, hecho que podría abrir nuevos caminos para el estudio de la producción intelectual y discursiva de otros autores. Este estudio realiza un corte transversal del texto, una especie de cirugía que nos

* **Artículo recibido el 27 de junio de 2010 y aceptada su publicación el 20 de septiembre de 2010.**

** Licenciado en Letras, Mención Literatura Hispanoamericana y Venezolana, egresado de la Universidad de Los Andes (1996). Obtuvo el diploma de Maîtrise d'Espagnol y el DEA en Literaturas Nacionales Comparadas francesas en la Universidad François Rabelais de Tours, Francia. Actualmente realiza estudios de doctorado en la Universidad Sophia Antipolis de Niza, Francia. Profesor Agregado del Departamento de Francés de la Escuela de Idiomas Modernos. Universidad de Los Andes, Mérida Venezuela. Dirección electrónica:: josegparada@hotmail.com; josegparada@ula.ve

permitirá conocerlo desde el interior. El programa Hyperbase permite este acercamiento con resultados bastante precisos. En este artículo abordamos un reducido número de nociones que pueden desprenderse de Hyperbase: Riqueza lexical, hápax, crecimiento lexical, frecuencias, progresión y regresión de palabras y distancia lexical.

PALABRAS CLAVE: Logometría, Julio Verne, estadística lexical.

RÉSUMÉ

Plus de 6 millions d'occurrences constituent le corpus Jules Verne que nous avons établi. Grâce aux outils informatiques dont nous disposons dans l'actualité, nous pouvons faire pour les gros corpus des approches statistiques très fiables et, par conséquent, des analyses impartiales et objectives. Dans cette démarche nous faisons appel à la logométrie comme science de la statistique lexicale afin de faire l'état de lieu de la production littéraire du célèbre écrivain français et l'analyser sous le regard des statistiques, ce qui pourrait ouvrir de nouvelles voies pour l'étude de la production intellectuelle et discursive d'autres auteurs. Cette étude fait une coupure transversale du texte, une sorte de chirurgie qui nous permettra de le connaître depuis l'intérieur. Le logiciel Hyperbase permet cette approche avec des résultats assez précis. Cet article traite d'un petit nombre de notions qui peuvent résulter d'Hyperbase : la richesse lexicale, les hapax, la croissance lexicale, la fréquence, la progression et al régression de mots et la distance lexicale.

MOTS CLÉS: Logométrie, statistique lexicale, Jules Verne.

ABSTRACT

The corpus Jules Verne that we have established is composed of more than 6 millions words. By using recent technology, new statistical and reliable approaches of large corpus can be made, and in consequence, impartial and objective analyses. "Logométrie", as the science of lexical statistics, is used in this study in order to provide an overview of the literary production of the famous French writer. This fact could open new ways into the studies of the intellectual and discursive production of other authors. This study provides a cross-section of the text, a kind of surgery that will allow us to know it from the inside. The program Hyperbase allows this approach with fairly accurate results. This article deals with a small number of notions that may result from Hyperbase: lexical richness, hapax, lexical growth, frequency, progression and regression of words and lexical distance.

KEY WORDS: Textometry, Jules Verne, lexical statistics.

INTRODUCTION

La logométrie est une science qui permet le traitement statistique d'un corpus. Elle « *traite en effet de façon exhaustive et systématique le vocabulaire d'un corpus. L'analyse est automatisée et porte sur des critères quantifiés. Grâce à l'indexation et aux procédés informatisés, on associe à chacun des mots (occurrence et vocable) du corpus plusieurs valeurs chiffrées qui permettent l'analyse statistique*» (Kastberg, 2006, p. 33).

Pendant les trois dernières décennies, les études logométriques ont expérimenté un développement inusité à tel point de permettre dans l'actualité une approche « millimétrique » du texte.

Un bon nombre d'auteurs ont été réétudiés sous cette nouvelle dimension, notamment par le propre concepteur du programme Hyperbase Étienne Brunet.

Dans cet article, nous présentons, dans un premier temps, les difficultés trouvées lors de la constitution du corpus Jules Verne. Une synthèse de ce corpus permet de comprendre l'étendue de la production vernienne. Dans un deuxième temps, le lecteur trouvera une briève description du logiciel

Hyperbase comme outil capital pour la démarche logométrique. Finalement, nous montrons toute une exposition des résultats possibles à obtenir avec Hyperbase avec les explications pertinentes.

Notre objectif a été de rassembler un maximum possible de textes de Jules Verne et d'en faire une analyse statistique avec des outils technologiques assez performants. Ce travail n'a pas encore été fait de façon exhaustive. Etienne Brunet, a établie une base de données Jules Verne pour de travaux de comparatistique avec d'autres auteurs français sans arriver à des résultats ponctuels et exclusifs sur notre auteur cible.

PROBLÈME

L'ÉTABLISSEMENT D'UN CORPUS ASSEZ COMPLEXE

Etablir le corpus d'un auteur aussi prolifique que Jules Verne (1828-1905) nous a obligé à fouiller toute source informatisée afin de conformer un fichier global contenant la plupart de textes possibles en format numérisé pour leur exploitation statistique. Tombés dans le domaine public, les livres de Jules Verne se trouvent assez facilement répandus sur le web, fait qui a facilité la tâche pendant la collecte des données. Néanmoins, pour certains textes comme *Voyage à Reculons en Angleterre et en Ecosse*, roman qui paraîtra pour la première fois en 1989 chez Le Cherche midi éditeur et dont la version électronique n'était pas disponible dans le web, nous nous sommes servi des moyens divers pour en produire une version numérisée. Plus de quatre-vingts textes ont été récupérés et organisés en ordre chronologique. Il fallait, nonobstant, en faire un choix en fonction des exigences du logiciel à employer pour le traitement statistique et de notre intérêt pour la prose vernienne. En ce sens, nous avons mis appart un livre de poèmes et le peu des pièces de théâtre dont nous disposions, faible échantillon par rapport à la vaste œuvre théâtrale qui a marqué le début de la carrière littéraire de l'auteur. D'autre part, afin d'harmoniser la taille des textes, nous avons regroupé les nouvelles en deux fichiers correspondant à deux périodes différentes et les essais, articles et souvenirs dans un autre fichier classé dans une date moyenne dans la répartition chronologique. Le corpus ainsi constitué apparaît dans le tableau No. 1 proposant l'année supposée ou prouvée d'écriture, le titre du texte, un mot clé pour le reconnaître dans l'ensemble, le genre auquel il appartient et un code de reconnaissance pour les analyses arborées.

Trois genres cohabitent donc dans ce corpus : nouvelles, essais et romans, constituant ceux derniers un pourcentage très représentatif de l'œuvre vernienne.

Le corpus ainsi reconstitué contient 59 fichiers dont 56 représentent des romans, et les trois restants 18 nouvelles et 9 essais. Ce corpus ne contient pas les œuvres dites « remaniées » par son fils Michel, d'autres écrites en collaboration ou celles parues sous sa signature appartenant intégralement à d'autres auteurs.

Tableau No 1. Le corpus Jules Verne

ANNEE	TITRE	MOT	GENRE	CODE
	NOUVELLES 1850-1870 *Un Drame dans les airs -Un drame au Mexique -Martin Paz -Pierre Jean -Maître Zacharius -Hivernage dans les glaces -Le mariage de Mr Anselme des Tilleuls -Joyeuses misères de trois voyageurs en Scandinavie -Le Comte de Chanteleine -Les Forceurs de blocus-Le Humbug.*	NOUV_1	NOUVELLES	01
1859	*Voyage à reculons en Angleterre et en Ecosse*	RECULONS	ROMAN	02
	Cinq semaines en ballon	BALLON	ROMAN	03
	Voyages et Aventures du capitaine Hatteras	HATTERAS	ROMAN	04

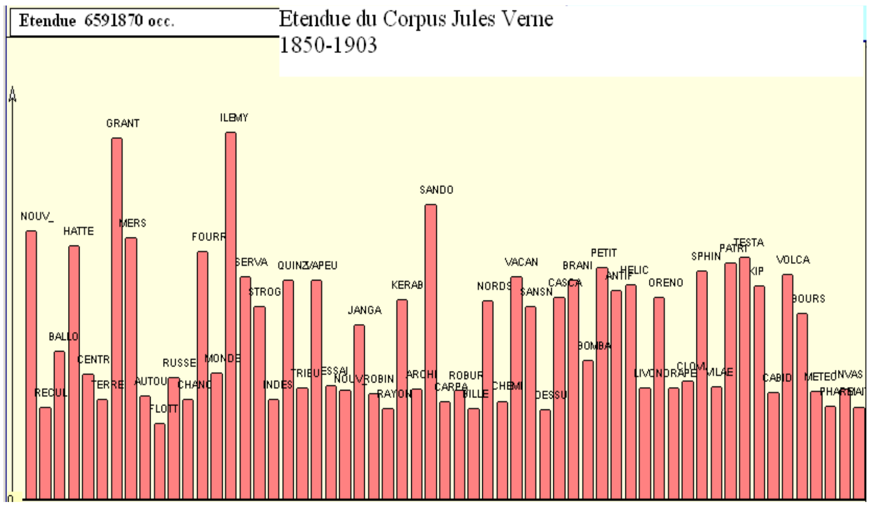
	Voyage au centre de la Terre	TERRE	ROMAN	05
	De la Terre à la Lune. Trajet direct en…	TERRELUNE	ROMAN	06
	Les Enfants du capitaine Grant (R)	GRANT	ROMAN	07
	Vingt mille lieues sous les mers	MERS	ROMAN	08
	Autour de la Lune	AUTOURLU	ROMAN	09
	Une ville Flottante	FLOTTANTE	ROMAN	10
	Les Aventures de trois Russes et de trois Anglais dans l'Afrique australe	3RUSSES	ROMAN	11
	Le Chancellor	CHANCELL	ROMAN	12
	Le pays des fourrures	FOURRURE	ROMAN	13
	Le tour du monde en 80 jours	MONDE	ROMAN	14
	L'Île mystérieuse	ILEMYST	ROMAN	15
	Les Aventures d'Hector Servadac autour du monde solaire	SERVADAC	ROMAN	16
	Michel Strogoff. De Moscou à Irkoutsk	STROGOFF	ROMAN	17
	Les Indes noires	INDES	ROMAN	18
	Un capitaine de quinze ans	15ANS	ROMAN	19
	Les Tribulations d'un Chinois en Chine	TRIBULAT	ROMAN	20
	La Maison à vapeur. Voyage à travers l'Inde septentrionale	VAPEUR	ROMAN	21
	ESSAIS *Salon de 1857 -Edgar Poe et ses œuvres -A propos du Géant -Géographie illustrée de la France et des ses colonies (extrait) -Les Méridiens et le calendrier- Vingt quatre minutes en ballon -Une ville idéale (1875b) -Inauguration du Cirque municipal d'Amiens -Souvenirs d'enfance et de jeunesse.*	ESSAIS	ESSAIS	22

		NOUV_2	NOUVELLES	23
	La Jangada, Huit cents lieues sur l'Amazone .	JANGADA	ROMAN	24
	Ecole de robinsons	ROBINSON	ROMAN	25
	Le rayon vert	RAYON	ROMAN	26
	Kéraban le Têtu	KERABAN	ROMAN	27
	L'Archipel en feu	ARCHIPEL	ROMAN	28
	Mathias Sandorf	SANDORF	ROMAN	29
	Le Château des Carpathes	CARPATH	ROMAN	30
	Robur le Conquérant	ROBUR	ROMAN	31
	Un billet de loterie. Le numéro 9672	BILLET	ROMAN	32
	Nord contre Sud	NORSUD	ROMAN	33
	Chemin de France	CHEMIN	ROMAN	34
	Deux ans de vacances	VACANC	ROMAN	35
	Famille Sans-Nom	SANSNOM	ROMAN	36
	Sans dessus dessous	DESSUS	ROMAN	37
	César Cascabel.	CASCABEL	ROMAN	38
	Mistress Branican	MISTRESS	ROMAN	39
	Claudius Bombarnac	CLAUDIUS	ROMAN	40
	P'tit Bonhomme	PETITBON	ROMAN	41
	Mirifiques aventures de Maître Antifer	ANTIFER	ROMAN	42
	Île à Hélice	HELICE	ROMAN	43
	Un Drame en Livonie	LIVONIE	ROMAN	44
	Le superbe Orénoque	ORENOQ	ROMAN	45

	Face au Drapeau	DRAPEAU	ROMAN	46
	Clovis Dardentor	CLOVIS	ROMAN	47
	Sphinx de Glaces	SPHINX	ROMAN	48
	Le village aérien	VILAERIEN	ROMAN	49
	Seconde patrie	PATRIE	ROMAN	50
	Le testament d'un excentrique	TESTAM	ROMAN	51
	Les Frères Kip	KIP	ROMAN	52
	Les Histoires de Jean-Marie Cabidoulin	CABIDOUL	ROMAN	53
	Le Volcan d'Or	VOLCAN	ROMAN	54
	Bourses de Voyage	BOURSES	ROMAN	55
	La chasse au météore	METEORE	ROMAN	56
	Le Phare du bout du monde	PHARE	ROMAN	57
	L'Invasion de la mer	INVASION	ROMAN	58
	Maître du monde	MAITRE	ROMAN	59

La figure No. 1 présente l'étendue du corpus Jules Verne. Les barres représentent la taille approximative de chaque texte. Par exemple, le texte No 10, *Une ville flottante*, se situe comme le texte le plus court avec 50 556 occurrences, devant *L'Ile mystérieuse* avec 245 927 occurrences. Il est évident de constater que les œuvres les plus longues sont constituées par le trio: *l'Ile mystérieuse*, *Les enfants du Capitaine Grant* et *Mathias Sandorf*. Les plus courtes sont représentées par : *Une ville flottante*, *Sans dessus dessous*, *Un billet de loterie*, *Le rayon vert*, *Le phare du bout du monde* et *Maître du monde*.

Figure No. 1. L'étendue du Corpus Jules Verne



Méthodes et outils informatiques utilisés

Le logiciel que nous utilisons, Hyperbase, a été conçu par Etienne Brunet du laboratoire *Bases, Corpus et Langage* (CNRS-Université de Nice Sophia-Antipolis), associé à l'étiqueteur Cordial (Mayaffre, 2004), et permet quatre traitements de façon intégrale et simultanée du texte brut : mots traités tels qu'ils ont été écrits ; du texte lemmatisé : les mots sont ramenés à leur canon ("vient" = "venir") ; des codes grammaticaux : les mots reviennent à leur catégorie ou fonction grammaticale ("vient" = verbe à la troisième personne du singulier au présent ; et des structures syntaxiques : le discours est ramené à ses enchaînements syntagmatiques ("le ballon s'éleva" = déterminant+nom+verbe…). De ce logiciel, nous pouvons ajouter de manière synthétique qu'il possède deux grandes fonctions : une dite « documentaire » et l'autre statistique.

La fonction documentaire nous donne la possibilité de naviguer autour du texte, de classer ses composants en lemmes et codes et de repérer facilement les passages liés par des traits caractéristiques à la manière d'un moteur de recherche.

La deuxième fonction nous met sur la portée des données statistiques et ses représentations graphiques qui incluent, entre autres, la richesse lexicale, la distance intertextuelle, les fréquences d'utilisation du vocabulaire, et les analyses arborées, méthode conçue par Xuan Luong du laboratoire cité plus haut.

Données et résultats statistiques

La richesse lexicale et les hapax

La richesse lexicale fait le dénombrement des formes différentes relevées dans chaque texte (Brunet, 2006). Le terme richesse lexicale est synonyme de variété lexicale.

Si l'on considère N le nombre des mots d'un texte, et V celui de ses vocables (Muller, 1992), nous pouvons établir un rapport mathématique entre V et N (V/N) dont un résultat proche de l'unité nous montrerait une richesse lexicale majeure par rapport à un résultat proche de zéro.

Le tableau No. 2 montre une partie de ces calculs réalisés par le logiciel avec les écarts réduits. Il faut se rappeler que l'écart réduit augmente en proportion avec la richesse lexicale (Magri, 1995). Le tableau présente donc une première colonne avec les vocables réels de chaque texte, une deuxième avec le nombre théorique que le texte devrait avoir par rapport à sa longueur, les écarts pertinents et le nombre de hapax, c'est-à-dire les vocables ayant une fréquence 1 dans l'ensemble.

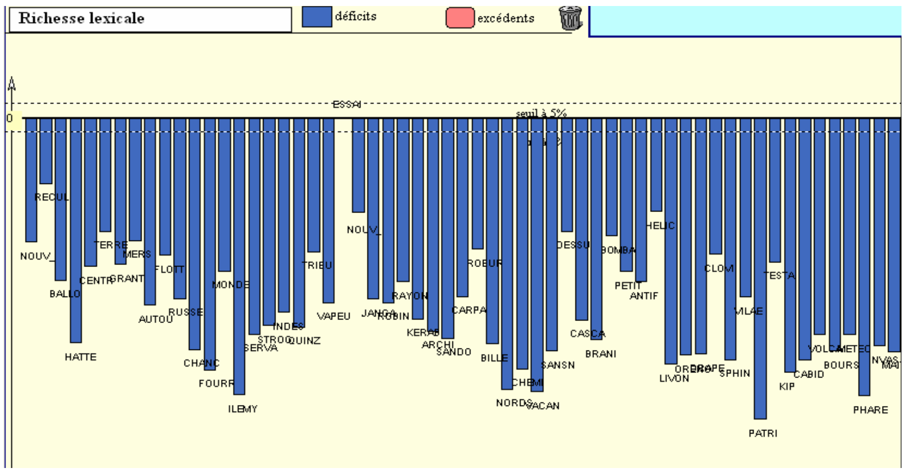
Tableau No 2. Richesse lexicale

n°	réel	théo	écart	réduit	Hapax	réduit	Titre
1	15149	17410	-2261	-17.14	1051	12.02	NOUV_1
2	9009	10325	-1316	-12.95	577	20.53	RECULONS
3	10059	12942	-2883	-25.34	534	6.48	BALLON
4	12489	16915	-4426	-34.03	588	-3.92	HATTERAS
5	9547	11910	-2363	-21.65	522	9.82	CENTRE
6	8564	10686	-2122	-20.53	440	10.24	TERRELUN
7	17146	20162	-3016	-21.24	1023	1.31	GRANT
8	14834	17185	-2351	-17.93	1175	17.59	MERS
9	7612	10850	-3238	-31.09	264	-1.02	AUTOURL
10	7261	9424	-2163	-22.28	227	1.52	FLOTTANT
11	8707	11740	-3033	-27.99	297	-1.91	RUSSES
12	6564	10705	-4141	-40.02	176	-5.90	CHANCEL
13	11860	16743	-4883	-37.74	382	-11.45	FOURRURE
14	8883	11994	-3111	-28.41	266	-4.36	MONDE
15	14318	20326	-6008	-42.14	581	-13.48	ILEMYST
16	11247	15827	-4580	-36.41	475	-5.32	SERVADAC
17	10527	14724	-4197	-34.59	476	-2.10	STROG
18	7405	10708	-3303	-31.92	213	-3.66	INDES
19	11371	15742	-4371	-34.84	538	-2.48	QUINZE
20	8698	11232	-2534	-23.91	353	2.93	TRIBULAT
21	12191	15728	-3537	-28.20	643	1.91	VAPEUR
22	10477	11363	-886	-8.31	975	38.01	ESSAIS
23	9297	11146	-1849	-17.51	576	16.24	NOUV_2
24	10321	14021	-3700	-31.25	622	6.80	JANGADA

A partir des notions et données précédentes, le texte No 1 (Nouv_1), recueil de nouvelles de 1850 à 1870, se présente comme un texte « moins riche » part rapport au texte No. 2 (Reculons) même si le nombre des hapax est plus important dans le premier. L'explication se trouve dans la longueur de deux textes car « *plus un texte est long, plus il a de chances de remployer les mêmes mots* » (Magri, 1995, p.64). D'après ce tableau et ce que nous pouvons constater plus bas dans la figure No. 2, les textes que l'on peut considérer les plus riches sont: *Voyages à reculons en Angleterre et en Ecosse*, *De la Terre à la lune*, *Vingt mille lieus sous les mers*, *Tribulations d'un chinois en Chine*, *Essais*, *Nouvelles 2*, *Robur le Conquérant*, *Sans dessus dessous*, *Claudius Bombarnac*, *Ile à hélice* et *Clovis Dardentor*.

Parmi ces textes, *Essais* constitue le plus varié de tous. Il faut se rappeler que ce fichier contient une variété importante de textes de courte longueur qui se baladent entre l'art, les voyages, les souvenirs, les expériences scientifiques et la critique littéraire. Dans l'ordre inverse, le corpus manifeste la présence de textes « peu variés » comme *L'Ile mystérieuse*, *Nord contre Sud*, *Deux ans de vacances*, *Seconde patrie* et *Le phare du bout du monde*. Même si Seconde patrie n'est pas le texte le plus long, nous constatons pour ce roman et pour une bonne partie d'autres appartenant à la fin de la vie l'auteur un appauvrissement du vocabulaire, facilement détectable par rapport à la longueur des barres, exception faite de *Testament d'un excentrique*, dans lequel l'auteur nous ramène dans un voyage autour des Etats-Unis. *Seconde patrie* se veut une sorte de répétition de la « robinsonnade » qu'il a bien réussi avec *L'Ile mystérieuse*. Le sujet du naufrage et l'île déserte modifiée par la main de l'homme étant un thème presque épuisé depuis le début de sa carrière littéraire. Non sans raison, *L'Ile mystérieuse* se présente comme un texte ayant un lexique de même peu varié à la suite d'autres qui ont fait incursion dans le mêmes champs lexicaux et sujets : *Vingt mille lieues sous les mers* et *Les enfants du Capitaine Grant*, chargés tous les deux du vocabulaire de la mer et la navigation, et le sujet pérenne de l'île déserte, le (s) naufragé(s) et la conformation d'une nouvelle colonie.

Figure No. 2. La richesse lexicale



L'accroissement lexical

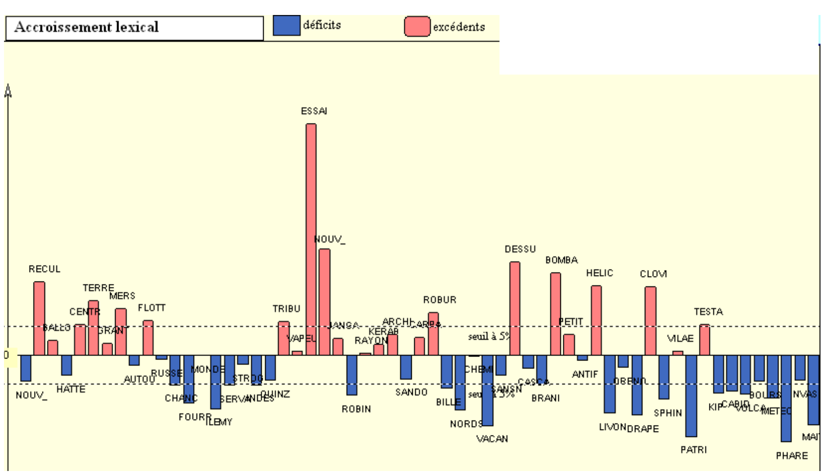
Une autre notion dérivée d'Hyperbase c'est l'accroissement lexical, calcul sous-entendu d'une distribution chronologique des textes. Cette notion met en évidence le nouveau vocabulaire qui apparaît au fur et à mesure d'une lecture linaire supposée du corpus.

Nous remarquons trois tranches chronologiques de « florescence lexicale ». La première va depuis le commencement (1850) jusqu'en 1869, année qui précède une série de romans répétitifs du sujet africain et russe (*Aventures de trois russes et de trois anglais*, *Un capitaine de quinze ans*, *Michel Strogoff*), des naufragés (*Le Chancellor*), du voyage (*Le Tour du monde*), du voyage spatial (*Hector Servadac*), tous exploités auparavant. Certainement, cette période de « faiblesse lexicale », entre 1870 et 1878, a été touchée par des événements troublants dans la vie de Verne dont l'influence ne peut être négligée : mort de son cher cousin Henri Garcet (1870), mort de son père Pierre Verne (1871), accusation de plagiat (1875), difficultés avec son fils Michel (1876, 1877), maladie de son épouse Honorine (1876, 1877), Lettre de protestation contre Verne du grand rabbin de Paris, embarquement forcé de Michel pour les Indes (Soriano, 1978).

A partir de 1878 l'accroissement lexical redevient positif avec *Tribulations d'un chinois en Chine* et se maintient, rares exceptions, jusqu'en 1885, année qui inaugure une nouvelle série de difficultés pour l'auteur dont les gros besoins d'argent se font sentir dès cette même année. Cette courte nouvelle période de baisse s'étend jusqu'en 1888. Comme événements à souligner nous y trouvons : la vente de son yacht le Saint Michel III (1886), l'accident dans lequel son neveu Gaston tire deux coups de révolver contre lui (1886), mort de sa mère Sophie Verne (1887) et vente de sa maison de Chantenay (1887). A partir de cette année, le vocabulaire donne signes de renouvellement jusqu'en 1897, bien évidemment avec certaines exceptions dont la plus importante serait *Seconde patrie* de 1896, année du procès « Turpin » contre Verne.

Une année après la mort de son frère Paul, commence la dernière étape de décroissement lexical qui accompagne le déclin physique et moral de l'auteur jusqu'à sa mort en 1905. Cette courte période est marquée par des événements lamentables : Dégradation de sa santé, vieillissement accéléré, mort de la fidèle gouvernante de la famille (1900), Verne brûle de façon inattendue une bonne partie de ses lettres et documents privés (1900), et mort de Caroline Tronson (son « premier » amour) (1903). (Soriano, 1978).

Figure No. 3. L'accroissement lexical



Les fréquences les plus élevées

Les données proportionnés par Hyperbase pour les 100 premiers mots les plus fréquents, se présentent dans l'ordre suivant : Le rang occupé par le mot (1, 2, 3…), total de la fréquence et le mot proprement dit (Nous n'avons relevé que quelques formes).

Pour les FORMES nous avons :

Tableau No. 3. Les formes

Les signes de ponctuation		
1	508671	-
2	243543	-
9	100222	-
Les Mots		
3	238652	de
4	138652	la
5	136470	le
6	124553	à
7	118141	et
8	103019	l'
10	93151	les
11	91781	il
26	39747	est
31	31447	était
40	24398	je
59	14227	deux
82	7642	capitaine

Ces fréquences deviennent plus claires dans le classement fait pour les lemmes, où l'ordinateur ne tient plus compte de la forme graphique mais de la fonction du mot dans la phrase. Les codes numériques employés par le logiciel sont les suivants : verbe 1, substantif 2, adjectif 3, numéral 4, pronom 5, adverbe 6, déterminant 7, conjonction 8, préposition 9, interjection et autres 0 (ces codes apparaissent à la fin de chaque ligne, juste après le lemme en question).

Pour les LEMMES :

Tableau No. 4. Les lemmes

rang	frq	mot
1	508717	,
4	243543	.
10	100468	-
17	57972	!
5	146512	être_1
11	91883	avoir_1
32	26787	pouvoir_1
33	25646	faire_1
64	9810	heure_2
67	8578	jour_2
74	7739	capitaine_2
49	13801	tout_3
83	7311	grand_3
100	5873	quel_3
52	13032	deux_4
9	113081	il_5
14	86300	se_5
21	46318	qui_5

rang	frq	mot
13	87165	ne_6
22	42664	pas_6
31	28306	plus_6
2	428249	le_7
7	122954	un_7
12	90780	de_le_7
15	74118	ce_7
16	60838	son_7
8	118140	et_8
3	303631	de_9
6	124453	à_9
18	50983	en_9

Il suffit d'ajouter que pour cette liste le rang occupé par certains lemmes diffère de celui proposé par l'ordinateur pour les mots du premier classement. « Être » occupe le 5ème rang grâce à une accumulation de toutes ses formes verbales, suivi d' « avoir » et bien de loin des verbes « pouvoir » et « faire ». Les prépositions « de » et « à », quant à elles, gardent les mêmes places. « Le » dans les lemmes regroupe tous les déterminants articles ce qui le positionne en 2ème rang des lemmes les plus fréquents. Pour les substantifs « heure » et « jour » qui dépassent maintenant le substantif « capitaine », la lemmatisation a dépouillé ces mots de leur accompagnateurs habituels : ce jour-là, ce jour-ci, cette heure-là, cette heure-ci, l'heure.

La progression des mots

L'ordinateur est capable de nous proposer une liste de mots qui sont utilisés de plus en plus dans la ligne du temps de notre auteur (Tableau No. 5). Le graphique qui en dérive est plus parlant évidemment pour les mots présentant un coefficient élevé. Comme cas de figure, des mots tels que « dès » (cf.

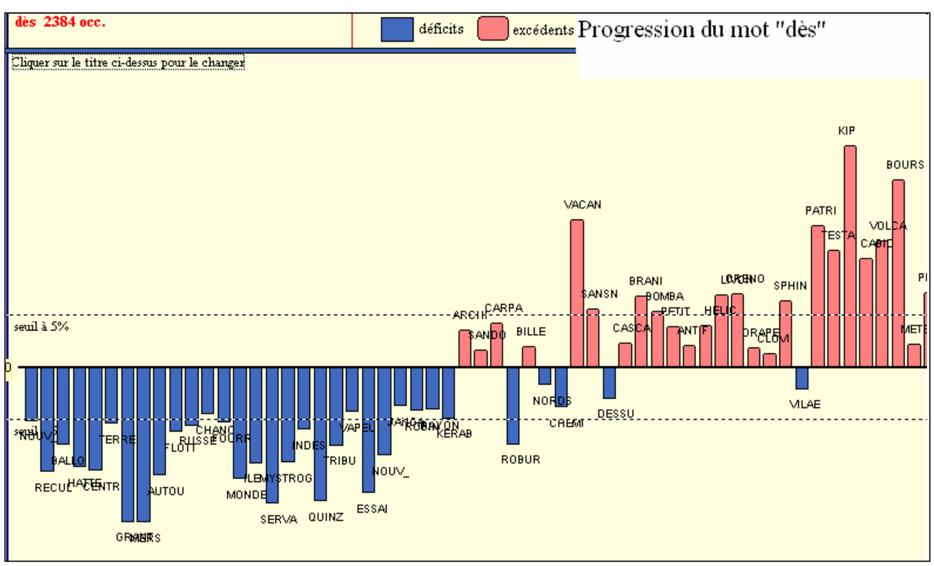
Figure No. 4), « lorsque », « aurait », « n' », « assurément », etc., d'après la liste qui suit, ont été très peu employés en début de carrière pour se manifester, au fur et à mesure, plus fréquemment jusqu'à atteindre un usage presque abusif. Dans cette liste, qui n'est pas exhaustivement représentée ici, nous remarquons des mots correspondant nettement à la mesure du temps (matinée, après, durant, semaine, midi) et un bon nombre de participes passés. La lemmatisation a dévoilé que « début », « matinée » et « lieu » sont les substantifs montrant la progression la plus forte, pendant que pour les verbes l'honneur correspond à « assurer », « concerner » et « effectuer ». Pour le mot « être » présent dans la liste, le lemmatiseur a séparé clairement le verbe du substantif, raison pour laquelle le nombre d'effectifs baisse de manière considérable dans la liste de lemmes.

Tableau No. 5. Mots en progression

Coeff.	Fréqu.	Mot
+ 0.812	2384	dès
+ 0.782	3668	lorsque
+ 0.776	4720	aurait
+ 0.770	36574	n'
+ 0.747	397	assurément
+ 0.739	5087	serait
+ 0.729	17783	si
+ 0.729	1663	vrai
+ 0.722	369	début
+ 0.718	412	matinée
+ 0.717	2369	lieu
+ 0.711	147	tarderait
+ 0.710	9982	après
+ 0.710	2066	lorsqu'
+ 0.695	121	effectuer
+ 0.690	218	quinzaine
+ 0.689	3586	ailleurs
+ 0.685	1121	seraient
+ 0.682	14671	être
+ 0.681	30316	…
+ 0.681	1541	sait
+ 0.680	213	admettant
+ 0.676	288	surplus

Coeff.	Fréqu.	Mot
+ 0.673	2191	eu
+ 0.673	576	attendant
+ 0.671	199	suffirait
+ 0.665	533	déclara
+ 0.664	257	durant
+ 0.663	202	concerne
+ 0.654	666	semaines
+ 0.650	182	concernait
+ 0.645	102	imposait
+ 0.645	99	préférable
+ 0.642	6329	moins
+ 0.639	296	soi
+ 0.637	895	occasion
+ 0.634	63310	que
+ 0.631	15670	y
+ 0.628	1631	mettre
+ 0.620	154	resterait
+ 0.619	844	entendu
+ 0.612	134	multiples
+ 0.611	50537	ne

Figure No 4. Progression du mot « dès »



Mots en régression

Contrairement à ce qui se passe pour certains mots dont un échantillon a été proportionné précédemment, d'autres se comportent de façon régressive. En d'autres termes, ces mots trop employés au début rentrent progressivement en désuétude. Le mot « mais » se trouve à la tête des termes abandonnés graduellement par l'auteur accompagné d'un nombre important d'adverbes en « _ment ». Plus bas nous insérons la liste pour les premiers 40 mots en régression (Tableau No.6). A gauche nous avons placé la liste des formes et à droit celle des lemmes afin de permettre une comparaison.

Tableau No. 6. Mots en régression dans le corpus Verne

Les formes			Les lemmes		
Coeff.	Fréqu.	Mot	Coeff.	Fréqu.	Mot
- 0.791	21190	mais	- 0.791	21190	mais_8
- 0.760	60834	un	- 0.738	2012	bientôt_6
- 0.738	2012	bientôt	- 0.716	245	subitement_6
- 0.716	245	subitement	- 0.695	1610	fort_6
- 0.685	316	tranquillement	- 0.682	5054	peu_6
- 0.680	544	parfaitement	- 0.680	544	parfaitement_6
- 0.669	7560	peu	- 0.675	1086	précipiter_1
- 0.658	788	immense	- 0.662	392	suspendre_1
- 0.634	1110	rapidement	- 0.660	1029	immense_3
- 0.632	2252	air	- 0.640	2483	regarder_1
- 0.621	336	semblaient	- 0.636	18843	dire_1
- 0.611	214	singulièrement	- 0.634	1110	rapidement_6
- 0.610	177	destinée	- 0.629	768	suivant_9
- 0.609	1253	suivant	- 0.611	214	singulièrement_6
- 0.609	823	oeil	- 0.602	310	spectacle_2
- 0.603	303	spectacle	- 0.597	3415	oeil_2
- 0.602	592	terrible	- 0.597	2454	air_2
- 0.602	263	auprès	- 0.597	1456	marcher_1
- 0.597	4279	cependant	- 0.594	497	destiner_1
- 0.595	305	résolut	- 0.593	251	auprès_9
- 0.588	2576	pieds	- 0.588	1329	moyen_2
- 0.587	717	instants	- 0.582	3568	quand_8

Les formes Coeff.	 Fréqu.	 Mot	Les lemmes Coeff.	 Fréqu.	 Mot
- 0.584	50753	une	- 0.577	1675	élever_1
- 0.584	1165	moyen	- 0.577	89	accoutumé_3
- 0.583	3571	quand	- 0.573	218	humide_3
- 0.575	330	apparut	- 0.572	4157	cependant_6
- 0.571	11198	dit	- 0.571	706	ombre_2
- 0.563	692	beaucoup	- 0.564	766	couvrir_1
- 0.559	830	paroles	- 0.563	692	beaucoup_6
- 0.555	127	ému	- 0.560	620	composer_1
- 0.554	1287	corps	- 0.556	106	sinistre_3
- 0.553	202	attentivement	- 0.555	396	agiter_1
- 0.551	593	ombre	- 0.554	1287	corps_2
- 0.549	230	précipita	- 0.553	202	attentivement_6
- 0.545	837	longtemps	- 0.550	175	accent_2
- 0.545	241	immenses	- 0.547	792	briser_1
- 0.542	347	ceci	- 0.544	246	aigu_3
- 0.542	173	promptement	- 0.543	490	particulier_3
- 0.537	133	torrents	- 0.542	173	promptement_6
- 0.536	568	passa	- 0.542	130	mobile_3

La distance lexicale

Pour finir avec cette approche générale de l'œuvre de Jules Verne, nous présentons le panorama de ses textes distribués à partir des calculs proportionnés par l'Analyse Factorielle de Correspondances afin de permettre une vision de l'ensemble et de la distance qui sépare les textes. La figure No. 4 dont les commentaires suffiraient à écrire un article en entier, nous montre quatre axes de distribution des textes : inférieur, supérieur, gauche et droit. Une distribution uniforme et d'une certaine façon logique, expose les textes en montrant leur chronologie. Les titres en rouge étant les premiers, ceux en mauve appartenant à la période moyenne, et finalement ceux en bleu ayant été écrits à la fin. L'axe inférieur est riche donc en textes de la première étape, pendant que l'axe supérieur se réserve de manière écrasante le reste. Les deux recueils de nouvelles, séparés dans le temps se rapprochent nonobstant dans le quadrant inférieur droit, de la même façon que beaucoup de textes attirés par le vocabulaire : *Les Indes noires*, *Le rayon vert*, *Le Tour du monde* et un peu plus loin *Voyage à reculons en Angleterre*, se regroupent autour d'un pôle commun, le Royaume Uni. Les grands voyages en mer se distribuent plus vers l'axe de gauche, et ceux qui se tiennent dans les airs se sont positionnés dans l'axe inférieur, même distribution pour ceux dont l'histoire se passe dans les profondeurs de la terre (*Les Indes noires* et *Voyage au centre de la Terre*). Dans l'axe supérieur se localisent, certaines exceptions faites, les romans dont les voyages se font notamment par la terre. Les décors glaciaux ont été attirés vers la gauche (*Sphinx*, *Fourrures*, *Hatteras*, *Mers*). *Maître du monde*, juste au milieu de la distribution, peut être considéré un texte neutre dans le sens où il n'est pas attiré vers les extrémités bien probablement par la fusion de sujets qu'il implique ; en effet, dans cette histoire un nouveau véhicule inventé par Robur-le-Conquérant sert d'automobile, avion et sous-marin. Sans vouloir approfondir dans le sujet, la distribution présentée dans le tableau qui suit, peut nous montrer le rapprochement des textes par rapport au thème traité. A guise d'exemple, sur le quadrant supérieur droit nous observons l'attirance entre *P'tit Bonhomme*, *Archipel en feu* et *Mathias Sandorf*. Pourquoi ne pas se demander la raison pour laquelle ces textes se cherchent entre eux ? La réponse nous la trouvons dans *Jules Verne, un regard sur le monde* (Chesnaux, 2001) lorsque l'auteur assure que Verne réaffirme dans ces textes une sympathie pour les mouvements de libération nationale. En effet, *Petit Bonhomme*, *Archipel en feu* et *Mathias Sandorf* ont un penchant pour les causes irlandaise, grecque et hongroise respectivement. Il en faut ajouter d'autres textes assez proches : *Famille-sans-nom*, histoire d'un héros canadien français qui se bat contre les Anglais pour l'indépendance de son pays ; *Un drame en Livonie*, dans lequel un patriote balte de la communauté slave est inculpé d'un crime qu'il n'a pas commis ; *Chemin de France* qui raconte l'histoire d'un jeune picard qui se bat à Valmy. Les mouvements antitsaristes dans *Michel Strogoff* ; l'insurrection de Méhémet Ali contre les turcs dans *Maître Antifer* ; le conflit entre turcs conservateurs et modernistes dans *Kéraban le Têtu* et le mouvement national norvégien dans *Billet de loterie*, toutes ces actions trouvent-elles, dans le même ordre d'idées, une place dans le quadrant supérieur droit. Les comparaisons étant si multiples méritent, comme nous l'avons signalé plus haut, d'être traitées dans un article appart.

L'analyse arborée (Figure No. 6) rend compte de la proximité des textes. Nous y avons signalé simplement une distribution des éléments par tranche chronologique d'après les résultats de l'algorithme. Nous devons néanmoins faire quelques remarques concernant quelques textes situés en dehors de la chronologie supposée. *Autour de la lune* et *De la Terre à la lune* ont été attirés par des textes rapportant à des voyages aériens ; *Michel Stroggof* et *Les Tribulations d'un chinois en Chine* se sont situés près de textes en relation avec l'Asie ; *Mistress Branican*, roman dont l'histoire se passe en Australie a été attiré par un autre texte dont le décor géographique est similaire (*Les enfants du Capitaine Grant*). *Robur le Conquérant* a eu la même tendance des romans de l'air.

Figure No. 5. La distance Lexicale

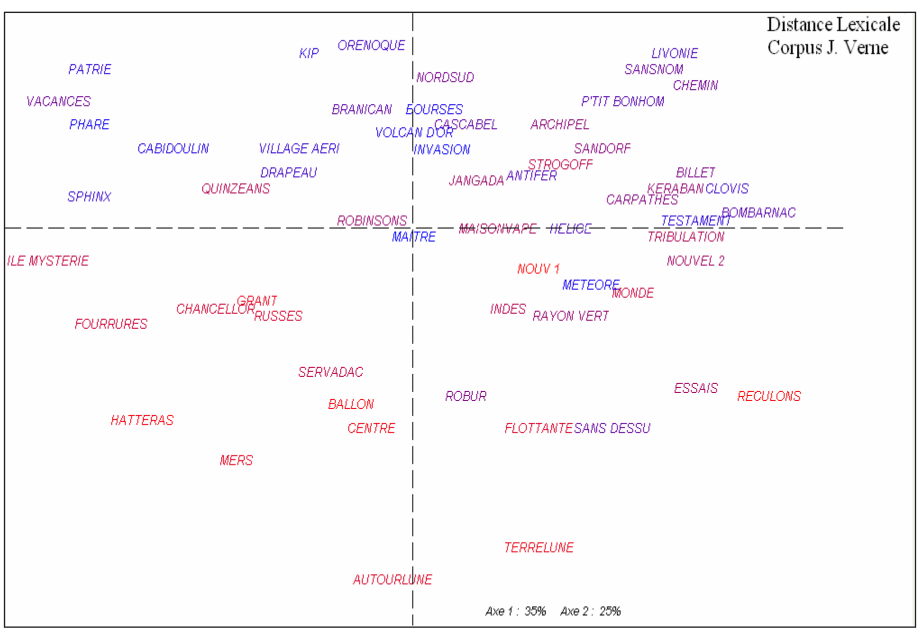
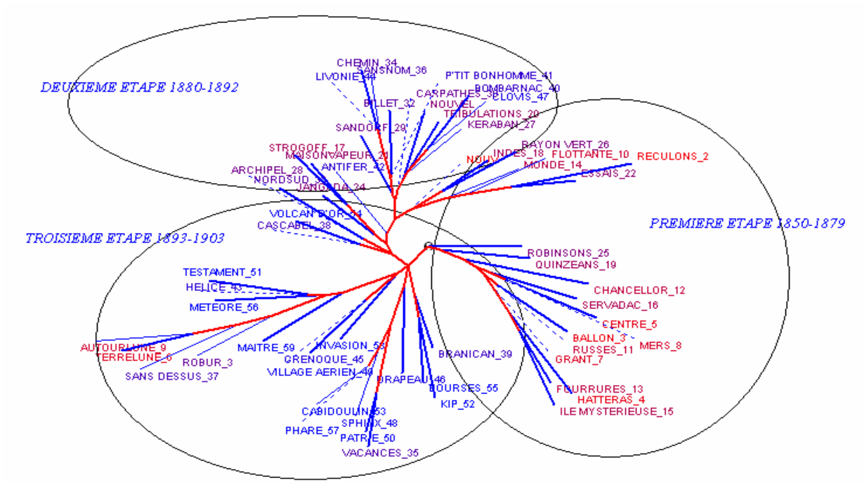


Figure No. 6. Analyse arborée



CONCLUSIONS

Cette étude préliminaire ne constitue qu'une introduction à la statistique textuelle sur Jules Verne. Nous avons atteint, nonobstant, une série de résultats concernant la richesse des œuvres et la croissance du vocabulaire employé. Jules Verne se détache facilement d'autres auteurs par rapport à la nature de sa production littéraire ayant notamment un composant scientifique qui varie de roman en roman. Des périodes bien définies ont été établies en faisant des comparaisons explicites entre le vocabulaire employé et les événements marquants de la vie de l'auteur.

NOTES

Brunet, E. (1981). Le vocabulaire français de 1789 à nos jours. Genève-Paris : Slaktine-Champion.

Année moyenne d'écriture des nouvelles de cette période : 1853. Nous avons. Cette chronologie émule le plus fidèlement possible celles établies par Jean-Paul Dekiss (1999) et par la Société Jules Verne.

Produit par la Société Synapse Développement (Toulouse).

Par exemple « que » peut accomplir des fonctions adverbiales, pronominales ou conjonctives.

BIBLIOGRAPHIE

Brunet, E. (2009). *Comptes d'auteurs*. Paris : Honoré-Champion.

Brunet, E. (2006). *Hyperbase, Manuel de référence. Version standard 6.0*. Nice : BCL.

Brunet, E. (1988). *Le vocabulaire de Victor Hugo*. Genève-Paris : Slaktine-Champion.

Brunet, E. (1981). *Le vocabulaire français de 1789 à nos jours*. Genève-Paris : Slaktine-Champion.

Chesneaux, J. (2001). *Jules Verne, un regard sur le monde*. Paris : Bayard.

Dekiss, J.-P. (1999). *Jules Verne l'Enchanteur, Biographie*. Paris : Kiron Edit. du Félin.

Kastberg S., M. (2006). *L'écriture de J. M. Le Clézio. Des mots aux thème*s. Paris : Honoré Champion.

Magri, V. (1995). *Le discours sur l'autre*. Paris : Honoré Champion.

Mayaffre, D. (2004). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République*. Paris : Honoré Champion.

Muller,Ch. (1992). *Principes et Méthodes de Statistique lexicale.* Paris : Champion.
Soriano, M. (1978). *Jules Verne, le cas Verne, Biographie*. Paris : Julliard.