L'évolution du lexique. Approche statistique.

Étienne Brunet

▶ **To cite this version:**

Étienne Brunet. L'évolution du lexique. Approche statistique.. CNRS-Editions. Histoire de la langue française 1914-1945, pp.95-124, 1995, 2-271-05386-0. hal-01558757

HAL Id: hal-01558757
https://hal.science/hal-01558757

Submitted on 9 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Étienne Brunet

L'évolution du lexique: approche statistique

Lorsque fut traitée la période 1880-1914 de la présente *Histoire de la langue*, ni les données, ni les outils ne manquaient véritablement pour une approche statistique du lexique. Les matériaux amassés à Nancy étaient disponibles à l'état brut mais aussi des synthèses plus élaborées, comme le *Dictionnaire des fréquences,* paru dès 1971[1]. Les méthodes s'étaient affinées sous l'action de Charles Muller[2]. Enfin les ordinateurs avaient perdu de leur barbarie primitive et leur accès commençait à s'ouvrir aux chercheurs littéraires. Et profitant de ces opportunités, nous nous étions lancé, à la même époque, dans l'exploitation systématique du corpus des XIXe et XXe siècles[3].

Pourtant l'heure présente est plus propice à une rétrospective qui engloberait non seulement les trente ans qui ont suivi la première guerre mondiale et le siècle qui l'a précédée, mais aussi l'ensemble de la littérature française depuis 1600. La démarche historique s'accommode mieux en effet d'une grande profondeur de champ et aime à distinguer les plans, des plus proches aux plus éloignés. C'est le propre aussi de la démarche statistique, qui est nécessairement comparative, aucune mesure, si précise soit-elle, ne fournissant le moindre enseignement si elle n'est rapportée à quelque autre. Comme dans le domaine linguistique aucune norme ne s'impose, ni naturelle comme le niveau de la mer, ni artificielle comme le mètre-étalon, toute observation chiffrée n'y a de valeur que relative et ne prend sens que par rapport à d'autres mesures. Plus les jalons sont denses et les repères assurés, plus grande est la fiabilité des résultats.

Or FRANTEXT fournit de tels jalons en abondance. Près de 3000 textes s'y trouvent rassemblés, avec un étiquetage assez discriminant pour permettre tous les regroupements, selon l'auteur, selon le genre littéraire ou - ce qui nous importe plus ici - selon la chronologie. Chaque occurrence, parmi 160 millions, peut y être repérée, chaque forme, parmi 500 000, y être relevée. Nul besoin de se déplacer à Nancy ou de demander l'envoi de

---

[1] Chez Didier, Paris. L'ouvrage, dont Robert Martin a été le maître d'oeuvre, se distribue ainsi: I - *Table alphabétique* (4 volumes); II - *Tables des fréquences décroissantes;* III - *Table des variations de fréquence*; IV - *Table de répartition des homonymes*.

[2] Charles Muller, *Initiation aux méthodes de la statistique linguistique,* Hachette Université, 1973, et *Principes et méthodes de statistique lexicale,* Hachette Université, 1977. Ces deux ouvrages viennent d'être réédités chez Champion, dans la collection Unichamp.

[3] Etienne Brunet, *Le vocabulaire français de 1789 à nos jours,* Slatkine-Champion, Genève-Paris, 1981, 3 tomes, préface de Paul Imbs.

bandes magnétiques ou d'états imprimés: grâce à la télématique et au logiciel Stella créé par Jacques Dendien, l'interrogation de cette immense base de données est devenue simple, souple, immédiate, puissante et peu coûteuse.

- I - L'inflation lexicale

On aurait tort pourtant d'y chercher ingénument, et plus encore d'y trouver, la réponse à la question première qui vient à l'esprit: Comment évolue le lexique? Car le lexique appartient à la langue, à une réalité virtuelle dont les réalisations écrites n'épuisent pas, loin s'en faut, les possibilités. Relevés et calculs ne peuvent se faire que dans le discours, c'est-à-dire dans un corpus, nécessairement limité, qui est pris pour témoin et dont la composition importe grandement, puisque de la qualité de l'échantillon dépend la portée des conclusions qu'on projette sur la population. À l'inverse des sondages électoraux qui peuvent espérer du vote réel la confirmation de leurs prévisions, nul espoir jamais d'atteindre dans son intégralité la population des mots et de l'amener devant les urnes. Quelle que soit l'étendue de l'enquête, il y aura toujours des recoins inexplorés, des lacunes imprévisibles et, ce qui est pire, des régions inaccessibles par définition: ces limbes indécis où naissent et flottent les mots qui attendent le baptême. Plutôt que de lexique il est donc avantageux de parler de vocabulaire, en entendant par là le registre des éléments lexicaux rencontrés dans les textes, sans exclusion ni extrapolation. Il est facile de voir que le vocabulaire ainsi défini ne recouvre pas la nomenclature d'un dictionnaire, quoique l'un et l'autre tendent à rejoindre, de façon asymptotique, la perspective fuyante du lexique[4].

Mais un vocabulaire est en principe constitué de vocables, c'est-à-dire de regroupements, les formes fléchies s'effaçant derrière leur chef de file qui est traditionnellement l'infinitif des verbes et le masculin singulier de la classe nominale. Cela suppose une coûteuse opération de lemmatisation qui n'a été entreprise, avec des raccourcis approximatifs, que pour les textes les plus récents. S'engager dans cette voie - comme nous l'avons fait

---

[4] L'étude que nous abordons s'écarte donc dans son principe de celle qui fut conduite, pour la même période, par J.Dubois, L.Guilbert, H.Mitterand et J. Pignon: *Le mouvement général du vocabulaire français de 1905 à 1960, d'après un dictionnaire d'usage*. Quand on étudie le détail des additions et des suppressions relevées dans deux éditions du *Petit Larousse*, on court le risque de poursuivre un objet de seconde main, soumis à des facteurs extralinguistiques, considérations commerciales ou préférences personnelles, qui expriment moins directement les mouvements de la langue que la conscience, nécessairement discutable, que les rédacteurs du *Larousse* ont de ces mouvements.

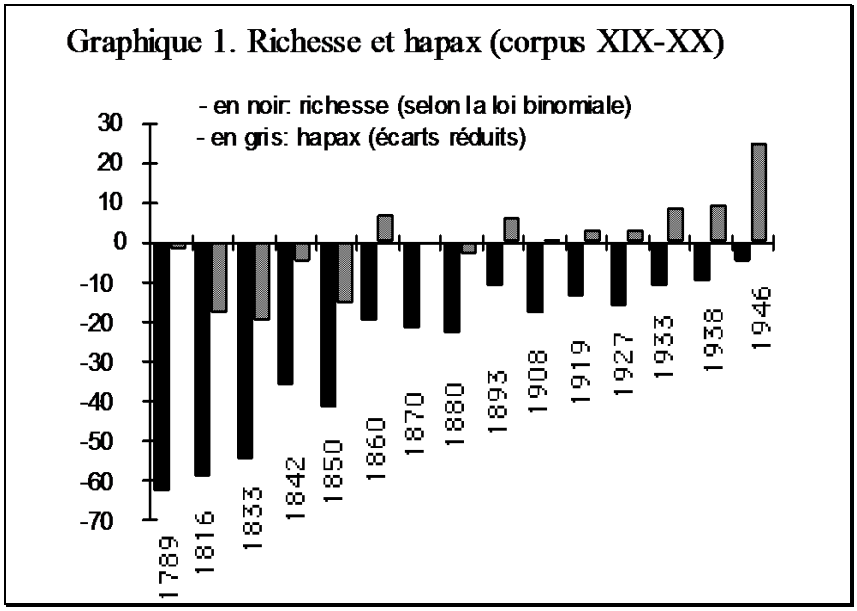
dans notre *Vocabulaire français de 1789 à nos jours* - eût été héroïque, vu l'énormité du corpus, dont la moitié restait à traiter, d'autant que la lemmatisation est plus complexe quand l'orthographe n'est pas fixée - comme c'est le cas des textes les plus anciens. On a donc cru devoir renoncer - au moins provisoirement - au traitement des vocables, en se contentant des formes dans leur plus simple appareil. Au reste il n'est pas sans intérêt de faire converger les approches. Quel résultat plus probant que celui qu'on obtient par des voies différentes? Voyons donc si l'examen des formes confirme celui des vocables, entrepris naguère.

1 - Il convient tout d'abord de rappeler les résultats antérieurs, obtenus à partir du corpus XIX^{e}-XX^{e}. Ce corpus a été divisé en 15 tranches chronologiques, dont les dernières correspondent à la période qui nous intéresse (tranches commençant respectivement en 1919, 1927, 1933 et 1938). Les données lemmatisées se présentent comme suit (après exclusion des signes de ponctuation, des chiffres, des noms propres et des mots étrangers):

tranche	1789	1816	1833	1842	1850	1860	1870	1880	1893	1908	1919	1927	1933	1938	1946	total
occurr.	5857336	5081449	5045419	4082572	4212666	4350647	4033535	4875409	5045345	4227531	4819111	4097582	4304089	4793038	5447822	70273551
vocabl.	24731	24213	24911	26453	25763	29939	28902	30402	32780	30067	31747	30009	31311	32464	34551	71640
richesse	*-63*	*-59*	*-55*	*-36*	*-42*	*-20*	*-22*	*-23*	*-11*	*-18*	*-14*	*-16*	*-11*	*-10*	*-5*	
hapax	1702	852	782	1072	735	1550	1207	1346	1748	1296	1566	1337	1592	1788	2620	21193
z	*-1,6*	*-18*	*-19,7*	*-4,7*	*-15,5*	*6,8*	*-0,3*	*-3,4*	*6*	*0,6*	*3,1*	*3*	*8,4*	*9,3*	*25.1*	

La représentation graphique des lignes 4 et 6 rend tout-à-fait clair le mouvement inflationniste qui se développe dans la monnaie lexicale, de 1789 à nos jours, qu'il s'agisse du vocabulaire pris dans son ensemble, ou de cette frange exclusive qu'on appelle hapax (ou mots employés une seule fois)[5].

[5] Précisons une fois pour toute que les calculs font appel systématiquement à l'écart réduit, qu'on obtient comme suit (en prenant pour exemple le cas des hapax):
Si *p* est le poids relatif de la dernière tranche dans le corpus: *p = 5447822 / 70273552 = 0,0775* (et *q* la probabilité complémentaire: *q = 1 - p = 0,9225*), le nombre théorique des hapax qu'on devrait rencontrer dans cette tranche est de: 21193 * p = 1642,95 (21193 est le nombre total d'hapax dans le corpus). Or on relève en réalité *2620* hapax dans la tranche considérée, soit un excédent de *2620-1642,95 = 977,05*. Il suffit de pondérer (de réduire) cet écart absolu par la racine carrée de l'effectif théorique pour obtenir la valeur de l'écart réduit, soit *977,05 / √1642,95 = 24,60*. En fait la formule plus exacte fait intervenir la probabilité *q* et aboutit à *z = 977,05 / √(1642,95 * q) = 25,1*.



2 - D'autres méthodes permettent de mesurer à travers le temps l'accroissement lexical et nous renvoyons le lecteur aux développements plus circonstanciés de notre ouvrage (pp. 51-80). Deux choses nous intéressent ici: tout d'abord la progression observée à partir de 1789 prend-elle son essor plus tôt dans l'histoire de la langue? et en second lieu l'afflux des naissances verbales qui touche les vocables touche-t-il aussi les formes? Le recours à FRANTEXT donne la réponse à ces deux questions, en même temps qu'à une troisième, qui est relative au genre littéraire et aux perturbations possibles que le dosage variable des genres peut créer dans la chronologie.

On a procédé par siècle et par genre, en ne détachant que les grandes masses. Cela suffit pour mettre en valeur la loi du genre: les essais techniques sont plus favorables à l'extension du lexique et à la multiplication des formes que les textes proprement littéraires - ce qui s'observe aussi bien dans le cas des hapax que dans l'ensemble des formes dont rend compte le calcul binomial[6]. Quant à la loi du temps, son action paraît contradictoire: pourquoi le XVII^{e} siècle échappe-t-il au mouvement et comment expliquer les privilèges qui sont les siens? L'explication, assez tri-

[6] La loi binomiale suppose la connaissance du tableau de distribution des fréquences, pour le corpus entier. En voici les premiers éléments:

classe	effectif	classe	effectif	classe	effectif	classe	effectif	classe	effectif
1	195548	10	5431	19	2272	28	1261	37	939
2	57811	11	4798	20	2091	29	1307	38	852
3	30448	12	4164	21	1993	30	1139	39	879
4	19962	13	3814	22	1846	31	1119	40	804
5	14328	14	3318	23	1742	32	1110	41	777
6	11223	15	1991	24	1602	33	1062	42	747
7	8736	16	2828	25	1516	34	1030	43	721
8	7318	17	2521	26	1498	35	902	44	694
9	6158	18	2399	27	1373	36	944	45	661

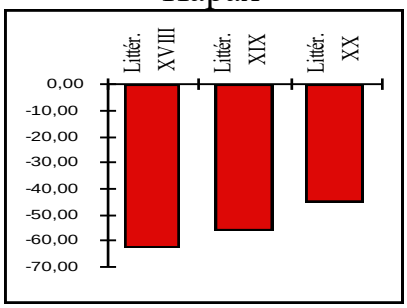
viale, tient à l'instabilité de l'orthographe à cette époque, la même forme canonique se présentant sous divers déguisements au hasard de l'accentuation. Comme les accents sont omis ou ajoutés alors avec une assez grande fantaisie, les mots qui comportent quelque diacritique apparaissent sous des variantes diverses, ce dont fait foi un seul exemple parmi des milliers d'autres: *eleve* 28 occurr. dans le corpus littéraire du XVII^{e}, *elevé* 29 occurr., *éleve* 70 occurr., à côté des formes régulières *élevé* 70 occurr., et *élève* 53 occurr. qui sont les seules connues dans le corpus XX^{e}-XX^{e} (respectivement 838 et 1445 occurr.). La comparaison n'est donc justifiée que lorsque l'orthographe tend à se stabiliser, au mieux à partir du XVIII^{e}. Quand cette condition est remplie, l'inflation lexicale se vérifie, comme le montrent les deux histogrammes du tableau 2.

Observons en passant que la notion de *créativité lexicale* utilisée par L. Guilbert ne traduit qu'un aspect du phénomène, celui des naissances dans la population des mots. Pour apprécier le taux de renouvellement, il faut tenir compte aussi bien des pertes. Or trois siècles d'observation font renoncer à la symétrie attendue entre les gains et les pertes, entre les rejetons et les rejets. Beaucoup de mots naissent, beaucoup vieillissent aussi, mais peu meurent. Est-on d'ailleurs certain de leur mort définitive? Un mot abandonné, qui cesse d'être employé dans le discours contemporain, ne cesse pourtant pas d'être compris si le lecteur le rencontre dans un texte plus ancien. Il arrive aux mots vieillis ce qui arrive aux objets vieillis ou usagés: on les met à la cave ou au grenier aussi souvent qu'à la poubelle. Délaissés, presque oubliés, on les retrouve pourtant avec surprise, parfois avec plaisir. L'usage littéraire du langage conduit ainsi à l'accumulation, voire à l'encombrement lexical. Ce jeu est d'autant plus complexe qu'il met en cause les écrivains autant que leurs lecteurs. Non seulement un lecteur cultivé doit être capable de secouer la poussière qui recouvre les mots chaque fois qu'il aborde un texte des siècles passés, mais encore certains auteurs archaïsants ne dédaignent pas de puiser eux-mêmes au magasin des accessoires et des antiquités et de recycler les mots anciens dans une prose moderne.

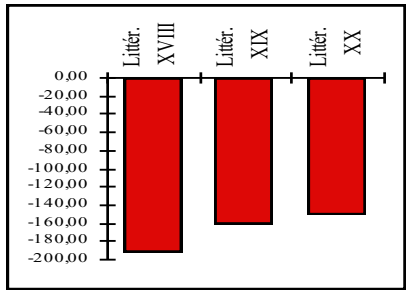
Tableau 2. Nombre de formes et d'hapax dans FRANTEXT

	Occurrences (sans chiffres ni ponctuations)	Hapax (réel)	Hapax (théo)	Écart (réduit) (hapax)	Formes (réel)	Formes (théo)	Écart (réduit) (formes)
Littérature XVII	11976466	32544	17076	123,93	136995	175770	-92,49
Littérature XVIII	18982239	17417	27065	**-63,21**	123735	212892	**-193,23**
Littérature XIX	33683023	37176	48026	**-57,05**	184683	269554	**-163,47**
Littérature XX	27782939	31422	39614	**-46,12**	173296	249075	**-151,84**
Essais XVII-XVIII	14152432	22734	20179	19,00	132117	188466	-129,80
Essais XIX-XX	29869351	53255	42588	58,48	214556	256592	-82,99
Total	136446450	194548	194548		477227	477227	

Hapax



Richesse lexicale



3 - Reste à mesurer non plus le taux global de renouvellement, mais la carte démographique des mots, en distinguant les catégories. Nous n'avons pas les moyens de mesurer les mouvements de population aux frontières, c'est-à-dire la part de l'immigration, de l'emprunt extérieur et par exemple des anglicismes. Seuls sont vérifiables les recensements internes, et notamment ceux qui portent sur la composition et la dérivation. Là aussi nous renvoyons le lecteur à notre étude antérieure qui porte sur 44 variétés de suffixes et 39 de préfixes[7]. La dérivation nous était apparue comme la ressource principale du renouvellement lexical, comme une planche à billets toujours disponible, apte à augmenter la masse monétaire avec une encaisse-or constante. On avait remarqué sans surprise que les besoins terminologiques sont beaucoup plus intenses dans les textes techniques et que certaines variétés suffixales leur sont propres (par exemple les suffixes *-ose, -ite, -ène, -ine, -ol, -one, -ate, -on* qui fleurissent en chimie ou en biologie). Mais plus intéressant nous avait paru le phénomène de mode qui à certains moments de l'histoire donne la faveur ou la retire à certaines formes de dérivation. Encore faut-il distinguer entre la créativité d'un suffixe, qu'on mesure en nombre de vocables mis en circulation, et le destin indépendant des mots suffixés, dont le cours peut se maintenir ou se développer au niveau des occurrences, même si le moule qui les a produits a cessé de fonctionner - et c'est ce qui arrive au suffixe *-tude* (corrélation chronologique[8] positive pour les occurrences: 0,46 et négative pour les vocables: -0,56). L'évolution est également indépendante dans le cas des suffixes *-ie, -ise, -eur,* ou *-té* dont les occurrences se font plus rares dans le

---

[7] Voir notre *Vocabulaire français ..,* p.415-685

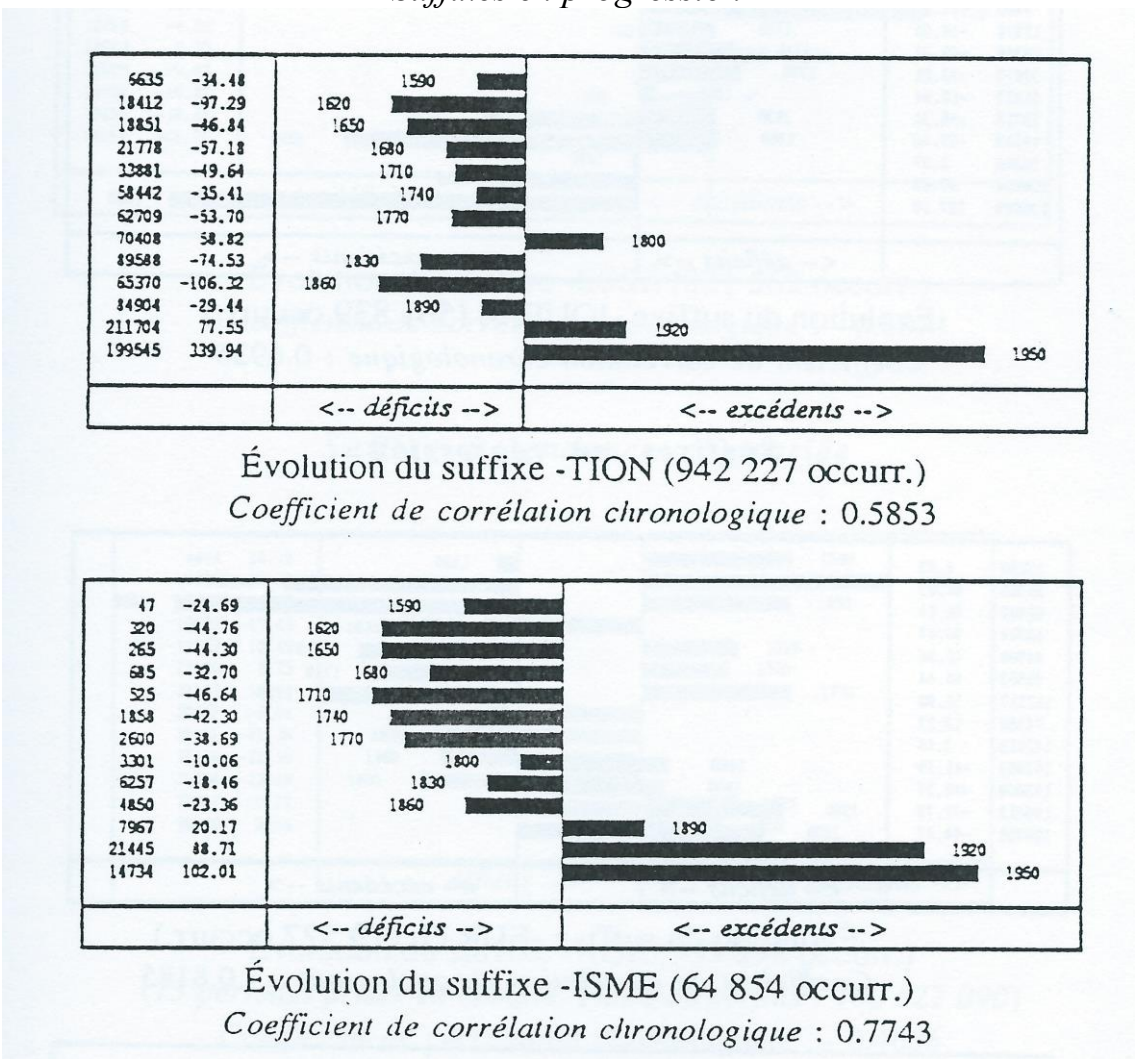
[8] Rappelons à l'usage des lecteurs peu familiers des techniques statistiques que le coefficient des rangs, ou coefficient de Spearman, s'obtient en comparant deux classements, selon la formule: $r = \frac{6 \sum d^2}{n(n^2-1)}$

(*d* étant l'écart entre les deux classements d'un même élément et *n* le nombre d'éléments classés). L'une des deux séries reproduit l'ordre chronologique, l'autre l'abondance relative de la classe ou catégorie étudiée, comme ici tel ou tel suffixe. Un résultat voisin est obtenu avec le coefficient de Bravais-Pearson qui est établi sur les valeurs et non sur les rangs.

discours contemporain mais dont la fécondité en créations nouvelles n'est pas encore tout-à-fait tarie. Le cas le plus général est pourtant celui d'une progression parallèle des vocables et des occurrences (on l'observe pour les suffixes *-tion, -ment*, *-age*, *-isme* , *-ence* ou pour les adjectifs correspondants: *-iste, -ique, -el, -al* et *-if*), ou d'une régression parallèle (*-at, -eux, -esse).* Il arrive aussi que de 1789 à nos jours la chronologie ait des mouvements contrariés et que le coefficient de corrélation - qui est linéaire - soit inopérant. Tantôt il s'agit d'une courbe en cloche (progrès au XIXe, régression au XXe): beaucoup de suffixes descriptifs, pittoresques ou concrets suivent ce profil qui convient au réalisme et au naturalisme*: -ise, -ure, ée, ien* (substantif), *-ier* (substantif), *-oir* (subst.), *-ant, -ante, -ette, -ard, -esque, -ois.* Tantôt la courbe est celle d'une cuvette, dont les extrémités relevées correspondent aux deux périodes extrêmes: celle de la Révolution et la tranche contemporaine, et c'est là qu'on rencontre les variétés les plus abstraites.

Graphique 3. Évolution de quelques suffixes

Suffixes en progression

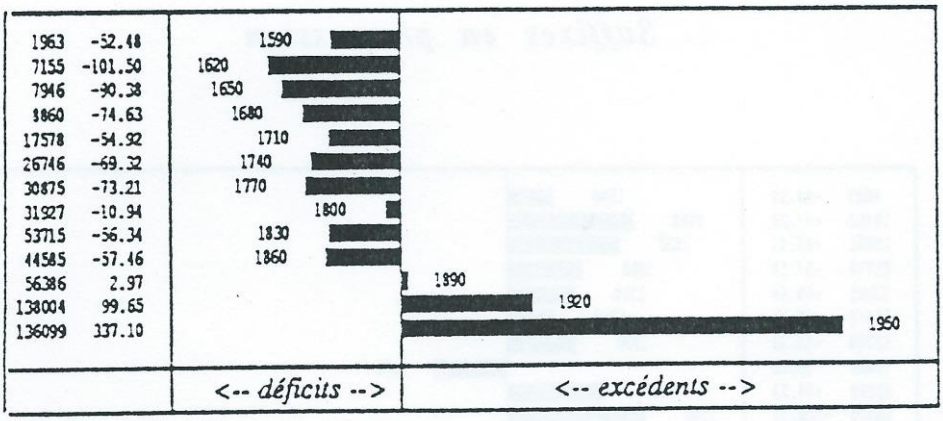


Évolution du suffixe -TION (942 227 occurr.)

Coefficient de corrélation chronologique : 0.5853

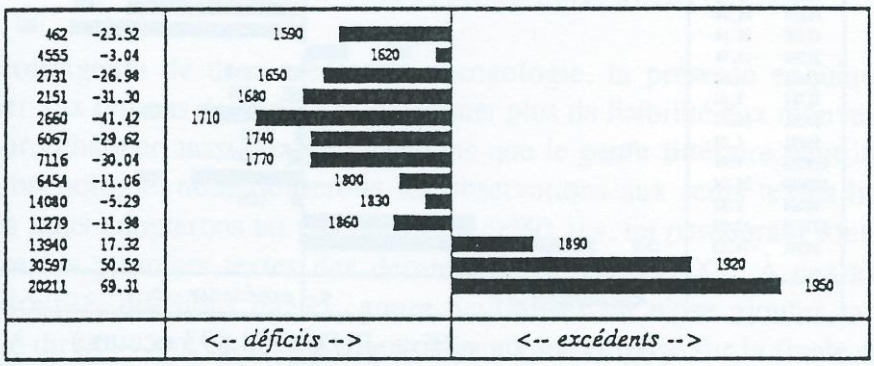
Évolution du suffixe -ISME (64 854 occurr.)

Coefficient de corrélation chronologique : 0.7743



Évolution du suffixe -IQUE(S) (561 839 occurr.)

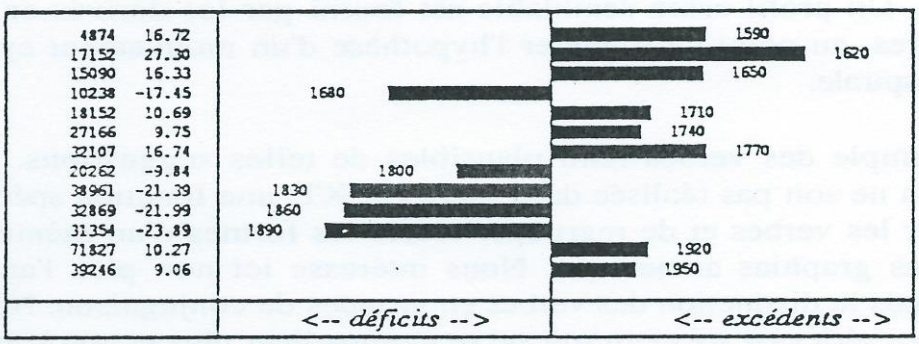
Coefficient de corrélation chronologique : 0.6935



Évolution du suffixe -ISTE et -ISTES (122 303 occurr.)

Coefficient de corrélation chronologique : 0.7408

Suffixe à tendance contrariée



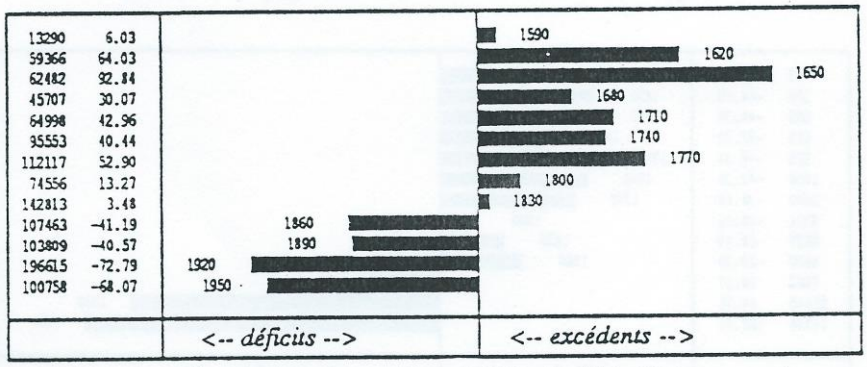
Évolution du suffixe -AGE (359 298 occurr.)

(13 périodes prises en compte. Taille du corpus : 160 822 090)

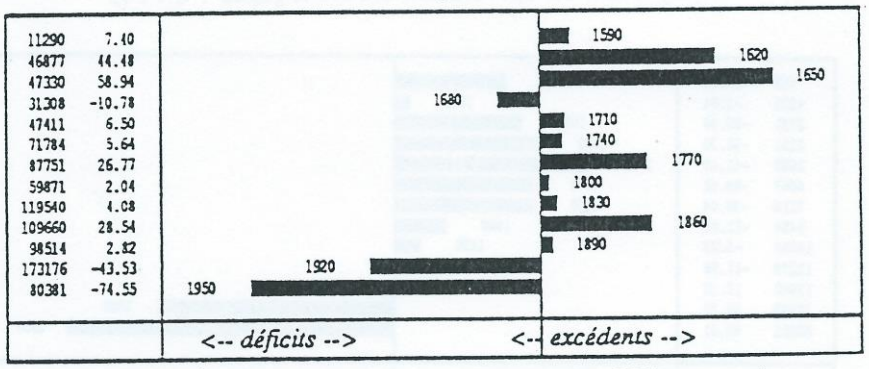
Coefficient de corrélation chronologique : - 0.4656

(Seuil à 5 % : 0.5529 pour 13 paires d'observations)

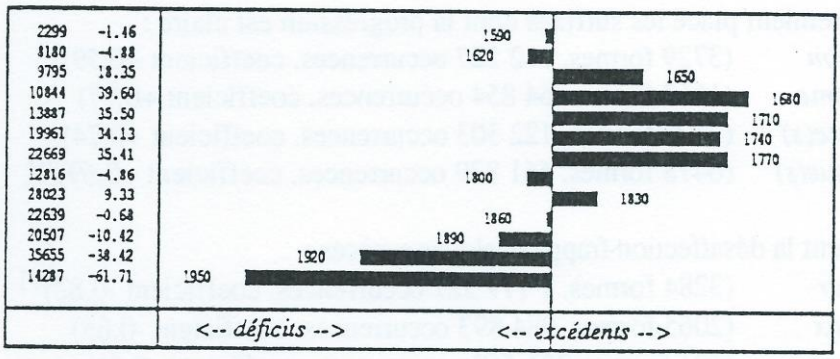
Suffixes en régression



Évolution du suffixe -EUR (1 179 527 occurr.)
Coefficient de corrélation chronologique : - 0.8185



Évolution du suffixe -EUX (984 893 occurr.)
Coefficient de corrélation chronologique : - 0.6455



Évolution du suffixe -ESSE (221 562 occurr.)
Coefficient de corrélation chronologique : - 0.6074

En prolongeant de deux siècles la chronologie, la présente enquête permet d'échapper aux remous de surface et de donner plus de lisibilité aux mouvements de fond. Pour échapper aussi aux perturbations que le genre littéraire peut introduire dans la chronologie, nous limiterons les observations aux seuls textes littéraires. Cette fois nous adopterons un pas

uniforme de 30 ans, en parcourant tout l'espace qui sépare les premiers textes des derniers dans FRANTEXT. À ces avantages (empan doublé, divisions égales, genre neutralisé) on n'ose ajouter la parfaite objectivité du tri - qui est ici purement mécanique et repose sur la finale des mots. Car beaucoup de suffixes traités ainsi sont si mêlés d'impuretés qu'ils en deviennent inexploitables[9]. Aussi bien nous contenterons-nous de quelques variétés, plus facilement isolables, représentées dans le graphique 3.

À gauche prennent place les suffixes dont la progression est claire:

-tion (3729 formes, 942 227 occurrences, coefficient +0,59)
-isme (1667 formes, 64 854 occurrences, coefficient +0,77)
-iste(s) (2402 formes, 122 303 occurrences, coefficient +0,74)
-ique(s) (6478 formes, 561 839 occurrences, coefficient +0,69)[10]

À droite la désaffection frappe quelques espèces:

-eur (3284 formes, 1 179 527 occurrences, coefficient -0,82)[11]
-eux (2065 formes, 984 893 occurrences, coefficient -0,65)
-esse (591 formes, 221 562 occurrences, coefficient -0,61)

Dans ces cas privilégiés, l'évolution est linéaire et confirme les résultats antérieurs, obtenus sur une période plus courte. Mais il est des faits plus complexes, dont nous donnons un exemple au bas du graphique 3. Il s'agit du suffixe *-age* qui est fort en faveur au XVII^e^ et XVIII^e^ siècles et dont la décrue est sensible au XIX^e^. Un retournement de tendance s'y manifeste pourtant dans la période 1914-1945 qui nous occupe. Un profil assez semblable est fourni par les dérivés en *-ance*[12] et quelques autres, au point de suggérer l'hypothèse d'un mouvement cyclique et le modèle de la spirale.

---

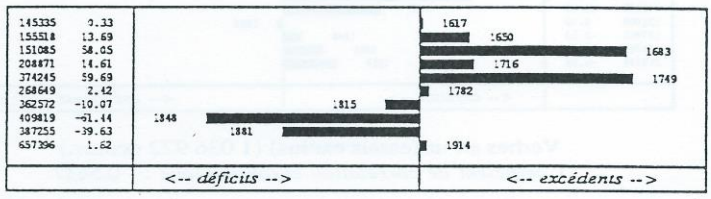
[9] Imaginons qu'on veuille isoler le suffixe *-ier*: comment opérer la décantation entre les substantifs (*plombier*), les adjectifs (*premier, fier)*, les verbes (*lier, fier*), les adverbes (*hier*)?

[10] Il y a cumul du singulier et du pluriel pour les formes en *-ique* et en *-iste*. Le pluriel n'est pas pris en compte dans les autres suffixes, d'autant qu'il est plus rare lorsqu'il s'agit de termes abstraits et que certaines confusions y devenaient indésirables (notamment avec les formes verbales en *-tions*).

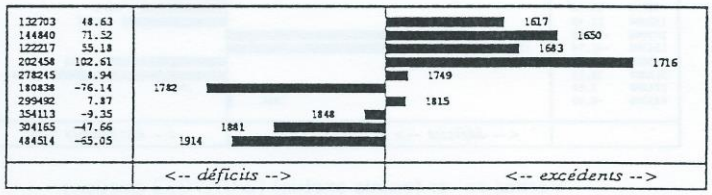
[11] Le suffixe *-eur* (il en est ainsi également de *-esse*) recouvre deux catégories distinctes, soit des notions abstraites (*fureur, tendresse)*, soit des agents, animés ou non (*censeur, ascenseur, maîtresse, tigresse*). C'est la première espèce qui domine lorsqu'on considère - comme ici - les occurrences. La seconde espèce - dont la créativité n'est pas émoussée - a un profil assez différent si l'on en juge par l'exemple du suffixe *-teur* qui ne comprend que des agents ou des outils et qui est en nette progression (1162 formes, 138028 occurrences, coefficient +0,71).

[12] La concurrence entre *-ance* et *-ence* qui a longtemps été profitable au premier tourne de plus en plus à l'avantage du second.

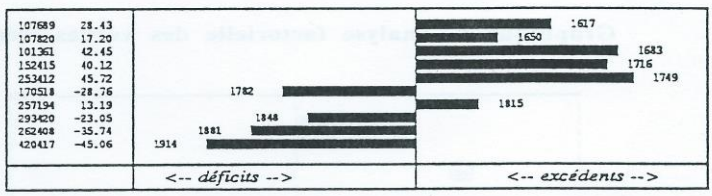
Graphique 4. Distribution des verbes



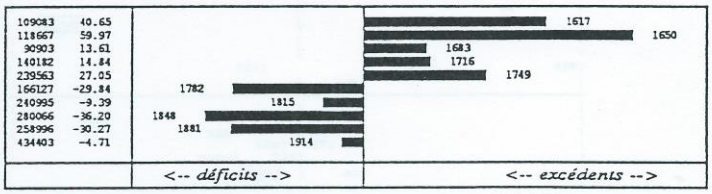
Être (essais exclus) (3 120 745 occurr.)
Coefficient de corrélation chronologique : - 0.5340



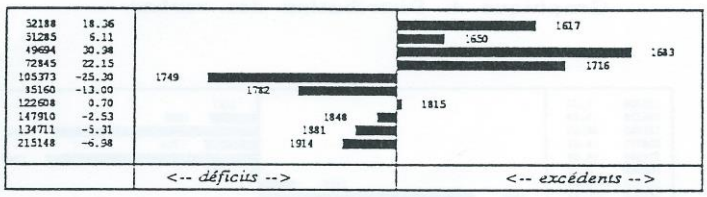
Verbes en -re (essais exclus) (2 503 585 occurr.)
Coefficient de corrélation chronologique : - 0.7766



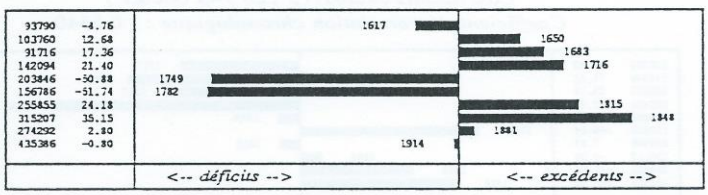
Avoir (essais exclus) (2 126 654 occurr.)
Coefficient de corrélation chronologique : - 0.7897



Verbes en -oir (essais exclus) (2 078 985 occurr.)
Coefficient de corrélation chronologique : - 0.8052



Verbes en -ir (essais exclus) (1 036 922 occurr.)
Coefficient de corrélation chronologique : - 0.5637



Verbes en -er (essais exclus) (2 072 732 occurr.)
Coefficient de corrélation chronologique : - 0.0605

4 - L'exemple des verbes (graphique 4) rend plausibles de telles suggestions. Quoique la lemmatisation ne soit pas réalisée dans FRANTEXT, une fonction spéciale permet d'y conjuguer les verbes et de regrouper toutes les formes d'un même paradigme (y compris les graphies anciennes). Nous intéresse ici non plus l'analyse de la dérivation, mais la distinction des verbes en groupes de conjugaison. Nul besoin de vérifier que les séries en *-re* et en *-oir* ne se renouvellent plus et que la naissance de nouveaux verbes n'est possible que dans les paradigmes en *-ir* et en *-er*. Mais la question se pose de savoir si l'érosion n'attaque pas les verbes sans descendance comme elle use les volcans sans activité.

Quoique l'enquête ne porte pas ici sur la totalité des verbes, mais sur les plus fréquents d'entre eux, elle enveloppe plus de la moitié des occurrences de la catégorie (soit 13 millions dans le seul corpus littéraire, de 1617 à 1945)[13]. Les résultats apparaissent dans le graphique 4 [14]. L'érosion s'y manifeste avec évidence dans les reliefs anciens: verbes en *-re* (r= -0,78), en *-oir* (r= -0,81) et en *-ir* (r= -0,56), auxiliaires *avoir* (r= -0,79) et *être* (r= -0,53). Seul le massif plus jeune des verbes en *-er* résiste à l'affaissement, et montre un regain de vitalité au XIXe siècle (r = +0,06). La conclusion est que sur quatre siècles la catégorie verbale est moins

---

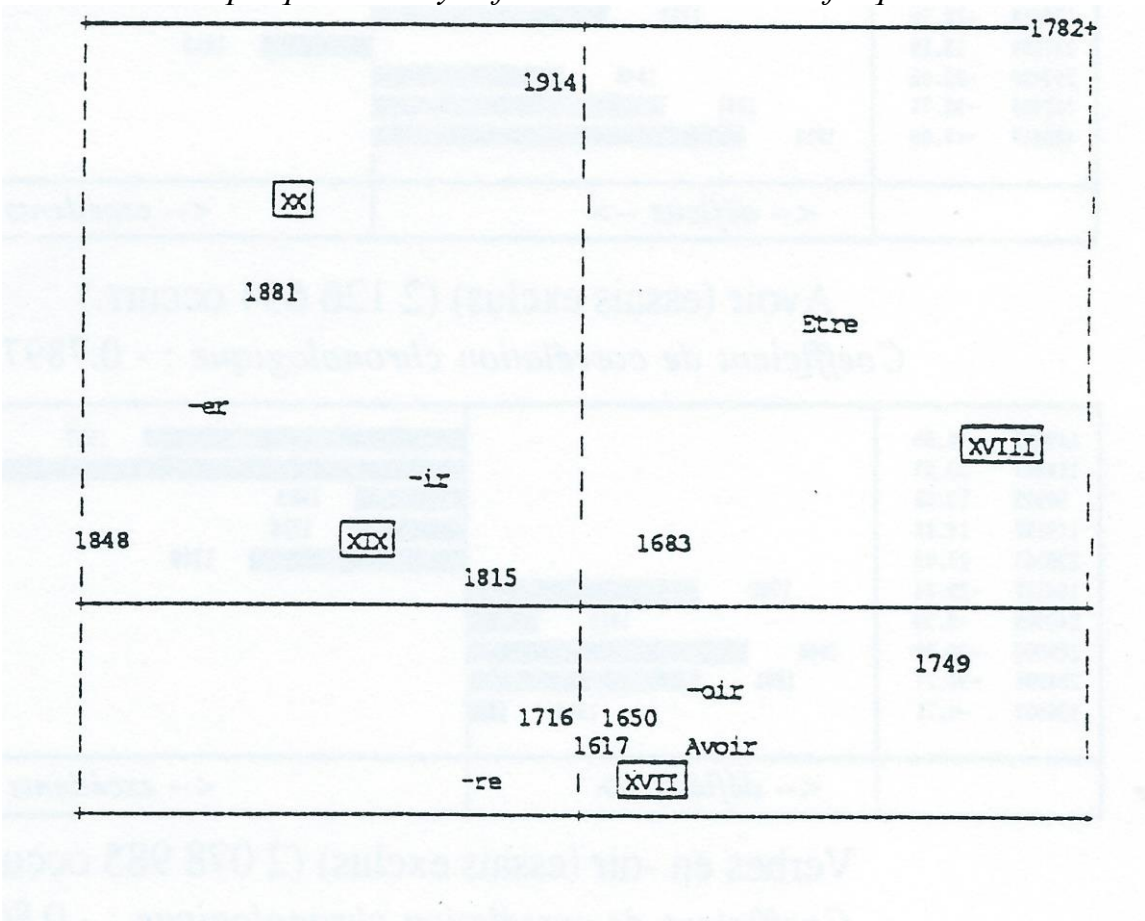
[13] Voici les verbes retenus:

- groupe en *-oir: asseoir, seoir, choir, valoir, falloir, vouloir, voir, savoir, recevoir, concevoir, percevoir, apercevoir, devoir, revoir, entrevoir, pourvoir, pleuvoir, mouvoir, pouvoir*, soit 2313 formes et 2 millions d'occurrences
- groupe en *-re*: *faire, dire, prendre, croire, mettre, rendre, vivre, écrire, entendre, reprendre, paraître, comprendre, sourire, perdre, connaître, lire, attendre, suffire,* soit 1883 formes et 2,5 millions d'occurrences
- groupe en *-ir: venir, sortir, mourir, souvenir, finir, partir, tenir, agir, sentir, devenir, servir, revenir,* soit 843 formes et 1 million d'occurrences (la distinction traditionnelle en deux groupes est ici négligée).
- groupe en *-er: parler, donner, aller, sembler, trouver, passer, demander, chercher, penser, laisser, arriver, regarder, appeler, marcher, aimer, importer, avancer, rester, quitter, porter, tomber, manger, exister, commencer, tirer,* soit 2179 formes et 2 millions d'occurrences.

[14] Les limites du corpus pour cette étude et les suivantes ont été réajustées afin que la dernière tranche corresponde exactement à la période 1914-1945 qui nous intéresse. Le pas de progression est ainsi de 33 ans, la date de départ 1617 et le point d'arrivée 1946. On a eu le souci, en ignorant les textes postérieurs à la Libération, de laisser vierge le terrain à explorer en vue du prochain volume de l'Histoire de la langue, d'autant que pour cette période de nombreux textes sont ajoutés chaque année dans le corpus de FRANTEXT et que la stabilisation n'y est pas assurée.

sollicitée[15]. Cela ne s'accorde guère avec les remarques que nous avions formulées jadis à propos des catégories grammaticales. Il nous avait semblé alors que les verbes regagnaient un peu de terrain aux dépens de la catégorie nominale.

Graphique 5 Analyse factorielle des verbes fréquents



C'est effectivement ce qui se passe si l'on prend pour point de départ l'époque de la Révolution, qui de toute la série est la plus défavorable au verbe (tous les histogrammes montrent un effondrement à cet endroit). S'agit-il là d'un léger soubresaut qui ne met pas en cause le déclin général de la catégorie ou est-ce l'amorce d'un sursaut et d'un renversement de tendance? L'analyse factorielle de la figure 5, réalisée à partir des données de la figure 4, laisse l'esprit perplexe. Sans doute y voit-on les groupes verbaux défiler dans l'ordre attendu (*-oir* et *-re*, puis *-ir* et *-er*) et l'axe du temps parcourir l'espace du graphique de gauche à droite et de bas en haut. Mais au lieu du dessin en croissant qu'on rencontre habituellement dans les

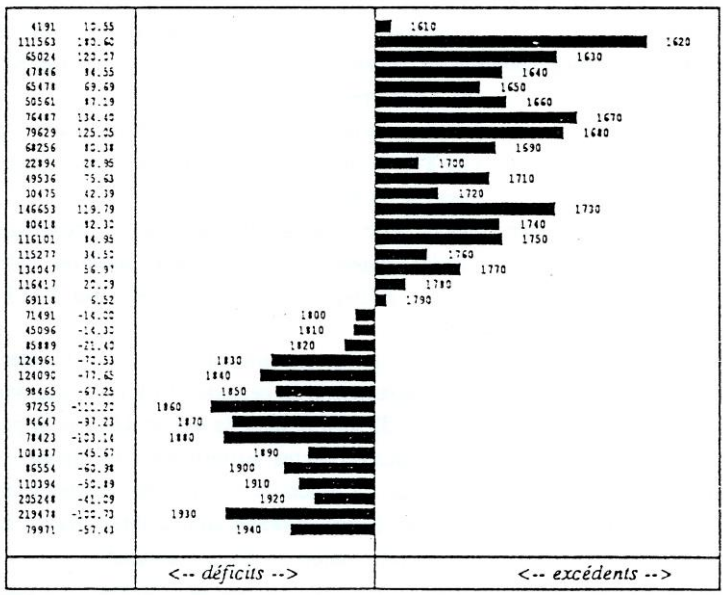
[15] Comme les classes nominale et verbale ont coutume de s'opposer, on peut en déduire que les substantifs ont tendance à envahir le discours. Cette tendance a été controuvée à court terme, au moins dans les textes proprement littéraires. Mais sur une longue distance de quatre siècles on devrait en retrouver la trace, si la catégorisation complète de FRANTEXT pouvait être entreprise.

données sérielles, c'est une ligne sinusoïdale que propose l'analyse, comme si le XX^e^ siècle était au XVIII^e^ ce que le XIX^e^ est au XVII^e^.

- II - Les mots grammaticaux

On a parfois pensé que les mots grammaticaux constituent un milieu homogène, une sorte de pâte indifférenciée, ou de ciment grisâtre destiné à accointer les mots pleins - qui seuls supporteraient les charges positives ou négatives du discours. Il n'en est rien: ces particules, si menues soient-elles, sont aimantées et s'orientent différemment selon le registre, ou l'époque. La lemmatisation n'est plus ici un préalable, car on l'applique rarement à ces unités élémentaires dont la masse est trop écrasante pour être l'objet d'un tri manuel. Certes l'ambiguïté demeure pour certains de ces mots-outils qui servent à plusieurs usages et qui ont parfois l'ambivalence d'un couteau suisse. Que faire, par exemple, de *le*, de *si*, de *quand*, de *comme* ou de *que*? Pourtant cette confusion des fonctions cache souvent une origine simple et unifiée, et elle n'interdit pas un comportement cohérent dans les situations de discours. Quoi de plus ambigu que le mot *que*? A-t-on affaire au relatif, à la conjonction, à l'interrogatif, à l'adverbe?

Figure 6. Évolution de *que* et *qu'*



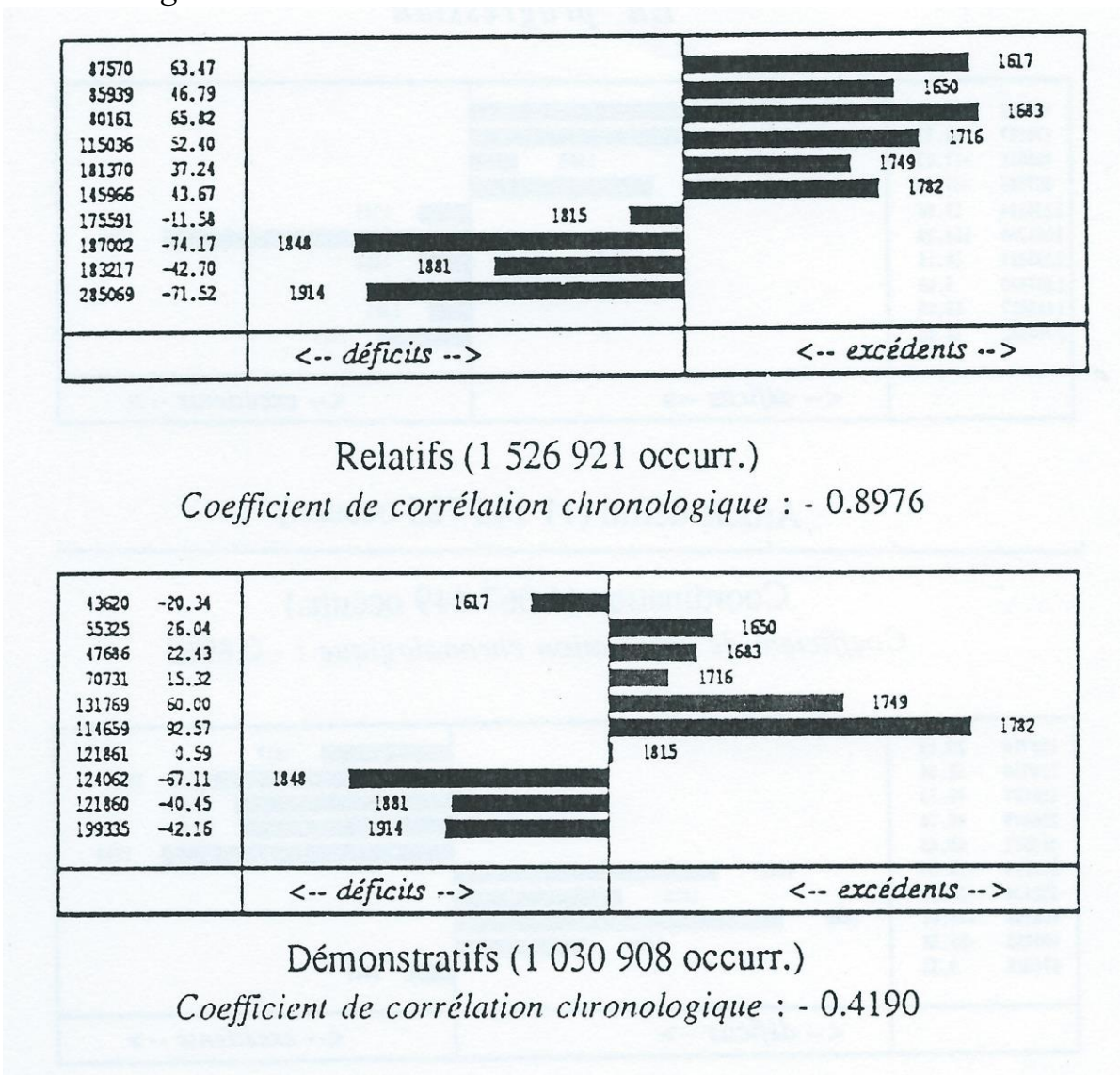
que et qu' (3 070 320 occurr.)
Coefficient de corrélation chronologique : - 0.8580

La courbe chronologique (graphique 6) de ce Janus à multiples visages a pourtant une pureté de profil assez singulière, tout-à-fait incompatible avec le jeu du hasard. Quand le déclin a une pente aussi nette

et aussi régulière, et qu'il porte sur trois millions d'observations, on ne peut manquer de s'interroger sur sa signification et sur les implications qu'une telle évolution entraîne dans les structures de la phrase. On sait la prédilection de Proust pour la phrase longue et les articulations complexes de la syntaxe et l'on a observé que - relatif ou subordonnant - la charnière *que* y joue un rôle primordial. C'est aussi ce qu'on constate à l'intérieur de FRANTEXT, quand on isole la période 1914-1945. Proust y vient en tête pour l'abondance des *que*, suivi de Gide, Du Bos, Valéry, Claudel et une pléiade de philosophes: G. Marcel, Alain, Gilson, Ruyer, Bergson. Or cette sensibilité stylistique n'est pas propre à *que*. On la retrouve dans la plupart des mots grammaticaux, même les plus effacés et les plus insignifiants, ce qui justifie une étude systématique.

Figure 7. Distribution des mots grammaticaux

En régression

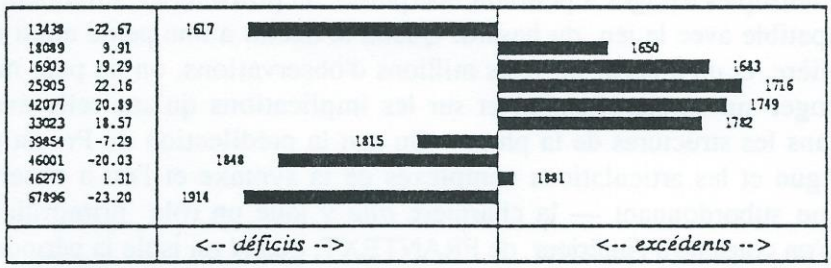


Relatifs (1 526 921 occurr.)

Coefficient de corrélation chronologique : - 0.8976

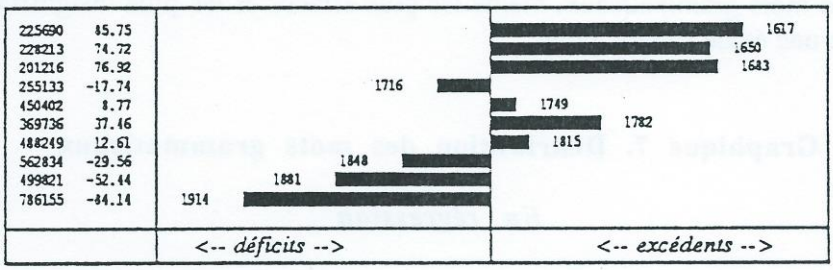
Démonstratifs (1 030 908 occurr.)

Coefficient de corrélation chronologique : - 0.4190



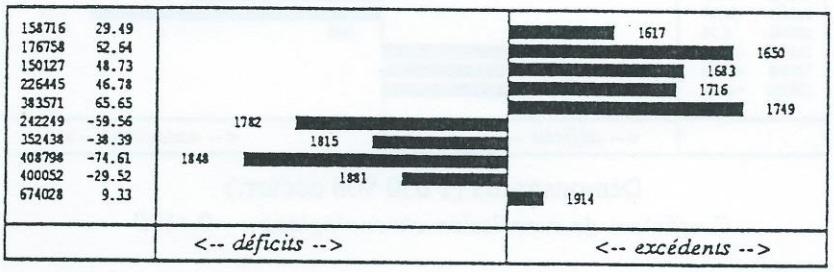
Subordination (349 465 occurr.)

Coefficient de corrélation chronologique : - 0.3419



Coordination (4 067 449 occurr.)

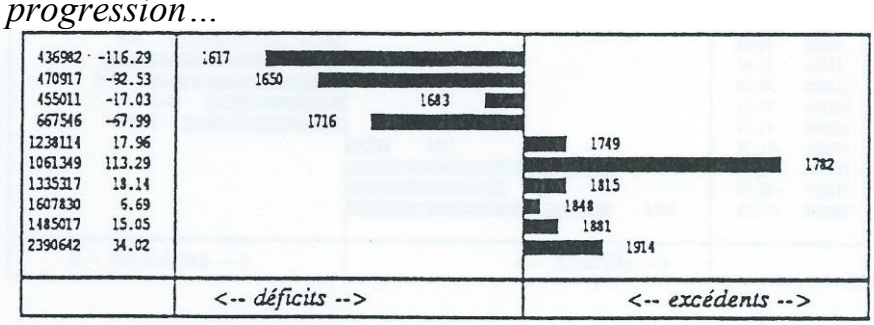
Coefficient de corrélation chronologique : - 0.8963



Négations (3 173 182 occurr.)

Coefficient de corrélation chronologique : - 0.6370

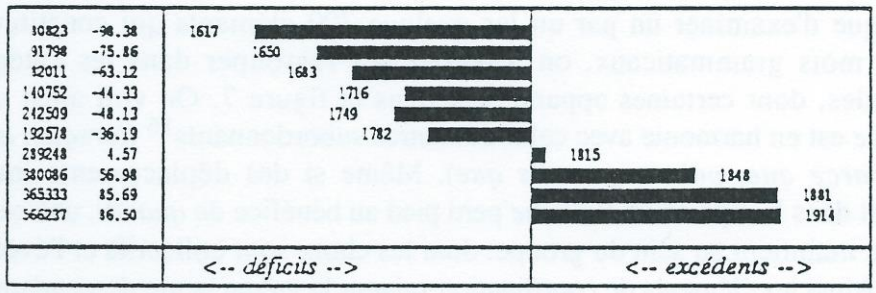
En progression…



Article défini (11 148 725 occurr.)

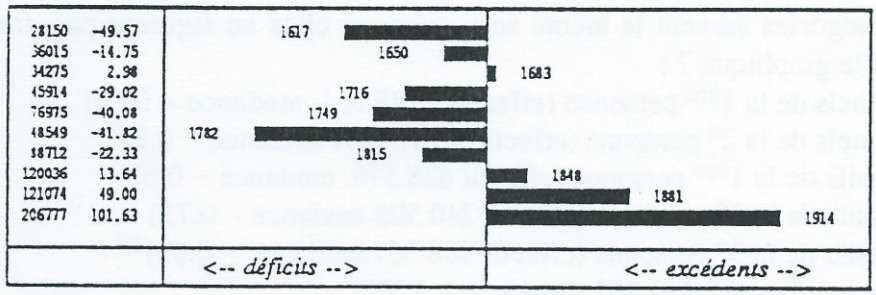
Coefficient de corrélation chronologique : 0.6998

En progression



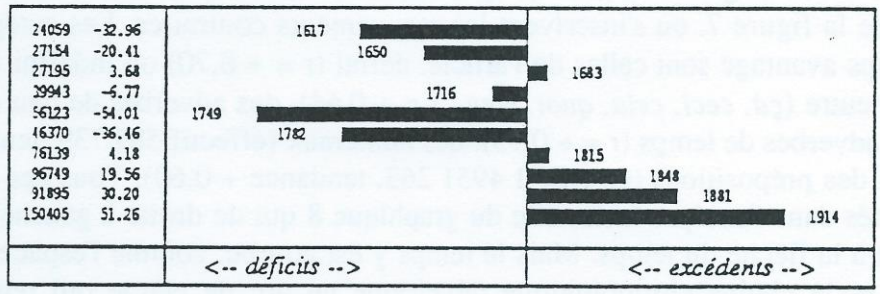
Article indéfini (2 431 375 occurr.)

Coefficient de corrélation chronologique : 0.9621



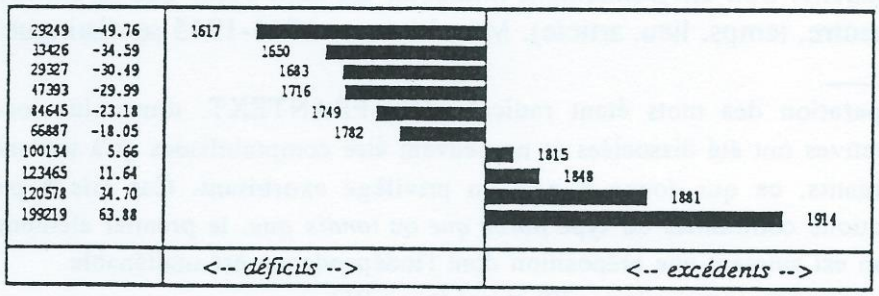
Neutre (806 477 occurr.)

Coefficient de corrélation chronologique : 0.6434



Lieu (636 032 occurr.)

Coefficient de corrélation chronologique : 0.6988



Temps (834 323 occurr.)

Coefficient de corrélation chronologique : 0.9469

Plutôt que d'examiner un par un les quelque 300 éléments qui constituent la classe des mots grammaticaux, on a préféré les regrouper dans les catégories traditionnelles, dont certaines figurent dans le tableau 7. On voit ainsi que le profil de *que* est en harmonie avec celui des autres subordonnants[16] (*lorsque, quand, quoique, parce que, puisque, tandis que*). Même si des déplacements internes apparaissent dans la répartition (*lorsque* perd pied au bénéfice de *quand)*, une certaine cohésion se maintient au sein du groupe, dont les choix sont collectifs et l'évolution cohérente. Ainsi les éléments de coordination suivent la même pente descendante que ceux de la subordination (la corrélation chronologique y est négative avec des coefficients de -0,90 et -0,34 respectivement). Même déclin des adverbes de négation[17] (r = -0,64), des relatifs (r = -0,90), et des démonstratifs[18] (r = -0,42). D'autres catégories suivent le même sort, même si elles ne figurent pas, faute de place, dans le graphique 7:

- les personnels de la 1ère personne (effectif 2 778 651, tendance -0,60)
- les personnels de la 2ème personne (effectif 1 147 447, tendance -0,86)
- les possessifs de la 1ère personne (effectif 828 579, tendance -0,62)
- les possessifs de la 2ème personne (effectif 240 505 tendance -0,72)
- les possessifs de la 3ème personne (effectif 908 361, tendance -0,63)[19]
- les indéfinis (effectif 2 114 301, tendance -0,32)
- les adverbes de quantité ou appréciation (effectif 2 511 756, tendance -0,47)

Avant de s'interroger sur le principe qui anime ces reflux, examinons les derniers éléments de la figure 7, où s'inscrivent les mouvements contraires. Les catégories que le temps avantage sont celles de l'article, défini (r=+0,70) ou indéfini (r=+0,96), du neutre (*çà, ceci, cela, quoi, rien*; r=+0,64), des adverbes de lieu (r=+0,70), des adverbes de temps (r=0,95), des numéraux (effectif 582739, tendance +0,32), et des prépositions (effectif 14951263, tendance +0,60). Tous ces choix sont résumés dans l'analyse factorielle de la figure 8 qui de droite à gauche obéit fidèlement à la flèche du temps. Mais le temps y est courbe, comme l'espace de la

---

[16] La séparation des mots étant radicale dans FRANTEXT, toutes les locutions conjonctives ont été dissociées et ne peuvent être comptabilisées qu'à travers leurs constituants, ce qui donne à *que* un privilège exorbitant. Car mis à part les associations contraintes du type *parce que* ou *tandis que*., le premier élément de la locution est souvent une préposition dont l'indépendance est inaliénable.

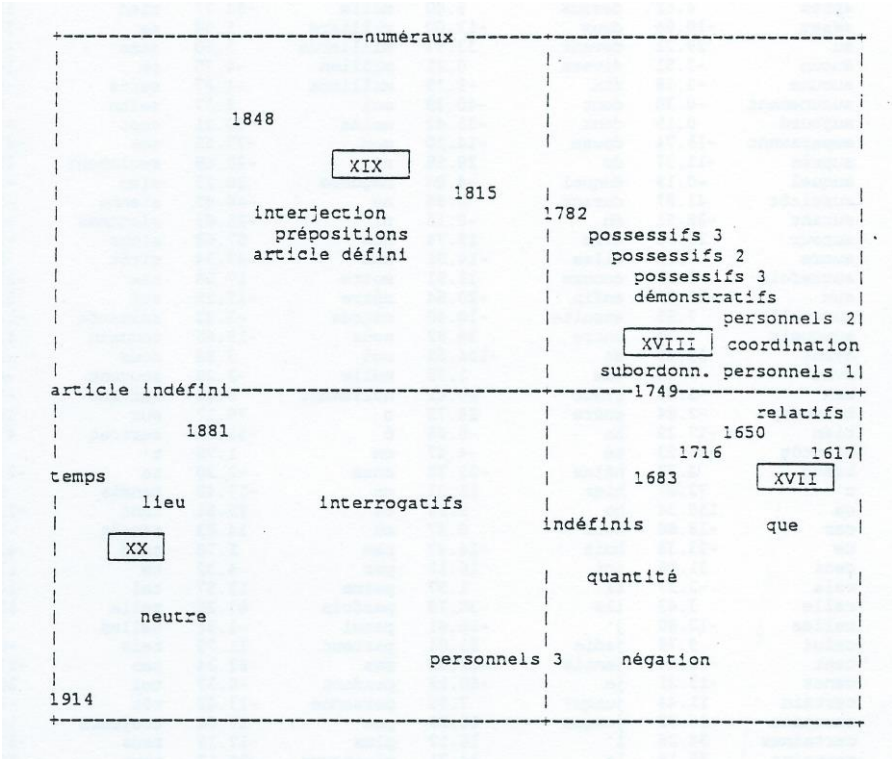
[17] Il y a opposition entre *ne* (z=-30) et *pas* (z=+91).

[18] Démonstratifs et relatifs sont d'ailleurs souvent associés dans les syntagmes *ce qui, celui qui, celle qui, ceux qui.*

[19] Les personnels de la 3ème personnels restent indécis (coefficient non significatif -0,10). Ils sont en effet trop mêlés à l'article (*le, la, les)* pour dégager leur personnalité propre.

relativité, et la chaîne chronologique développe un arc de cercle qui parcourt successivement les quatre quadrants du graphique, en réservant un siècle à chacun. L'époque qui nous occupe s'inscrit à la pointe ultime du parcours, dans le coin inférieur gauche. Son environnement immédiat est bien celui que la chronologie favorise (neutre, temps, lieu, article). Mais l'époque 1914-1945 se distingue aussi par des goûts particuliers qui prennent la tendance générale à rebrousse-poil et qui se déclarent au bas de la figure: personnels de la 3ème personne (z=+6,75), indéfinis (z=+13), adverbes appréciatifs (z=+25,08). Elle affirme aussi son indépendance à l'égard du XIXe siècle en repoussant les numéraux (z = -62,35) et les interjections (z=-19,02) et en montrant de la tiédeur à l'égard des prépositions.

Graphique 8. Analyse factorielle des classes de mots grammaticaux



Au total il semble bien que l'on ait face à face deux univers. À droite c'est l'ancien régime, où comptent surtout les personnes, les relations, la hiérarchisation dans la phrase comme dans la société. À gauche le cadre a éclaté, l'individu se disperse et la phrase se dilue dans les choses, dans le milieu, dans les circonstances, dans le temps, dans l'espace. Deux acteurs cachés tirent les ficelles en coulisse. À droite c'est le verbe qui pousse sur la scène les personnels, les négations, et toutes les articulations de la phrase: relatifs, subordonnants et coordinations. À gauche le substantif a partie liée avec les articles, les prépositions et les circonstances qui fixent le lieu et le temps. Si l'époque 1914-1945 a choisi son camp, des indices

laissent à penser que l'évolution n'est pas close et qu'un retour vers le verbe n'est pas à exclure, du moins dans les textes littéraires[20].

Tableau 9. Écarts réduits des mots grammaticaux dans la période 1914-1945

(Certains éléments se retrouvent dans le tableau 13, avec une valeur légèrement différente. Ce décalage tient au fait que la base FRANTEXT grossit d'année en année et que les calculs qui s'appuient sur cette norme évolutive sont soumis à d'imperceptibles variations s'ils sont exécutés à de longs intervalles).

à	13.63	des	2.32	mien	-15.50	quiconque	-9.05
afin	-29.38	dès	23.52	mienne	-12.13	quinze	-9.85
ah	-13.83	desquelles	-1.01	miennes	-8.51	quoi	20.00
ailleurs	42.01	desquels	-6.91	miens	-11.38	quoiqu'	-35.91
alors	42.92	dessous	4.10	mieux	-9.75	quoique	-42.01
après	4.62	dessus	6.80	mille	-54.77	rien	12.22
assez	-10.96	deux	-17.00	milliard	1.08	sa	-30.48
au	29.21	devant	33.97	milliards	3.60	sans	-8.34
aucun	-2.51	divers	0.21	million	-4.25	se	14.42
aucune	-2.18	dix	-5.39	millions	-4.87	seize	-6.10
aucunement	-0.38	donc	-40.38	moi	9.37	selon	6.15
aujourd·	0.19	dont	-33.42	moins	-10.41	sept	-7.74
auparavant	-15.74	douze	-14.20	mon	-77.63	ses	-47.05
auprès	-11.37	du	28.55	n'	-20.49	seulement	28.50
auquel	-0.19	duquel	-4.06	naguère	20.23	sien	-9.47
aussitôt	41.87	durant	0.86	ne	-48.65	sienne	-5.87
autant	-26.91	éh	-0.16	ni	-25.63	siennnes	-0.16
autour	29.96	elle	19.74	non	57.68	siens	-3.66
autre	7.19	elles	-14.01	nos	-47.34	sitôt	5.63
autrefois	-5.13	encore	11.91	notre	17.28	six	-23.45
aux	-22.93	enfin	-20.54	nôtre	-17.25	soi	53.28
auxquelles	7.55	ensuite	-10.60	nôtres	-3.42	soixante	-14.26
auxquels	1.25	entre	39.82	nous	-19.65	soudain	63.48
avant	32.85	et	-104.62	nul	3.98	sous	-0.82
avec	7.48	eux	1.75	nulle	-2.29	souvent	-4.21
bah	-2.27	grâce	20.41	nullement	8.48	suivant	-7.62
beaucoup	-2.84	guère	24.72	o	79.17	sur	18.01
bien	-17.29	ha	-6.65	ô	-52.35	surtout	43.65
bientôt	-20.23	hé	-4.47	oh	1.75	t'	7.33
bravo	4.78	hélas	-23.78	onze	-2.30	ta	-23.03
c'	72.67	hier	15.01	or	-17.49	tandis	5.67
ça	150.34	ho	3.26	ou	25.61	tant	-33.84
car	-18.80	hors	0.47	où	14.03	tantôt	-5.32
ce	-23.38	huit	-14.47	pan	5.70	tard	43.58
ceci	31.06	ici	16.17	par	4.32	te	13.39
cela	-3.27	il	1.37	parce	12.57	tel	14.92
celle	3.43	ils	-38.79	parfois	83.25	telle	12.77
celles	-12.80	j'	-16.61	parmi	-1.31	telles	7.22
celui	9.78	jadis	21.01	partout	11.72	tels	-0.37
cent	-30.94	jamais	-12.09	pas	83.34	tes	-33.98
cents	-13.21	je	-60.12	pendant	-0.37	toi	20.53
certain	11.44	jusqu'	7.59	personne	-13.42	tôt	-6.03
certaine	26.92	jusque	10.29	peu	17.04	toujours	14.64
certaines	34.26	l'	36.17	plus	-12.19	tous	-57.78
certains	55.16	la	34.71	plusieurs	-28.17	tout	20.36
ces	-36.58	là	37.58	plutôt	22.02	toute	3.26
cet	-25.92	laquelle	-2.94	pour	-43.54	toutes	-43.06
cette	-10.44	le	17.82	pourquoi	28.46	travers	33.92
ceux	-33.82	lequel	-4.54	près	14.42	treize	-1.33
chacun	0.34	les	-44.90	puis	27.92	trente	-8.83
chacune	5.89	lesquelles	-2.61	puisqu'	-7.27	très	38.34
chaque	27.84	lesquels	-5.51	puisque	-6.14	trois	-6.65
chez	14.68	leurs	-63.12	qu'	-76.93	trop	3.32
chut	-0.91	loin	11.38	quand	-0.93	tu	58.73
cinq	-9.37	longtemps	30.76	quant	-6.59	un	44.02
cinquante	-7.52	lorsqu'	-22.86	quarante	-7.33	une	81.45
combien	-19.48	lorsque	-19.96	quatorze	-2.25	uns	-19.89
comment	2.71	lui	4.93	quatre	-16.27	vers	49.86
contre	-1.08	ma	-68.27	que	-103.87	vingt	-11.46
dans	-10.40	maintenant	47.08	quel	-23.18	vingts	-7.45
davantage	-5.62	mais	36.22	quelle	-9.65	vos	-83.21
de	-31.08	mal	-5.45	quelles	-1.68	votre	-85.85
dedans	-10.01	malgré	1.25	quelque	-34.66	vôtre	-33.04
dehors	17.64	me	-58.11	quelquefois	-40.78	vôtres	-11.83
déjà	56.61	même	72.29	quelques	-17.94	vous	-201.07
demain	0.17	mêmes	-2.81	quels	-16.19	vraiment	38.52
depuis	2.03	mes	-64.61	qui	-64.23	TOTAL	-86.03

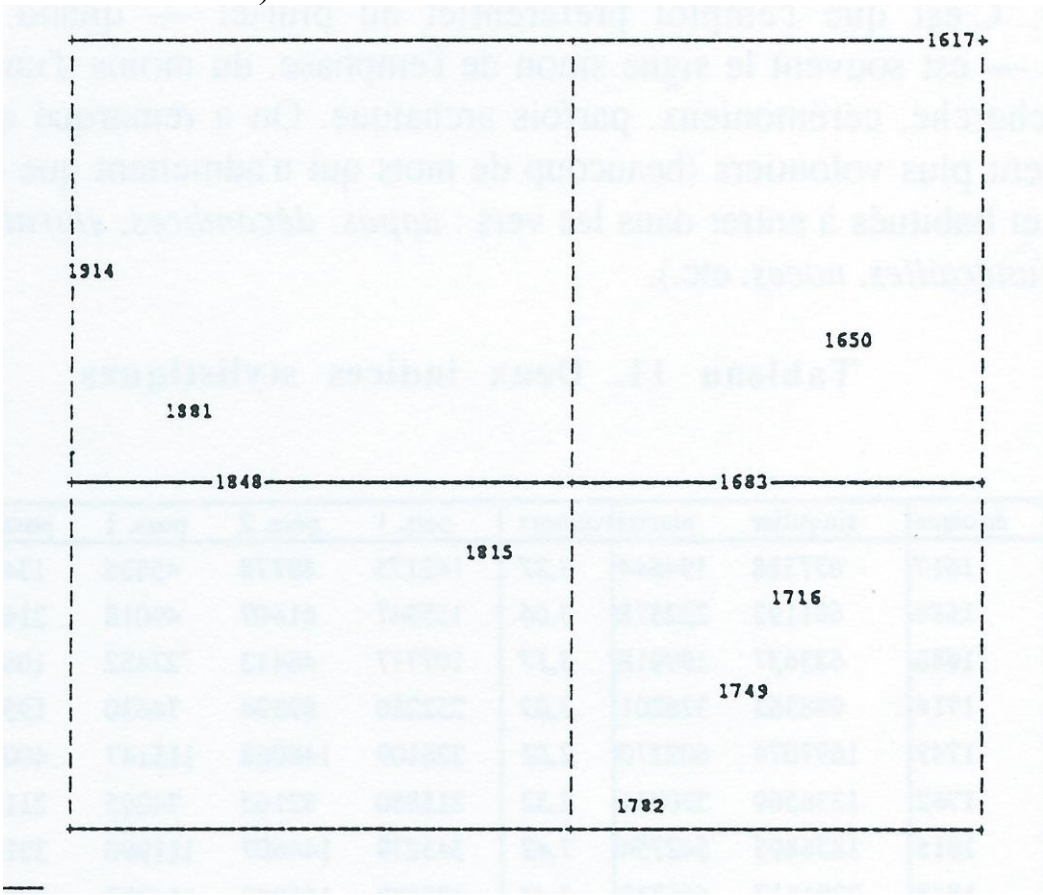
Si l'on craint quelque brutalité dans la catégorisation qu'on vient d'opérer, rien n'empêche de considérer isolément chacun des 267 mots grammaticaux étudiés. Tous se trouvent séparés et réunis dans la même

[20] Les textes techniques sont définitivement liés à la classe nominale, et les verbes n'y jouent plus guère que le rôle de copule.

liste (tableau 9) et la même analyse factorielle (figure 10). Il n'est guère utile de s'appesantir sur les détails de ces graphiques puisqu'ils confirment ceux qu'on vient de commenter[21].

On pourrait aussi soupçonner l'influence perturbatrice des genres littéraires, dont le dosage n'est pas constant au cours des siècles. Ainsi le théâtre - qui est favorable au verbe, comme la langue parlée - y occupe respectivement 14%, 9%, 5% et 8% des textes littéraires. Inversement le roman voit croître sa part avec le temps. Pour neutraliser ce facteur, nous avons renouvelé l'analyse, à genre constant, en isolant précisément le roman. Or les résultats restent stables[22].

Figure 10. Analyse factorielle de 267 mots grammaticaux (considérés individuellement)



[21] Le croissant caractéristique que décrit la chronologie dans la figure 10 est le même que celui de la figure 8, même si le haut et le bas ont été inversés, cette inversion n'ayant aucune signification.

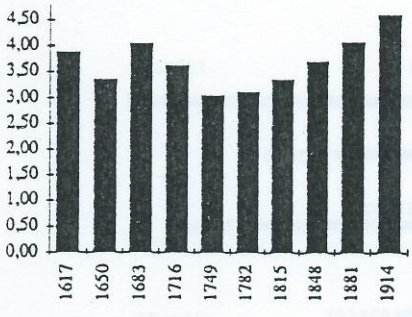
[22] Comme l'air du soupçon fouette la moindre certitude, on a essayé de mettre en cause l'inconstance de l'orthographe. L'analyse a été refaite en balayant tous les mots accentués. Les résultats ont été corroborés et les doutes balayés.

Avant de quitter les mots grammaticaux pour les mots sémantiques, et la structure du vocabulaire pour son contenu, deux indices peuvent être calculés à partir des présentes données. Le premier est dû à Charles Muller et porte le nom d'indice pronominal. Fondé sur le rapport entre personnels et possessifs des deux premières personnes, il mesure la familiarité du style. Rien ne justifie théoriquement son efficacité, mais, appliqué à un grand nombre de corpus, il n'a jamais été controuvé jusqu'ici. Et il passe brillamment l'épreuve du grand corpus de FRANTEXT. Observons en effet la dernière colonne du tableau 11 et le graphique associé: si l'indice tend à prouver la familiarité croissante des lettres française depuis 1750, il n'invite pas à conclure pour autant à une plus grande noblesse de style au siècle classique. C'est le pseudo-classicisme qui tend vers le ton soutenu et la haute distinction. Il y a plus de naturel au XVIIe siècle.

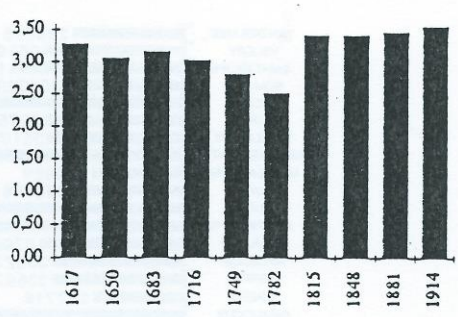
Tableau 11. Deux indices stylistiques

époque	singulier	pluriel	*rapport*	pers. 1	pers. 2	poss. 1	poss. 2	*indice*
1617	637388	194844	*3,27*	142173	88778	45835	13477	*3,89*
1650	681193	222878	*3,06*	155947	81407	49018	21658	*3,36*
1683	633637	199918	*3,17*	107717	46413	27452	10628	*4,05*
1716	988363	326201	*3,03*	252280	89594	74630	19540	*3,63*
1749	1697074	602270	*2,82*	326109	146068	115147	40070	*3,04*
1782	1336569	530835	*2,52*	215850	82165	74295	21154	*3,12*
1815	1856895	542754	*3,42*	343279	144607	111990	33177	*3,36*
1848	2281317	666732	*3,42*	388988	165942	114789	34865	*3,71*
1881	2132752	613953	*3,47*	312820	120784	85704	20623	*4,08*
1914	3402530	952156	*3,57*	533488	181689	129719	25313	*4,61*
total	15647718	4852541	*3,22*	2778651	1147447	828579	240505	*3,67*

Indice pronominal



Rapport singulier/pluriel



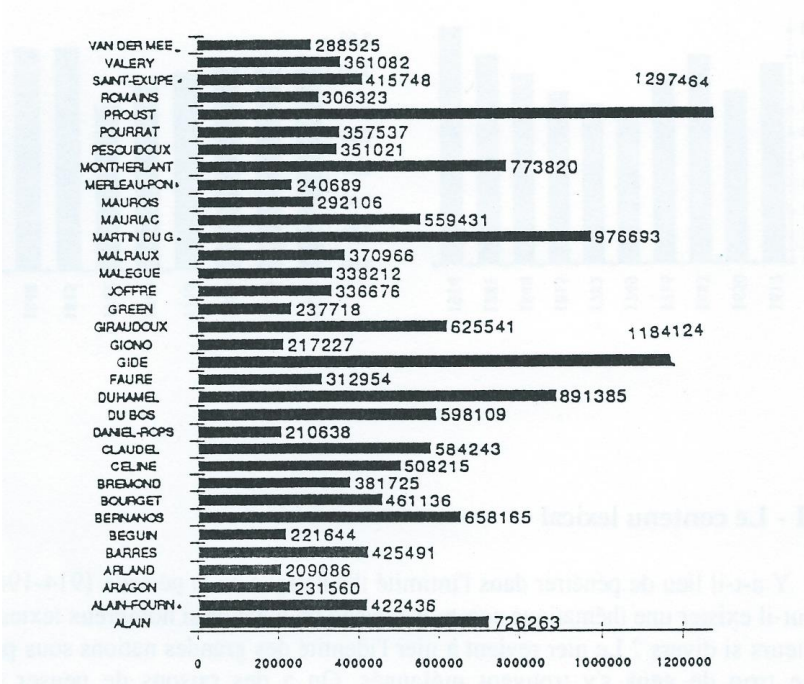
Le second rapport met en jeu le singulier et le pluriel, dont les marques sont assez faciles à distinguer dans un choix raisonné de mots grammaticaux. Là non plus l'évolution n'est pas linéaire. Si l'époque contemporaine tend vers le singulier (c'est dans la période 1914-1945 que le rapport est le plus haut), la période classique le privilégie aussi à l'origine,

avant que le pluriel ne s'impose progressivement jusqu'à la Révolution. On peut se demander si les deux indices ne sont pas liés. Effectivement le coefficient qui mesure leur corrélation est largement significatif: r = 0,75. C'est que l'emploi préférentiel du pluriel - quand le choix est possible - est souvent le signe sinon de l'emphase, du moins d'une tendance au style recherché, cérémonieux, parfois archaïque. On a remarqué que les poètes l'emploient plus volontiers (beaucoup des mots qui n'admettent que le pluriel sont anciens et habitués à entrer dans les vers: *appas, décombres, entrailles, ténèbres, vêpres, funérailles, noces*, etc.).

- III -
Le contenu lexical

Y a-t-il lieu de pénétrer dans l'intimité thématique de la période 1914-1945? Et peut-il exister une thématique commune quand on mêle de si nombreux textes et des auteurs si divers? Le nier revient à nier l'identité des grandes nations sous prétexte que trop de gens s'y trouvent mélangés. On a des raisons de penser que la communauté risque d'être moins sûre dans la famille que dans la nation, parce que trop dépendante de l'excentricité d'un seul membre. Les écarts individuels sont au contraire noyés dans les grands nombres, les signes contraires s'annulant quand les autres se renforcent.

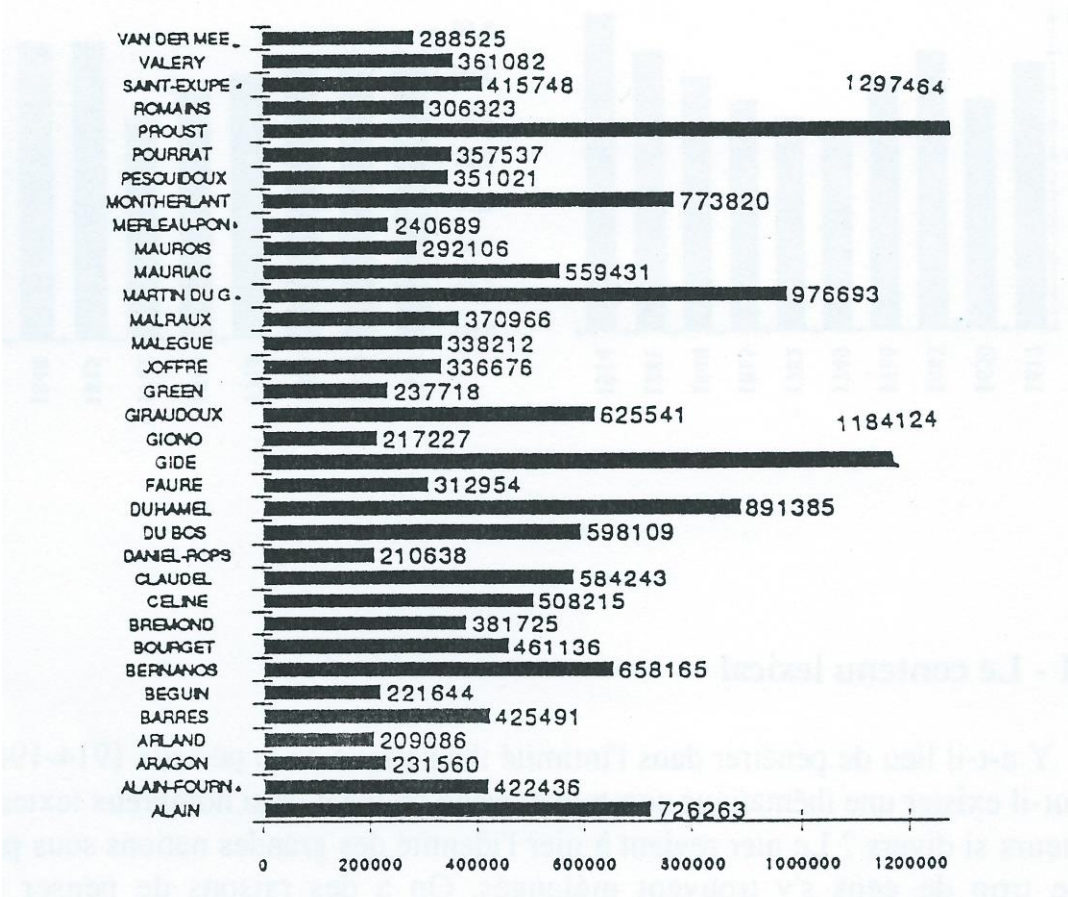
Figure 12. La période 1914-1945. Importance relative des écrivains (étendue > 200 000 occurr.)



Il importe pourtant, avant de s'engager dans cette étude, de circonscrire précisément la période qui nous occupe et de distinguer les individualités qui entrent dans la composition du corpus. La figure 12 reproduit graphiquement la part respective qu'y occupent les écrivains majeurs de cette époque, pour la plupart des romanciers. C'est Proust qui vient en tête avec l'intégralité de la *Recherche du temps perdu* (1297464 occurr.), suivi de Gide (1184124 occ.), Martin du Gard (976693 occ.), Duhamel, Montherlant, Alain, Bernanos, Giraudoux, Du Bos, Claudel, Mauriac, Céline, etc..

On reconnaît d'ailleurs sans peine la composition du sous-ensemble à la lecture des noms propres dans la liste des spécificités: *Albertine, Guermantes, Charlus, Balbec, Verdurin* dénoncent Proust, *Prouhèze* trahit Claudel, *Costals* Montherlant, et *Isabelle* tout à la fois Gide et Giraudoux. Ces noms propres ont cependant été expurgés du tableau 13, qu'ils encombraient sans grand profit: on les exclut généralement des calculs statistiques car on voit mal quelle conclusion générale peut s'attacher à leur valeur particulière et à leur création arbitraire.

Graphique 12. La période 1914-1945
Etendue relative des écrivains (seuil minimal : 200 000 occurrences)



La liste 13 délivre le vocabulaire spécifique (on dit parfois caractéristique ou significatif) de la période considérée. Elle souligne les reliefs (les excédents) de ce qui se présente comme le portrait collectif les Lettres françaises de l'époque (il en existe une autre pour les ombres ou déficits). En réalité ce portrait-robot est constitué de milliers de touches dont quelques centaines seulement apparaissent dans notre extrait. On sait que l'écart réduit permet des prévisions solides, dès lors que sa valeur atteint 2 en valeur absolue. Même en fixant la valeur du seuil fort au-delà de cette limite ($z > | 9 |$), on a recueilli près de 3000 formes spécifiques dans le sens positif, et plus de 1000 dans le sens négatif. C'est dire combien le vocabulaire est sensible aux conditions variables du discours et combien l'exercice naturel du langage est loin d'un saupoudrage aléatoire[23]. Ces écarts portent aussi bien sur les mots grammaticaux que sur les autres et l'on pourra vérifier que ceux qu'on a déjà rencontrés dans la figure 9 apparaissent bien dans la liste 13 [24].

Les premiers éléments du tableau 13 n'appartiennent pas au vocabulaire proprement dit, mais au système des ponctuations. On peut répugner à s'attarder sur les points de suspension qui occupent la première place, en considérant qu'il s'agit là d'un signe plus récemment introduit dans le système et que cette place est attendue. En réalité c'est Céline qui est surtout responsable de ce débordement. Mais l'élément suivant est d'un intérêt plus évident, car il s'agit du point et donc de la longueur des phrases. Comme les points d'interrogation sont aussi largement excédentaires, on ne prend aucun risque en affirmant que la phrase est en moyenne plus courte dans la période 1914-1945 que dans celles qui ont précédé[25]. Toutes les ponctuations fortes sont en forte croissance depuis 1600 (point +0,85, exclamation +0,76, interrogation +0,61, suspension +0,86), au contraire des ponctuations faibles (deux-points -0,33, point-virgule -0,51).

---

[23] Certains tirent argument de ces écarts, toujours constatés, pour mettre en cause le schéma d'urne, dont l'inadéquation, à leur dire, est trop patente pour servir de modèle. Mais comme rien n'a été proposé pour remplacer les lois classiques, on croit devoir continuer à les utiliser, non point tant comme modèle, mais comme référent stable pour les mesures comparatives. Le géomètre qui établit les courbes de niveau sur le terrain se soucie peu que la nature s'écarte du modèle géométrique.

[24] Les écarts ne sont pourtant pas strictement équivalents. Car la comparaison a été établie ici sur l'ensemble du corpus de FRANTEXT, alors que la précédente avait exclu les essais techniques.

[25] Il n'y a qu'une définition opératoire de la phrase dans un traitement automatique: l'espace compris entre deux ponctuations fortes. Les points d'exclamations - dont on note un léger déficit dans notre période - sont à ranger aussi parmi celles-ci.

Tableau 13. Spécificités positives

1914-1945	Corpus total		Écart réduit	1914-1945	Corpus total		Écart réduit
123723	383015	...	227.3	8373	32340	pouvait	36.1
1115270	5303763	. (point)	172.0	4199	14055	musique	36.1
22752	56489	ça	136.5	11159	45372	ailleurs	35.6
88631	331478	était	128.3	7769	29761	as	35.6
467928	2194549	-	122.0	3492	11269	simplement	35.4
74598	275366	avait	121.8	4008	13440	connaissance	35.1
198839	910074	pas	91.6	153583	780553	elle	35.0
111407	499245	c'	76.4	9581	38379	guerre	34.7
30407	116394	(	70.6	2068	5799	apparaît	34.6
14848	48773	avais	70.5	11205	46020	serait	34.5
30857	119467	)	68.9	12016	49861	oui	34.5
250379	1222284	une	67.0	2012	5609	appel	34.4
47240	197129	tu	67.0	1395	3419	mi	34.4
6488	18225	parfois	61.1	2231	6447	comprends	34.3
4121	10152	soudain	58.7	8349	32852	surtout	34.2
100710	473135	comme	56.0	3199	10279	compris	34.1
6974	21990	regard	52.2	4326	15045	gauche	33.8
7998	26655	visage	50.2	2325	6870	journaux	33.8
31601	134994	non	50.2	11499	47812	avaient	33.5
1070	1679	the	48.5	5379	19695	venait	33.4
1043	1617	offensive	48.4	3685	12413	sourire	33.4
5849	18415	soi	47.9	13530	57564	sens	33.3
40345	179903	là	47.1	1894	5304	contact	33.2
21653	89661	vers	46.6	1143	2665	kilomètres	33.1
3569	9909	geste	46.2	1147	2699	allemande	32.8
288754	1473067	est	45.8	5698	21245	travers	32.8
4326	12833	peux	45.7	1645	4433	chance	32.8
2964	7789	début	45.6	1723	4724	angoisse	32.7
8888	31780	faisait	45.4	4830	17563	fallait	32.2
315858	1617861	un	45.4	6240	23873	rue	32.0
3732	10666	journal	45.1	2475	7682	exactement	32.0
6842	23211	oeuvre	44.8	5627	21177	certains	31.8
5274	16768	aussitôt	44.7	92928	465966	tout	31.8
1081	1862	ben	44.7	1127	2705	cellules	31.7
9249	33606	étais	44.6	2074	6149	grand'	31.7
5076	16150	vite	43.8	1333	3431	pensais	31.5
3138	8824	sentait	42.4	1926	5633	savais	31.2
1903	4468	mètres	42.4	1979	5837	attitude	31.2
2349	6162	minute	40.7	26089	120741	très	31.2
11903	46896	sais	40.6	1598	4419	drame	31.1
2707	7527	phrase	40.1	9145	37635	forme	31.0
1337	2836	mystique	40.1	1593	4425	allemand	30.8
21536	93092	alors	39.5	10985	46461	soir	30.7
6210	21866	tard	39.3	2552	8212	poésie	30.4
5094	17172	semblait	39.2	7214	28821	autour	30.3
3420	10411	train	38.9	11003	46768	pourquoi	30.2
4747	15795	comprendre	38.8	2700	8852	oeuvres	30.2
9847	38179	aurait	38.8	1500	4160	absurde	30.0
16209	68065	étaient	38.4	1528	4268	papa	29.9
3835	12154	savait	38.3	2928	9839	vide	29.9
8853	33881	pourtant	38.1	2778	9243	parlait	29.7
2115	5617	pensait	37.9	15760	70140	avant	29.7
1389	3149	gare	37.8	11237	48151	nuit	29.5
5899	20969	allait	37.5	9905	41767	abord	29.5
15522	65235	petit	37.4	2082	6442	sud	29.5
3525	11076	éléments	37.4	8289	34118	nouveau	29.5
4796	16342	vraiment	37.2	2342	7502	odeur	29.4
1839	4720	gaz	37.1	3508	12388	voyait	29.4
7862	29766	disait	37.0	12214	52964	voix	29.3
1254	2781	rythme	36.9	22381	103310	fois	29.3
100369	498738	?	36.2	4772	17972	pourrait	29.2

1914-1945	Corpus total		Écart réduit	1914-1945	Corpus total		Écart réduit
4765	17966	aurais	29.1	1127	3246	loup	24.5
4184	15410	droite	29.0	3202	11955	importance	24.5
3856	27608	longtemps	28.8	3243	12143	nord	24.5
1030	2577	russe	28.7	1059	2998	métal	24.4
3811	13832	connaître	28.7	25130	120498	entre	24.4
22919	106496	puis	28.6	1259	3762	signification	24.4
1904	5853	notion	28.6	1593	5086	gestes	24.4
11583	50258	presque	28.5	2480	8809	lignes	24.4
1897	5856	essence	28.3	96517	495426	lui	24.4
2418	7972	brusquement	28.2	1039	2928	divisions	24.3
2570	8603	type	28.2	2326	8156	aimait	24.3
3666	13294	ceci	28.2	1753	5749	complètement	24.3
2126	6783	culture	28.2	2336	8212	critique	24.2
1664	4974	dirait	28.0	1326	4039	accent	24.2
1320	3675	paysage	28.0	1745	5733	vas	24.1
2868	9928	tenait	27.8	1143	3344	visages	24.1
1374	3908	vision	27.6	7074	30130	savoir	24.0
1002	2560	piano	27.6	46150	229893	moi	24.0
2248	7365	retrouver	27.6	2536	9126	faudrait	23.9
11548	50495	toi	27.5	13887	63712	seulement	23.9
2878	10028	courant	27.4	2572	9295	minutes	23.8
1229	3388	patron	27.4	11388	51298	quoi	23.8
5215	20428	existence	27.4	1293	3964	cherchait	23.6
3323	11957	ligne	27.4	1712	5657	attaque	23.6
5648	22430	face	27.3	1082	3148	aube	23.6
10602	46020	petite	27.2	2856	10587	fenêtre	23.6
5024	19609	es	27.2	1879	6363	prenait	23.6
1578	4735	souffrance	27.1	3132	11837	théâtre	23.5
1229	3419	élan	27.0	3926	15450	effort	23.4
3846	14350	expérience	26.9	1554	5042	arrivait	23.4
3733	13855	midi	26.9	1669	5516	entendait	23.3
5165	20357	conscience	26.7	4320	17288	front	23.3
1503	4485	revue	26.7	1910	6524	artiste	23.3
2858	10073	clair	26.6	26581	128753	vie	23.3
3154	11354	impression	26.6	2004	6933	bureau	23.2
1682	5198	cuisine	26.6	1352	4241	poser	23.2
9468	40821	eau	26.5	1252	3848	albert	23.2
6533	26818	silence	26.4	1923	6604	roman	23.1
2560	8841	groupe	26.4	2101	7366	souvenirs	23.1
6089	24744	question	26.4	1714	5737	sensation	23.1
2298	7739	justement	26.4	2521	9211	donnait	23.0
1334	3878	sentais	26.3	392602	2092858	il	23.0
1823	5802	disais	26.2	1065	3141	lourde	22.9
3172	11511	garçon	26.2	1544	5056	robert	22.9
336320	1779214	d'	26.1	1053	3111	hein	22.7
1126	3131	liquide	25.9	2414	8783	regardait	22.7
2527	8787	connais	25.8	1845	6331	attendait	22.7
1033	2798	intellectuelle	25.8	1303	4096	souci	22.7
2472	8582	restait	25.6	1833	6284	pouvais	22.7
6765	28186	sait	25.5	4308	17397	sûr	22.6
12101	54097	va	25.5	1358	4327	lampe	22.6
4064	15714	entendu	25.1	7787	34069	matin	22.6
1596	5034	note	25.0	3257	12588	signe	22.5
4972	19921	penser	24.9	1627	5445	texte	22.5
2068	6983	connaît	24.9	1506	4968	pose	22.3
3463	13051	expression	24.9	7342	32011	travail	22.2
4259	16656	paraît	24.9	3566	14070	rouge	22.2
4347	17071	dû	24.8	1261	3991	commandant	22.0
1620	5149	connaissait	24.8	1425	4659	formule	22.0
2523	8928	certes	24.8	2216	8035	reconnaître	21.9
3596	13674	importe	24.7	2674	10063	voudrais	21.9
2417	8489	mur	24.7	12691	58680	chaque	21.9
3870	14928	image	24.7	10780	49121	coup	21.8
2331	8134	images	24.6	3231	12612	existe	21.8
1958	6561	café	24.6	1392	4558	instants	21.7

Après les ponctuations, les premiers éléments de la liste 12 ne permettent pas encore d'aborder la thématique. Car il s'agit de formes verbales (*était, avait, avais,* et plus loin *faisait, étais, sentait, semblait, étaient, savait, pensait*, etc..) dont le temps - et non le contenu - est déterminant: c'est l'imparfait de la narration, ce "temps cruel" où Proust voyait une des caractéristiques stylistiques de Flaubert et du roman. La liste suggère d'autres observations qui se situent entre le style et le thème: par exemple l'afflux des mots anglais dont la trace est laissée visible par les particules *the, of, to, that,* ou encore l'émergence de termes familiers comme *gars, gosse, papa, boches, type* (pour certains de ses emplois). D'autres notations semblent décrire le cadre, qui n'est plus exactement le "milieu" de Zola. Déjà les éléments de la vie moderne sont en place: c'est le *quartier* (ou le *village*), avec ses *rues*, son *café*, son *église*, son *hôtel* et ses notables (*curé*, *médecin* ou *docteur*, *directeur* ou *patron*) et l'ameublement commercial ou privé dont l'époque a le goût: *bureau, salle, fauteuil, machine, cuisine*. Ce sont les transports et les communications modernes: *train, voie, voies, gare, tramway, auto, chauffeur, essence, avion, téléphone*. L'obsession du corps, si fréquente à l'époque naturaliste, a disparu: ne reste que le *visage (face, voix, regard, joue, nez, peau*). Cela ne veut pas dire qu'on en revienne au sentiment et qu'on rejoigne la sensibilité romantique: les termes où s'exprimait l'amour au XIX^{e} figurent parmi les déficits spécifiques: *coeur, amour, plaisirs, charme, amitié, amant, voeux, larmes, pleurs, peines, douleur, maux.* On ne revient pas davantage aux valeurs traditionnelles, qui se maintiennent dans la zone négative: *honneur, justice, vertu, gloire, fortune, estime, bonté, principes*.

Ce qui est nouveau c'est la plus-value dont jouit l'exercice de l'écriture sous toutes ses formes, du journalisme à la poésie (*journal, presse, revue, critique, texte, écrivain, roman, poésie, lyrisme, langage, phrase, mot*) et l'extension du vocabulaire esthétique à l'ensemble des arts (*création, art, artiste, théâtre, cinéma, peinture, image*[26]*, musique, sonate, piano, fugue, chant, danse*). Ce qui est nouveau, ce n'est pas tant l'attention particulière prêtée aux choses militaires, même si la guerre de 1914 laisse des traces indélébiles dans les hantises et le vocabulaire de l'époque: *guerre, soldat, ennemi, divisions, capitaine, commandant, chef, offensive, attaque, obus, mitrailleuses, carte, sac,* etc. C'est - après l'affaire Dreyfus et la grande Guerre - le débat politique qui oppose la *droite* et la *gauche.* C'est, après Freud, une exploration des profondeurs de la conscience[27], et déjà une interrogation existentialiste et la problématique des sciences humaines (*conscience, question, expérience, existence, absurde, geste, attitude, sens, signification, signe, expression, angoisse, drame, souffrance, appel, vision, civilisation, culture, groupe, habitude, travail, mode,* etc..). C'est le

---

[26] La couleur, déjà fort prisée au temps du naturalisme, se ravive encore. Le mot *couleur* est significatif ainsi que la plupart des teintes de la palette : *bleu, vert, rose, rouge, blanc noir, gris.*

[27] Le mot *conscience* supplante le mot *âme,* tombé en désuétude.

temps qui n'est plus un temps à longue portée, orienté vers l'horizon et l'éternité, comme chez Chateaubriand, mais un temps éclaté, écourté qui se disperse dans les *instants* ou les *moments* et se mesure en *minutes* et en *heures* et non en années ou en siècles.

Tout cela constitue-t-il une thématique? À quelques signaux lointains on croit reconnaître parfois les thèmes de Proust, Breton ou Bernanos, dans la dérive qui emporte l'époque d'une guerre à l'autre. Mais il serait plus juste de parler d'un climat, d'une sensibilité diffuse qui s'accroche à certains mots, comme des pans de brume suspendus aux reliefs.