Marcel PROUST A la recherche du temps perdu

Cyril Labbé, Dominique Labbé

Semaine Data-SHS
"Traiter et analyser des données en sciences humaines et sociales"
Plateforme Universitaire de Données (TIR-PROGEDO)
Université de Grenoble-Alpes
Maison des sciences de l'homme
9-14 décembre 2019

Humanités numériques
Données et méthodes

Marcel PROUST
A la recherche du temps perdu

jeudi 12 décembre 14h-16h

Cyril Labbé
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
(cyril.labbe@imag.fr)

Dominique Labbé
Univ. Grenoble Alpes, PACTE, Institut d'Etudes Politiques de Grenoble
(dominique.labbe@umrpacte.fr)

Résumé
A l'occasion de la semaine "data SHS" de Grenoble, cette conférence montre comment les sciences humaines peuvent tirer profit de l'informatique, de la statistique et des bases de données. Elle utilise un grand corpus étiqueté composé de plus d'une centaine de romans du XIXe et du début du XXe siècle (de Balzac à Zola) et présente une analyse approfondie de la *Recherche du temps perdu* : vocabulaire de l'oeuvre, singularités par rapport aux autres romans contemporains ; style de Proust, notamment la question classique de la longueur de ses phrases. Toutes les réponses sont apportées par le calcul, elles sont vérifiables et reproductibles. Ces méthodes fournissent de nouvelles données pour les "humanités numériques".

Les corpus présentés lors de cette conférence sont disponibles sur demande auprès des auteurs

En novembre 1919, le prix Goncourt est attribué à Marcel Proust (1871-1922) pour son roman *A l'ombre des jeunes filles en fleurs*, roman qui compose le deuxième volume de *A la Recherche du temps perdu.* Cette *Recherche* compte cinq tomes (annexe 2) et plus de 1,3 million de mots, ce qui en fait le plus long roman en langue française, et une sorte de révolution dans les lettres françaises, si l'on en croit l'abondance des analyses qui lui ont été consacrées depuis près d'un siècle.

Pour célébrer cet anniversaire, nous mettons à disposition de la Plateforme Universitaire de Données Grenoble Alpes (PUD-GA) un extrait de notre bibliothèque électronique du français moderne (voir annexe 1) comportant, outre de Proust la *Recherche* et les *Plaisirs et les jours*, plus d'une centaine de romans du XIXe et du début du XXe que Proust est susceptible d'avoir lus afin de fournir une base de comparaison qui permettra de juger de la singularité de cette oeuvre. Il ne s'agit pas simplement de rassembler ces textes déjà disponibles en ligne – notamment sur Wikisource – mais de fournir à leur propos une série de données lexicographiques et statistiques originales avec comme objectif de montrer en quoi la statistique appliquée au langage (lexicométrie) peut être un outil utile pour les sciences humaines et tout particulièrement la lexicographie[1].

Ces données originales permettront de reprendre quelques questions concernant :

- le vocabulaire de Proust : quels sont ses mots préférés et qu'est-ce qui le singularise par rapport aux écrivains de son temps ?

- comment caractériser le style de Proust et, particulièrement ses phrases ?

Le statisticien ne peut répondre à ces deux questions qu'à deux conditions.

1. Etablir le texte avec précision.

Cela pose le problème de l'édition de référence. En effet, pour la *Recherche du temps perdu*, ce choix existe et introduit une légère incertitude concernant quelques mots et la ponctuation (discussion dans Ferré 1957, Milly 1985, Mauriac-Dyer 2005, Serça 2010), spécialement pour les trois derniers livres parus après la mort de Proust. Nous nous sommes tenus au principe général selon lequel l'édition de référence est l'ultime version révisée par l'auteur ou, à défaut, la plus proche de sa mort. Il s'agit ici de l'édition originale chez Gallimard (annexe 2 et références bibliographiques à la fin de cette conférence). De plus, cette édition originale s'impose puisqu'elle est dans le domaine public et peut être communiquée librement aux chercheurs soucieux de reproduire nos résultats et d'aller plus loin dans cette analyse.

[1] Science et techniques de rédaction des dictionnaires et par extension, description empirique du composant lexical d'une langue ou d'un sous-ensemble de celle-ci (lexique de "spécialité").

En second lieu, un dépouillement rigoureux, contrôlable et reproductible. Nous évoquerons succinctement ces méthodes car elles ont déjà été présentées dans des publications antérieures qui sont consultables en ligne.

2. Disposer d'un étalon de comparaison.

Pour juger de la singularité d'une oeuvre, selon certains caractères statistiques spécifiques, il faut disposer de ces mêmes paramètres concernant une population de référence à laquelle appartient l'individu étudié (non seulement les valeurs moyennes mais aussi les variations "normales" au sein de cette population pour savoir si l'œuvre est ou non singulière).

Nous proposons une expérience à l'aide d'un corpus de 116 romans – par 33 auteurs (ou associations d'auteurs) différents – dont la parution s'étale de 1800 à 1922 (date de la mort de Proust). Ce corpus est présenté en annexe 3. Il comporte au total 11 457 253 de mots et un vocabulaire de 54 898 vocables différents. Pour chaque auteur, un dossier comporte, outre les textes originaux et la présentation du corpus, une version lemmatisée et deux index décrivant le vocabulaire des oeuvres.

I. VOCABULAIRE DE PROUST

Le mot est défini comme l'occurrence d'un vocable, c'est-à-dire une entrée dans le lexique de la langue française selon la norme présentée par Muller (1963). Selon cette norme lexicographique, la *Recherche* compte 1 327 850 mots (N dans la suite) et 21 836 vocables différents. Le vocabulaire de l'œuvre est décrit à l'aide de deux index (alphabétique et hiérarchique).

Index alphabétique

L'index alphabétique range chacun des vocables employés dans le texte dans l'ordre du dictionnaire avec, pour chacun d'eux, les flexions sous lesquelles ces vocables apparaissent et leurs effectifs absolus (occurrences). Le tableau 1 ci-dessous donne un extrait de cet index en se limitant à la colonne total.

Tableau 1. Extraits de l'index alphabétique de la Recherche *(colonne des totaux)*

aura (nf)	*2*	*être (n. m.)*	*717*
avion (n. m.)	*15*	*être*	*438*
avion	*6*	*êtres*	*279*
avions	*9*	*pouvoir (v)*	*5827*
avoir (v.)	*31 861*	*pouvoir*	*307*
aura	*116*	*puis*	*31*
avions	*184*	*pouvoir (nm)*	*107*
est (n.m.)	*9*	*pouvoir*	*1014*
Est	*4*	*pouvoirs*	*6*
est	*5*	*puis (adv)*	*716*
été (n.m.)	*81*	*soit (cj)*	*382*
été	*78*	*somme (nf)*	*173*
étés	*3*	*somme*	*168*
être (v.)	*33 593*	*sommes*	*5*
est	*9065*	*somme (nm)*	*4*
été	*2258*	*somme*	*3*
être	*2685*	*sommes*	*1*
soit	*343*	*suivre*	*390*
sommes	*211*	*suis*	*1*
suis	*542*		

L'index alphabétique complet de la *Recherche*, disponible sur demande auprès des auteurs, comporte 56 199 lignes - correspondant aux 21 787 vocables différents qui constituent le vocabulaire de la *Recherche* avec leurs 34 412 flexions - et 11 colonnes comportant les effectifs totaux de chacun de ces vocables (et de leurs flexions) puis le nombre de leurs apparitions (occurrences) dans chacune des 8 parties de la *Recherche*.

Pour les sciences humaines, ce type de document présente de multiples intérêts. Par exemple, le mot est-il utilisé par l'auteur étudié ? Dans ce cas, avec quelles flexions et dans quels textes ? Dans l'affirmative, l'étape suivante consiste à rechercher les contextes dans lesquels l'auteur a utilisé ces mots ou ces flexions particulières (concordances). Par exemple, le tableau 2 ci-dessous présente le contexte des deux occurrences du substantif féminin "aura" qu'il faut distinguer des 116 occurrences de "aura" (verbe *avoir* indicatif futur).

Tableau 2. Concordances du vocable *aura* (n. f.) dans la *Recherche*

situation militaire, il reconnaît la perception par le peuple de cette "**aura**" qui entoure les grands événements et qui peut être visible (*A l'ombre des jeunes filles en fleurs*).

meurt pas tout de suite pour nous, il reste baigné d'une espèce d' **aura** de vie qui n'a rien d'une immortalité véritable (*La fugitive*).

Les exemples de "collisions" présentées dans le tableau 1 ne sont pas anecdotiques. Dans tout texte en français, *être* et *avoir* sont les verbes les plus fréquents, *pouvoir* figure toujours parmi les verbes les plus utilisés et les substantifs usuels, etc. En moyenne, dans un texte en

français, plus d'un mot sur trois peut être rattaché à plusieurs entrées de dictionnaire (homographies). Il est donc indispensable de lever ces ambiguïtés D'ailleurs, imagine-t-on qu'un statisticien puisse répondre à un littéraire, spécialiste de Proust, qu'il ne peut rien lui apprendre sur les substantifs "aura", "avion", "être" – l'un des mots préférés de Proust -, "est", "été" – la saison préférée de Proust -, "pouvoir", "devoir", l'adverbe "puis" ou la conjonction "soit" (et plusieurs milliers d'autres ambiguïtés représentant plus du tiers des mots de la *Recherche*), car leurs occurrences sont inextricablement mélangées à celles des formes verbales homographes[2] ?

La solution réside dans la définition d'une "norme de dépouillement" applicable à l'ensemble des textes en français et dans sa programmation informatique (Labbé 1990). Il s'agit d'épouser au plus près les conventions lexicographiques de cette langue. D'une part, chaque mot, dans le texte même, est doté d'une étiquette qui le rattache à son entrée de dictionnaire - par exemple, l'infinitif du verbe ou le masculin singulier de l'adjectif – et à sa catégorie grammaticale. D'autre part, le "mot" ne correspond pas à la "forme graphique" telle qu'elle est délimitée par les logiciels d'analyse "textuelle". Dans ces logiciels, les caractères séparateurs de "formes" sont l'espace, l'apostrophe et le trait d'union. Ceci aboutit à couper certains mots en plusieurs formes. Par exemple, "aujourd'hui", "parce que" "grand-mère" ou "Saint-Loup" sont comptés comme deux formes graphiques – "aujourd'" et "hui", "grand" et "mère", etc. - alors qu'il s'agit d'un seul mot. Il y a 1 449 "parce que" dans la *Recherche*, soit plus d'un mot pour dix mille ; 787 fois "Saint-Loup" : ami d'enfance du narrateur, c'est l'un des personnages les plus importants de la *Recherche* ; 681 fois "grand-mère" (autre personnage essentiel du roman). Il ne faut donc pas les couper en deux ! A l'inverse, les formes graphiques "l'", "le", "la", "les" sont des flexions de "le" qui peut être pronom ou articles ; "du" ou "des" sont la contraction de deux entrées du lexique : préposition "de" et article "le". En fonction de la norme retenue (unités lexicales ou formes graphiques), le nombre de "mots" (en anglais "token") dans un texte peut varier de près de 10%.

Index hiérarchique

Dans ce second type d'index, les vocables sont classés en fonction de leur nombre d'apparitions (occurrences) (Tableau 3). Ce tableau donne les fréquences exprimées en "pour mille mots" (‰) pour permettre les comparaisons entre corpus de longueurs différentes.

[2] C'est ce que fait Brunet (1983), ce qui rend ce livre inutilisable. Le titre de l'ouvrage est d'ailleurs trompeur puisque le dépouillement ne porte pas sur le vocabulaire mais sur les formes graphiques.

Tableau 3. Début de l'index hiérarchique de la *Recherche*

Rang	Lemme et catégorie grammatic	Effectifs	Fréquence (‰)
1	le (dét)	101 689	76,58
2	de (pré)	96 497	72,67
3	à (pré)	35 575	26,79
4	être (v)	33 594	25,30
5	je (pro)	33 417	25,17
6	avoir (v)	31 862	24,00
7	il (pro)	31 781	23,93
8	un (dét)	30 692	23,11
9	que (cj)	27 365	20,61
10	et (cj)	25 171	18,96
11	ne (adv)	23 635	17,80
12	qui (pro)	15 633	11,77
13	que (pro)	15 016	11,31
14	pas (adv)	14 341	10,80
15	son (dét)	12 562	9,46
16	ce (pro)	11 813	8,90
17	ce (dét)	11 753	8,85
18	le (pro)	11 627	8,76
19	se (pro)	11 087	8,35
20	en (pré)	10 953	8,25

On s'attend à ce que la tête de l'index renseigne sur les mots favoris de l'auteur or, dans les premiers rangs, ne figurent aucun substantif ou adjectif et aucun verbe (en dehors de *être* et *avoir*). Les mots les plus fréquents jouent un rôle essentiellement syntaxique et n'acquièrent un contenu qu'avec le contexte de l'énonciation. Les premiers mots non-outils sont loin dans la liste. Ainsi, le substantif le plus utilisé - "madame" (souvent écrit "Mme") – se trouve au 47e rang avec 3 370 occurrences (soit 2,54 ‰ de la surface des textes). Avec une fréquence inférieure à 0,25%, "madame" est donc trente fois moins fréquent que l'article "le", bien que beaucoup plus importante pour l'analyse de la *Recherche*. De même, le nom propre le plus utilisé (*Albertine*) figure au 60^{e} rang (1,79 pour mille mots).

Dans le vocabulaire de la *Recherche*, 46 des 50 vocables les plus fréquents sont des mots outils. A eux seuls, ils couvrent plus de la moitié de la surface des textes (55,9%) bien qu'ils ne constituent que 1,6 % du vocabulaire. A l'opposé, 95,2% des vocables apparaissent moins de 100 fois et ne couvrent que 13,2% de la surface des textes. Ce sont pourtant eux qui véhiculent l'essentiel du message.

Autrement dit, le vocabulaire d'un corpus est une grande collection d'événements rares (les occurrences d'un vocable) très inégalement distribués, dont les plus fréquents ne sont pas les

plus facilement interprétables. Ce phénomène d'inégale distribution des fréquences a été mis en lumière par Zifp 1935 (voir également Mandelbrot, 1957).

Pour éviter une sélection, toujours subjective, les vocables sont regroupés par catégories grammaticales et les effectifs (absolus) sont convertis en fréquences (relatives), ce qui permet la comparaison de corpus de longueurs différentes. Les tableaux 4 à 8 donnent ces renseignements pour les mots à majuscules ("noms propres"), les substantifs, les adjectifs, les verbes.

Tableau 4. Les vingt noms propres les plus utilisés dans la *Recherche*

Rang	Vocable	Effectif	Fréquence (‰)
1	Albertine	2 376	1,79
2	Guermantes	1 754	1,32
3	Swann	1 647	1,24
4	Charlus	1 290	0,97
5	Verdurin	1 175	0,88
6	Françoise	795	0,60
7	Saint-Loup	787	0,59
8	Balbec	766	0,58
9	Gilberte	704	0,53
10	Odette	701	0,53
11	Paris	538	0,41
12	Morel	511	0,38
13	Bloch	486	0,37
14	Combray	429	0,32
15	Cambremer	402	0,30
16	Villeparisis	397	0,30
17	Andrée	392	0,30
18	Robert	374	0,28
19	Cottard	373	0,28
20	Brichot	310	0,23

Cette liste montre quels sont les principaux personnages de la *Recherche* – outre le narrateur évoqué plus bas – mais aussi les principaux noms de lieux qui ont une grande importance dans cette œuvre.

Les personnages sont évidemment beaucoup plus nombreux. Par exemple, le musicien Vinteuil arrive en 21^{e} position, l'écrivain Bergotte en 22^{e}, le peintre Elstir en 23^{e} et M. de Norpois en 24^{e}, tous quatre avec une fréquence de 0,23‰. Enfin, Dreyfus est le 38^{e} mot à majuscule initiale. Cité plus d'une centaine de fois, sa présence marque le rôle considérable que

joue l'affaire dans la *Recherche*, alors que ce roman n'évoque pratiquement aucun autre fait historique (à part la première guerre dans le dernier volume) ni personnage contemporains.

Quant aux noms de villes, il est intéressant de voir que *Balbec* (station balnéaire où le narrateur passe ses vacances et où se place l'essentiel des *Jeunes filles en fleurs* et une partie de *Sodome et Gomorrhe*) est plus souvent cité que *Paris*, *Combray*, *France* ou Venise (32^{e} avec une fréquence de 0,11‰). Cette dernière est la seule ville étrangère à avoir marqué le narrateur. *Parme* (ex-aequo avec Venise) ne désigne pas la ville mais la *princesse de Parme* comme l'indique le dépouillement sur les combinaisons de mots (syntagmes) les plus fréquentes. Ces combinaisons éclairent également les substantifs les plus utilisés dans la Recherche.

Tableau 5. Les vingt substantifs les plus utilisés

Rang	Vocable	Effectif	Fréquence (‰)
1	madame	3 370	2,54
2	monsieur	3 152	2,37
3	jour	2 112	1,59
4	femme	2 031	1,53
5	chose	2 021	1,52
6	vie	1 756	1,32
7	fois	1 722	1,30
8	temps	1 645	1,24
9	moment	1 602	1,21
10	homme	1 520	1,14
11	air	1 480	1,11
12	monde	1 318	0,99
13	heure	1 236	0,93
14	oeil	1 194	0,90
15	personne	1 135	0,85
16	nom	1 101	0,83
17	plaisir	1 098	0,83
18	gens	1 066	0,80
19	fille	989	0,74
20	amour	915	0,69

Le tableau 5 donne déjà un aperçu des principaux thèmes de la *Recherche*. Par exemple, les deux premiers indiquent que certains personnages sont désignés par leurs noms précédés de Mme ou M. – *M. de Charlus* (1 182 fois), *Mme Verdurin* (672), *Mme de Guermantes* (533), *Mme de Villeparisis* (369), *M. de Norpois* (249), etc. C'est une manière de signaler la distance séparant ces personnages du narrateur. D'autres sont plus proches et sont simplement désignés par leurs noms, comme *Swann*, *Bergotte*, *Elstir* ou *Bloch*, voire par leurs seuls prénoms - *Albertine*, *Odette*, *Gilberte*, *Robert*, etc. Ainsi se dessine une manière de géographie sociale au

moins aussi importante dans la *Recherche* que les lieux et le temps (*jour, temps, moment, heure*) et les sentiments. Enfin, la présence au 5e rang du mot "chose" est inhabituelle. Plus ce mot est fréquent, plus cette présence donne une tournure orale ou familière au texte.

En prolongeant la liste, on trouve *mère* (34e, 0,52‰) et *grand-mère* (35e, 0,51‰), ce qui souligne l'importance de ces deux personnages. A vrai dire, *mère* ne désigne pas seulement celle du narrateur (dans ce cas, c'est souvent "maman" : 184e, 0,16‰). Si l'on considère la fréquence d'apparition d'un mot comme un indice de son importance pour le narrateur, la grand-mère est un personnage au moins aussi important que la mère dans la *Recherche*.

Cette discussion permet de comprendre qu'il ne faut pas considérer les vocables isolément mais reconstituer les combinaisons dans lesquelles ils entrent et la famille (ou "paradigme") à laquelle ils appartiennent. Par exemple, "œil" s'associe à "regard" (40e substantif), au verbe "voir" mais aussi à la "beauté" et aux adjectifs "beau" et "jeune" pour former l'un des principaux thèmes de la *Recherche* (Erman 1988).

Tableau 6. Les vingt adjectifs les plus utilisés

Rang	Vocable	Effectif	Fréquence (‰)
1	grand	1 999	1,51
2	petit	1 713	1,29
3	seul	1 440	1,08
4	jeune	1 313	0,99
5	nouveau	935	0,70
6	beau	865	0,65
7	bon	816	0,61
8	différent	664	0,50
9	vrai	641	0,48
10	vieux	591	0,45
11	dernier	462	0,35
12	ancien	454	0,34
13	même	449	0,34
14	mauvais	412	0,31
15	simple	381	0,29
16	heureux	365	0,27
16	pareil	365	0,27
18	particulier	335	0,25
19	long	330	0,25
20	possible	327	0,25

Dans tout texte en français de la longueur d'un roman, les deux premiers adjectifs sont *grand* et *petit*, puis viennent, dans un ordre révélateur des préférences de l'auteur, les couples *jeune/vieux, nouveau/ancien, beau/laid, bon/mauvais, heureux/malheureux, vrai/faux* etc.

Toutefois la hiérarchie, entre les couples et à l'intérieur de ceux-ci, a son importance car elle indique les préférences de Proust et le terme valorisé : la jeunesse (plutôt que la vieillesse) puis la nouveauté (plutôt que l'ancienneté), la beauté, la gentillesse, le bonheur…

Comme pour le choix des substantifs, les différences entre les auteurs sont importantes et dépendent des thèmes choisis. Par exemple, Balzac place *petit* légèrement devant *grand* et, juste après *jeune* et *beau* vient l'adjectif *pauvre*, nettement plus employé que "*riche*". Ce dernier couple est pratiquement absent chez Proust…

Le même constat peut être fait à propos des verbes (tableau 7).

Tableau 7. Les vingt verbes les plus utilisés

Rang	Vocable	Effectif	Fréquence (‰)
1	être	33 593	25,30
2	avoir	31 861	23,99
3	faire	8 632	6,50
4	dire	7 284	5,49
5	pouvoir	5 827	4,39
6	voir	4 306	3,24
7	aller	3 247	2,45
8	savoir	3 014	2,27
9	venir	2 584	1,95
10	vouloir	2 360	1,78
11	croire	2 286	1,72
12	trouver	2 097	1,58
13	donner	2 033	1,53
14	connaître	1 936	1,46
15	devoir	1 891	1,42
16	parler	1 734	1,31
17	aimer	1 689	1,27
18	prendre	1 643	1,24
19	demander	1 556	1,17
20	sembler	1 522	1,15

Dans tout texte en français, long d'au moins 10 000 mots, les trois premiers verbes sont, dans cet ordre : *être, avoir, faire* et pour les romans du corpus : "dire". Ensuite viennent les préférences de l'auteur : le possible (*pouvoir* presque deux fois plus employé par Proust que par la moyenne des romanciers), le regard (*voir*) qui est donc au centre d'un thème très important dans la *Recherche*, la connaissance - *savoir* et surtout *connaître, trouver* (chez Proust dans le sens de "estimer que" -, la croyance (*croire*), puis "devoir", "parler" et "aimer". En revanche, Proust ne se singularise pas pour "aller", "venir", "donner" ou "demander" et utilise

moins "vouloir" que la moyenne des autres auteurs, ce qui ne surprendra pas les connaisseurs de son œuvre où le manque de volonté du narrateur est une dimension importante.

Un choix fondamental : les personnes

Pour dire en quoi Proust se singularise, on utilise donc comme point de comparaison, les romans du XIXe et du début du XXe. Etant donné les différences de longueur, les effectifs absolus n'ont pas d'intérêt. La comparaison s'appuie sur le rang et les fréquences. Le tableau 8 donne cette comparaison sur les pronoms et illustre un choix fondamental de Proust.

Tableau 8. Densité des pronoms dans la *Recherche* et dans le corpus des romans du XIXe

Pronoms	Proust	A	Romans	B	A/B
	Rang	Fréquence (‰)	Rang	Fréquence (‰)	(%)
je	1	25,17	2	18,02	+39,7
il	2	23,93	1	25,93	-7,7
qui	3	11,77	4	9,24	+27,4
que	4	11,31	7	7,13	+58,6
ce	5	8,90	6	7,22	+23,3
le	6	8,76	5	9,19	-4,7
se	7	8,35	3	12,28	-32,0
on	8	5,17	10	4,47	+15,7
lui	9	4,99	9	5,15	-3,1
nous	10	4,98	12	3,26	+52,8
vous	11	4,52	8	7,00	-35,4
ils	12	3,89	11	3,40	+14,4
y	13	2,83	13	2,91	-2,7
moi	14	2,63	16	2,02	+30,2
en	15	2,47	14	2,82	-12,4
celui	16	2,22	21	1,13	96,5
lequel	17	1,87	24	0,92	103,3
cela	18	1,75	22	1,13	54,9
dont	19	1,68	18	1,38	21,7
autre	20	1,48	23	1,13	31,0
Total		138,67		125,73	10,3

Lecture : Les pronoms personnels de la première personne (je, j', me, m') sont les plus employés dans la *Recherche* avec une densité supérieure à 25 ‰ mots alors qu'en moyenne dans le corpus de référence (« romans »), ils n'occupent que le second rang - derrière la troisième personne du singulier (il, elle) – avec une densité moyenne de 18,2 ‰. Chez Proust la fréquence d'emploi du "je" est supérieure de +39,7% par rapport à la moyenne des romans.

Le rang marque clairement les préférences : la *Recherche* privilégie les premières personnes du singulier (le narrateur[3]) et du pluriel (le narrateur et d'autres personnes), voire le "on" (qui est souvent un *nous* familier). En revanche, la troisième personne, caractéristique du récit, est nettement moins importante que dans le roman type : "il" (qui regroupe aussi *elle*) passe du premier rang au deuxième et, surtout le réfléchi (*il se*) de la troisième à la septième).

Les deux premières lignes du tableau reflètent le choix devant lequel se trouve tout auteur de fiction : utiliser ou non la première personne ? L'écrasante majorité des romans de la période (1800-1920) sont écrits à la troisième personne : le narrateur s'efface du récit et présente une histoire dont il serait simple témoin. Cela ne signifie pas que le "je" est absent de ces romans : il est utilisé notamment quand le récit rapporte les propos des personnages en ouvrant les guillemets (discours rapporté). Souvent, il s'agit de dialogues : dans ces romans, la seconde personne ("tu" et "vous") est donc utilisée en association avec le "je" et leurs densités relatives sont plus ou moins proportionnelles. Naturellement, si le livre contient beaucoup de dialogues (comme chez Balzac ou Dumas), la densité des première et seconde personnes peut être importante sans altérer le choix fondamental (qui se marque par une densité élevée de guillemets mais aussi par le temps des verbes, comme nous le verrons plus bas).

Proust a donc choisi la première personne : le narrateur est le personnage principal du texte (ce qui ne signifie pas que ce narrateur puisse être identifié à l'auteur : Tadié 1971, chapitres 1 et 2). En conséquence, la troisième personne (*il, se, lui*) est nettement sous-employée.

La *Recherche* utilise également beaucoup moins la seconde personne (*tu, vous*) que la moyenne des romans. Cela peut sembler paradoxal puisqu'elle comporte de nombreuses scènes de conversations (notamment dans les salons des Verdurins ou de la duchesse de Guermantes) ou des dialogues, notamment avec sa maîtresse (Albertine), sa mère ou sa grand-mère, sa bonne (Françoise), etc. Dans la *Recherche*, beaucoup de ces conversations sont rapportées au style indirect, sans utilisation de guillemets (et ne contiennent donc pas de "je/tu" ou "je/vous" comme dans un récit classique).

Le tableau 8 signale deux autres caractéristiques stylistiques fondamentales de la *Recherche*. D'une part, comparée à l'usage littéraire du temps, elle contient un net excédent de pronoms relatifs (*qui, que, dont, lequel*) qui permettent d'établir, au sein d'une même phrase, des relations entre des personnes ou des choses ou entre des événements (par exemple, des causalités ou des hiérarchies). D'autre part, Proust utilise beaucoup les pronoms démonstratifs (*ce, celui, ceci, cela*) qui sont typiques du français oral, un peu comme si l'auteur montrait familièrement du doigt au lecteur les personnes et les choses dont il parle.

Cependant, le tableau soulève au moins deux questions auxquelles la statistique lexicale permet également d'apporter des réponses.

[3] Il y a assez peu d'énoncés rapportés dans la *Recherche* et la plupart de ces énoncés ne contiennent pas "je".

- Les écarts sont-ils significatifs du point de vue statistique ? Par exemple : *le, lui, y*. En effet, tout phénomène naturel est soumis à de faibles variations aléatoires. Un test statistique permet de déterminer à partir de quel moment on peut considérer que le phénomène sort des limites "normales" (Labbé 2019).

- La dernière ligne du tableau suggère que la *Recherche* contient une densité plus forte de pronoms par rapport à la moyenne des autres (+9,33%). Est-ce que cela ne concerne que les pronoms les plus usuels ou bien Proust marque-t-il une préférence pour l'ensemble de la catégorie grammaticale ?

Ces deux questions renvoient au style de Proust, notamment à sa préférence pour le verbe et pour les phrases longues.

III. STYLE DE PROUST

Le style singulier de la *Recherche* occupe les analystes pratiquement autant que son contenu. Parmi les nombreux indices de cette singularité stylistique, deux sont particulièrement notables : une préférence marquée pour le verbe et des phrases singulières.

Préférence pour le verbe

Dans les fichiers lemmatisés, chaque mot du texte est rattaché à son entrée de dictionnaire qui comporte un mot vedette (par exemple l'infinitif des verbes) et une catégorie grammaticale. En recensant ces catégories et en rapportant les effectifs absolus à la longueur du corpus, on obtient le poids relatif de chacune ("densité" ou proportion de la surface du texte qu'elle occupe). La même opération est effectuée sur chaque texte du corpus de référence afin de déterminer la densité moyenne des mêmes catégories. On peut alors comparer ces densités dans la *Recherche* et dans le corpus de référence, puis à l'aide du test statistique évoqué plus haut, déterminer si les différences de densité entre la *Recherche* et les autres romans est ou non significative (tableau 9).

Cette comparaison aboutit à un contraste frappant entre la *Recherche* et les autres romans, contraste que l'on peut résumer par deux dimensions essentielles.

Tableau 9 Poids des catégories grammaticales dans la *Recherche* comparée aux autres romans

Catégories	A (Corpus-Proust) ‰	B Proust (‰)	(B-A)/B (%)	Indice
Verbes	163,0	168,9	+3,6	1,0
Futurs	*3,5*	*2,0*	*-43,8*	*0,0*
Conditionnels	*3,5*	*5,2*	*+51,0*	*1,0*
Présents	*39,7*	*34,9*	*-12,1*	*0,0*
Imparfaits	*38,5*	*48,3*	*+25,3*	*1,0*
Passés simples	*24,1*	*11,5*	*-52,0*	*0,0*
Participes passés	*18,6*	*24,3*	*+30,4*	*1,0*
Participes présents	*7,0*	*7,0*	*-0,1*	≈
Infinitifs	*28,1*	*35,6*	*+26,6*	*1,0*
Noms propres	27,0	23,8	-11,9	0,0
Noms communs	182,8	159,0	-13,0	0,0
Adjectifs	58,7	54,2	-7,6	0,0
Adj. participe passé	*11,7*	*8,9*	*-23,4*	*0,0*
Pronoms	135,1	148,1	+9,6	1,0
Pronoms personnels	*85,3*	*85,9*	*+0,9*	≈
Déterminants	164,1	142,7	-13,1	0,0
Articles	*112,6*	*99,7*	*-11,4*	*0,0*
Nombres	*9,6*	*4,1*	*-57,1*	*0,0*
Possessifs	*23,6*	*19,3*	*-18,2*	*0,0*
Démonstratifs	*10,1*	*8,9*	*-12,2*	*0,0*
Indéfinis	*8,3*	*10,7*	*+28,8*	*1,0*
Adverbes	67,4	87,3	+29,5	1,0
Prépositions	144,4	148,7	+3,0	1,0
Coordinations	31,7	30,8	-2,6	0,0
Subordination	21,9	34,6	+58,1	1,0
Mots étrangers	0,5	0,3	-39,8	0,0

Lecture : dans l'ensemble de référence (dont on a enlevé la *Recherche*), il y a 163 verbes pour 1000 mots et 168,9 ‰ dans la *Recherche*, soit 3,6% de plus. Le 1.0 indique que, quoiqu'apparemment assez faible, cette différence est significative avec moins de une chance sur 10 000 de se tromper. En dehors des participes présents et des pronoms personnels (différences non significatives au seuil de 1‰), toutes les différences sont significatives à ce seuil.

1. La préférence de Proust pour le verbe s'accompagne de différences très importantes concernant les temps et les modes.

- Le futur et le présent sont sous-employés. Le contenu de la *Recherche* est donc tourné vers le passé, fidèle en cela à son titre. Cependant, l'imparfait et le passé composé – où le plus-que-parfait, tous deux décomptés grâce au participe passé - sont préférés au passé simple, lui aussi nettement sous-employé (Proust en utilise moitié moins que les autres). En français, le passé simple situe l'événement dans un instant précis, clairement délimité dans le temps du récit.

L'imparfait et plus encore le passé composé – encore appelé "parfait" - donnent une durée à cet événement passé et abolissent, au moins partiellement, la frontière entre passé et présent :

> « Le parfait à la première personne est la forme autobiographique par excellence.» « Il établit un lien vivant entre l'événement passé et le présent où son évocation trouve place. C'est le temps de celui qui relate les faits en témoin, en participant ; c'est donc aussi le temps que choisira quiconque veut faire retentir jusqu'à nous l'événement rapporté et le rattacher à notre présent » (Benveniste, 1966, p. 244).

Comment mieux définir le projet de Proust dans la *Recherche* ?

- L'infinitif se combine avec un autre verbe et il est signe de tension, d'autant plus que chez Proust, il est souvent construit avec « ne… pas ».

- Il y a dans la *Recherche* plus de pronoms et surtout d'adverbes que dans les autres romans. L'excédent des pronoms vient des démonstratifs (*ce*, *cela*, *celui*) et surtout des relatifs (*que*, *qui*, *dont*, *lequel*) qui ont pour fonction d'établir des liens au sein de phrases syntaxiquement complexes.

2. Proust emploie significativement peu de noms propres, de substantifs, d'adjectifs et de déterminants.

- Il est spécialement réticent envers les adjectifs issus d'un participe passé (comme "perdu", dans le "temps perdu"). En supprimant le verbe et son sujet, voire le complément d'agent, ces constructions donnent un tour accompli à la situation qu'elles décrivent alors que Proust cherche à donner une impression de mouvement, de développement et d'inaccompli, du moins jusqu'à la dernière page.

- Parmi les déterminants, les nombres (chiffres et dates) sont ceux pour lesquels la différence est la plus grande (-57%). Comme les patronymes et les toponymes qui ancrent le récit dans l'espace, social ou géographique, les dates et les chiffres lui fournissent un ancrage temporel, économique, etc. Cette sobriété de Proust, quant aux dates et aux chiffres, a le même effet que la faible utilisation du passé simple : le récit est peu situé dans l'histoire. A part l'affaire Dreyfus et la grande guerre dans le dernier tome, la *Recherche* mentionne très peu d'événements historiques.

- Il n'y a pas plus de pronoms personnels que dans la moyenne du corpus de référence. En face de la présence considérable du narrateur (*je* mais aussi *nous* ou *on*), les pronoms personnels de la seconde personne (*tu* et *vous*) sont quasiment absents, ce qui donne à la *Recherche* l'aspect d'un monologue.

Enfin, le tableau appelle une remarque : les pronoms, les adverbes et les conjonctions de subordination varient dans le même sens que les verbes (il y en a plus dans la *Recherche* que dans la moyenne des autres romans du XIXe). A l'inverse, les noms propres, les substantifs, les

adjectifs, les déterminants et les prépositions varient dans l'autre sens (il y en a moins dans la *Recherche* par rapport au corpus de référence). Les mêmes mouvements sont observés dans toutes les expériences comparables réalisées sur le français moderne. Dès lors, on peut rassembler toutes ces catégories en deux groupes : celui du verbe et celui du nom.

Certes, le partage n'est pas absolu : on trouve des adverbes dans le groupe nominal (notamment devant l'adjectif) ; il y a des prépositions dans les groupes verbaux, etc. Cette réserve admise, le regroupement révèle des tendances intéressantes (tableau de synthèse ci-dessous)

Tableau 10 Poids des groupes du verbe et du nom dans la *Recherche* comparée aux autres romans du XIXe.

Catégories	A (Corpus-Proust) ‰	B Proust (‰)	(B-A)/B (%)	Indice
Groupe du verbe	387,5	438,9	+13,3	1,0
Groupe du nom	608,7	559,1	-8,1	0,0

Dans la *Recherche*, les éléments du groupe du verbe pèsent 13,3% de plus que dans le reste du corpus des romans du XIXe et ceux du nom, 8,1% de moins. Ces différences sont considérables et s'expliquent de deux manières.

Premièrement, du point de vue stylistique, les caractéristiques de la *Recherche* la placent assez loin des romans traditionnels et à proximité de genres comme la correspondance (Labbé et Labbé 2013) qui est le genre écrit le plus proche de l'oral car il mime une conversation à distance. Mais dans la *Recherche* le destinataire est absent de la surface du texte ; elle serait donc une sorte de monologue, un peu comme si l'auteur était à côté du lecteur et lui livrait ses confidences.

A propos de Proust, J. Cocteau avait confié qu'il lui était « difficile de lire son œuvre au lieu de l'entendre. Presque toujours sa voix s'impose, et c'est à travers elle que je regarde les mots »[4]. Ce trait stylistique est sans doute celui qui singularise le plus la *Recherche* et pourrait expliquer en partie la fascination qu'elle exerce sur certains lecteurs : ils entendent Proust leur parler… (Sur ce point voir aussi : Ferré 1957 et Héron 2010).

Deuxièmement, du point de vue linguistique, le verbe a une double fonction : la fonction "cohésive" qui organise "en une structure complète les éléments de l'énoncé" et la fonction "assertive" qui "dote l'énoncé d'un prédicat de réalité" car l'élément verbal implique une référence à un ordre qui n'est plus simplement celui du discours mais aussi celui de la réalité (Benveniste 1966, 1, p. 154). Autrement dit, cette densité importante de verbes cimente ensemble les éléments disparates du souvenir, comme, au début de Combray, le dormeur qui

[4] Hommage à Marcel Proust. Cité par Tadié 1971, p. 57.

rêve rassemble autour de lui, dans son lit, les objets familiers et, dans son esprit, les souvenirs disparates qu'il organise progressivement.

Phrases de Proust

La phrase est l'empan de texte dont le premier mot comporte une majuscule initiale et qui se termine par une ponctuation majeure (point, points d'interrogation et d'exclamation, points de suspension) suivie d'un mot avec une majuscule initiale (ou de la fin de texte). Ce n'est pas tout à fait la définition retenue par Milly (1975 et 1986), ce qui explique de très légères différences entre son décompte et les nôtres. Cette opération ne peut pas être entièrement automatisée car aucun des quatre signes de ponctuation majeure ne marque forcément une fin de phrase :

- le point dans « M. Verdurin » ne termine pas une phrase bien qu'il soit suivi d'un mot à majuscule initiale. Il y a dans la *Recherche* 3 152 "monsieur" écrits "M." et beaucoup sont suivis d'un patronyme commençant par une majuscule. *Monsieur* est le deuxième substantif le plus fréquent dans la *Recherche* (juste derrière "Mme"), soit 2,4 pour mille mots. Ce point "non-terminal" se retrouve également dans les initiales que Proust utilise pour "anonymer" certains noms (Mme X.) ou derrière des abréviations (etc.).

- dans la *Recherche*, plus de trois points d'interrogation sur 10 sont internes à la phrase (721). Par exemple, « tout le monde aussitôt se demandait : « Une visite, qui cela peut-il être ? » mais on savait bien que cela ne pouvait être que Monsieur Swann »(Combray). Le mot suivant (mais) en minuscule indique que c'est la même phrase qui continue.

- il y a 1 201 points d'exclamation internes à la phrase. Par exemple : « Quand ils parlent de choses ou de gens qui nous intéressent ! » enchérit ma tante. » (Combray). Comme dans l'exemple précédent, l'énoncé rapporté est inséré dans la phrase sans en interrompre le cours.

- 190 points de suspension également dans cette situation. Proust a plusieurs fois déclaré son hostilité envers ces derniers mais il les utilise. Par exemple : « La duchesse émit très fort, mais sans articuler : « C'est l'... i Eon l... b... frère à Robert. » (*la Prisonnière*).

Cette rapide discussion permet de comprendre qu'on ne peut rapporter le nombre de mots d'un ouvrage à l'effectif de ses ponctuations "majeures" – comme le fait Brunet (1981) - pour obtenir une longueur moyenne des phrases et pourquoi ce type d'estimation peut tomber loin de la réalité. D'ailleurs ce procédé ne permet pas d'associer à la moyenne un écart-type empirique mesurant la dispersion des valeurs observées autour de cette moyenne qui perd ainsi pratiquement toute utilité. Par conséquent, il y a lieu de localiser précisément ces fins de phrases dans le texte même afin de mesurer la longueur de chacune des phrases. Pour ce faire, un automate place, dans le texte, des balises localisant les fins de phrase et, en cas de doute, l'opérateur choisit : fin de phrase ou ponctuation interne ? A condition que cet opérateur suive

toujours la même norme, le dépouillement est fait sans erreur et les résultats obtenus sur un auteur sont comparables à ceux des autres.

Ce recensement établit le nombre de phrases de la *Recherche* (annexe 2). Il permet surtout de caractériser ces phrases à l'aide de deux ensembles d'indices.

1. Les longueurs de phrases sont analysées à l'aide d'un certain nombre de valeurs centrales : mode, médiane, moyenne, médiale ; déciles et quartiles. Mais aussi grâce à des indices de dispersion des longueurs : étendue, écart-type de la moyenne (et coefficient de variation), rapports entre valeurs centrales, comme la médiane et la médiale, et indice de concentration du caractère sur les individus les plus grands (Gini)[5]. Deux tableaux récapitulatifs présentent ces résultats en annexes 4 et 5.

Pour un caractère très dispersé comme les longueurs de phrases, la valeur centrale la plus caractéristique n'est pas la moyenne mais la médiale (ou "seconde médiane") : la moitié du texte est couverte par des phrases de longueurs inférieures à cette valeur et l'autre moitié par des longueurs supérieures à cette valeur. Ainsi, le lecteur idéal de la *Recherche* – qui lirait toujours à la même vitesse – se trouve, la moitié du temps, confronté à des phrases de longueurs égales ou supérieures à 50 mots, ce qui est considérable et fort éloigné de ce que ce lecteur a l'habitude de rencontrer dans les journaux ou les livres contemporains (une vingtaine de mots). Même en se reportant un ou deux siècles en arrière – à une époque où l'on avait l'habitude d'une expression plus complexe - une médiale aussi élevée se rencontrait très rarement (annexe 5) et jamais sur des ouvrages aussi longs que les différents volumes de la *Recherche*.

Dans son premier ouvrage (*Les plaisirs et les jours*), Proust en était assez loin, même s'il présentait déjà une propension à concentrer une proportion importante du texte dans des phrases relativement longues (indice de concentration dit de Gini).

L'annexe 4 montre qu'aucun des trois écrivains – Barrès, Bourget et France qui dominaient la scène littéraire à l'époque où Proust a entrepris la rédaction de la *Recherche*, et dont deux sont susceptibles d'avoir servi de modèle à l'écrivain Bergotte (Levaillant 1952) - ne présente des caractéristiques semblables. L'annexe 5 indique qu'il en va de même pour les auteurs et les livres mentionnés implicitement ou explicitement, dans la *Recherche* (Nathan 1968), voire dans le reste de son œuvre et dans sa correspondance (Chantal 1967, Compagnon 2009, Mauriac-Dyer et Al. 2013), notamment Mme de Sévigné, Balzac (Bouillaguet 2000), Saint-Simon (Jullien 1989), Chateaubriand, Hugo, Musset, Sand, Vigny ou encore Barbey d'Aurevilly (Rogers 2000), Flaubert (Naturel 2007), Régnier (Milly 2000).

La *Recherche* se situe dans la partie haute pour pratiquement tous les indices – sauf le mode qui est la valeur la plus fréquente : il y a donc des phrases courtes dans la Recherche ! - et notamment pour l'étendue, la médiale ou la propension à concentrer une proportion importante

[5] Pour une définition de ces notions : Labbé et Labbé 2010, et pour leur application à Proust : Labbé et Labbé 2018.

du texte dans les phrases les plus longues (rapport médiane/médiale et indice de Gini). Cependant, on observe certaines caractéristiques égales ou supérieures à celle de Proust dans quelques œuvres : Huysmans (*A rebours*), les frères Goncourt (*Mme Gervaisais*), Barbey d'Aurevilly (*Le Chevalier des Touches, les Diaboliques*). Mais ces textes sont beaucoup plus brefs que la *Recherche* et les autres livres des mêmes auteurs ne présentent pas ces caractéristiques singulières comme s'il s'agissait d'accidents dans leurs œuvres, ou d'expérimentations. De même Dumas (père), Hugo, Vallès ou Vigny présentent la même propension à concentrer une proportion importante du texte dans des phrases relativement longues mais celles-ci n'en demeurent pas moins nettement plus brèves, en valeur absolue, que celles de Proust.

Seuls deux auteurs (Hugo et les frères Goncourt) ont approché les dimensions des phrases les plus longues de Proust. L'annexe 6 permet de comparer quatre de ces phrases remarquables et d'observer la singularité du style proustien : alors que Hugo ou les frères Goncourt procèdent par empilement de petites propositions faiblement hiérarchisées et utilisent beaucoup les coordinations, les virgules et points virgules, Proust imbrique ces propositions souvent assez longues en constructions complexes et solidement reliées entre elles. Pour comprendre cette singularité, on peut également se reporter à l'analyse de C. Bureau (1977) et au schéma qu'il donne pour représenter la construction de la phrase sur les chambres (deuxième dans l'annexe 6).

F. Richaudeau (1988) étudiant les phrases de la *Recherche*, a découvert que la majorité de celles de très grande longueur contiennent un thème (parmi d'autres puisque généralement plusieurs thèmes sont imbriqués) : les mécanismes de la pensée et de la mémoire. En effet, les phrases "hors normes" (du point de vue statistique) dévoilent les préoccupations de l'écrivain. A force de penser à une question, les termes lui en sont devenus familiers et il veut en rendre compte dans toute sa complexité. Tout naturellement, il se laisse entraîner par son sujet et produit des énoncés disproportionnés. Dès lors, il suffit de relever ces phrases hors normes pour connaître les préoccupations qui occupent l'esprit de celui qui parle ou écrit. L'annexe 6 donne quatre exemples assez clairs à ce sujet, notamment la dette de V. Hugo envers Louis-Philippe qui l'avait nommé pair de France en 1845 et la fascination des frères Goncourt pour un quartier populaire de Rome.

Au total, dans l'état actuel de nos dépouillements, la singularité des phrases de Proust apparaît nettement dans la littérature de fiction du XIX et du début du XXe siècle. Et ceci d'autant plus qu'il se distingue également sur un second plan.

2. La "structuration" de la phrase mesurée à l'aide de deux indices : la ponctuation interne, les connecteurs.

La dernière colonne de l'annexe 4 indique le nombre de ponctuations internes à la phrase selon une longueur standardisée. Par exemple, dans la *Recherche*, il y a en moyenne 10,3

ponctuations pour 100 mots. Cette densité est plus faible que celles observées dans les œuvres de France (10,6), de Barrès (10,9) et surtout de Bourget (13,2) alors que l'on s'attend à ce que ces ratios soient inférieurs puisque les phrases de ces trois auteurs sont beaucoup moins complexes que celles de Proust.

Chez ce dernier la cohésion de la phrase (et des phrases entre elles) est obtenue non pas par la ponctuation interne mais grâce aux constructions relatives (tableau 8) et conjonctives ainsi qu'à l'aide de divers autres mots outils, comme certaines prépositions ou adverbes (tableau 11). Ainsi la *Recherche* contient 58% de conjonctions de subordination de plus que la moyenne des autres romanciers, ce qui est tout à fait remarquable et ne peut s'expliquer seulement par la préférence de l'auteur envers le verbe.

Tableau 11 Rangs et poids des prépositions et des conjonctions dans la *Recherche* comparée aux autres romans du XIXe.

Proust *Recherche*		A	Romans	B	
Rang	Vocable	Fréquence (‰)	Rang	Fréquence (‰)	A/B (%)
1	de	72,67	1	70,30	+ 3,4
2	à	26,79	2	25,66	+ 4,4
3	que	20,61	4	13,19	+ 56,3
4	et	18,96	3	22,46	-15,6
5	en	8,25	5	8,34	-1,1
6	pour	7,81	7	6,22	+ 25,6
7	dans	7,65	6	7,97	-4,0
8	mais	6,44	10	4,31	+ 49,4
9	comme	5,76	11	4,25	+ 35,5
10	par	5,00	9	4,39	+ 13,9
11	avec	4,42	12	4,20	+ 5,2
12	si	3,72	13	2,57	+ 44,7
13	sur	3,06	8	4,49	-31,8
14	ou	2,68	16	1,61	+ 66,5
15	sans	2,55	14	2,32	+ 9,9
16	quand	2,53	15	1,70	+ 48,8
17	chez	1,71	17	1,05	+ 62,9
18	car	1,38	21	0,90	+ 53,3
19	parce que	1,09	33	0,41	+ 165,9
20	après	0,95	18	1,04	-8,7

Pour la lecture, voir la légende sous le tableau 8.

Alors que Proust n'utilise pas plus de coordinations que la moyenne, sa préférence pour la subordination est massive : *parce que* (+166%, à mettre en relation avec "car"), *que* (+56%) qui

est l'outil principal de la subordination, *si* (+48%). Sont également significatifs les suremplois de : *chez, ou, mais, quand.*

En prolongeant la liste on trouve également : "même" (adverbe, +88%), "peut-être" et "ailleurs" (+70%), "seulement" (+68), "tard", "où", etc. En revanche, Proust évite autant que possible *et* ou *sur.*

Pour compléter ce portrait, il faudrait encore analyser les figures de rhétorique employées dans ces phrases longues. La parenthèse et l'incidente sont les préférées : plus de la moitié des phrases de longueur supérieure au dernier décile en contiennent. Elles consistent à interrompre le fil du discours pour introduire une autre proposition plus ou moins en rapport avec la principale. Curtius (1928)[6] parle à leur propos d'une sorte de suspension qui doit maintenir le lecteur en haleine et mettre la chute en valeur en la retardant. On trouve ensuite l'*inversion* (consistant à placer le complément d'objet avant le verbe ou le sujet après le verbe), l'*épanalepse* (répétition d'un mot ou d'un groupe de mots dans des sens plus ou moins semblables), etc. Pour l'instant, ce type de recensement ne peut être fait que manuellement et il n'aurait véritablement de portée que si l'on disposait de dépouillements semblables sur d'autres auteurs ce qui permettrait de mettre en lumière les singularités de Proust.

Conclusions

Le but est de fournir aux chercheurs des données utiles à propos des œuvres et non de se substituer à eux pour les conclusions et les commentaires. Nous nous permettons cependant de livrer un résumé des principales caractéristiques de la *Recherche*.

La *Recherche* apparaît, statistiquement, comme une œuvre singulière dans notre histoire littéraire, tant du point de vue du vocabulaire - marqué par une prépondérance remarquable du verbe, de la première personne du singulier, et des mots de liaison (pronoms relatifs et subordinations) - que du point de vue stylistique. C'est un long monologue du narrateur avec des caractéristiques de l'oral soutenu et des phrases sans commune mesure avec celles des autres romans du XIXe et du début du XXe. Ces phrases combinent des propositions incidentes, relatives et subordonnées dont la diversité évite la monotonie de l'empilement et surprennent le lecteur.

Les données qui viennent d'être présentées succinctement dans cette communication ont peut-être paru banales ou évidentes aux lecteurs familiers de Proust. Mais qui les avait déjà mesurées précisément ? Certes, certaines des conclusions ont déjà été énoncées par des critiques ou des universitaires. Mais leur statut change de nature : nos conclusions sont vérifiables et

[6] Voir les deux extraits qu'en donne Tadié 1971 "Etude de Lilas. Le rythme des phrases" (p. 68-72)

reproductibles au lieu d'être fondées sur une lecture érudite où l'intuition du critique joue la part essentielle.

Ajoutons que cette communication ne présente qu'une petite partie des résultats obtenus. On peut aussi examiner la répartition des mots (et des thèmes), le sens spécifique que donne l'auteur à ses mots favoris, les principales combinaisons de mots, approfondir les thèmes, détecter les ruptures et les continuités dans la *Recherche*, mesurer la richesse de son vocabulaire et ses singularités par rapport aux romans contemporains, etc. Toutes ces données seront présentées ultérieurement.

Il s'agissait ici d'une brève présentation suggérant combien ces données sont riches et quelles perspectives nouvelles elles peuvent ouvrir aux "humanités" classiques.

Il s'agit aussi de permettre la reproductibilité de nos expériences, car en science, les chercheurs doivent disposer des moyens de refaire les calculs des autres et de parvenir aux mêmes résultats.

Des travaux ultérieurs, utilisant les mêmes méthodes appliquées à ces corpus littéraires, permettront d'affiner les portraits lexicaux et stylistiques des principaux auteurs qui ont dominé la littérature française de ces quatre derniers siècles. En plaçant certains de nos fichiers à la disposition des chercheurs, nous espérons susciter d'autres recherches en ce domaine.

L'ensemble présenté aujourd'hui forme un embryon de ce que pourrait être la section «romans du XIXe-XXe siècles » d'une bibliothèque électronique du français moderne.

Les grandes bibliothèques électroniques, ainsi conçues, pourraient apporter une aide précieuse aux linguistes et aux lexicographes. L'outil trouverait également des applications dans de nombreuses activités allant de la terminologie à la critique littéraire, en passant par l'enseignement des langues, la traduction assistée par ordinateur, l'histoire de la littérature, la science de la communication ou la recherche d'informations sur la toile.

Remerciements et crédit.

La majorité des textes utilisés ont été déchargés sur Wikisource dont nous remercions les contributeurs anonymes.

Nous remercions également les organisateurs de cette semaine DATA-SHS et les responsables de la PUD - Grenoble-Alpes.

Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Gaétan Péaquin, Jacques Picard, André Pibarot, Mathieu Ruhlman et Jacques Savoy ont collaboré à la mise au point des outils de lexicométrie.

Les logiciels d'analyse des corpus lemmatisés sont disponibles auprès de Cyril et Dominique Labbé.

Toutes nos recherches ont été réalisées sans financement public ni mécénat.

Références

Les éditions utilisées :

- *Les Plaisirs et des jours*. Paris : Calmann-Lévy, 1896. Sans la préface d'A. France.
- *Du côté de chez Swann*. Paris : Gallimard, 1919.
- *A l'ombre des jeunes filles en fleurs*. Paris : Gallimard, 1919.
- Le côté de Guermantes. Paris : Gallimard, 1920 et 1921.
- Sodome et Gomorrhe. Paris : Gallimard, 1921 et 1922.
- *La Prisonnière*. Paris : Gallimard, 1923.
- Albertine disparue. Paris : Gallimard, 1925.
- *Le Temps retrouvé*. Paris : Gallimard, 1927.

Bibliographies et biographies

Bonnet Henri (1976). *Marcel Proust de 1907 à 1914. Bibliographie complémentaire. Index général des bibliographies*. Paris : Nizet

Bouillaguet Annick & Rogers Brian G (dir) (2004). *Dictionnaire Marcel Proust*. Paris : Champion.

Graham Victor F. (1976). *Bibliographie des études sur Marcel Proust et son œuvre*. Genève : Droz.

Tadié Jean-Tves (1996). *Marcel Proust*. Biographie. Paris : Gallimard.

Deux bulletins :

Bulletin Marcel Proust (https://www.amisdeproust.fr/index.php/fr/)

Bulletin d'informations proustiennes de l'Institut des textes et manuscrits modernes (CNRS) https://www.presses.ens.fr/527

Ouvrages cités dans la communication.

Benveniste Emile (1966 & 1970). *Problèmes de linguistique générale*. Paris: Gallimard (réed. 1980).

Bouillaguet Annick (2000). *Proust lecteur de Balzac et de Flaubert. L'imitation cryptée*. Paris : Champion.

Brunet Etienne (1981). La phrase de Proust. Longueur et rythme. *Travaux du cercle linguistique de Nice*, p. 97-117.

Brunet Étienne (1983). *Le vocabulaire de Proust avec l'Index complet et synoptique de "A la recherche du temps perdu", d'après les données de L'Institut National de la langue française (CNRS)*. Paris : Slatkine-Champion.

Bureau Conrad (1976). Marcel Proust ou le temps retrouvé par la phrase. *Linguistique fonctionnelle et stylistique objective*. Paris : PUF, p. 178-231.

Chantal René de (1967). *Marcel Proust, critique littéraire*. Montréal : Presses de l'Université de Montréal.

Compagnon Antoine (dir) (2009). *Proust, la mémoire et la littérature*. Paris : Odile Jacob

Curtius Ernst-Robert (1928). *Marcel Proust.* Paris : La Revue nouvelle.

Erman Michel (1988). *L'œil de Proust. Ecriture et voyeurisme dans* A la recherche du temps perdu. Paris : Nizet,

Ferré André (1957). La ponctuation de M. Proust. *Bulletin de la Société des Amis de Marcel Proust*, 7, p 171-192.

Héron Pierre-Marie (2010). Littérature et conversation au XXe siècle : Proust (encore). *Revue d'Histoire Littéraire de la France*. 110/1, p. 93-111.

Jullien Dominique (1989). *Proust et ses modèles. Les* Mille et Une Nuits *et les* Mémoires *de Saint-Simon*. Paris : José Corti.

Labbé Cyril, Labbé Dominique (2010). Ce que disent leurs phrases. In Bolasco S., Chiari I., Giuliano L. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto. Vol 1, p. 297-307.

Labbé Cyril, Labbé Dominique (2013). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. In Banks David (Ed). *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85.

Labbé Cyril, Labbé Dominique (2018). Les phrases de Marcel Proust. In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data.* Roma: UniversItalia, 2018, p. 400-410.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Levaillant J. (1952). Note sur le personnage de Bergotte. *Revue des sciences humaines*. Janvier-Mars 1952, p 33-48.

Mandelbrot Benoît (1957). Étude de la loi d'Estoup et de Zipf. Fréquences des mots dans le discours. Apostel Léo et al. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

Mauriac-Dyer Nathalie, Yoshikawa Kazuyoshi et Robert Pierre-Edmond (éds) (2013). *Proust face à l'héritage du XIXe siècle. Tradition et métamorphose*. Paris : Presses de la Sorbonne nouvelle.

Mauriac-Dyer Nathalie (2005). *Proust inachevé. Le dossier* Albertine disparue. Paris : Champion.

Milly Jean (1970). *Proust et le style*. Minard-Lettres Modernes.

Milly Jean (1975). *La phrase de Proust. Des phrases de Bergotte aux phrases de Vinteuil*. Paris : Larousse.

Milly Jean (1985). *Proust dans le texte et l'avant-texte*. Paris : Flammarion.

Milly Jean (1986). *La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine.

Milly Jean (2000). Proust et Henri de Régnier : modes proustiens de l'intertextualité. *Revue d'Histoire littéraire de la France*. 2000-1, p 27-44.

Monière Dominique, Labbé Cyril & Labbé Dominique (2008). Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, p. 43-69.

Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Nathan Jacques (1969). *Citations, références et allusions de Marcel Proust dans* A la recherche du temps perdu. Paris : Nizet (Première édition : 1953).

Naturel Mireille (1986). La phrase longue dans le *Temps retrouvé* : fonctions et limites. *Bulletin d'Informations Proustiennes*, 22, p 57-65.

Naturel Mireille (2007). *Proust et Flaubert : un secret d'écriture*. Amsterdam-Atlanta : Rodopi.

Richaudeau François (1988). *Ce que révèlent leurs phrases*. Paris : Retz.

Rogers Brian (2000). *Proust et Barbey d'Aurevilly. Le dessous des cartes*. Paris : Champion.

Serça Isabelle (2010). *Les coutures apparentes de la Recherche. Proust et la ponctuation*. Paris : Champion.

Spitzer Léo (1970). *Le style de Marcel Proust. Etudes de style*. Paris : Gallimard, p 397-473.

Tadié Jean-Yves (1971). *Lectures de Proust*. Paris ; A. Colin.

Tadié Jean-Yves (1971). *Proust et le roman. Essai sur les formes et les techniques du roman dans* A la recherche du temps perdu. Paris : Gallimard.

Zipf George K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Annexe 1. Bibliothèque électronique du français moderne (novembre 2019).

	Nombre de documents	Volume (mots)
Discours politique (1867-2018)		
France	11 447	22 291 261
Amérique du nord	2 861	5 764 227
Autres pays francophones	101	577 377
Total discours politique	**14 409**	**28 632 865**
Littérature (XVIIe-XXe)		
Romans et nouvelles	172	15 089 592
Théâtre	395	5 415 069
Poésie	28	663 931
Correspondance	5	599 343
Divers	28	1 712 962
Total littérature	**597**	**23 480 897**
Presse	1 513	2 039 631
Scientifique	158	774 514
Oral (entretiens)	438	3 637 009
Total bibliothèque	**17 115**	**58 564 916**

Annexe 2. Corpus *A la Recherche du temps perdu* (Marcel Proust. Paris : Gallimard, 1919-1927)*

Livres	Longueur	Vocabulaire	N phrases
Combray	79 904	6 497	1 727
Un amour de Swann	84 141	5 852	2 226
Noms de pays : le nom	19 434	2 819	374
Du côté de chez Swann (1919)**	**183 481**	**9 344**	**4 327**
Autour de Mme Swann	91 451	6 529	2 511
Noms de pays : le pays	134 192	8 276	3 334
A l'ombre des jeunes filles en fleur (1919)	**225 643**	**10 388**	**5 845**
Le côté de Guermantes 1	75 492	6 279	1 903
Le côté de Guermantes 2, chapitre 1	84 354	6 366	2 781
Le côté de Guermantes 2, chapitre 2	89 723	6 699	2 700
Le côté de Guermantes (1920-21)	**249 569**	**11 172**	**7 384**
Sodome et Gomorrhe 1	13 512	2 475	271
Sodome et Gomorrhe 2, chapitre 1	63 788	5 566	2 082
Sodome et Gomorrhe 2, chapitre 2	84 676	6 688	3 056
Sodome et Gomorrhe 2, chapitre 3	57 604	5 303	1 811
Sodome et Gomorrhe 2, chapitre 4	8 137	1 371	250
Sodome et Gomorrhe (1921-22)	**227 717**	**10 961**	**7 470**
La Prisonnière (1923)	173 408	9 056	5 124
La Fugitive (1925)	115 865	6 447	3 255
Le Temps retrouvé (1927)	152 157	8 698	3 930
Dernier volume (posthume)***	**441 430**	**13 513**	**12 309**
Total général (*A la recherche du temps perdu)*	**1 327 838**	**21 780**	**37 335**

* Ce tableau remplace celui en annexe de Labbé et Labbé 2018.

** Paru initialement chez Grasset en novembre 1913 et repris chez Gallimard en 1919.

*** Publié en trois tomes séparés. Dans la première édition, *la Fugitive* porte le titre : *Albertine disparue*. Proust a relu les épreuves de *la Prisonnière*. A sa mort (novembre 1922), le manuscrit était achevé mais les trois derniers livres ont été édités par son frère. Cette édition est considérée comme fautive en plusieurs passages (erreurs qui ont été rectifiées dans les éditions ultérieures, notamment pour la Pléiade). Nous conservons cependant cette première édition car c'est la seule dans le domaine public. Elle peut donc être communiquée librement sans risquer de violer le droit des éditeurs de ces versions ultérieures.

Annexe 3. Les romans du XIXe : 1800-1920

Auteurs	N Romans	Mots	Vocables
Balzac Honoré de (1799-1850)	17	1 158 871	19 530
Barbey d'Aurevilly Jules (1808-1889)	2	158 282	9 053
Barrès Maurice (1862-1923)	2	217 087	11 369
Bourget Paul (1832-1935)	5	381 266	12 103
Chateaubriand François René de (1768-1848)	3	105 801	7 181
Daudet Alphonse (1840-1897)	3	167 385	8 746
Dumas Alexandre (1802-1870)	3	932 472	15 085
Dumas Alexandre fils (1824-1895)	1	69 648	3 496
Erckmann (1822-1899) et Chatrian (1826-1890)	1	68 528	3 649
Flaubert Gustave (1821-1880)	6	503 259	16 925
Alain-Fournier (1886 - 1914)	1	69 302	4 732
France Anatole (1844-1924)	7	490 796	16 492
Fromentin Eugène (1820-1876)	1	80 238	5 629
Gautier Théophile (1811-1872)	3	123 725	9 276
Goncourt Jules de (1830-1870) et Edmond de (1822-1896)	2	134 813	8 931
Hugo Victor (1802-1885)	2	749 776	20 562
Huysmans Karl Joris (1848-1907)	2	97 021	11 046
Lamartine Alphonse de (1790 - 1869)	2	115 743	5 858
Loti Pierre (1850-1923)	2	108 203	6 701
Maupassant Guy de (1850-1893)	7	471 235	12 489
Musset Alfred de (1810-1857)	5	146 178	6 462
Nerval Gérard Labrunie (1808-1855)	3	131 361	8 639
Proust Marcel (1871-1922)	9	1 386 377	22 242
Régnier Henri de (1864-1936)	2	172 925	8 449
Sainte-Beuve Charles-Augustin (1804-1869)	1	132 777	7 727
Sand George (1804-1876)	4	246 856	8 261
Staël-Holstein Anne-Louise (1866-1817)	1	110 628	3 674
Stendhal Henri Beyle (1783-1842)	2	385 786	9 725
Sue Eugène (1804-1857)	1	578 931	12 318
Vallès Jules (1832-1885)	1	92 294	5 928
Verne Jules (1828-1905)	2	131 876	8 498
Vigny Alfred de (1797-1863)	2	198 147	8 250
Villiers de l'Isle Adam (1838-1889)	1	68 098	7 338
Zola Emile (1840-1902)	10	1 518 313	17 968
34 auteurs	116	11 457 253	54 898

Annexe 4. Principaux indices statistiques concernant les phrases de Proust comparées à celle de Bourget, France et Barrès.

Auteurs, textes	Etendue	Mode	Médiane (Me)	Moyenne	Médiale (Ml)	V%	Ml/Me	Concentration	Ponctuations internes (%)
Proust									
Plaisirs	250	7	21,30	27,87	37,20	84,68	1,75	543	10,4
Combray	542	15	33,70	46,27	67,10	90,99	1,99	569	9,7
Swann	387	9	29,00	39,84	58,40	91,37	2,01	574	10,2
Ombre	352	12	28,30	38,60	55,40	87,55	1,96	561	9,3
Guermantes	346	12	24,98	33,80	47,14	86,54	1,89	552	10,1
Sodome	931	10	22,52	30,49	42,61	90,47	1,89	546	11,2
Prisonnière	430	14	24,88	33,84	47,32	86,99	1,90	551	11,2
Fugitive	315	17	27,80	35,60	48,55	82,62	1,75	536	9,5
Retrouvé	299	16	29,69	38,71	52,90	80,89	1,78	531	10,1
Recherche	931	11	25,88	35,20	**49,57**	89,27	1,92	554	10,3
Bourget									
Idylle	161	10	17,95	22,75	31,80	78,45	1,77	540	12,5
Fantôme	201	10	17,63	22,73	31,77	81,43	1,80	544	11,2
Deux sœurs	123	10	16,54	20,67	27,78	76,24	1,68	524	10,5
Albani	129	6	13,24	17,32	24,16	77,89	1,82	536	14,3
Ecuyère	192	7	16,80	21,34	29,29	76,30	1,74	527	12,7
Total Bourget	201	7	16,62	21,34	29,58	79,19	1,78	539	12,2
France									
Bonnard	125	11	15,35	18,73	24,25	70,09	1,58	497	9,2
Thaïs	139	11	16,74	19,90	24,55	65,53	1,47	463	9,6
Pédauque	135	8	16,69	20,68	26,44	70,99	1,58	489	10,0
Lys rouge	152	6	11,81	15,30	19,31	77,06	1,64	499	10,5

Bergeret	394	9	11,26	15,35	19,93	92,09	1,77	512	10,4
Pingouins	195	8	17,50	22,31	28,79	74,72	1,65	502	10,0
Dieux	142	15	17,51	21,72	29,26	74,67	1,67	520	11,1
Anges	204	11	16,46	21,07	27,68	78,47	1,68	513	10,6
Total France	394	8	15,79	19,10	26,07	79,67	1,65	504	10,6
Barrès									
Déracinés	139	13	18,16	22,08	28,34	68,47	1,56	488	11,5
Colline inspirée	195	12	17,96	21,96	28,53	71,69	1,59	496	9,8
Total Barrès	195	12	18,08	22,03	28,41	69,71	1,57	493	10,9

Annexe 5 Longueur des phrases chez quelques écrivains antérieurs ou contemporains de Proust

	Etendue	Mode	Médiane	Moyenne	Médiale	Me/Ml	Gini
Recherche	**931**	**11**	**26,28**	**35,57**	**49,93**	**0,900**	**554**
Balzac*	391	10	17,27	21,88	29,00	0,680	511
Barbey d'A. (*Chevalier*)	192	7	21,92	29,40	43,00	**0,964**	**557**
Barbey d'A (*Maîtresse*)	171	7	17,60	22,60	30,39	0727	517
Barbey d'A. (*Diaboliques*)	189	7	20,10	27,10	39,33	**0,956**	553
Barrès*	195	8	17,86	21,94	28,59	0,601	497
Bourget*	201	7	16,62	21,34	29,58	0,780	539
Chateaubriand (*Mémoires*)	195	**22**	24,46	28,5	34,28	0,401	437
Daudet*	203	5	13,14	17,84	25,26	0,923	549
Dumas	243	7	14,90	20,28	29,00	**0,947**	**557**
Flaubert*	231	7	13,75	18,37	25,24	0,837	528
France	394	8	15,79	19,98	26,06	0,651	504
Gautier*	282	**18**	27,11	33,07	41,90	0,546	493
Giraudoux*	466	4	18,60	25,77	37,76	**1,031**	**580**
Goncourt (*Lacerteux*)	133	11	15,58	21,30	29,95	0,922	547
Goncourt (*Gervaisais***)**	**670**	8	24,17	34,05	**51,47**	**1,130**	**597**
Goncourt (*Journal*)	373	3	19,80	25,37	37,62	0,900	**580**
Hugo*	**828**	6	11,39	16,89	23,68	**1,079**	**561**
Huysmans (*Marthe*)	177	8	19,84	26,22	37,69	0,899	**571**
Huysmans (*A rebours*)	254	**28**	**44,24**	**51,49**	**65,82**	0,488	**558**
Lamartine*	375	9	20,53	26,97	36,02	0,754	532
Maupassant*	168	6	14,44	18,98	26,39	0,828	542
Musset*	197	**16**	19,56	23,82	29,57	0,512	485
Nerval*r	136	12	19,93	24,21	31,27	0,569	499
Régnier	158	9	17,02	22,83	30,73	0,804	517
Saint-Simon	361	**18**	27,89	34,15	44,14	0,523	506
Sainte-Beuve	224	**16**	29,71	35,45	42,69	0,437	460
Sand (*Champi*)	117	**21**	22,11	26,19	32,56	0,473	477
Sévigné (*Lettres*)	307	11	25,72	31,99	40,96	0,593	490
Stael	178	2	29,12	33,90	44,74	0,536	517
Stendhal*	235	**18**	20,18	23,92	29,79	0,477	463
Sue	200	5	14,54	19,62	28,43	**0,956**	**567**
Vallès	117	5	11,28	16,08	23,56	**1,089**	553
Vigny*	315	**17**	20,82	27,47	37,41	0,797	538
Zola*	153	8	15,80	19,91	25,66	0,624	491

Annexe 6. Phrases remarquables

1. Les deux plus longues phrases de Proust

Sans honneur que précaire, sans liberté que provisoire, jusqu'à la découverte du crime ; sans situation qu'instable, comme pour le poète la veille fêté dans tous les salons, applaudi dans tous les théâtres de Londres, chassé le lendemain de tous les garnis sans pouvoir trouver un oreiller où reposer sa tête, tournant la meule comme Samson et disant comme lui : « Les deux sexes mourront chacun de son côté » ; exclus même, hors les jours de grande infortune où le plus grand nombre se rallie autour de la victime, comme les Juifs autour de Dreyfus, de la sympathie — parfois de la société — de leurs semblables, auxquels ils donnent le dégoût de voir ce qu'ils sont, dépeint dans un miroir qui, ne les flattant plus, accuse toutes les tares qu'ils n'avaient pas voulu remarquer chez eux-mêmes et qui leur fait comprendre que ce qu'ils appelaient leur amour (et à quoi, en jouant sur le mot, ils avaient, par sens social, annexé tout ce que la poésie, la peinture, la musique, la chevalerie, l'ascétisme, ont pu ajouter à l'amour) découle non d'un idéal de beauté qu'ils ont élu, mais d'une maladie inguérissable ; comme les Juifs encore (sauf quelques-uns qui ne veulent fréquenter que ceux de leur race, ont toujours à la bouche les mots rituels et les plaisanteries consacrées) se fuyant les uns les autres, recherchant ceux qui leur sont le plus opposés, qui ne veulent pas d'eux, pardonnant leurs rebuffades, s'enivrant de leurs complaisances ; mais aussi rassemblés à leurs pareils par l'ostracisme qui les frappe, l'opprobre où ils sont tombés, ayant fini par prendre, par une persécution semblable à celle d'Israël, les caractères physiques et moraux d'une race, parfois beaux, souvent affreux, trouvant (malgré toutes les moqueries dont celui qui, plus mêlé, mieux assimilé à la race adverse, est relativement, en apparence, le moins inverti, accable qui l'est demeuré davantage) une détente dans la fréquentation de leurs semblables, et même un appui dans leur existence, si bien que, tout en niant qu'ils soient une race (dont le nom est la plus grande injure), ceux qui parviennent à cacher qu'ils en sont, ils les démasquent volontiers, moins pour leur nuire, ce qu'ils ne détestent pas, que pour s'excuser, et allant chercher, comme un médecin l'appendicite, l'inversion jusque dans l'histoire, ayant plaisir à rappeler que Socrate était l'un d'eux, comme les Israélites disent de Jésus, sans songer qu'il n'y avait pas d'anormaux quand l'homosexualité était la norme, pas d'antichrétiens avant le Christ, que l'opprobre seul fait le crime, parce qu'il n'a laissé subsister que ceux qui étaient réfractaires à toute prédication, à tout exemple, à tout châtiment, en vertu d'une disposition innée tellement spéciale qu'elle répugne plus aux autres hommes (encore qu'elle puisse s'accompagner de hautes qualités morales) que de certains vices qui y contredisent, comme le vol, la cruauté, la mauvaise foi, mieux compris, donc plus excusés du commun des hommes ; formant une franc-maçonnerie bien plus étendue, plus efficace et moins soupçonnée que celle des loges, car elle repose sur une identité de goûts, de besoins, d'habitudes, de dangers, d'apprentissage, de savoir, de trafic, de glossaire, et dans laquelle les membres mêmes qui souhaitent de ne pas se connaître aussitôt se reconnaissent à des signes naturels ou de convention, involontaires ou voulus, qui signalent un de ses semblables au mendiant dans le grand seigneur à qui il ferme la portière de sa voiture, au père dans le fiancé de sa fille, à celui qui avait voulu se guérir, se confesser, qui avait à se défendre, dans le médecin, dans le prêtre, dans l'avocat qu'il est allé trouver ; tous obligés à protéger leur secret, mais ayant leur part d'un secret des autres que le reste de l'humanité ne soupçonne pas et qui fait qu'à eux les romans d'aventure les plus invraisemblables semblent vrais, car dans cette vie romanesque, anachronique, l'ambassadeur est ami du forçat ; le prince, avec une certaine liberté d'allures que donne l'éducation aristocratique et qu'un petit bourgeois tremblant n'aurait pas, en sortant de chez la duchesse s'en va conférer avec l'apache ; partie réprouvée de la collectivité humaine, mais partie importante, soupçonnée là où elle n'est pas étalée, insolente, impunie là où elle n'est pas devinée ; comptant des adhérents partout, dans le peuple, dans l'armée, dans le temple, au bagne, sur le trône ; vivant enfin, du moins un grand nombre, dans l'intimité caressante et dangereuse avec les hommes de l'autre race, les provoquant, jouant avec eux à parler de son vice comme s'il n'était pas sien, jeu qui est rendu facile par l'aveuglement

ou la fausseté des autres, jeu qui peut se prolonger des années jusqu'au jour du scandale où ces dompteurs sont dévorés ; jusque-là obligés de cacher leur vie, de détourner leurs regards d'où ils voudraient se fixer, de les fixer sur ce dont ils voudraient se détourner, de changer le genre de bien des adjectifs dans leur vocabulaire, contrainte sociale légère auprès de la contrainte intérieure que leur vice, ou ce qu'on nomme improprement ainsi, leur impose non plus à l'égard des autres mais d'eux-mêmes, et de façon qu'à eux-mêmes il ne leur paraisse pas un vice.
(*Sodome et Gomorrhe*, 1, 931 mots).

Mais j'avais revu tantôt l'une, tantôt l'autre, des chambres que j'avais habitées dans ma vie, et je finissais par me les rappeler toutes dans les longues rêveries qui suivaient mon réveil ; chambres d'hiver où quand on est couché, on se blottit la tête dans un nid qu'on se tresse avec les choses les plus disparates : un coin de l'oreiller, le haut des couvertures, un bout de châle, le bord du lit, et un numéro des Débats roses, qu'on finit par cimenter ensemble selon la technique des oiseaux en s'y appuyant indéfiniment ; où, par un temps glacial, le plaisir qu'on goûte est de se sentir séparé du dehors (comme l'hirondelle de mer qui a son nid au fond d'un souterrain dans la chaleur de la terre), et où, le feu étant entretenu toute la nuit dans la cheminée, on dort dans un grand manteau d'air chaud et fumeux, traversé des lueurs des tisons qui se rallument, sorte d'impalpable alcôve, de chaude caverne creusée au sein de la chambre même, zone ardente et mobile en ses contours thermiques, aérée de souffles qui nous rafraîchissent la figure et viennent des angles, des parties voisines de la fenêtre ou éloignées du foyer, et qui se sont refroidies ; — chambres d'été où l'on aime être uni à la nuit tiède, où le clair de lune appuyé aux volets entr'ouverts jette jusqu'au pied du lit son échelle enchantée, où on dort presque en plein air, comme la mésange balancée par la brise à la pointe d'un rayon ; — parfois la chambre Louis XVI, si gaie que même le premier soir je n'y avais pas été trop malheureux, et où les colonnettes qui soutenaient légèrement le plafond s'écartaient avec tant de grâce pour montrer et réserver la place du lit ; parfois au contraire celle, petite et si élevée de plafond, creusée en forme de pyramide dans la hauteur de deux étages et partiellement revêtue d'acajou, où, dès la première seconde, j'avais été intoxiqué moralement par l'odeur inconnue du vétiver, convaincu de l'hostilité des rideaux violets et de l'insolente indifférence de la pendule qui jacassait tout haut comme si je n'eusse pas été là ; — où une étrange et impitoyable glace à pieds quadrangulaires barrant obliquement un des angles de la pièce se creusait à vif dans la douce plénitude de mon champ visuel accoutumé un emplacement qui n'y était pas prévu ; — où ma pensée, s'efforçant pendant des heures de se disloquer, de s'étirer en hauteur pour prendre exactement la forme de la chambre et arriver à remplir jusqu'en haut son gigantesque entonnoir, avait souffert bien de dures nuits, tandis que j'étais étendu dans mon lit, les yeux levés, l'oreille anxieuse, la narine rétive, le cœur battant ; jusqu'à ce que l'habitude eût changé la couleur des rideaux, fait taire la pendule, enseigné la pitié à la glace oblique et cruelle, dissimulé, sinon chassé complètement, l'odeur du vétiver, et notablement diminué la hauteur apparente du plafond.
(*Combray*, 542 mots)

2. La plus longue phrase de V. Hugo

Fils d'un père auquel l'histoire accordera certainement les circonstances atténuantes, mais aussi digne d'estime que ce père avait été digue de blâme ; ayant toutes les vertus privées et plusieurs des vertus publiques ; soigneux de sa santé, de sa fortune, de sa personne, de ses affaires ; connaissant le prix d'une minute et pas toujours le prix d'une année ; sobre, serein, paisible, patient ; bonhomme et bon prince ; couchant avec sa femme, et ayant dans son palais des laquais chargés de faire voir le lit conjugal aux bourgeois, ostentation d'alcôve régulière devenue utile après les anciens étalages illégitimes de la branche aînée ; sachant toutes les langues de l'Europe et, ce qui est plus rare, tous les langages de tous les intérêts, et les parlant ; admirable représentant de " la classe moyenne ", mais la dépassant, et de toutes les façons plus grand qu'elle ; ayant l'excellent l'esprit, tout en appréciant le sang dont il sortait,

de se compter surtout par sa valeur intrinsèque, et, sur la question même de sa race, très particulier, se déclarant Orléans et non Bourbon ; très premier prince du sang tant qu'il n'avait été qu'altesse sérénissime, mais franc bourgeois le jour où il fut majesté ; diffus en public, concis dans l'intimité ; avare signalé, mais non prouvé ; au fond, un de ces économes aisément prodigues pour leur fantaisie ou leur devoir ; lettré, et peu sensible aux lettres ; gentilhomme, mais non chevalier ; simple, calme et fort ; adoré de sa famille et de sa maison ; causeur séduisant, homme d'état désabusé, intérieurement froid, dominé par l'intérêt immédiat, gouvernant toujours au plus près, incapable de rancune et de reconnaissance, usant sans pitié les supériorités sur les médiocrités, habile à faire donner tort par les majorités parlementaires à ces unanimités mystérieuses qui grondent sourdement sous les trônes ; expansif, parfois imprudent dans son expansion, mais d'une merveilleuse adresse dans cette imprudence ; fertile en expédients, en visages, en masques ; faisant peur à la France de l'Europe et à l'Europe de la France ; aimant incontestablement son pays, mais préférant sa famille ; prisant plus la domination que l'autorité et l'autorité que la dignité, disposition qui a cela de funeste que, tournant tout au succès, elle admet la ruse et ne répudie pas absolument la bassesse, mais qui a cela de profitable qu'elle préserve la politique des chocs violents, l'état des fractures et la société des catastrophes ; minutieux, correct, vigilant, attentif, sagace, infatigable, se contredisant quelquefois, et se démentant ; hardi contre l'Autriche à Ancône, opiniâtre contre l'Angleterre en Espagne, bombardant Anvers et payant Pritchard ; chantant avec conviction la Marseillaise ; inaccessible à l'abattement, aux lassitudes, au goût du beau et de l'idéal, aux générosités téméraires, à l'utopie, à la chimère, à la colère, à la vanité, à la crainte ; ayant toutes les formes de l'intrépidité personnelle ; général à Valmy, soldat à Jemmapes ; tâté huit fois par le régicide, et toujours souriant ; brave comme un grenadier, courageux comme un penseur ; inquiet seulement devant les chances d'un ébranlement européen, et impropre aux grandes aventures politiques ; toujours prêt à risquer sa vie, jamais son oeuvre ; déguisant sa volonté en influence afin d'être plutôt obéi comme intelligence que comme roi ; doué d'observation et non de divination ; peu attentif aux esprits, mais se connaissant en hommes, c'est-à-dire ayant besoin de voir pour juger ; bon sens prompt et pénétrant, sagesse pratique, parole facile, mémoire prodigieuse ; puisant sans cesse dans cette mémoire, son unique point de ressemblance avec César, Alexandre et Napoléon ; sachant les faits, les détails, les dates, les noms propres, ignorant les tendances, les passions, les génies divers de la foule, les aspirations intérieures, les soulèvements cachés et obscurs des âmes, en un mot, tout ce qu'on pourrait appeler les courants invisibles des consciences ; accepté par la surface, mais peu d'accord avec la France de dessous ; s'en tirant par la finesse ; gouvernant trop et ne régnant pas assez ; son premier ministre à lui-même ; excellant à faire de la petitesse des réalités un obstacle à l'immensité des idées ; mêlant à une vraie faculté créatrice de civilisation, d'ordre et d'organisation on ne sait quel esprit de procédure et de chicane ; fondateur et procureur d'une dynastie ; ayant quelque chose de Charlemagne et quelque chose d'un avoué ; en somme, figure haute et originale, prince qui sut faire du pouvoir malgré l'inquiétude de la France, et de la puissance malgré la jalousie de l'Europe, Louis-Philippe sera classé parmi les hommes éminents de son siècle, et serait rangé parmi les gouvernants les plus illustres de l'histoire, s'il eût un peu aimé la gloire et s'il eût eu le sentiment de ce qui est grand au même degré que le sentiment de ce qui est utile.

(*Les Misérables*, Tome IV, Livre 1, 828 mots)

3. La plus longue phrase des frères Goncourt

Un quartier sauvagement populacier, rejeté, isolé sur la rive droite du Tibre, le quartier ouvrier de la manufacture des tabacs, des fabriques de bougies et de cierges pour les centaines d’églises de la ville ; le faubourg lointain, perdu, arriéré, qui garde le vieux sang de Rome dans ces mains d’hommes promptes au couteau, dans ces lignes graves de la beauté de ses femmes ; cette espèce de banlieue où semble commencer la barbarie d’un village italien, mêlant à un aspect d’Orient des souvenirs d’antiquité ; - des angles de rues étayés avec des morceaux de colonnes, des assises où les blocs sont des Minerves entières

; à côté d'une porte blanchie à la chaux, surmontée d'un morceau de natte et de l'ombre d'un moucharaby, des maisons frustes, effacées, rabotées par le temps, des façades où, sous un cintre à moitié bouché et maçonné, se dégage l'élancement d'une fine colonnette au chapiteau ionique ; - à tout moment, du plâtre déchiré, craquelé sur des briques du temps d'Auguste, des hasards de couleurs pareils à ces palettes de tons qu'un peintre garde à son mur, et d'où un bout de passé, un profil, une esquisse des grands os de Rome reparaît et reperce ; - souvent un vaste palais, noir de vieillesse, qui de sa splendeur délabrée n'a gardé qu'un vol d'oiseau de proie soutenant toujours en l'air un balcon disparu ; là-dedans, la primitivité d'une civilisation qui commence, d'une humanité crédule aux commerces naïfs : les boutiques de barbiers phlébotomistes, avec leurs enseignes, sur leurs carreaux, de jambes et de bras dont le sang jaillit rougement dans un verre ; des boucheries où le prix de la viande d'agneau est affiché sur une sorte de tambour de Basque, des boutiques de loterie, avec les numéros sortis écrits à la craie sur leurs volets ; les spaccio di vino à deux baïoques et demie, les magasins aux dessus de porte enfumés, aux ouvertures d'écurie, laissant le marchand et les marchandises au jour et à l'air de la rue, les trous béants du petit trafic où se détache, sur un fond de cave, le cuivre brillant de la balance des pays chauds ; l'étal, l'industrie, le travail, à l'état de nature, sur de petites places, au-dessus du voltigement des lessives pendues, où le moindre souffle met en passant des bruits de voiles qui se gonflent ; de grands ateliers de grossiers charronnages, le remisage sous le ciel de charrues rappelant Cincinnatus, et de robustes chars, aux roues pleines, qui pourraient encore porter les fardeaux de la République et les vieux fers de Caton ; - sur le pavé, des passages de troupeaux de chèvres blanches, se bousculant, se montant l'une sur l'autre, ou bien des repos d'attelages de buffles noirs, à l'œil de verroterie bleue, à la fade odeur de musc, immobiles dans leur ruminement méditatif ; - sur les ordures, sur les fumiers d'herbes potagères, par les rues, au fond des impasses haillonneuses, un grouillement d'animalité domestique, de volailles, de chiens quêtant, la canaille errante des bêtes ; et au milieu de tout cela, des femmes travaillant sur des chaises, au soleil qui leur marque sur la joue l'ombre de chacun de leurs cils ; le fourmillement d'une marmaille vivace jetée là à poignée par la procréation chaude, enfance aux yeux ardents, monde de petites filles, porteuses et berceuses des plus petites qu'elles, que l'on voit vaguer le long des maisons, dégrafées par derrière, la chemise passant au dos, ou, dans le sombre d'un escalier de bois, vêtues comme d'une robe de jour, descendre en s'appuyant de la main au mur, - c'est ce qu'on nomme le Transtevere à Rome.

(*Madame Gervaisais*, chapitre LXXIX, 670 mots).