Ce que disent les chiffres

Étienne Brunet

Etienne Brunet
Institut National de la langue française (CNRS)

Ce que disent les chiffres

Aperçu statistique sur le vocabulaire français

L'histoire de la langue a-t-elle besoin de chiffres? La langue est une réalité familière, une donnée immédiate, à laquelle la conscience a un accès direct. De même qu'on a conscience du temps qui passe, même sans horloge, ou du temps qu'il fait, même sans thermomètre, ainsi le sentiment linguistique se fait jour sans aide extérieure, sans qu'un appareil enregistreur soit nécessaire pour capter les représentations intérieures. Peut-on d'ailleurs imaginer quelque détecteur de mensonge ou de langage qui ne soit lui-même un mensonge? S'il s'agit du monologue intérieur que secrète la pensée en action ou en sommeil, on peut être rassuré sur le viol improbable de la conscience. Les techniques les plus audacieuses de la résonance magnétique nucléaire peuvent localiser les évènements dont le cerveau est le siège, mais non les définir et les analyser. Reste qu'assez souvent ces phénomènes mentaux se manifestent au-dehors par le langage. Dès qu'est franchi le mur du son ou de l'écriture, la technique peut installer ses micros, ses appareils, ses pièges à paroles. Des montagnes de documents enregistrés s'accumulent chaque jour dans les agences de presse, les salles de rédaction, les studios de radio ou de télévision et même dans les entreprises où l'information compte autant que l'argent et où s'impose l'obligation de la veille technologique. Mais cette masse hétéroclite et sans cesse renouvelée de documents écrits,

iconographiques ou sonores produit de redoutables engorgements, dont se dégagent difficilement les méthodes les plus sophistiquées et les plus puissantes de l'informatique et de la statistique. Qu'on songe par exemple que le réseau *Internet* par où tend de plus en plus à circuler l'information s'accroît à une vitesse exponentielle. Le serveur *Lycos* qui possède un moteur de recherche impressionnant gère un parc de 60 millions de documents répartis dans le monde entier, alors que sa base de données atteignait à peine 14 millions, il y a un an. Il faut une ou deux secondes pour isoler un mot parmi des milliards et restituer l'environnement qui est le sien et l'information qu'il véhicule.

Ce n'est pourtant pas à *Lycos* qu'il faut s'adresser, non plus qu'à d'autres services comme *Yahoo*, *AltaVista* ou *Infoseek*, pour l'étude de la langue française - dont, au reste, on se sert peu sur *Internet*[1]. Les informations qu'on en tirerait n'auraient guère d'intérêt linguistique, sinon pour constater que l'orthographe française y est souvent malmenée et les signes diacritiques sacrifiés[2].Et surtout la faible profondeur dans le temps, sinon dans l'espace, du champ d'observation interdirait tout point de vue comparatif et historique. Car la démarche statistique - et celle de l'historien a des exigences semblables - suppose que les faits soient étudiés à intervalles réguliers et confrontés les uns aux autres, aucune mesure, si précise soit-elle ne fournissant la moindre indication tant qu'elle n'est pas rapportée à quelque autre. Dans le domaine de la langue, on ne peut compter sur aucune norme, ni naturelle comme le niveau de la mer, ni artificielle comme le mètre-étalon. Les observations n'y ont qu'une portée relative et ne prennent sens que par rapport à d'autres mesures.

Frantext

Or depuis quelques mois *Internet* autorise de telles mesures, si l'on consulte, par ce canal, la meilleure des bases de données linguistiques qui existe au monde: *Frantext*. Avec près de 3000 textes de la littérature nationale, engrangés méthodiquement depuis trente ans, *Frantext* n'a guère

---

[1] On fait partie des optimistes si l'on estime à 5% la part du français sur Internet. Pour mesurer cette présence française sur la "toile", on peut s'appuyer sur un indice, que beaucoup d'autres confirment: le nombre d'occurrences du mot *homme* dans Internet (3673 le 16 août 1996), rapporté à la fréquence du correspondant anglais *man* (65111 occurrences). Les optimistes feront valoir aussi que Internet s'est développé d'abord aux États-Unis, où l'informatique est plus dense qu'en France dans les foyers et les bureaux, qu'il y a donc lieu d'espérer un rattrapage dans les années qui viennent et qu'ainsi les parts de marché peuvent s'améliorer pour le français.

[2] Ce sacrifice des accents, imposé naguère dans le courrier électronique par l'incohérence et l'incomplétude du code ASCII, n'a plus de raison d'être, les caractèrers accentués étant prévus et standardisés sur la "toile" du WEB.

d'équivalent dans les autres langues, ni pour l'étendue, ni pour l'homogénéité des données, ni même - cela est nouveau - pour leur accessibilité. *Frantext* est certes disponible depuis des années à la communauté scientifique et deux versions successives du logiciel d'exploitation ont été mises en oeuvre sur les réseaux existants, principalement *Transpac*. Mais l'interrogation de la base supposait un certain apprentissage qui a rebuté plus d'un chercheur, au lieu que l'ergonomie de la "Toile" est d'une telle facilité (il suffit de "cliquer" les choix proposés) et d'une telle généralité que l'obstacle technique a disparu, même pour le chercheur le plus craintif. On peut s'en assurer avec l'exemple de consultation donné ci-après (figure 1).

Figure 1. L'accueil de Frantext sur Internet, à l'adresse:
http://www.ciril.fr/~mastina/FRANTEXT

Figure 2. Sélection du corpus

Figure 3. Recherche d'une expression

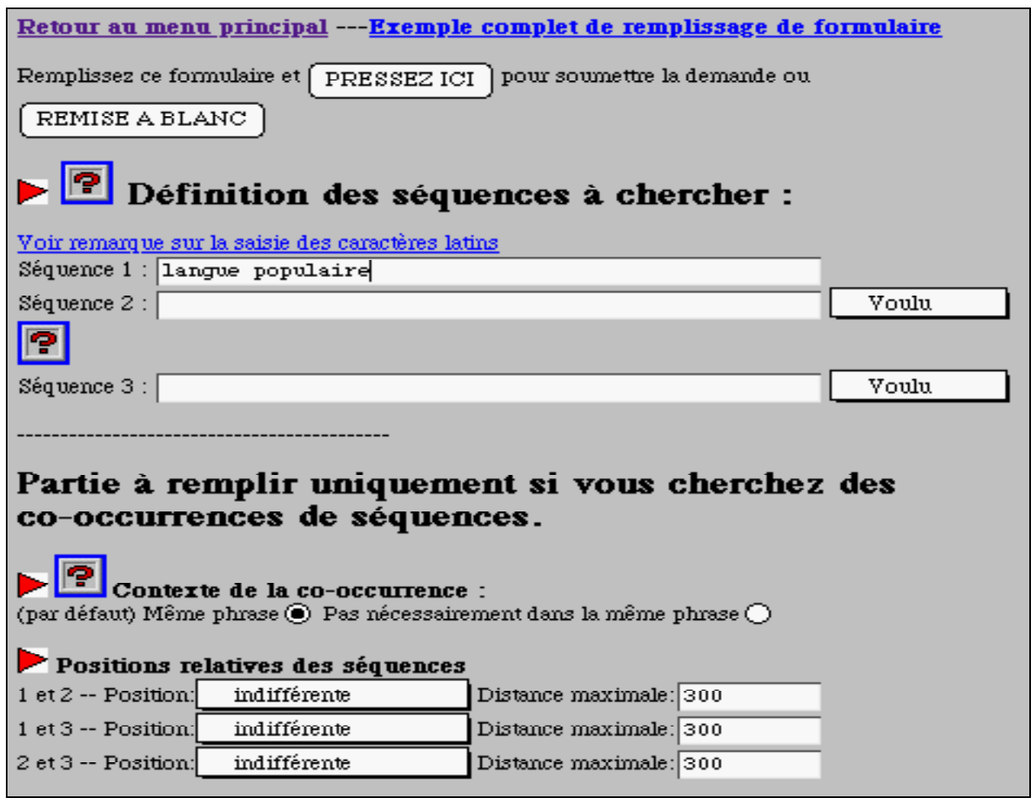


Figure 4. Résultat de la requête

On a affaire dans cet exemple à une question simple. Le langage d'interrogation (qui porte de nom de *Stella* et a été réalisé par Jacques Dendien) permet des consultations plus complexes et des réponses plus précises ou plus étendues. S'agissant de l'ensemble des faits linguistiques dont se nourrit l'histoire de la langue, le recours à *Frantext* exige des opérations particulières qui font appel aux fonctions statistiques offertes par *Stella*. Sans s'appesantir sur le détail technique, on se bornera à indiquer qu'on a puisé dans *Frantext* des relevés bruts, dont le traitement a été assuré par le logiciel *Thief*, créé par nos soins (menu principal ci-dessous, fig. 5).

Figure 5. L'exploitation statistique de *Frantext*

En particulier on a constitué à partir de *Frantext* un dictionnaire des fréquences, en relevant toutes les formes du corpus littéraire et les sous-fréquences de chacune dans 12 tranches chronologiques distinguées du XVI^{e} siècle à nos jours. Une fois constitué, ce dictionnaire, devenu disponible en local (sans liaison extérieure), a fait l'objet d'une exploitation intensive dont sont issus la plupart des résultats qui vont suivre. Les limites des tranches n'ont pu être établies sur un pied d'égalité, car les textes dépouillés sont très inégalement répartis selon les siècles. Afin d'équilibrer la taille des sous-ensembles, l'empan chronologique a été élargi là où les textes étaient rares, c'est-à-dire au XVIe siècle, et resserré là où ils abondaient, aux XIXe et XXe siècles. La première tranche s'étend ainsi sur un siècle (on l'a représentée pas son année médiane: 1550) tandis que les plus proches ne recouvrent guère que deux décennies. Voir tableau 6.

Tableau 6. Limites des 12 tranches

	Nb.mots	Nb.Formes	prob.p	prob.q	époque
1	1719178	67014	0.014625	0.985375	1550
2	8346862	101892	0.071006	0.928994	1630
3	6087533	69612	0.051786	0.948214	1692
4	9380093	77841	0.079796	0.920204	1735
5	11946384	99028	0.101627	0.898373	1780
6	11124272	98905	0.094633	0.905367	1820
7	16184517	124845	0.137680	0.862320	1855
8	13780168	116085	0.117227	0.882773	1885
9	8695375	98488	0.073971	0.926029	1910
10	11361661	109218	0.096653	0.903347	1928
11	10083262	106498	0.085777	0.914223	1942
12	8842284	112367	0.075220	0.924780	1960
TOT	117551589	393848			

Même ainsi, l'égalité dans l'étendue des tranches n'est pas respectée et les calculs de pondération sont inévitables. Ils s'appuient tous sur les probabilités indiquées dans le tableau précédent. On renvoie le lecteur aux ouvrages de Charles Muller pour tout ce qui concerne les opérations techniques de la statistique linguistique[3]. On s'en fera toutefois une idée suffisante si l'on sait que toute observation réelle (pour un mot donné dans une tranche donnée) est comparée à une fréquence théorique, obtenue par une règle de trois, sur la base de l'étendue respective des tranches. Le résultat de cette comparaison est un nombre négatif ou positif dont le signe indique s'il s'agit d'excédent ou de déficit et dont la valeur absolue mesure l'importance de l'écart (quand l'écart est faible, entre -2 et + 2, le hasard peut être invoqué).

Avant de rendre compte des observations chiffrées une précaution est à prendre. On évitera un écart ou un abus de langage en s'abstenant de parler de la langue ou même du lexique, pour s'en tenir au seul vocabulaire. Car langue et lexique sont des réalités virtuelles dont les réalisations écrites n'épuisent pas 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. On se bornera donc à tenir 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 et, de façon plus molle encore, la trajectoire incertaine de la langue.

En outre une plus grande prudence s'impose en l'absence de lemmatisation. Cette opération, qui consiste à regrouper les formes fléchies derrière leur chef de file, qui est traditionnellement l'infinitif des verbes et le masculin singulier de la classe nominale, exige de coûteux et longs efforts qu'on ne peut guère entreprendre, avec des raccourcis approximatifs, que pour les textes les plus récents. S'engager dans cette voie - comme nous l'avons fait dans notre *Vocabulaire français de 1789 à nos jours* - eût été

---

[3] Charles Muller, *Initiation aux méthodes de la statistique linguisttique,* 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.

héroïque, vu l'énormité du corpus, 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. Faute de pouvoir isoler sûrement les vocables, on s'est donc contenté, à regret, des formes dans leur plus simple appareil.

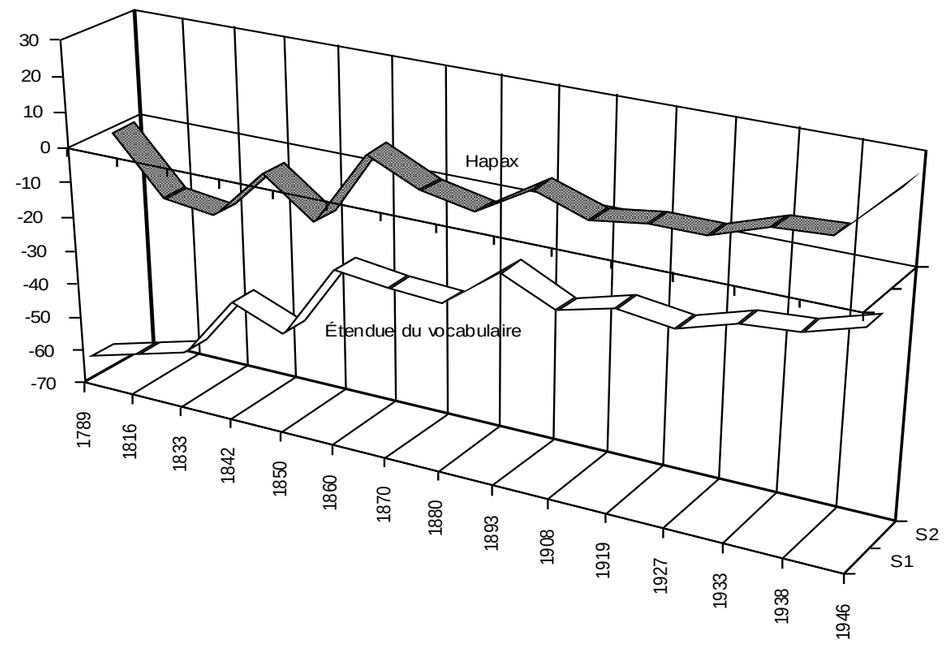
L'inflation lexicale

Rappelons d'abord les résultats antérieurs, qui abordent le problème bien connu de la richesse lexicale et que l'on a obtenus, en traitant des données lemmatisées, à partir du corpus XIXe-XXe, divisé en 15 tranches chronologiques. Voici le tableau des effectifs et des écarts, 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 écarts réduits (lignes 4 et 6 du tableau) montre la progression en deux siècles du flux verbal. Cette planche à billets qu'est la créativité lexicale a bien fonctionné depuis la Révolution, et l'inflation du vocabulaire s'observe au niveau général, quand toutes les unités lexicales sont prises en compte, et au niveau particulier mais révélateur des hapax (ou mots employés une seule fois).

Figure 7. L'inflation lexicale depuis 1789



La courbe correspond au sentiment qu'on peut avoir naïvement des mouvements du vocabulaire. La masse lexicale, comme la masse monétaire, s'accroît sans cesse, pour répondre aux besoins de la technologie qui invente

des objets nouveaux, qu'il faut bien nommer, pour répondre aussi à l'usure des mots et à la surenchère naturelle qui s'exerce dans le commerce des mots comme dans le commerce des biens et des marchandises. La loi des échanges donne toujours une plus-value à ce qui est neuf. Mais dans le domaine linguistique ce qui est vieux n'est pas pour autant perdu et oublié. Les mots vieillis vivotent longtemps et quand on les rencontre dans un texte ancien, on les reconnaît encore, avec surprise et plaisir, comme les objets abandonnés au grenier. Il n'y a pas équilibre entre les morts et les naissances verbales et cela conduit sinon à la surpopulation, du moins à un certain encombrement des communications.

Pourtant la courbe prolongée jusqu'au XVI^{e} siècle semble en désaccord avec les observations faites jusqu'ici. Ce sont les premières tranches qui l'emportent, tant pour l'étendue du vocabulaire (figure 8) que pour le nombre des hapax, ce que tend à montrer aussi l'accroissement dynamique du vocabulaire. Dans cette dernière perspective, on se déplace dans le temps, d'une tranche à l'autre, en notant l'apport lexical de chacune. Là encore le XVI^{e} siècle garde une part de ses prérogatives, sans masquer tout-à-fait la tendance invincible au renouvellement (figure 9).

Figure 8. Courbe de la richesse lexicale depuis 1500
(par la méthode de la loi binomiale)

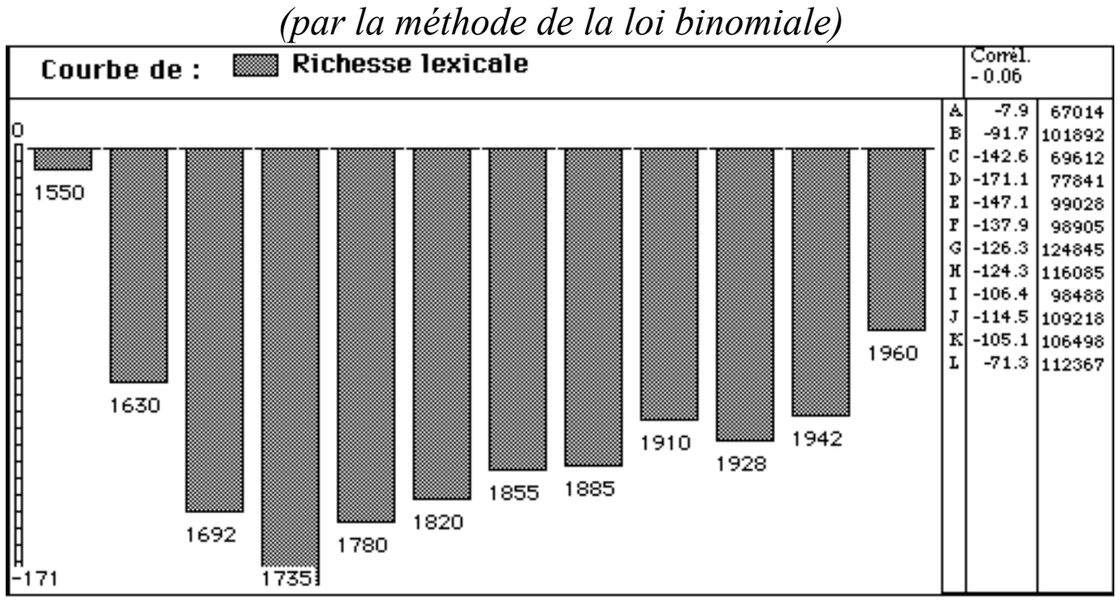
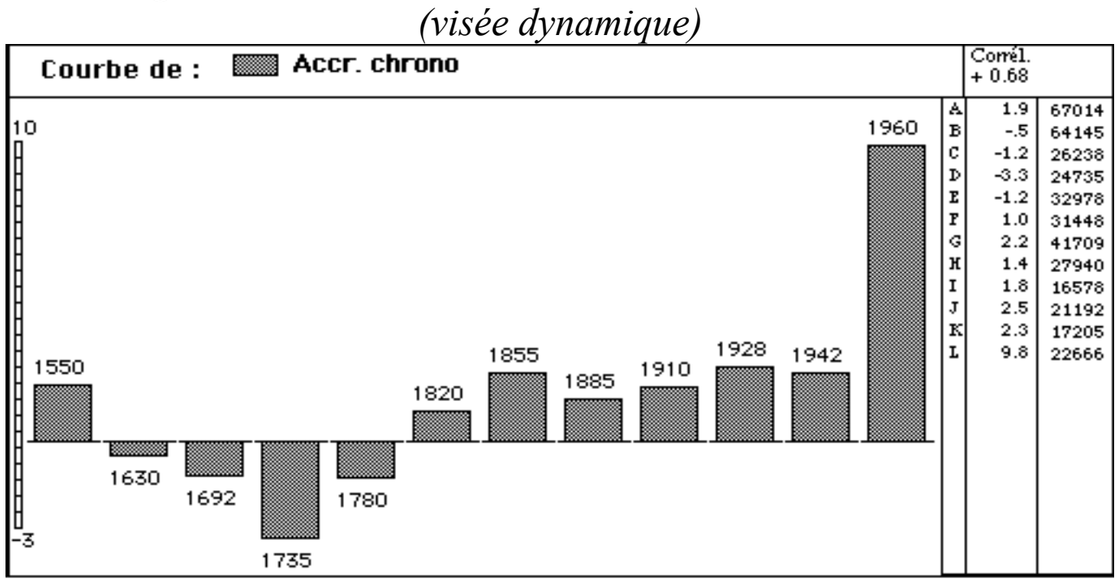


Figure 9. Courbe de l'accroissement du vocabulaire
(visée dynamique)



L'explication de cette anomalie est assez triviale. Point n'est besoin d'invoquer Malherbe pour opposer l'esthétique sobre et sévère des classiques à la luxuriance lexicale de la Renaissance, à qui ni les archaïsmes, ni les néologismes ne faisaient peur. Il s'agit tout bonnement de variations orthographiques. Si le corpus avait été lemmatisé, cet artefact aurait disparu, les doublons et les variantes rejoignant la vedette de regroupement. Mais les formes brutes ne peuvent échapper aux perturbations d'une orthographe non normalisée et le gonflement des effectifs tient à ce qu'un même mot est comptabilisé plusieurs fois dès que l'ajout ou le retrait d'un accent lui donne une identité nouvelle. Ainsi on a compté jusqu'à neuf variantes orthographiques de l'*évêque* dans les premières tranches. L'époque moderne a naturellement supprimé les sièges surnuméraires et les titres *in partibus*.

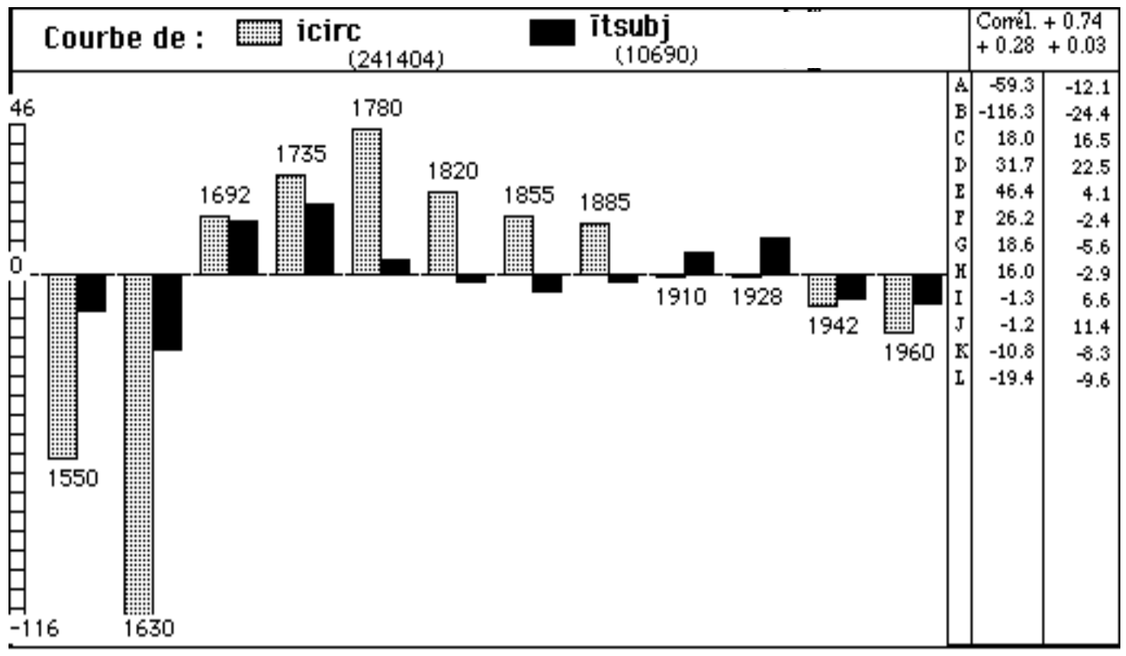
L'orthographe

La tendance à l'inflation lexicale ne peut donc être mise en doute[4] et son effet se fait sentir dans les courbes, même au XVIIe siècle, dès lors que l'orthographe commence à se stabiliser. Quand l'attention se fixe précisément sur les signes diacritiques, on observe en effet de grands flottements, et ici l'étude des formes représente un avantage méthodologique. La récente réforme de l'orthographe a porté sur la place publique le fameux accent circonflexe. Les historiens de la langue savent que ce symbole apparaît pour la première fois dans la première moitié du XVIe siècle et que ses fonctions ont été diversifiées à l'origine (notation d'une diphtongue, amuissement d'un *e*, signalisation d'un *s* effacé) et sont devenues assez incohérentes au fil du temps. La réforme n'a pas touché aux deux voyelles *a* et *o*, que nous écarterons de nos relevés. Restent le *î* et le *û* qui sont devenus facultatifs, sauf dans les désinences verbales et là où quelque homographie serait à craindre (*mûr, sûr, dû, fût*). Il ne nous appartient pas de décider si cette rectification est bonne ou mauvaise, mais force est de constater qu'elle va dans le sens de l'histoire et qu'elle officialise une désaffection progressive des écrivains, depuis deux siècles. La courbe en grisé de la figure 10 est relative à la lettre *i,* quand elle se coiffe d'un circonflexe. Cela se produit 241404 fois, soit dans un mot sur 500 en moyenne. Mais le destin de ce signe est celui d'un feu de paille, qui couve un siècle, se consume brusquement au siècle suivant et s'éteint par la suite.

---

[4] N'oublions pas que nous ne considérons que le corpus littéraire, qui est moins perméable aux apports nouveaux que le corpus technique. La comparaison des genres a été faite et les calculs confirment cette évidence.

Figure 10. Le *i* circonflexe
(en gris la courbe d'ensemble, en noir le subjonctif imparfait)



Cependant certains emplois résistent mieux que d'autres. C'est le cas du subjonctif (courbe en noir de la figure 10). C'est qu'en cette circonstance, le circonflexe joue un rôle pleinement distinctif, en opposant *fit* et *fît*[5], *prit* et *prît* et tous les éléments du paradigme aux formes correspondantes du passé simple. Il apparaît donc que les auteurs de la réforme ont été bien inspirés en confortant le circonflexe là où il est utile et où le maintient l'usage littéraire.

Il n'en va ainsi du passé simple où le circonflexe est associé aux formes en *îmes* et en *îtes*. Ici les homographes restent rares[6] et sont le fait de rencontres accidentelles avec des substantifs ou des participes (*mîmes* et *mimes*, *prîmes* et *primes*, *frîtes* et *frites*, *dîmes* et *dimes*). Cette moindre motivation explique en partie l'effondrement de la série, dont rend compte également l'abandon, de plus en plus accusé, du passé simple. On remarquera l'écart qui se creuse entre la première personne (*îmes* 6404 occurrences) et la seconde (*îtes* 616 occurr.). Les formes en *îmes* trouvent encore à s'employer dans certains récits autobiographiques, où l'archaïsme des formes accompagne volontiers la nostalgie du souvenir et le recul du temps. Les formes en *îtes* supposent au contraire un dialogue où n'entrent guère ces ingrédients[7].

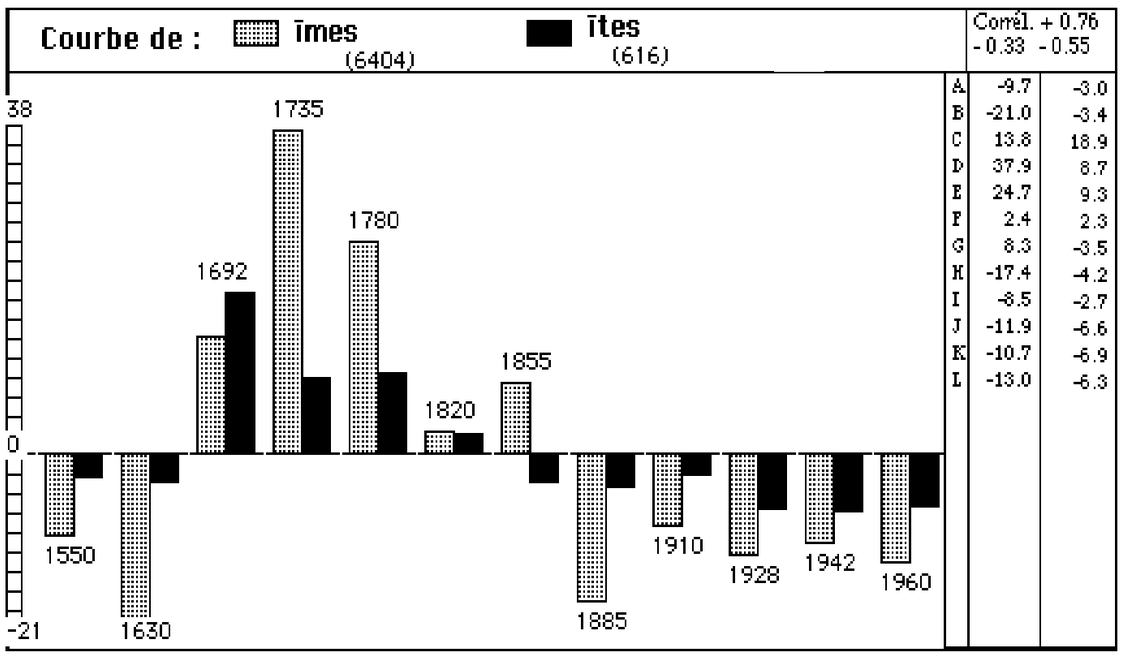
---

[5] Sur plus de 10000 occurrences du subjonctif en *ît*, la forme *fît* en accapare le quart (2427). Viennent ensuite *prît* (762), *mît* (488), *rendît* (362), *dît* (312), *permît* (301), *entendît* (277), *comprît* (147), etc.

[6] On a cependant *dîtes* et *dites* (indicatif et participe).

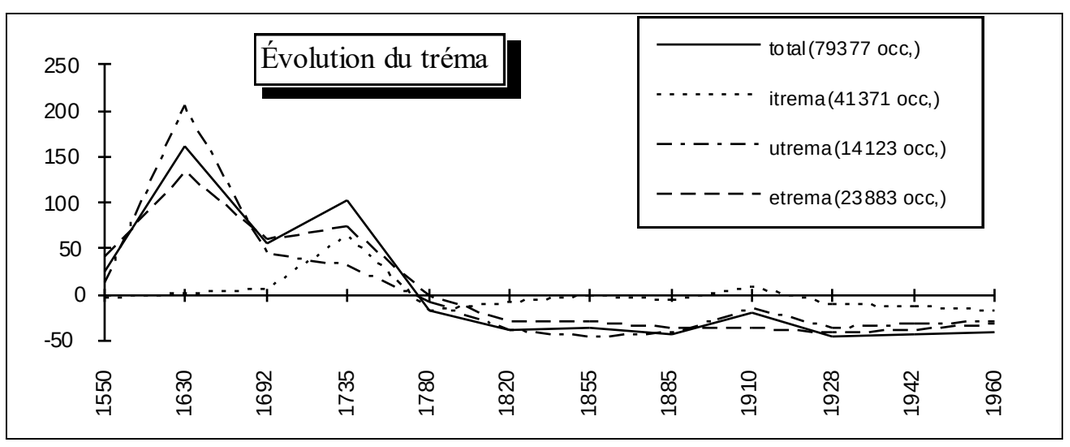
[7] Le verbe faire est ici devancé par le verbe voir: *vîmes* 1118 contre *fîmes* 814. Mais il garde ici sa suprématie à la seconde personnes: *fîtes* 199, devant *vîtes* 88 et *dîtes* 56.

Figure 11. Les formes en î du passé simple



Les faits qu'on relève pour le *u* circonflexe sont tout à fait semblables. Le parallélisme est dans une fréquence équivalente (222 618 occurrences) et dans la répartition des emplois. Quant au tréma, les chiffres montrent un embarras des usagers lorsqu'il faut choisir entre le *e* et le *u* dans les formes féminines du type *aiguë* (ou *aigüe*). La réforme tente bien de mettre fin à cette hésitation, mais il semble que ce soit trop tard, le Français optant pour l'abstention, à partir de la Révolution. La courbe 12 ne laisse guère de chances de survie à ce signe promis à l'oubli. Le moins menacé des trémas est celui qui accompagne la lettre *i*. C'est aussi le moins rare, puisqu'il compte autant d'occurrences que les deux autres réunis. Son coefficient de corrélation chronologique (-0,36) indique une descente moins abrupte que celle des deux autres (respectivement -0,63 et -0,81)[8].

Figure 12. Le tréma



[8] Rappelons pour ceux à qui la statistique est peu familière que le coefficient de corrélation évolue entre deux limites -1 et +1. Dans le cas présent (avec des séries de 12 éléments) les valeurs qui dépassent 0,57 sont significatives et échappent au hasard.

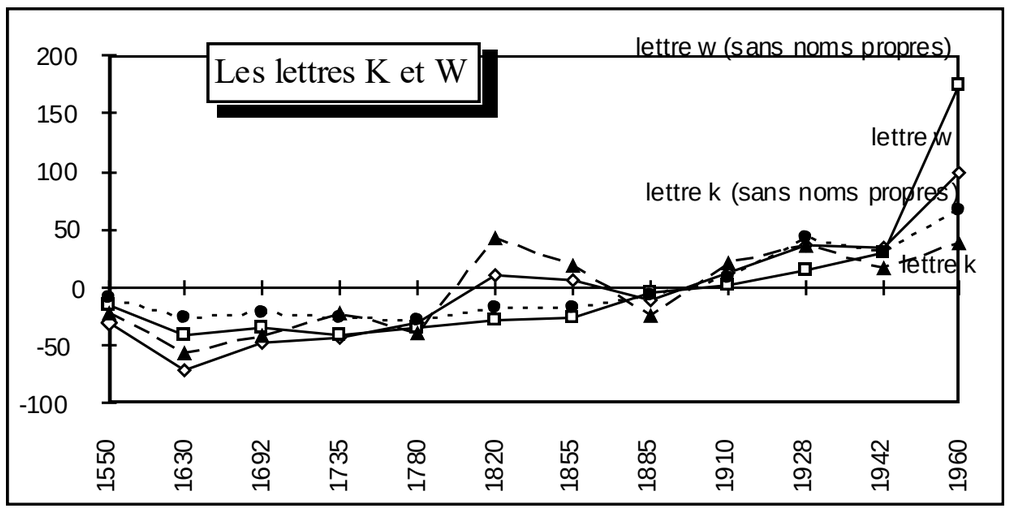
On a mesuré précédemment le taux global de renouvellement lexical, et quelques traits accessoires dans la physionomie orthographique du vocabulaire. Ces signes diacritiques peuvent flotter au cours des siècles, sans modifier grandement la structure de la langue. Au reste on accepte que les majuscules perdent parfois leurs accents dans les documents dactylographiés, même si la tradition typographique recommande de les maintenir - et la richesse du clavier des traitements de texte modernes permet de satisfaire cette exigence mieux que les anciennes machines à écrire. Mais si l'alphabet français contient 26 lettres, accentuées ou non, toutes n'ont pas le même statut et la même légitimité historique. Et cela est plus vrai encore de certaines combinaisons de lettres, qu'on reconnaît tout de suite comme d'importation étrangère. On a là une manière indirecte de dresser la carte démographique des mots, en distinguant les nationalités, et 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.

Les emprunts: l'anglais

Deux graphèmes sont moins bien intégrés au système français et servent au transit des étrangers, dont certains ne sont que de passage: ce sont les noms propres. Ces touristes-là gardent leur passeport d'origine et ne sont pas soumis au recensement. On les trouve pourtant de plus en plus souvent dans les Lettres françaises, et il est probable que l'internationalisation des échanges a le même effet dans les autres langues. Ces toponymes ou anthroponymes ont toujours circulé plus ou moins librement dans la langue nationale, mais la tradition classique accueillait plus volontiers les noms antiques de l'histoire ou de la mythologie. Comme on puisait à la même source gréco-latine que le vocabulaire commun, la francisation allait d'elle-même. Il en va autrement maintenant, les noms propres étrangers s'habillent de façon plus étrange, en multipliant les *k* et les *w*. C'est aussi le cas des noms communs qui franchissent les frontières. On ne s'intéressera qu'aux frontières anglo-saxonnes puisque c'est là que se produisent surtout les mouvements de population et qu'en reniflant les *k* et les *w* on a toute chance de repérer les intrus qui viennent de ce côté. On a lancé l'ordinateur sur la piste, en lui ordonnant de rapporter tous les mots où l'on trouvait un *k* (il y en a 72252) ou un *w* (41607 occurrences). Puis dans un second temps on n'a gardé que les formes sans majuscule (respectivement 23780 et 9071 occurrences). Le filtrage montre la part prépondérante des noms propres dans la catégorie: 2/3 pour le *k*, 3/4 pour le *w* - ce qui laisse entendre que la tentative de naturalisation est le fait d'une minorité. Mais, même timide, ce mouvement est croissant, comme le montre le tableau ci-dessous (et sa représentation graphique dans la figure 13).

lettres k et w	1550	1630	1692	1735	1780	1820	1855	1885	1910	1928	1942	1960
k (72252)	-29,7	-70,5	-46,7	-43,4	-28,9	10,2	6,1	-9,6	13,8	36,5	35,2	99,6
k2 (23780) sans N.propres	-15,5	-41,8	-33,5	-40,6	-33,4	-28,4	-25,1	-4,6	1,5	15,7	30,4	174,6
w (41607)	-21,6	-55,0	-41,0	-21,5	-39,3	42,6	19,8	-23,2	22,0	35,8	18,3	38,6
w2 (9071) sans N.propres	-8,5	-26,1	-21,5	-26,7	-28,4	-16,8	-16,5	-7,2	8,6	43,1	31,0	66,5

Figure 13. Importation croissante des mots empruntés
(à partir du critère des lettres k et w)



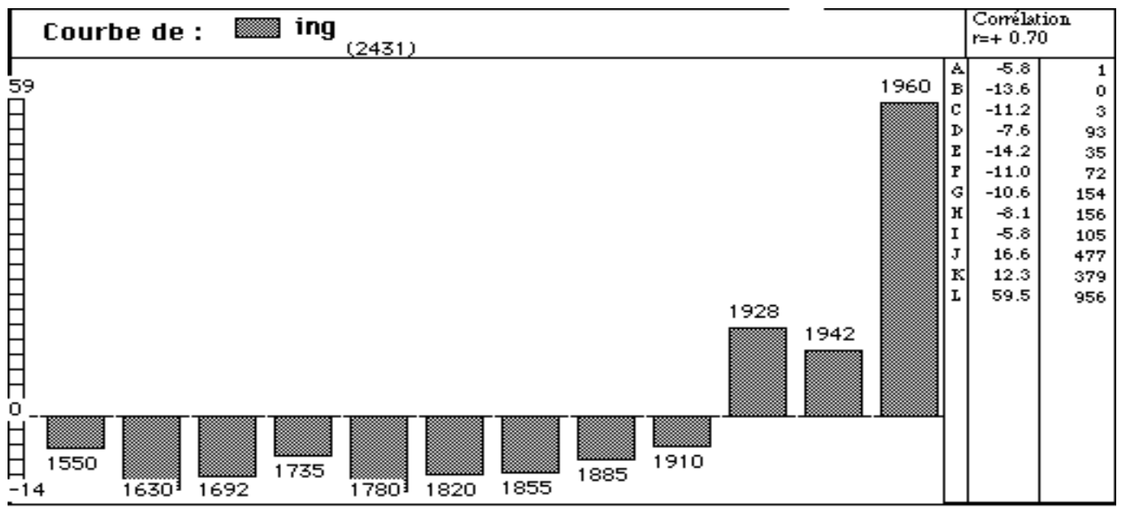
On peut craindre cependant que le filtrage soit trop grossier pour isoler les anglicismes. Le *w* n'est pas propre à l'anglais et les langues germaniques l'utilisent pareillement. Quant au *k*, ce peut-être l'indice d'un mot venu du Japon (*kaki*), de Chine (*kaolin*), de Russie (*knout*), d'Allemagne (*képi*) ou du monde arabe (*khan, khôl*) ou, plus simplement encore, du monde grec (*kilo*). Il faut donc recourir à une contrainte supérieure, qui exige certains assemblages de lettres, et par exemple la finale en *-ing*. Certes là aussi des sonorités chinoises peuvent occasionnellement se mêler à l'anglais ou plus souvent les graphies anciennes du type *besoing, gaing,* ou *loing*, mais la pureté y est mieux garantie, une fois dégagée, manuellement, la gangue des scories[9]. La progression de cette catégorie lexicale est en tout cas fort nette, sans qu'on puisse démêler s'il s'agit d'importations clandestines ou d'emprunts régularisés. On a tout lieu de croire que le flux croissant des entrées grossit tout à la fois l'égout des rejets et le bassin d'épandage qui recycle celles qui sont acceptées pour l'usage domestique.

Mais si les naturalisations vont en augmentant, on est loin d'un raz-de-marée. L'effectif obtenu (rappelons qu'il s'agit d'occurrences et non de formes différentes) est d'une modicité rassurante. Les 2400 exemples relevés pèsent peu dans une masse de 120 millions de mots. Il est vrai que le corpus est largement constitué d'oeuvres des siècles passés et qu'au surplus le goût des écrivains est plus sévère que celui des médias, lorsqu'il s'agit de préserver la pureté de la langue. Il est vrai aussi que la mode de l'anglais - qui envahit la rue, la publicité et les formes modernes de la communication - n'offre pas en elle-même des gages de pérennité. Les modes - il y en a d'autres, comme celle de l'argot ou du verlan - durent parfois le temps d'une vague. La langue se soulève à leur passage, un instant désorientée, pour revenir, apaisée, à son état initial.

---

[9] En voici quelques-unes, parmi une trentaine de formes: *besoing* (170 ocurrences), *coing* (106), *loing* (819), *poing* (2918), *seing* (100), *soing* (583), *tesmoing* (181). Sans ce nettoyage, un seul mot comme *poing* aurait à lui seul compté autant que le reste de la série.

Figure 14. La progression des formes anglaises en *-ing*



Afin qu'on se fasse une idée précise de ces emprunts, on a établi un extrait de la liste en se bornant à la fin de l'alphabet.

Tableau 15. Extrait de la liste des formes anglaises en *-ing*

sailing	4	singing	2	string	6	uncomforting	1
saying	2	sitting	17	stripling	1	underlying	1
schampoing	1	skating	7	summing	1	understanding	1
schampooing	3	sleeping	22	surprising	1	undoing	2
scheling	1	sling	1	sweating	1	unlearning	1
schmeling	1	slumming	2	swimming	1	ushing	1
scorning	1	smiling	2	swing	31	using	1
screeching	3	smilling	1	swinging	2	vling	1
screening	1	smoking	116	taking	2	vorfrühling	1
seeking	2	smushing	1	talking	2	vrring	1
seeming	1	sniviling	1	tating	1	vrrring	1
setting	1	something	11	teaching	2	walking	4
settling	1	somking	1	teasing	1	wallowing	1
shampoing	1	sounding	1	thanksgiving	1	wanting	2
shampooing	24	sparling	1	thing	22	watching	1
shaving	3	sparring	1	thinking	3	waterleiding	2
shelling	5	spelling	1	tooting	1	waving	2
shilling	8	spinning	1	touching	1	weaving	1
shipping	5	sponsoring	1	touring	5	weighing	2
shirking	2	sporting	10	towering	1	whirling	1
shirting	1	spring	36	training	2	willing	1
shocking	4	staging	1	translating	1	winning	1
shoking	2	standing	31	travelling	6	withdrawing	1
shooting	1	starling	1	trembling	2	working	5
shoping	1	starring	1	trespassing	1	wrestling	1
shopping	5	starting	6	tricoting	1	writing	1
shouting	2	stealing	1	tring	2	wuthering	7
signifying	2	sterling	143	trucking	1	yachting	4
silencing	1	stocking	1	trying	1	yearling	1
sing	11	stopping	2	turning	1		

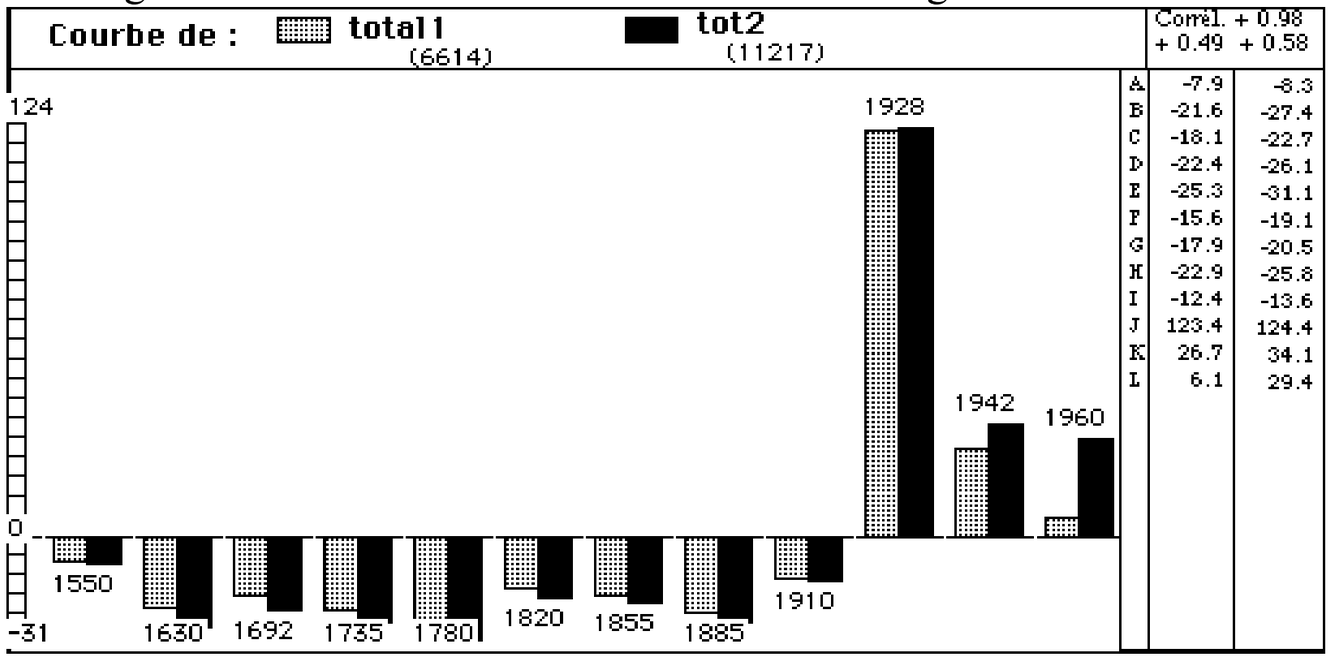
La question tant débattue du franglais impose encore une autre mesure. Il ne s'agit plus cette fois de démasquer les produits d'importation qui s'introduisent en ignorant la douane. Il s'agit des mots qui doivent se lire en anglais dans un environnement français et jouissent d'une sorte d'exterritorialité. Même le plus ignorant des français les reconnaît d'emblée, dès qu'ils apparaissent dans une expression ou une citation. On en a choisi une vingtaine, parmi les plus fréquents, en éliminant ceux qui avaient un homographe autochtone (*on, are, an, but, for*). En voici la liste.

Tableau 16. Quelques mots grammaticaux de la langue anglaise

its	19	the	1261	when	36		
am	111	make	18	their	17	where	33
and	564	not	176	then	13	which	64
be	191	of	904	there	56	who	68
been	18	off	36	they	50	why	23
by	90	one	108	this	77	with	132
does	11	other	17	to	607	you	285
had	35	our	56	was	105	your	45
has	57	she	42	we	47	TOTAL	6614
is	394	so	163	were	16		
it	259	that	331	what	79		

Comme on s'y attendait, ces menus outils anglais n'apparaissent guère avant le 20e siècle. Une faveur soudaine leur sourit entre les deux guerres, qui ne se maintient pas au même niveau dans les décennies qui suivent. Comme cette décrue relative pouvait surprendre ceux qui croient à l'invasion irrésistible de l'anglais, on a cru devoir la confirmer en renouvelant l'expérience sur une base plus large. Ici un détour est nécessaire qui fait appel à une fonction très puissante de *Frantext*. *Frantext* est en effet capable de trouver non seulement la fréquence d'un mot ou d'une liste de mots, mais aussi celle de tous les mots qui entourent le mot choisi (ou les mots de la liste choisie). La liste pure mais restreinte des mots qui précèdent peut servir à repérer l'environnement habituel qui les accompagne dans un discours français et qui donne une forte probabilité aux termes anglais. En somme ils servent d'appât ou d'appeau pour attirer les autres. Un test statistique permet ensuite de trier les accointances les plus fortes - qui bien évidemment s'exercent selon la nationalité. Les homographes, gênés par leur ambiguïté, sont tenus à l'écart et n'apparaissent pas en tête de liste, où ne figurent que les éléments purs (soit 150 au total). L'effectif des observations double alors (11217 contre 6614), mais sans apporter de grands changements aux conclusions précédentes - ce qu'on peut constater dans la figure 17 qui juxtapose les deux expériences.

Figure 17. Courbe de la distribution des mots anglais



Le latin

La même procédure peut s'appliquer au latin. Mais le latin imprègne plus intimement le français et sa mesure réclame des précautions supplémentaires. On peut certes se livrer à une enquête sur les mots qui sont passés d'une langue à l'autre, sans le moindre changement, au moins graphique. On peut ainsi considérer comme des emprunts des termes comme *album, alibi, alias, accessit, abdomen, aquarium, atlas* ou *argus*, mais le transfert est si ancien que le souvenir s'en est perdu dans la conscience des usagers. De tels mots sont en progrès au cours de cinq siècles, au moins dans

l'échantillon de la lettre a. Mais il serait abusif d'attribuer au latin cette progression. En réalité ces termes sont parfaitement intégrés au français et s'ils ont eu un démarrage lent du fait de leur bizarrerie orthographique, le handicap initial a disparu depuis longtemps.

Une autre mesure, partielle, a consisté à puiser un certain nombre de citations latines dans les pages roses du *Petit Larousse* et d'en relever les occurrences dans *Frantext*. Là aussi on constate une progression. Mais si cela peut donner des indications sur la diffusion du *Petit Larousse*, le fondement de l'enquête, trop restrictif, ne permet guère de conclusions sur le latin. Un relevé d'un millier de citations latines ne donne peut-être pas la mesure exacte de la présence du latin dans le français. Car les mots latins, même insérés dans un discours français, gardent leur liberté d'association et beaucoup échappent aux expressions figées dont rend compte le *Petit Larrousse.* On peut même soupçonner quelque artefact dans le progrès constaté et craindre que le relevé des pages roses ne soit qu'un florilège des tournures sauvées de l'oubli, dont le progrès illusoire cacherait la déroute de l'ensemble?

Il est donc nécessaire de partir d'un dictionnaire latin, et non d'un dictionnaire français. Va-t-on alors repérer dans la littérature française chacune des entrées qu'on trouve, par exemple, dans le *Gaffiot* ? La tâche paraît ardue, d'autant qu'une langue à déclinaison comme le latin multiplie les formes d'une même entrée. La difficulté majeure cependant, si l'on entreprend de faire l'intersection des deux langues, sera l'obligation de faire le tri de la partie commune et d'y expurger tous les homonymes qui ont un sens en français et un autre en latin sans rien qui soit réellement commun. I1 nous a paru plus habile de faire une première liste, limitée mais exempte de toute ambiguïté, et d'en rechercher les éléments dans la littérature française. Cette liste a été empruntée à un manuel d'initiation au latin et comprend 200 mots latins jugés essentiels. Encore a-t-il fallu éliminer quelques unités qui comme *ira, causa, arma* entraient en concurrence avec des formes françaises.

Tableau 18. Relevé des mots latins dans le corpus littéraire du TLF

(fréquence supérieure à 10)

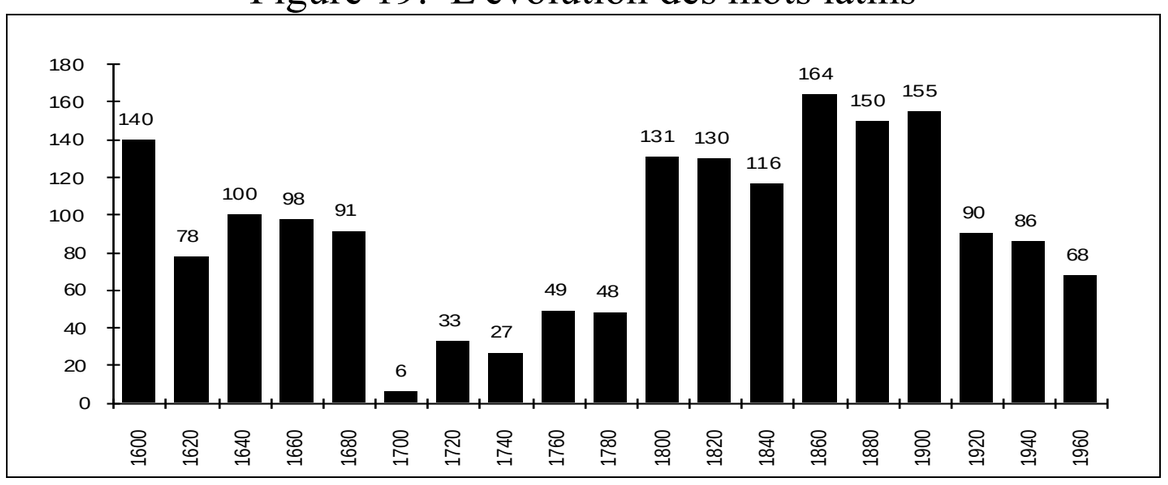
ac	70	bonum	31	dixit	17	fac	112	hanc	13
adsum	23	bonus	25	doctor	25	facere	21	hic	160
aeternam	19	cadaver	13	dolorosa	22	faciem	14	hinc	14
ager	11	caput	29	domini	71	facio	17	his	101
alta	80	certe	100	dominum	32	facit	21	hoc	159
alter	33	christi	43	dominus	101	factum	67	hodie	23
amat	17	christo	22	domum	17	feci	11	homine	12
amo	17	christum	13	domus	19	fecit	32	hominem	33
amor	68	christus	26	dulcis	17	fiat	87	homines	14
amore	31	civitas	8	dum	32	fides	14	homini	16
animam	21	contra	48	dux	14	fieri	20	hominibus	13
animum	12	corpora	9	ea	25	filii	33	hominis	21
animus	11	corpore	19	eam	21	filius	54	hominum	12
anni	12	cui	97	ecce	80	fortuna	21	hora	30
ante	67	cujus	15	ego	156	frater	47	huic	14
apud	17	cum	157	ei	39	fratres	24	hujus	13
aqua	37	cur	26	ejus	68	fructus	19	ibi	14
ars	97	dedit	17	eo	25	fuerit	10	igitur	14
atque	36	dei	186	eorum	11	genuit	12	ignis	16
aut	69	deus	224	erat	40	genus	19	illa	22
autem	58	dicere	21	ergo	74	gloriam	22	ille	61
beati	21	dico	22	erit	13	gratia	18	illi	14
beatus	13	diebus	12	esse	91	habens	16	illis	11
bellum	17	diem	19	estis	12	habet	26	imo	101
bene	93	dies	119	etiam	26	habitat	35	imperium	11
bona	25	dignus	36	eum	30	hac	21	injuria	18
bono	32	dimittis	16	excelsis	23	haec	40	inter	77

interpretes	72	modus	23	pax	65	servum	11	ubi	50
ipse	35	mori	30	perdere	12	seu	52	ultima	37
ipso	35	mortis	15	plena	17	sibi	16	umbra	12
ipsum	21	mulier	14	populi	23	sicut	72	una	65
irae	71	multa	14	populus	11	simul	15	unde	31
iste	16	mundi	34	potest	18	sine	93	uno	21
ita	15	mundo	23	propter	31	sinite	11	unum	34
iterum	15	mundus	15	puer	56	sint	15	unus	28
jacet	16	nam	20	pulvis	10	sinu	11	usque	31
jam	42	natura	25	qua	78	sit	47	utinam	15
jesu	53	natus	14	quae	81	solum	12	valete	12
jesum	13	nec	111	qualis	22	solus	16	valle	57
juvat	17	neque	25	quam	81	speculum	20	vel	24
lege	12	nescio	18	quantum	15	spes	30	velut	11
leo	23	nihil	57	quas	12	spiritu	24	veneris	19
lex	15	nisi	37	quem	39	spiritus	70	veniat	13
liber	16	nobis	100	quibus	27	stabat	32	venit	16
loco	44	nocte	14	quidem	13	sua	66	ventris	14
locum	11	nomen	16	quidquid	13	suam	23	verbum	14
locus	15	noster	62	quis	68	sub	81	vere	29
longa	10	nostra	42	quo	133	sui	93	veritas	16
lux	52	nostro	23	quod	147	sum	92	vero	30
magis	29	nostrum	20	quoniam	14	sumus	18	verum	13
magna	38	nova	50	quoque	20	sunt	123	vestris	26
magno	26	nox	14	ratio	20	suo	26	viam	15
magnum	38	num	14	regina	53	tamen	24	vir	43
magnus	146	nunc	93	regnum	21	tantum	33	virtus	15
mala	55	nunquam	11	rerum	62	tecum	18	virum	18
manus	41	oculus	14	res	78	tempore	23	vita	117
mater	208	odi	14	rex	45	tempus	13	vitae	37
mea	113	omnes	55	romani	11	terra	60	vitam	18
meae	16	omni	30	romanus	16	terram	21	vivere	20
meam	37	omnia	77	salus	22	tibi	98	vivit	11
mecum	22	omnis	34	sancti	20	tota	10	vixit	13
mei	39	omnium	30	sancto	22	totus	15	vobis	48
mendient	31	opus	34	satis	22	tuam	22	vobiscum	29
mentis	37	oratio	11	scientia	10	tuas	13	voce	30
meo	19	orbis	15	scilicet	8	tui	24	volo	17
meum	57	ordo	19	scio	25	tulit	13	voluntas	23
meus	37	ossa	62	secundum	14	tum	11	vult	19
mihi	88	pacem	16	sed	100	tunc	13	----------	
minus	56	patri	20	sede	10	tuo	21	Total	12934
miser	21	patria	51	semper	65	tuum	48		
modo	52	patris	27	sensu	12	tuus	83		

Nous ne montrerons pas le résultat de cette première moisson, parce que les grains récoltés ont servi de semence pour la deuxième. Et c'est la deuxième récolte que nous ferons fructifier. Pour l'obtenir on a soumis la première sélection à la procédure utilisée plus haut pour les mots anglais. On constate alors que les formes qui se rapprochent le plus souvent des formes témoins sont d'autres formes latines non initialement citées. Le latin attire le latin, comme l'anglais l'anglais. Or l'attirance, mesurée par l'écart réduit, est plus faible lorsqu'on a affaire à un homonyme, ce qui permet d'écarter de la liste définitive des éléments douteux comme *domino, cor, visa, salis, docte, malo, suave, dicta, fuit, jus, mens.*

La liste qui résulte de ce traitement ultime est reproduite dans le tableau 18, où n'ont pris place que les formes dont la fréquence est supérieure à 10. L'ensemble totalise plus de 12 934 occurrences. Comme aucun soupçon de subjectivité ne peut fausser la perspective, on peut se fier à la courbe chronologique sans craindre les effets de trompe-l'oeil.

Figure 19. L'évolution des mots latins



Le graphique 19 rend compte de cette évolution. Les tranches sont ici de vingt ans uniformément et l'histogramme est fondé sur les fréquences relatives (pondérées par l'étendue des tranches)[10]. On y distingue quatre paliers. Au 16e siècle, le latin est en faveur et le palier est haut. Le latin est encore le recours dès qu'une citation, un témoignage ou une garantie doivent accompagner le discours, et cette habitude est très visible chez Rabelais et Montaigne. Ce recours au latin se raréfie progressivement au siècle classique. Il est vrai que le théâtre y tient une place prépondérante et que le latin ne trouve guère lieu de s'employer sur les planches. L'étiage est au plus bas au XVIIIe, mais la source se ravive au XIXe pour atteindre son plein débit à la fin du siècle, et donner de nouveau des signes d'épuisement à l'époque contemporaine. On s'éloigne un peu des conclusions qu'avait suggérées l'étude des mots latins francisés et celle des expressions consacrées. L'avenir du latin dans les lettres françaises n'apparaît plus voué à une progression linéaire, non plus qu'à une chute inéluctable. Ballotté par les vagues de l'histoire, le latin semble soumis à un mouvement cyclique dont la phase périodique s'étend sur un siècle entier.

La longueur du mot

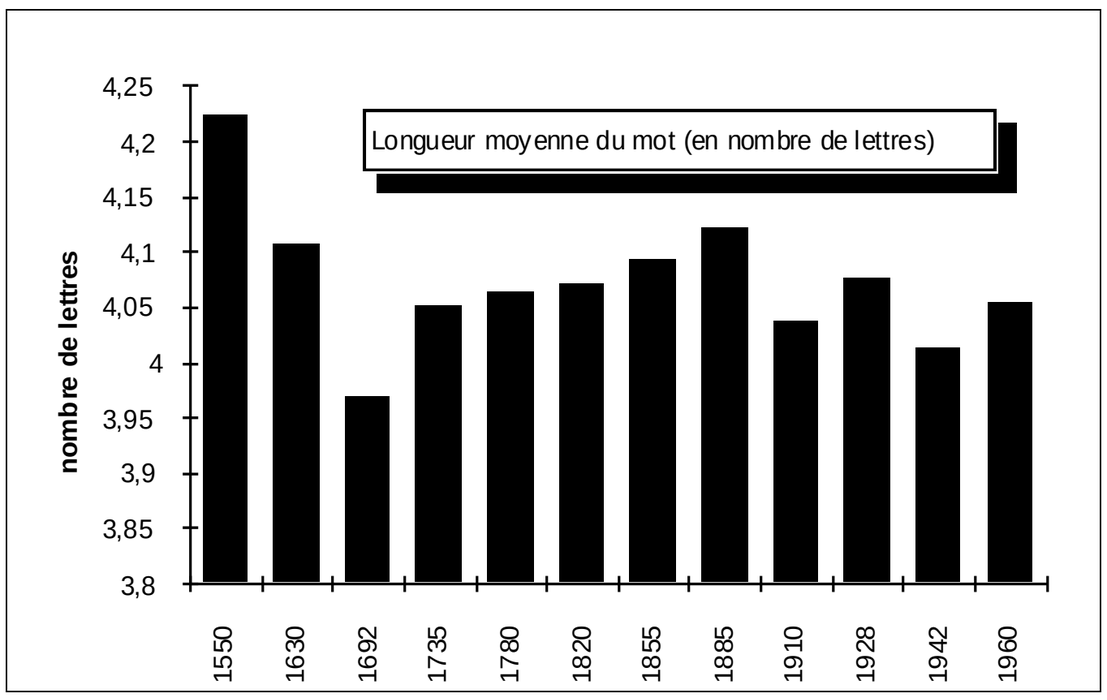
Il n'est guère possible d'aller plus avant dans l'exploration de l'origine des mots, sauf si l'on dispose d'un dictionnaire pourvu d'une telle information. Le *TLF* est doté d'une rubrique étymologique et ouvrira la voie à des recherches synthétiques de cette espèce, dès qu'il sera disponible sur le réseau (le tome 14 du *TLF* peut d'ores et déjà être consulté à travers *Internet*, à l'adresse électronique de *Frantext*). Mais un dictionnaire de fréquences, même réduit à sa plus simple expression (où seules subsistent la forme et la fréquence), offre encore des ressources à l'exploitation. La longueur du mot est un des attributs qu'on peut le plus facilement appréhender. On sait depuis longtemps qu'il faut en moyenne un peu plus de quatre lettres pour un mot français (et qu'il en faut moins pour l'anglais et plus pour l'allemand). En réalité ce paramètre ne suffit pas à mesurer l'efficacité d'une langue. Tout dépend du statut qu'on donne aux mots composés et du rôle que jouent le blanc, le trait d'union ou le collage pur et simple et plus généralement de la façon, analytique ou synthétique, dont on fait intervenir les éléments pré- ou postposés, les prépositions, les cas, les modalités, etc. Mais si ce critère est incertain pour la comparaison entre langues différentes, il garde une certaine valeur à l'intérieur d'une même langue. Le graphique 20 montre que le mot est le plus long au 16e siècle, et le plus court au 17e. Il reprend progressivement du volume au 18e, puis au 19e siècle, pour revenir à une valeur moyenne au 20e. On peut interpréter la chute brutale du début de la chronologie par deux sortes de raisons: d'une part, le mot ne s'est pas extirpé totalement de sa gangue au 16 siècle et sur son corps traînent encore des morceaux de la

---

[10] C'est sous cette forme que *Frantext* distribue l'information.

coquille latine dont beaucoup viennent de s'extraire. La mode est d'ailleurs d'en laisser quelque trace, même sans utilité pour la prononciation, comme on fait d'une griffe sur un produit de grande marque. Malherbe, Vaugelas et l'Académie mirent un peu d'ordre et de sobriété dans le régime, et le mot, soumis à la diète, en sortit purifié et amaigri (le dictionnaire aussi). Mais la maigreur relative du mot à l'époque classique a peut-être une explication plus banale qui tient à la composition du corpus. Le théâtre est mieux représenté en ce temps-là, et les mots sont généralement plus courts, plus simples et plus communs au théâtre, parce que le spectateur n'a pas le temps de débrouiller les expressions compliquées et le loisir de consulter un dictionnaire.

Figure 20. La longueur moyenne du mot



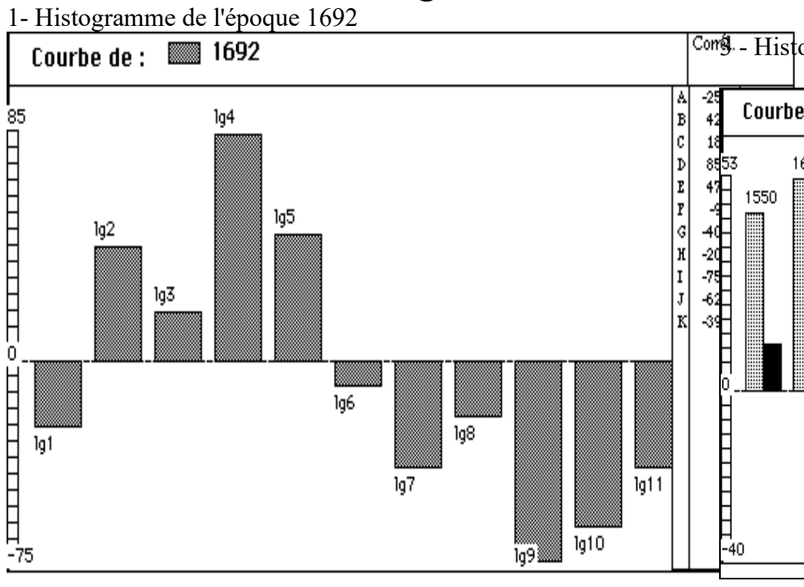
Ce choix, conscient ou non, du mot bref à la fin du 17e siècle apparaît plus clairement quand on répartit les mots en classes de longueur. On a ainsi un tableau à deux dimensions[11], dont chaque cellule inscrit l'effectif des mots qui ont *i* lettres dans la tranche *j*. Si on isole la 3e tranche, qui correspond à l'âge d'or du classicisme, on obtient la courbe ci-dessous (histogramme 1 de la figure 21), où les mots courts montrent un net excédent. Le fait est confirmé par l'histogramme 2 qui isole des lignes et non plus des colonnes et qui reproduit la distribution des classes de 3 ou 4 lettres, où 17e et 18e siècles sont en faveur. Inversement l'histogramme 4, consacré aux mots longs, donne l'avantage au 19e et 20e siècles, tandis que la courbe 3, qui rend compte des mots de longueur moyenne, montre le cours sinueux d'une classe de transition: faveur au 16e et au 18e siècles, discrédit au 17e (parce que déjà trop longs) et au 20e (parce que déjà trop courts).

Comment se fait-il donc que la longueur moyenne du mot n'ait pas dans la figure 20 le profil pur et progressif que les distributions de détail

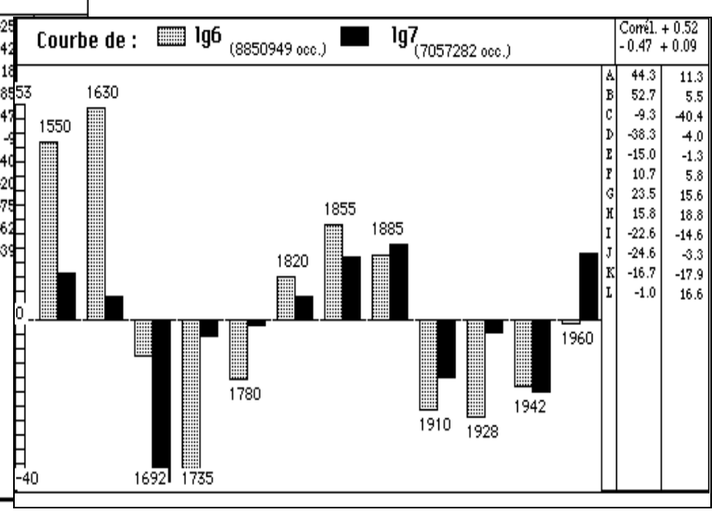
pouvaient laisser espérer? La perturbation vient des mots très courts, de 1 ou 2 lettres, qui sont des nécessités du discours et que la dernière période cultive autant et plus que les autres. Il aurait sans doute été prudent de les exclure du calcul, car leur statut d'outil grammatical leur permet d'échapper au choix, parfois conscient, qui porte sur la longueur. On peut rechercher ou fuir les mots de quatre syllabes (Jean Paulhan voulait les proscrire). Jamais personne ne s'est soucié de rejeter ou de privilégier les mots de 2 ou 3 lettres.

Figure 21. Les classes de longueur

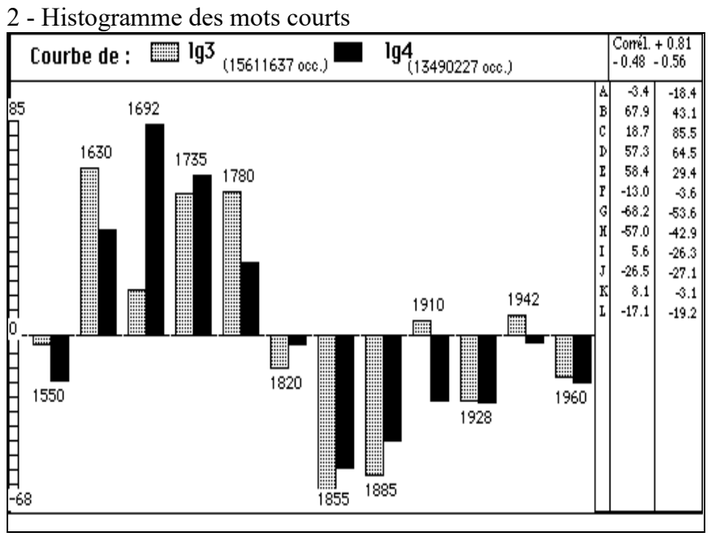
1- Histogramme de l'époque 1692



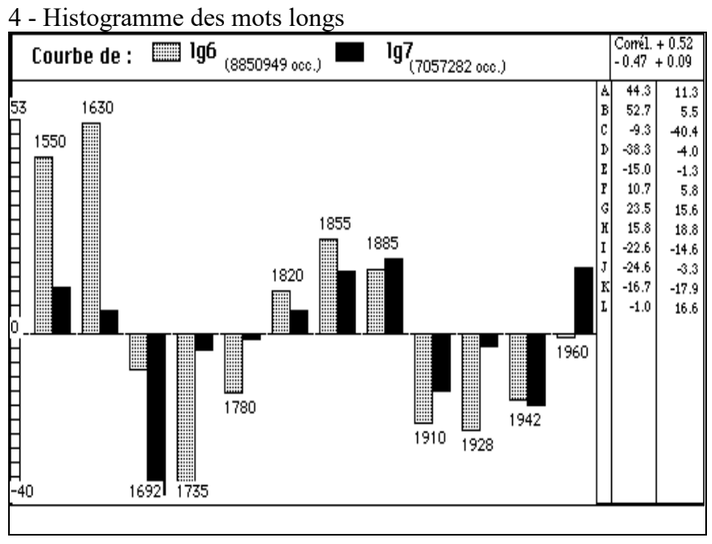
3 - Histogramme des classes moyennes



2 - Histogramme des mots courts



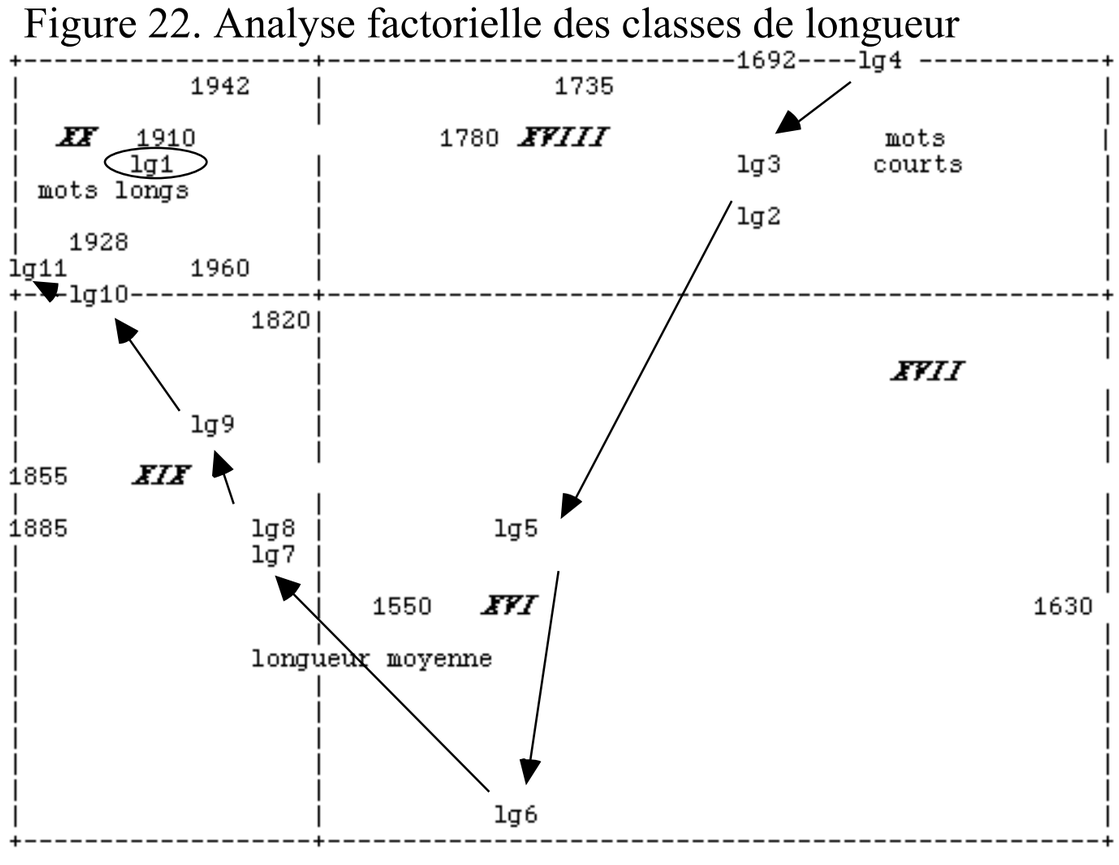
4 - Histogramme des mots longs



Cette perturbation est dénoncée par l'analyse factorielle représentée ci-dessous. Certes les tranches sont bien classées et les classes bien tranchées - les mots courts auprès du XVII, les mots longs auprès du XX, avec des positions intermédiaires pour le XVIII et le XIX. Mais une anomalie notoire se révèle dans la position de la classe 1 (mots d'une seule lettre) qui se porte à l'opposé des mots courts et rejoint le camp adverse.

[11] Afin d'équilibrer les effectifs on a regroupé certaines classes moins bien représentées: la classe 9 cumule les mots de 9 et 10 lettres, la classe 10 ceux de 11, 12 ou 13 lettres, la classe 11 les mots de 14 lettres ou plus.

Figure 22. Analyse factorielle des classes de longueur



Les classes de fréquence

L'expérience de la longueur du mot vient de nous alerter sur la fragilité des mesures trop réductrices qui rabotent les faits et nivellent les écarts pour aboutir à une moyenne sans grande signification. Quel sens aurait une mesure comme le poids moyen ou la taille moyenne des humains si l'on ne distinguait ni les âges, ni les sexes, ni les pays? Notre première approche de la structure lexicale peut être précisée et ne pas se contenter d'une mesure globale. Pour apprécier l'étendue du vocabulaire, on peut procéder par classes, comme on vient de le faire pour la longueur des mots, ou comme ferait un sociologue traitant des classes d'âge ou des catégories socio-professionnelles. La classe (f1) des hapax (ou mots de fréquence 1) a déjà été isolée. Isolons aussi les mots relativement rares qui ont une fréquence comprise entre 2 et 16 (classe f2), puis ceux qui se situent entre 17 et 64 (f3), 65 et 512 (f4), 513 et 1024 (f5), 1025 et 2048 (f6), 2049 et 8192 (f7), 8193 et 65536 (f8) et enfin au2delà de 65536 (classe f9). On obtient un tableau à deux dimensions qui, comme le précédent, a 12 colonnes correspondant aux 12 tranches. Il comporte cette fois 9 lignes où prennent place les 9 classes de fréquence qu'on vient de distinguer. On peut représenter la courbe de ces 12 tranches en étudiant dans chacune le dosage des classes de fréquence, ou bien, adoptant le point de vue inverse, on peut représenter l'évolution de chacune des classes, à travers la chronologie. En réalité, il s'agit d'un continuum: les classes voisines diffèrent peu et l'on passe insensiblement de

l'une à l'autre[11]. C'est ce qu'on voit dans les basses fréquences qu'on a représentées dans la figure 3 et où s'illustrent les premières tranches et la dernière. Inversement les tranches intermédiaires (figure 24) s'écartent des basses fréquences et préfèrent le haut de la gamme des fréquences.

Figure 23. Les basses fréquences

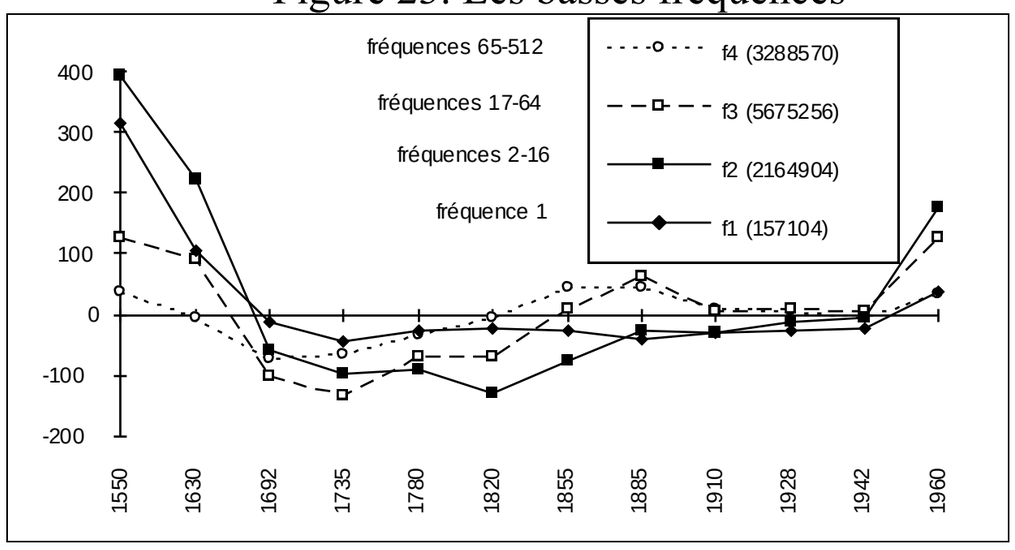
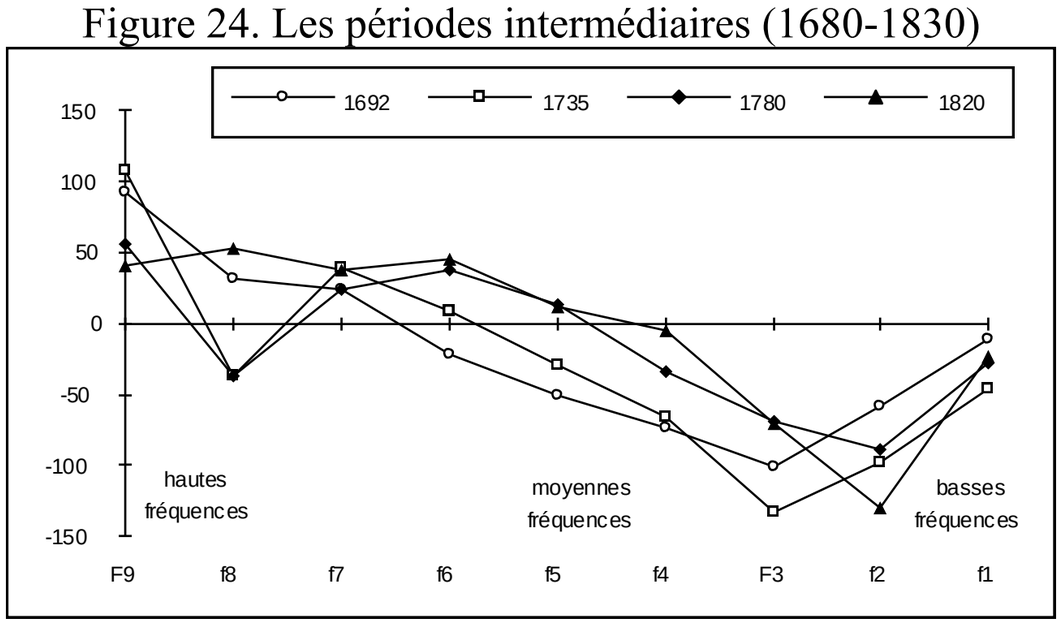


Figure 24. Les périodes intermédiaires (1680-1830)



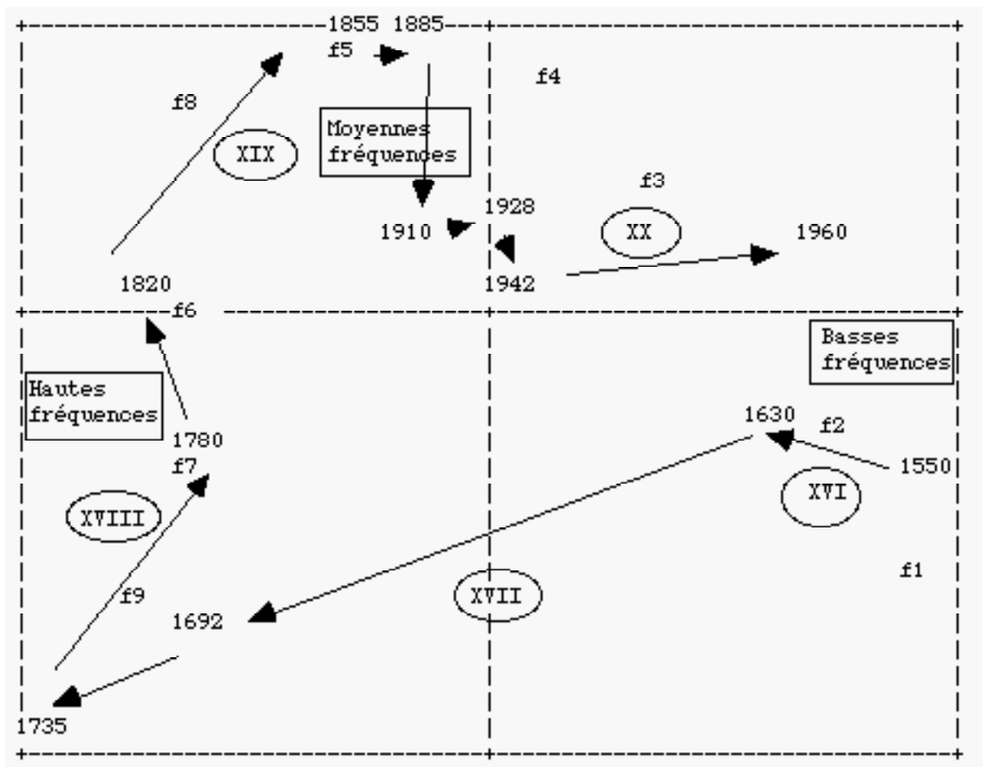
Les conclusions sont plus synthétiques dans l'analyse factorielle qui reprend toutes les données du tableau et les résume dans le même schéma. Les classes de fréquences font la ronde autour du centre, dans l'ordre attendu: f1, f2, f3, etc. Les tranches font de même et l'ordre chronologique est en gros respecté.? Le plus intéressant est la fusion des deux séries et l'entrelacement des deux rondes. Car le principe mathématique est ici celui des rondes enfantines bien ordonnées: chacun embrasse sa chacune. Les hautes fréquences rejoignent le XVII^{e} et le XVIII^{e} siècles, les moyennes se tournent vers le XIX^{e} , et les basses fréquences hésitent entre le XVI^{e} et le XX^{e}. En

---

[11] Ce continuum apparaîtrait mieux si l'on disposait d'une place suffisante pour juxtaposer les 9 histogrammes - ce que nous avons fait ailleurs, dans notre *Vocabulaire français de 1789 à nos jours,* Slatkine-Champion, Genève-Paris, 1981. Voir en particulier pp. 240-242.

réalité cette dernière classe est une collectivité factice où les individus (les mots) n'hésiteraient nullement, s'ils étaient libres, entre les deux pôles. Ils sont ensemble parce que les réunit une propriété paradoxalement commune: leur exclusivité.

Figure 26. Analyse factorielle des classes de fréquence



La connexion ou distance lexicale

On n'a envisagé jusqu'ici le matériau lexical que sous un angle assez limitatif, à partir de critères externes, comme la fréquence, le nombre des lettres qui composent un mot ou la nature de ces lettres. Ni les goûts propres de chaque mot, ni son individualité n'étaient pris en compte. Mais on peut faire mieux: au lieu de les concentrer de force dans un camp, parce qu'ils avaient un w et les cheveux roux, on peut leur demander leur avis et procéder à des élections générales où chaque mot est appelé à choisir son camp et son époque de prédilection. Le mot lui-même constitue le bulletin de vote. Quand on le trouve dans une tranche, il contribue à la coloration particulière de cette tranche. Mais il peut choisir deux tranches à la fois (ou davantage) et contribuer par sa présence au rapprochement des deux tranches. Dans le dépouillement final chaque tranche compte non seulement ses partisans, mais aussi les bulletins multiples où elle est associée à d'autres tranches. Au total deux tranches sont proches quand elles partagent plus de mots communs. Pour une paire donnée, ce rapport entre les mots exclusifs et les mots communs constitue la distance entre les deux éléments de la paire considérée. Ainsi au croisement des tranches 1550 et 1630 on compte 32664 mots qui sont dans la première sans être dans la seconde, 75521 qui sont dans la situation inverse et 39554 qui forment la zone commune du vocabulaire. La

distance entre les deux tranches s'obtient en cumulant la part exclusive de chacune, soit pour l'exemple choisi:

$$\frac{\text{exclusivités 1550}}{\text{vocabulaire 1550}} + \frac{\text{exclusivités 1630}}{\text{vocabulaire 1630}} = \frac{32664}{32664+39554} + \frac{75521}{75521+39554} = 1{,}109$$

On trouvera dans le tableau 24 l'ensemble des données qui permettent d'établir la carte électorale des mots au cours de cinq siècles de littérature.

Tableau 24. Distance lexicale des 12 tranches chronologiques

Nombre de formes privatives

	1550	1630	1692	1735	1780	1820	1855	1885	1910	1928	1942	1960
1550	0	32664	46289	47397	46252	48692	48182	48834	50438	50321	50657	51855
1630	75521	0	70289	71263	68986	74369	73253	75325	78260	78150	78600	81183
1692	54123	35266	0	33796	32063	36836	36106	38781	41487	40848	41745	44201
1735	62036	43045	40601	0	29727	36871	36869	39519	42742	42228	42984	45627
1780	87873	67750	65850	56709	0	51207	49266	53736	58817	58370	59443	63230
1820	90515	73335	70825	64055	51409	0	38526	44574	50821	50414	51635	55970
1855	125984	108198	106074	100032	85447	74505	0	63940	74929	73841	75788	80850
1885	111507	95141	93620	87553	74788	65424	48811	0	57984	57370	58857	63977
1910	90728	75693	73943	68393	57486	49288	37417	35601	0	38096	40037	44890
1928	102770	87742	85463	80038	69198	61040	48488	47146	50255	0	47560	52306
1942	97939	83025	81193	75627	65104	57094	45268	43466	47029	42393	0	45797
1960	97181	83652	81693	76314	66935	59473	48374	46630	49926	45183	43841	0

	1550	1630	1692	1735	1780	1820	1855	1885	1910	1928	1942	1960
						(Nombre de formes communes)						
1550		39554	25929	24821	25966	23526	24036	23384	21780	21897	21561	20363
1630	1,109		44786	43812	46089	40706	41822	39750	36815	36925	36475	33892
1692	1,317	1,051		46256	47989	43216	43946	41271	38565	39204	38307	35851
1735	1,371	1,115	0,890		57130	49986	49988	47338	44115	44629	43873	41230
1780	1,412	1,195	0,979	0,840		62632	64573	60103	55022	55469	54396	50609
1820	1,468	1,289	1,081	0,986	0,901		75515	69467	63220	63627	62406	58071
1855	1,507	1,358	1,158	1,091	1,002	0,834		86080	75091	76179	74232	69170
1885	1,503	1,360	1,178	1,104	1,026	0,876	0,788		76907	77521	76034	70914
1910	1,505	1,353	1,175	1,100	1,028	0,884	0,832	0,746		74412	72471	67618
1928	1,521	1,383	1,196	1,128	1,068	0,932	0,881	0,803	0,742		77107	72361
1942	1,521	1,378	1,201	1,128	1,067	0,931	0,884	0,800	0,749	0,736		73703
1960	1,545	1,417	1,247	1,175	1,125	0,997	0,950	0,871	0,824	0,804	0,756	
						(Distance globale des textes deux à deux)						
	1550	1630	1692	1735	1780	1820	1855	1885	1910	1928	1942	1960

Avant de commenter ce tableau et d'en faire apparaître les lignes de force, il convient de remarquer qu'ici seul compte le critère présence/absence, en dehors de toute considération de fréquence. En fait les mots fréquents, qu'on retrouve nécessairement dans toutes les tranches, se trouvent par là même empêchés de manifester leur préférence et n'exercent plus cette domination gênante qu'ils imposent dans d'autres calculs. La tranche 1910, représentée ci-dessous (à droite) correspond exactement à ce qu'on attend d'une évolution progressive et régulière. La proximité du vocabulaire a un lien direct avec la proximité chronologique et la distance lexicale est plus courte avec les tranches voisines. Et cela se constate dans presque toutes les tranches, notamment dans la dernière où la courbe prend la forme d'une diagonale parfaite. Mais on ne trouve pas la diagonale symétrique attendue au début de la chaîne. Au lieu d'une pente régulière, la figure 25, dans sa partie gauche, montre que la première tranche est sans lien avec les autres, sinon, faiblement, la seconde, et qu'un mur abrupt sépare le XVIe siècle de ceux qui suivent. L'instabilité de l'orthographe a déjà été invoquée pour expliquer ce phénomène. D'autres facteurs jouent sans doute aussi qu'il est malaisé de

circonscrire. En tous cas, le temps n'apparaît pas homogène et la figure 26 qui cumule les 12 profils n'est ni symétrique, ni réversible. Les premières tranches, isolées, servent de repoussoir quand les dernières apparaissent comme le lieu de la convergence. L'analyse factorielle appliquée au tableau 24 confirme cette dissymétrie. On y reconnaît le croissant caractéristique des données sérielles et toutes les tranches se suivent le long de la chaîne du temps, sans aucune permutation. Mais la distance entre elles est inégale. À gauche de longs espaces séparent les cinq tranches qui précèdent la Révolution. À droite les intervalles se rétrécissent et les dernières tranches se recouvrent presque. Est-ce là le reflet de la composition du corpus, où les limites temporelles sont inégalement réparties, de plus longs espaces ayant été alloués aux premières tranches parce que la densité des textes y était plus faible? Ou s'agit-il d'un ralentissement du mouvement de la langue, au moins dans son aspect lexical. On a coutume d'évoquer dans d'autres domaines l'accélération de l'histoire. Dans le domaine de la langue, cette accélération ne se fait pas sentir. On constate plutôt des effets de freinage, et les tentatives, anciennes ou récentes, de réforme de l'orthographe ont confirmé la force de l'inertie.

Figure 25. La distance lexicale dans les tranches 1650 et 1910

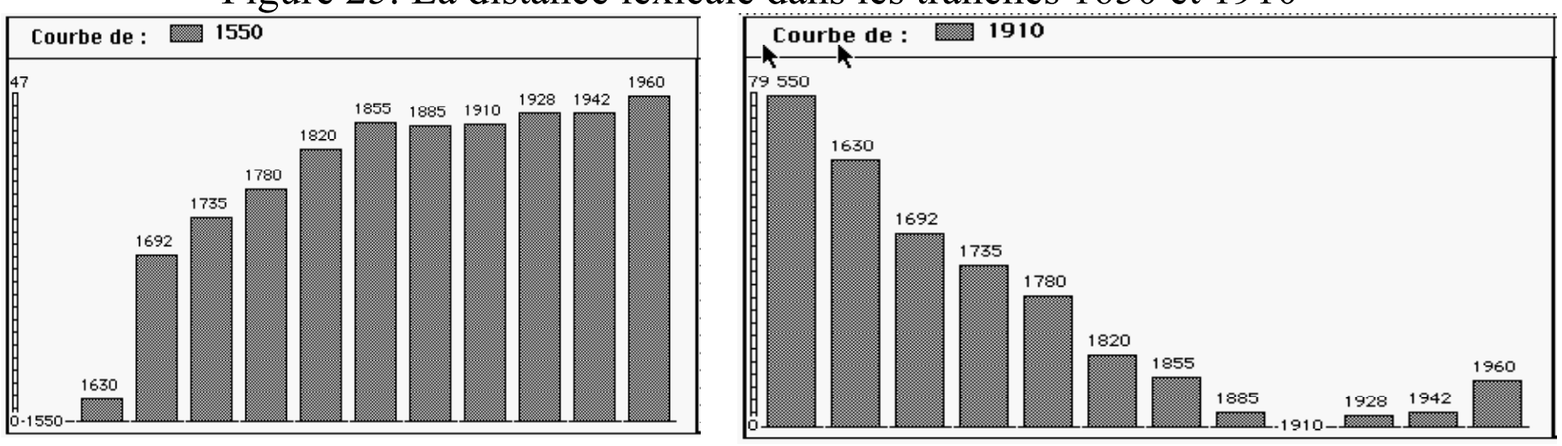
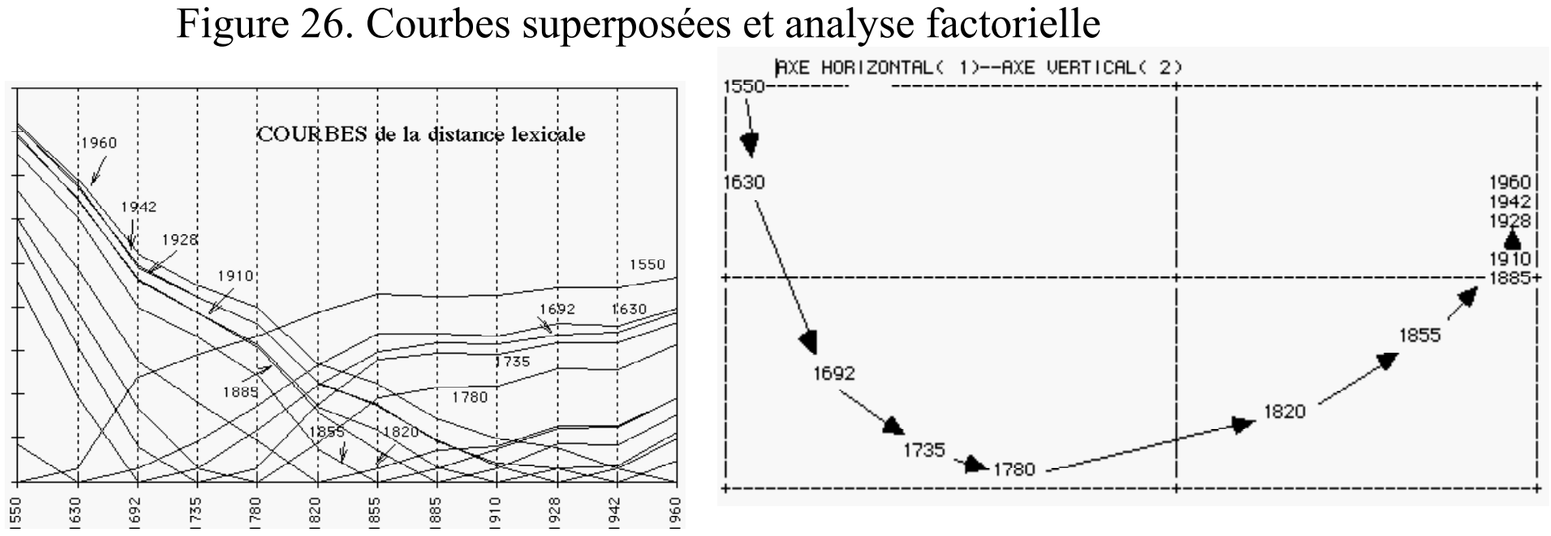


Figure 26. Courbes superposées et analyse factorielle



La suffixation

Le principe *Rien ne se perd, rien ne se crée,* qui a toujours cours en physique, s'applique-t-il aux sciences humaines et à la linguistique? En principe non. La créativité lexicale, comme l'invention artistique, n'a pas à se soumettre à une loi extérieure et chacun est libre de fabriquer des mots arbitraires et opaques et de les dire poétiques et surréalistes. Dans la réalité des faits, il n'y a guère de création *ex nihilo* dans le domaine du lexique. Les contraintes de la communication imposent au locuteur et au destinataire un terrain de rencontre où chacun puisse trouver et partager ses repères. Quand donc il y a nécessité de proposer un nouveau vocable, il suffit généralement d'une nouvelle combinaison des éléments anciens. C'est avec du vieux qu'on fait du neuf, parfois grâce à l'emprunt extérieur ou au recyclage interne, le plus souvent par des produits dérivés. Sur ce point nous renvoyons le lecteur à notre étude antérieure, assez détaillée, qui porte sur 44 variétés de suffixes et 39 de préfixes. 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. C'est évidemment dans le domaine technique que les besoins terminologiques sont les plus impérieux. Mais notre base littéraire n'est pas le meilleur observatoire pour des études de ce genre, même lorsqu'on y fait entrer les essais et les textes techniques - qui représentent un tiers du corpus total. On peut certes y observer que les suffixes *-ate, -ène, -ine, -ite, -ol, -on, -one, -ose* ont à voir avec la chimie et la biologie et que d'autres spécialités disciplinaires se sont arrogé d'autres particularités suffixales. Mais plutôt que les bras du fleuve et les canaux où l'on dévie son cours, c'est le lit principal du français qui nous intéresse ici et où se mêlent les eaux de la langue littéraire et de l'usage courant. Des remous s'y produisent au cours du temps qui, selon les modes, donnent la faveur ou la retirent à certaines formes de dérivation. Il faut toutefois distinguer entre les suffixes proprement dits, dont la créativité peut s'éteindre ou s'enflammer, et les mots suffixés qui peuvent poursuivre une évolution propre, indépendamment du moule dont ils sont issus. Il en est ainsi du suffixe *-tude* dont le modèle est quasi abandonné (on a pourtant *foultitude)* et dont les occurrences sont en progression. Le tableau 26 ci-dessous restitue pour les espèces recensées les deux sortes d'information: nombre de formes différentes (ligne 1) et nombre d'occurrences (ligne 2). On n'a isolé qu'un échantillon restreint mais représentatif de ce vaste pan du vocabulaire. Il n'était pas utile de reprendre dans le détail les études entreprises naguère. Et l'on s'est contenté d'une vingtaine de variétés. Encore s'agit-il des formes et non des vocables: les pluriels non plus que les féminins n'ont été pris en compte. Mais cela suffit pour s'assurer des tendances. Un coefficient de tendance (ou de corrélation chronologique) a précisément été calculé pour chaque espèce, qui permet de distinguer deux lots: les suffixes qui progressent, et ceux que la mode a abandonnés.

Tableau 26. Quelques relevés de suffixes

V/N	*corrél.*	*TOTAL*	*1550*	*1630*	*1692*	*1735*	*1780*	*1820*	*1855*	*1885*	*1910*	*1928*	*1942*	*1960*
able		1271	254	332	247	242	304	306	391	428	363	483	472	419
able	-0,71	161340	2227	14050	9456	15941	18993	15605	19591	18176	10516	14641	13351	8793
age		1377	130	213	170	163	243	244	427	536	384	445	514	576
age	-0,25	199642	2322	18219	9943	16682	22171	17759	25524	21030	13661	19388	18051	14892
al		835	169	156	109	124	177	227	303	276	271	332	319	330
al	+0,76	116429	1226	4827	4473	4967	9283	12711	18272	14162	9528	14647	9811	12522
ance		719	234	321	175	176	190	182	206	189	183	210	201	198
ance	-0,67	157018	2796	14774	7683	16129	19680	17019	18838	14124	10031	14004	12299	9641
el		536	73	100	78	76	115	127	160	140	145	173	183	242
el	-0,14	108649	3111	7503	4310	7382	10245	10864	15253	13241	8350	11627	9162	7601
ence		449	138	179	134	149	158	141	161	160	159	173	183	164
ence	-0,23	192587	2599	13963	9489	16811	21376	20000	24611	20391	13964	19724	17517	12142
esque		186	13	30	10	14	14	27	47	56	31	31	48	60
esque	+0,61	4881	21	273	50	132	273	503	1034	744	479	465	436	471
esse		330	104	130	91	94	106	109	123	115	105	106	96	120
esse	-0,76	128550	2357	9245	8660	14879	16160	12542	16129	15247	8175	10435	8513	6208
eur		2462	500	595	364	455	622	636	842	1011	782	836	852	981
eur	-0,90	418924	8200	3901	725406	39102	48954	45461	52987	44166	26198	33799	31004	24630
eux		1377	387	438	252	275	350	334	422	488	410	439	465	516
eux	-0,61	210220	3552	17493	9403	16752	25442	20879	28661	25241	15511	17468	16176	13642
ible		217	57	74	52	53	77	86	88	93	96	94	107	91
ible	+0,22	84971	955	5463	3041	6618	10430	8740	11381	9985	6002	9438	7675	5243
ien		774	108	140	81	133	153	151	217	199	178	224	216	189
ien	-0,33	56253	1399	3560	2024	4684	6866	5262	8580	6334	4338	5194	4606	3406
if		472	88	110	72	97	144	150	178	179	166	194	199	225
if	+0,76	25866	317	909	427	1416	2725	2853	3874	2953	1898	3384	2477	2633
tion		2297	516	717	566	647	908	839	973	1012	951	1078	1063	992
tion	+0,50	407010	3788	20513	14950	35077	44096	43050	60816	46326	29547	44421	34868	29558
ique		1631	147	371	206	302	381	441	616	607	536	622	648	618
ique	+0,82	155061	683	4447	4194	8841	13766	15693	26732	20623	14361	17300	15257	13164
isme		958	16	46	35	73	119	192	278	368	317	361	363	269
isme	+0,77	23521	24	343	292	466	1532	2526	3864	3498	3158	3290	2840	1688
iste		731	25	69	37	81	117	137	229	258	251	279	273	269
iste	+0,86	25427	126	780	298	876	1875	1761	4347	3445	2989	2874	3379	2677
teur		832	151	205	139	190	252	262	336	388	313	322	325	340
teur	+0,40	69802	1353	3986	2487	4840	7566	6254	10385	8838	5035	7161	6750	5147
tude		94	32	38	32	35	40	36	38	39	39	41	41	44
tude	+0,64	45579	357	1698	1452	3265	5547	4529	6320	5231	3986	5743	4268	3183

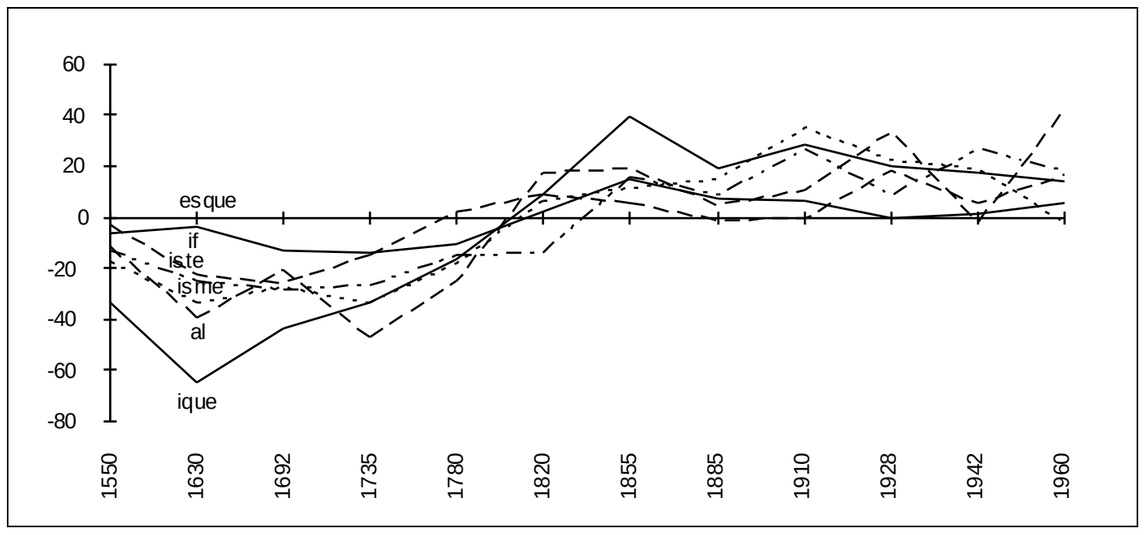
On a réuni quelques-uns des premiers dans le graphique 27. Au premier rang se trouve le couple *isme/iste* dont la pente montante est mesurée par un coefficient très significatif (respectivement +0,77 et +0,86). Cela confirme nos analyses antérieures, réalisées à partir d'une composition différente du corpus où les textes techniques étaient pris en compte[12] et où les jalons temporels avaient été placés différemment. On aurait aimé y joindre le couple *ie/ique*. Mais si l'adjectif est facile à isoler (progression de +0,82), la finale en *ie* est un critère insuffisant pour repérer le substantif correspondant, en dehors de tout code grammatical (on aurait récolté tous les adjectifs ou participes féminin qui partagent la même désinence). Il a fallu pareillement renoncer au suffixe en *-té*, difficile à dégager des participes passés, et au suffixe en *-ment*, trop mêlé aux adverbes (*absolument*), aux adjectifs (*clément*) et aux verbes (*allument*)[13]. Le suffixe *-tion* est le plus souvent

[12] *Histoire de la langue française 1914-1945*, sous la direction de Gérald Antoine et Robert Martin, CNRS Editions, 1995, p. 103.

[13] Précisons qu'une contrainte supplémentaire a été imposée au filtrage, afin d'écarter les intrus qui peuvent partager par hasard une finale propre à un suffixe sans rien partager d'autre: on a exigé que la base du suffixe ait au moins trois lettres, ce qui a permis d'écarter, par exemple, *prisme* ou *séisme* du suffixe en *-isme*; *coeur, leur* et *fleur* du suffixe en *-eur*; *deux, ceux, eux, yeux* de la série en *-eux*; et ailleurs *cesse,*

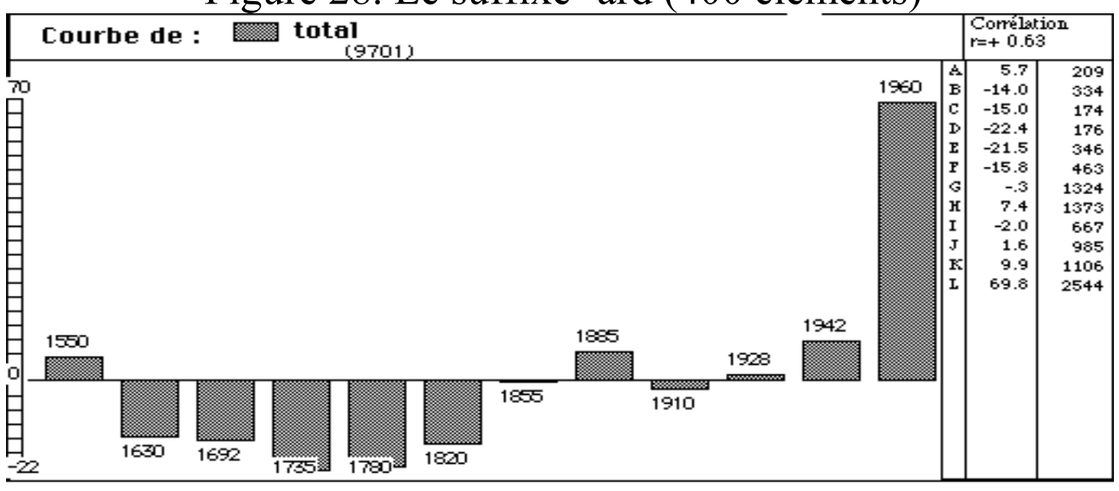
sollicité, tant pour la créativité (2297 variétés) que pour la fréquence dans le discours (407010 occurrences). C'est une valeur sûre dont le cours est régulièrement à la hausse (+0,50). Parmi les adjectifs, on signalera le succès croissant de *-al* (+0,76), de *-if* (+0,76) et de *-esque* (+0,61), ce dernier particulièrement en faveur au XIX^{e} siècle.

Figure 27. Quelques suffixes en progrès



À titre d'illustration, on ne résiste pas au plaisir d'exhiber l'histogramme exemplaire de la série en *-ard* qui a pourtant été difficile à isoler puisqu'il fallait éliminer une vingtaine d'intrus encombrants comme *boulevard, hasard, regard, ou retard*. Quoique ce suffixe soit ancien et nullement ignoré des premières tranches, son succès est récent et semble lié aux connotations péjoratives et à l'accent populaire que partagent la plupart des éléments du paradigme.

Figure 28. Le suffixe -ard (400 éléments)



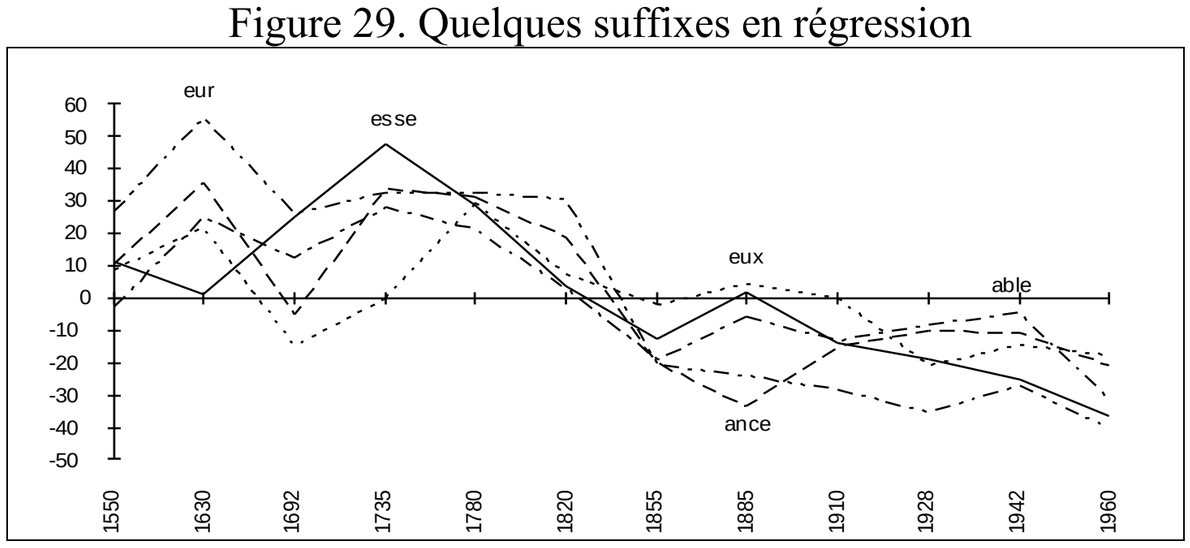
La figure 29 reproduit quelques-uns des suffixes dont le cours est à la baisse. Deux d'entre eux ne sont pas purs: sous les suffixes *-eur* et *-esse* des catégories distinctes s'interpénètrent, soit des notions abstraites (*douceur, tendresse)*, soit des agents, animés ou non (*penseur, moteur, négresse, tigresse*). C'est la première espèce qui domine lorsqu'on considère - comme ici - les occurrences. La seconde espèce - dont la créativité n'est pas émous-

---

messe, presse; avance, chance, lance; diable, sable, table; bible, faible, foible; ciel, duel, miel, quel, réel, sel; tel; bal, égal, mal; juif, soif, suif; evesque, fresque, presque; bien, chien, mien, rien, sien, etc.

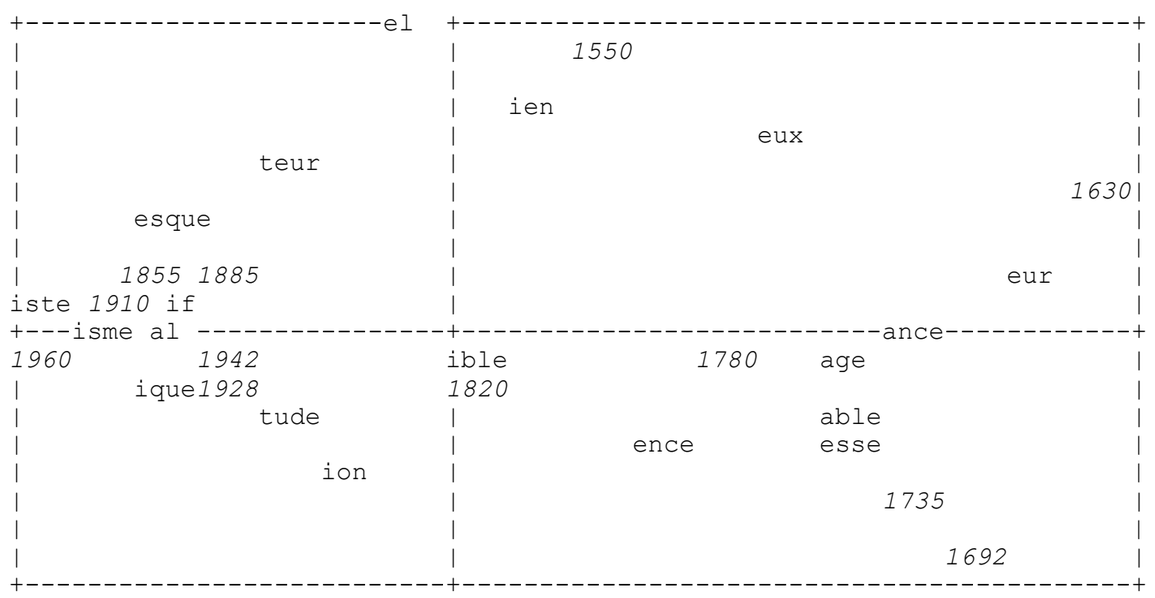
sée - a un profil assez différent si l'on en juge par l'exemple du suffixe *-teur* qui ne comprend que des agents ou des outils et qui, lui, est en progression (+0,40). Les adjectifs suffixés en *-eux*, comme il est naturel, accompagnent dans leur chute les noms abstraits en *-eur*, dont ils sont la doublure (sur le modèle *douleur/douloureux*). L'adjectif verbal en *-able* qui potentiellement peut être créé à partir de la plupart des verbes transitifs a effectivement une puissante fécondité (1271 espèces relevées). Mais le corpus littéraire n'en abuse pas, et tend même à l'utiliser moins (-0,71). Enfin la concurrence entre *-ance* et *-ence* qui a longtemps été profitable au premier lui est de moins en moins favorable (-0,67 pour *-ance* et -0,23 pour *-ence*).

Figure 29. Quelques suffixes en régression



Tous ces faits et mouvements sont résumés dans les analyses factorielles qui suivent. La première (figure 30) est fondée sur le relevé des occurrences (lignes paires du tableau 26).

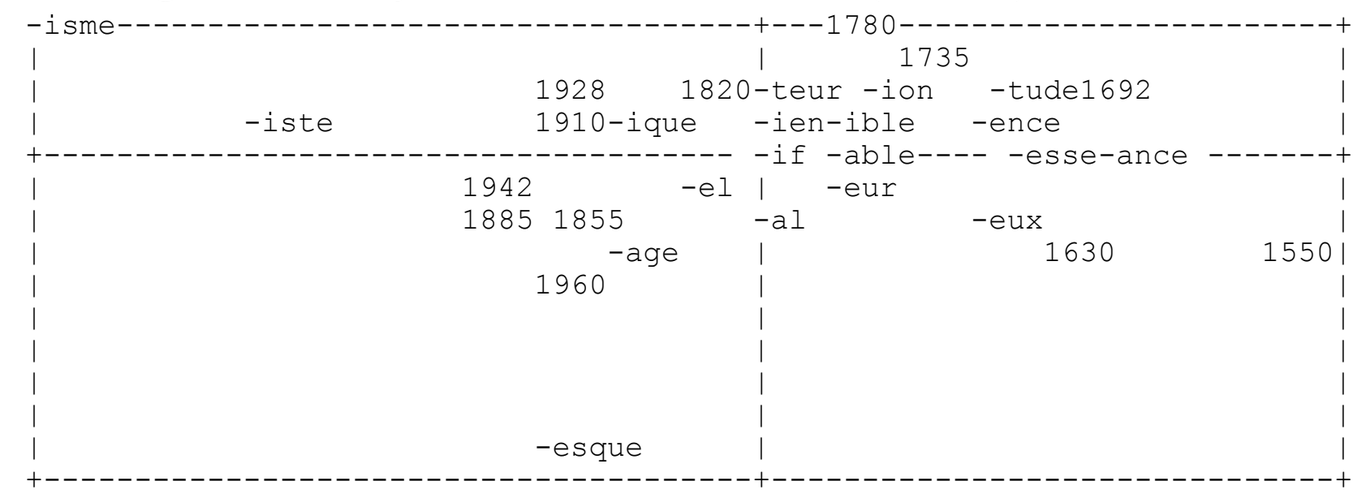
Figure 30. Analyse factorielle des suffixes (occurrences)



Deux camps s'y affrontent, à droite et à gauche de l'axe des x. D'un côté les tranches antérieures à la Révolution: tous les suffixes de la figure 28 se portent de ce côté-là, avec le renfort plus mou de *-ien*, de *-ence* et de *-ible*. De l'autre côté, XIXe et XXe siècles accueillent les éléments les plus jeunes et les plus dynamiques, même si des vétérans comme *-tude* ou *-tion* s'y trouvent aussi, mais non loin de la ligne médiane.

On attachera plus d'attention à l'analyse des variétés différentes (lignes impaires du tableau 26), parce que ce critère indique plus précisément la disponibilité d'une dérivation particulière et la direction où s'oriente la création des mots. On peut en effet imaginer le processus de la dérivation sur le modèle industriel: certaines usines fabriquent des produits de grande diffusion, à partir d'un catalogue restreint; d'autres au contraire préfèrent diversifier leur production, quitte à réduire les séries, en volume et en durée. Ce choix de l'adaptation au marché, qui peut aller jusqu'à la fabrication sur mesure, est celui que met en lumière la figure 31 qui ne considère que le catalogue de chaque unité de production[14]. En règle générale les deux objectifs, diversité et quantité, ne sont pas inconciliables et les deux analyses 30 et 31 sont largement superposables. Les tranches chronologiques respectent la même ligne de démarcation. Mais la position du XVIe siècle est beaucoup plus nette lorsqu'il s'agit de la diversité que de l'abondance. Et l'essoufflement, voire l'épuisement, de certaines productions s'y manifeste avec plus de vigueur que dans l'analyse des occurrences. On a déjà parlé de *-tude* qui appartient à cette espèce. Le graphique signale aussi le cas de *-tion*, qui semble courir sur son erre. Ce serait probablement le cas de *-té, -ure, -at, -ise, -ée,* si nous avions pu incorporer ces variétés à l'étude présente.

Figure 31. Analyse factorielle des suffixes (variétés)



Les groupes verbaux. Les temps. Les modes.

Nous n'avons considéré que les suffixes appartenant à la classe des sunstantifs et des adjectifs. Il en existe aussi pour les verbes. Mais leur identification est plus difficile à formaliser. Cependant les verbes sont traditionnellement répartis en plusieurs groupes, dont l'origine et la

---

[14] Faut-il jeter un coup d'oeil indiscret dans la cuisine où les chiffres sont assaisonnés? Les préparations diffèrent lorsqu'on épluche les occurrences ou les variétés. Dans le premier cas nous avons utilisé la pondération de l'écart réduit, dans le second celle des logarithmes.

conjugaison diffèrent. 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.

Or ce relevé est possible, au moins sur échantillon, dans *Frantext*. Quoique la lemmatisation n'y soit pas réalisée, une fonction spéciale permet d'y conjuguer les verbes et de regrouper les formes d'un même paradigme (y compris les graphies anciennes). En l'absence de codage grammatical, il est difficile d'épingler tous les verbes, mais on peut se contenter des plus fréquents, qui accaparent plus de la moitié des occurrences de la catégorie verbale. Les courbes montrent que l'érosion se manifeste dans les reliefs anciens: verbes en *-re* (coefficient de -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 sollicitée[15].

Elle est aussi moins variée dans l'expression des modes et des temps. Le simple indice du circonflexe nous a suffi pour illustrer le déclin du passé simple et du subjonctif imparfait, au moins dans les formes où cette graphie apparaît. On peut reprendre l'étude plus systématique des temps et des modes à partir de la même base qui vient de nous servir à distinguer les groupes de conjugaison[16]. Tous les verbes n'y figurent pas, encore moins toutes leurs formes, d'autant qu'ont été éliminées celles qui prêtaient à confusion, par exemple celles qui servent à la fois à l'indicatif, à l'impératif et au subjonctif présents. Mais quand on choisit les verbes les plus fréquents, les cas d'homographie se réduisent heureusement, comme il est facile de le vérifier pour *être* et *avoir*. L'embarras le plus grand vient du fait qu'on n'a accès qu'aux formes individuelles et qu'ainsi échappent tous les temps composés. Le présent ou l'imparfait des auxiliaires cachent ainsi des passés composés et des plus-que-parfaits que seule la présence ambiguë des participes passés permet de suspecter.

Mais ces réserves de méthode n'empêchent pas les résultats d'être très clairs, ce qui laisse supposer qu'on aurait obtenu une plus grande clarté

---

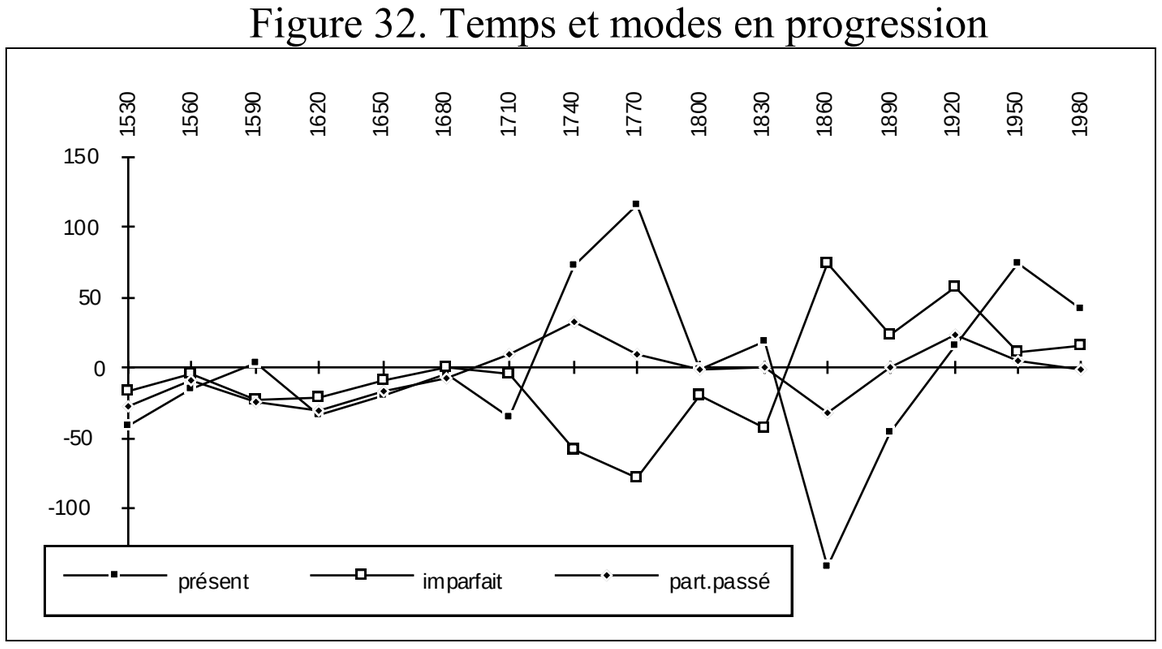
[15] Comme les classes nominale et verbale ont coutume de s'opposer, on peut en déduire que les substantifs ont tendance à envahir le discours. Sur une longue distance de quatre siècles on devrait retrouver la trace de cette tendance, si la catégorisation complète de *Frantext* pouvait être entreprise.

[16] Le corpus est toujours celui de *Frantext*, dans son aspect littéraire (textes techniques exclus). Mais la division en tranches est différente, car on a adopté ici la proposition du logiciel d'interrogation qui n'envisage que des tranches égale en durée. Le pas de progression est ici de 30 ans. Limitée aux verbes fréquents, l'observation portent sur 7000 formes verbales et 8 millions d'occurrences.

encore si la décantation avait pu être radicale et les relevés exhaustifs. Car il est rare que l'entropie à l'entrée produise l'ordre à la sortie, son effet le plus général étant de brouiller les résultats. Quoi qu'il en soit, l'échantillon a une base suffisante - plusieurs millions d'observations - pour permettre une extrapolation raisonnable et des conclusions solides, qu'on a explicitées dans les courbes 32 et 33.

La figure 32 est réservée aux temps que le temps n'attaque pas. Ils sont peu nombreux et se réduisent au présent (faible progression de +0,22) et à l'imparfait (+0,42). Le participe passé (+0,47) se joint au couple, avec une nette préférence pour le présent, auquel il est associé dans le passé composé et dont il suit la courbe avec des inflexions plus molles. L'imparfait se pose plutôt en rival du présent, et gagne des parts de marché au XIXe siècle, à partir de Balzac et de Flaubert.

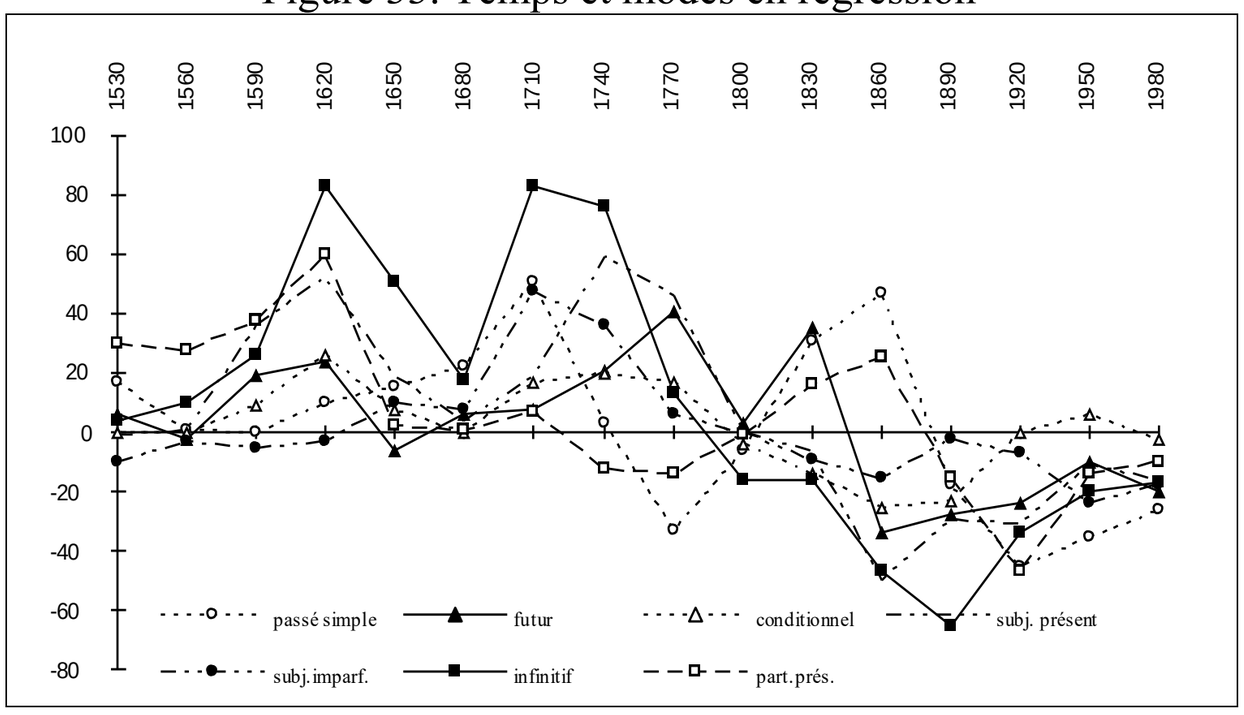
Figure 32. Temps et modes en progression



Tous les autres composants du système verbal sont en déclin: cela était attendu pour le passé simple (r=-0,45) et le subjonctif imparfait (r=-0,31), ce dernier passant pour le signe extérieur du beau langage, qu'on n'ose plus guère arborer dans la conversation et même l'écriture sinon par coquetterie ou dérision. Il est plus surprenant de voir le subjonctif présent entraîné dans la chute (-0,56), bien qu'aucune de ses formes ne suscite le sourire. Rien d'obsolète non plus dans celles du futur et du conditionnel. Et pourtant la décrue est là aussi accusée (-0,46 et -0,42). Même l'infinitif (-0,69) et le participe présent (-0,60) ne sont pas protégés par leur simplicité, si bien qu'on peut douter qu'il s'agisse seulement d'une réduction du système verbal et d'une simplification de la conjugaison - ce qui, à terme, mènerait du côté de l'anglais. On a peine à imaginer que, pour éviter la peine de conjuguer, les français parlent un jour "petit nègre". On croit plutôt que c'est le verbe, en tant que tel, qui voit son emploi diminuer sur une distance de 4 ou 5 siècles

et que les éléments les plus faibles et les plus rares de la classe verbale ont fait les frais de cette désaffection[17].

Figure 33. Temps et modes en régression



Approche de la syntaxe. Les mots grammaticaux

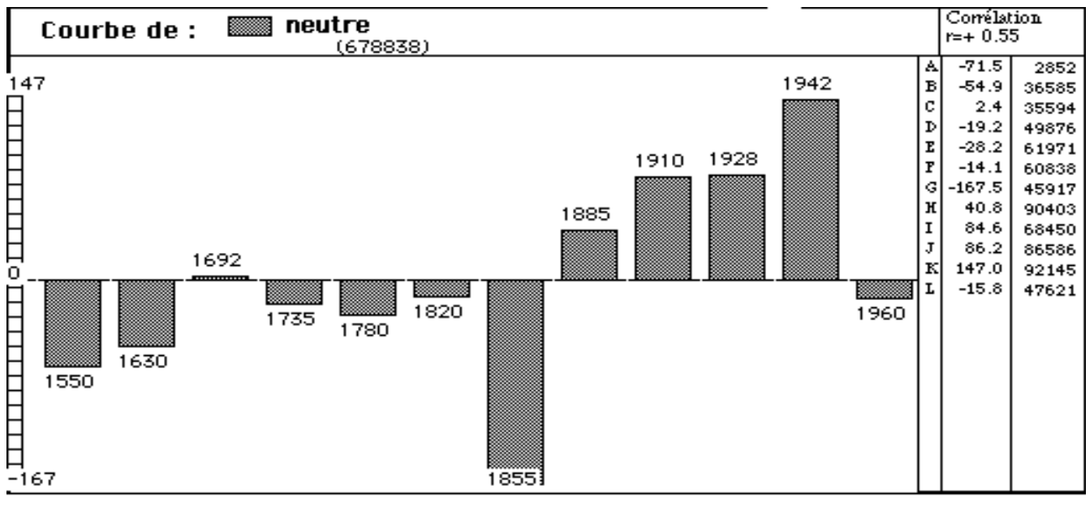
En abordant la dérivation, la catégorisation grammaticale et la flexion verbale, on a fait un pas timide vers la syntaxe. Les suffixes, les préfixes et les marques de la flexion constituent les éléments d'une sorte de combinatoire syntaxique qui se développe à l'intérieur du mot. Mais, en s'appuyant sur ces marques, l'ordinateur permet d'aller beaucoup plus loin dans la direction proprement syntaxique, si l'on emprunte la route de l'intelligence artificielle. Cette voie s'écarte sensiblement de la méthode statistique que nous avons choisie et nous préférons renvoyer le lecteur aux experts de cette discipline prometteuse. Nous ne ferons que deux remarques là-dessus, l'une vinaigrée, l'autre huileuse. La première est que les promesses s'ajoutant aux promesses, l'intelligence artificielle n'est pas encore en état de rendre compte des langages naturels de façon pleinement opérationnelle. Si la synthèse est satisfaisante, l'analyse laisse encore à désirer et la machine à traduire ne fonctionnera vraiment qu'au 21e siècle. La seconde est que la statistique est encore utile dans l'approche de l'intelligence artificielle. Quand la combinatoire grammaticale et sémantique devient en effet trop longue à explorer, la statistique offre à l'automate de précieux raccourcis qui font gagner du temps en proposant d'abord la solution la plus probable, quitte à faire marche arrière en cas d'impasse.

---

[17] On se gardera de faire des projections dans l'avenir, car il nous semble que les écrivains contemporains résistent mieux que d'autres au déclin de l'emploi du verbe. C'est dans les textes techniques, dans les milieux de l'information que le verbe se raréfie, non dans les romans, encore moins au théâtre ou dans les dialogues de films.

En s'en tenant aux seuls outils statistiques, il est encore possible d'apporter une contribution non négligeable à la description syntaxique du français et plus nettement encore à l'étude des changements qui s'opèrent dans ce domaine. Pour ce faire un matériau privilégié s'impose: les mots grammaticaux, qu'on appelle aussi, à juste titre, les mots de relation. Car leur rôle est de servir d'agent de liaison, non de définir une substance. Ce ne sont pas pourtant des mots vides[18], par opposition aux mots dits "pleins" qui seuls seraient porteurs de la charge sémantique. Car cette charge peut être très lourde dans une interjection, une négation, ou un pronom personnel. Et ce ne sont pas des mots neutres, qui seraient insensibles à la situation de l'énonciation, et aux mille nuances qui donnent au discours son sens et son style. On peut même soutenir que les marques signalétiques à quoi on reconnaît un écrivain ne sont pas nécessairement celles qui sont le plus claires à la conscience de l'auteur ou de son lecteur. On ne peut refuser tout crédit aux empreintes digitales sous prétexte qu'elles passent inaperçues et que jamais personne n'a eu l'idée de s'en plaindre ou de s'en glorifier. Rien n'est moins neutre, par exemple, que le neutre. La langue française n'en a gardé que des vestiges dont on pourrait croire qu'ils s'évanouiraient, à mesure qu'on s'éloigne du latin. Il n'en est rien, si l'on en croit l'histogramme 34, qui regroupe quelques-uns des représentants de cette classe: *rien, c', ça, ceci, cela, quoi,* en écartant les moins purs (*ce* et *tout* par exemple). Le progrès est sensible et ne doit rien au hasard (corrélation de +0,55). Que signifie-t-il? sans doute un relâchement du discours, la volonté de faire simple, de faire peuple, et aussi de faire court en empruntant le raccourci de l'anaphore.

Figure 34. Le neutre en français



Comme les mots grammaticaux se répartissent en espèces multiples, il serait opportun d'étudier chacune d'elles dans le détail. Les espèces sont distinctes mais il y a des zones limitrophes, comme celles qui se situent entre les relatifs et les interrogatifs, ou entre les prépositions et les subordonnants. Cela tient au fait que certains des mots-outils appartiennent à plusieurs classes et qu'ils remplissent plusieurs fonctions selon le contexte, comme ces

---

[18] B. Pottier le dit avec force dans *Systématique des éléments de relation*, (Klinksieck, p. 95): "Nous nous refusons à croire que la langue puisse posséder des mots vides."

couteaux suisses qui ont des ciseaux, des ouvre-boîtes, des tire-bouchons et qui ont aussi une lame. Et pourtant le mot *que* qui est bien la forme la plus ambiguë de la langue française montre un profil d'une pureté surprenante, lorsqu'on projette son évolution sur un plan[19]. On s'en assurera en considérant le graphique 9 de la figure 37, où on le montre en exemple dans la courbe déclinante des subordonnants. Mais, faute de place, il serait difficile de faire un sort à chaque forme individuelle[20], si essentielle soit-elle, et nous bornerons notre commentaire aux grandes catégories que reconnaît la grammaire traditionnelle. Les histogrammes réunis dans les figures 35 et 37 sont le reflet des collectivités, même si certaines sont réduites à quelques individus. À gauche, dans la figure 35, on a groupé les classes qui connaissent un succès croissant et à droite (figure 36) celles que l'usage littéraire tend à abandonner.

L'article fait l'objet de la première illustration (figure 35.1). Les 10 millions d'occurrences qui sont amoncelées là[21] ne forment nullement cette couche de neige étale qu'aurait produit le hasard, s'il n'y avait pas eu le vent de l'histoire et le relief accidenté des auteurs et des thèmes. Sans doute l'article imprègne tout discours et se glisse dans la moindre phrase, comme l'eau dans les corps vivants. Mais le dosage diffère et sur cinq siècles on voit qu'il augmente (coefficients de progression: +0,64 pour l'article défini et +0,90 pour l'indéfini). En cela la langue française poursuit une évolution qui l'éloigne de ses origines latines. On pouvait penser que les progrès de l'article n'avaient plus de raison d'être, au-delà du XVI^{e} siècle, quand son usage s'est généralisé, imposé par la perte des cas et l'affaiblissement dans la prononciation des *s* du pluriel et des *e* du féminin. Mais une fois lancés les grands mouvements d'une langue courent longtemps sur leur erre et la tendance analytique du français ne cesse de se renforcer. Cependant une tendance plus précise accompagne et explique ce recours croissant à l'article. C'est celle qui privilégie la classe nominale au détriment du verbe. Si l'état actuel de notre corpus, privé de code grammatical, ne nous permet que des mesures indirectes ou incomplètes, toutes sont convergentes et s'accordent d'ailleurs avec l'intuition que peuvent avoir là-dessus les usagers de la langue. On peut entendre au journal télévisé de longues suites de phrases quasiment dépourvues de verbe, comme celle-ci: *La question du pouvoir*

---

[19] Il n'est pas le seul. Qu'on songe à *si*, à *comme*, à *en*, à *le*, *la* ou *les*.

[20] De très nombreuses courbes individuelles se trouvent réunies dans le tome 3 de notre Vocabulaire français (Slatkine, 1981). La plupart des mots grammaticaux y figurent, en raison de leur grande fréquence. Deux variables, le genre littéraire et l'époque, s'y trouvent croisées, ce qui permet une interprétation fine. Précisons cependant que le corpus étudié est celui de la littérature contemporaine et ne s'étend pas au-delà de la Révolution.

[21] Il y a des impuretés dans cette neige. Les pronoms de rappel *le, la, les* s'y trouvent mêlés indissolublement, à raison d'un pronom pour sept articles. Mais inversement les articles qui sont incrustés dans les formes contractes (*au, aux*, *du* et *des*) n'ont pas été pris en compte pour ne pas trop diluer le matériau étudié.

d'achat des classes ouvrières a fait l'objet d'un débat houleux à l'Assemblée nationale. Tout le monde reconnaîtra un énoncé moderne dans cette phrase et les expressions datées comme *pouvoir d'achat*, *classes ouvrières* ou *Assemblée nationale* (et peut-être aussi *débat houleux*) sont des indices aussi certains que le carbone 14. Mais la tournure de la phrase l'est aussi, avec l'abondance des substantifs et la réduction du verbe à une simple copule logique (*a fait l'objet*).

L'histogramme (36.2) qui rend compte des prépositions confirme cette présomption. La fonction principale d'une préposition étant de relier un groupe nominal à la proposition, le sort de cette classe d'outils est évidemment lié à celui des substantifs. On objectera que certaines prépositions, en particulier celles qui portent sur le temps, n'ont pas de contrat d'exclusivité et peuvent offrir leurs services à un infinitif ou s'associer à *que* (ou *de*) pour introduire une proposition et donc un verbe. Aussi bien avons-nous mis à part cette catégorie et observé qu'elle évoluait en sens inverse (-0,63). Voir graphique 35 ci-dessous. Avant d'accepter qu'une liaison trop étroite avec le verbe explique cette régression, on peut se demander s'il ne s'agirait pas plutôt d'un désintérêt progressif pour la catégorie du temps. Le graphique 35.4 répond par la négative à cette hypothèse. Car les adverbes de temps dont il rend compte ont une évolution positive (+0,86), comme les adverbes de lieu (0,75, graphique 36.3). Et nous verrons plus loin que, dans le domaine du contenu lexical, tout ce qui a trait aux notations et aux mesures temporelles, de la seconde à l'éternité, occupe une place croissante dans le discours littéraire.

Graphique 35. Les prépositions conjonctives

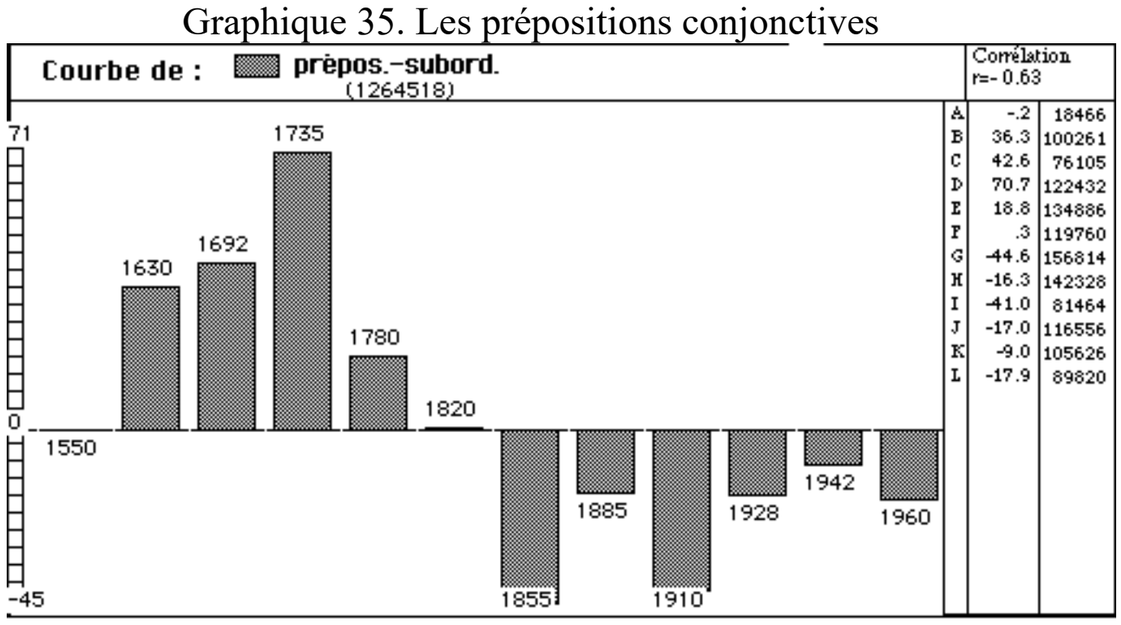
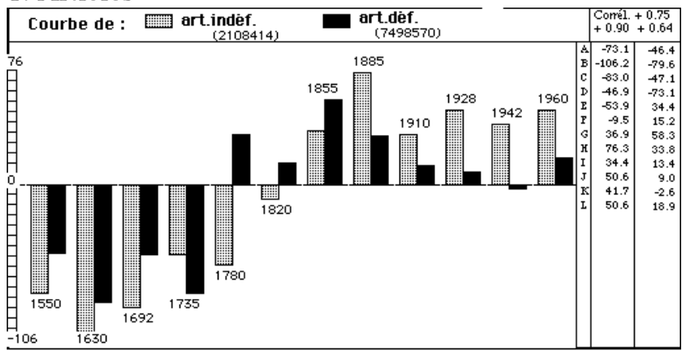
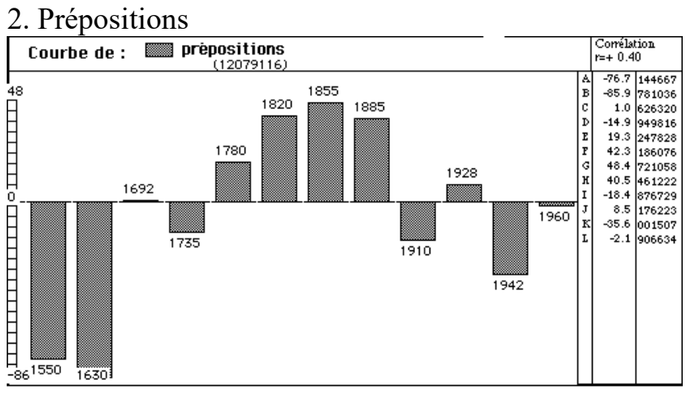


Figure 36. CATÉGORIES EN PROGRESSION

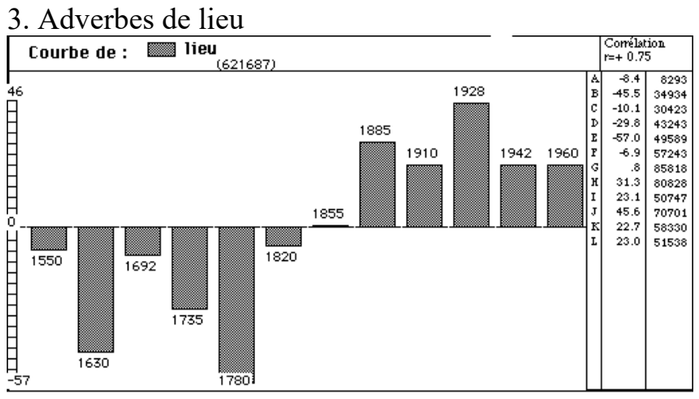
1. Articles



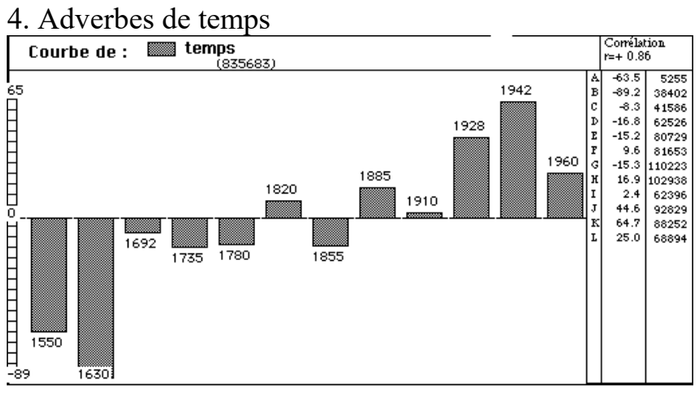
2. Prépositions



3. Adverbes de lieu



4. Adverbes de temps



5. Négation *ne* et *pas*

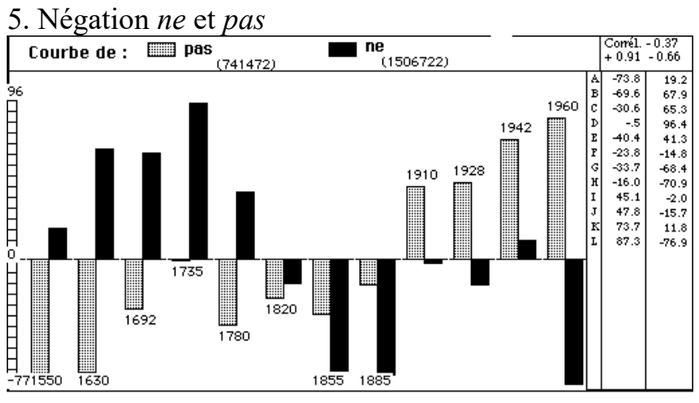
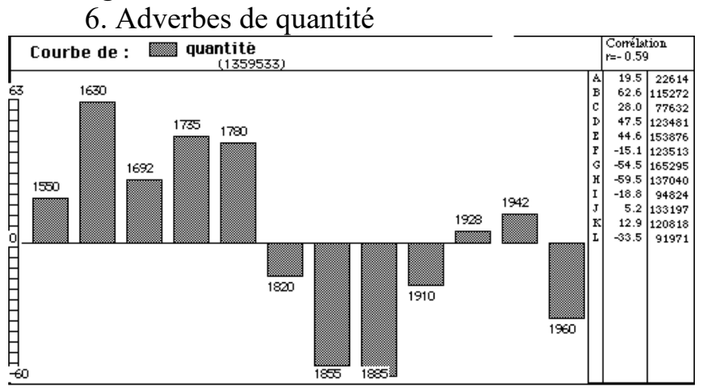
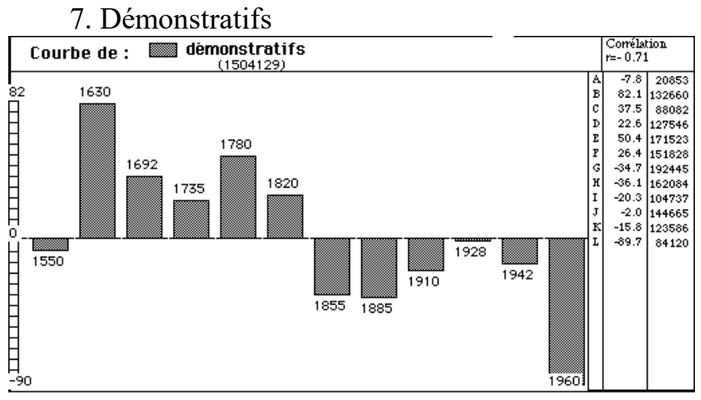


Figure 37. CATÉGORIES EN RÉGRESSION

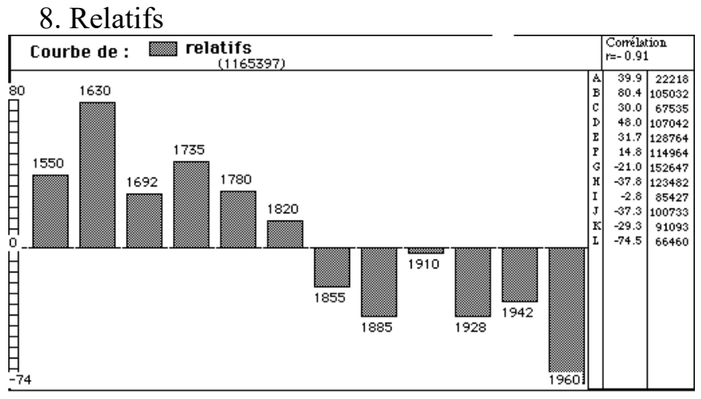
6. Adverbes de quantité



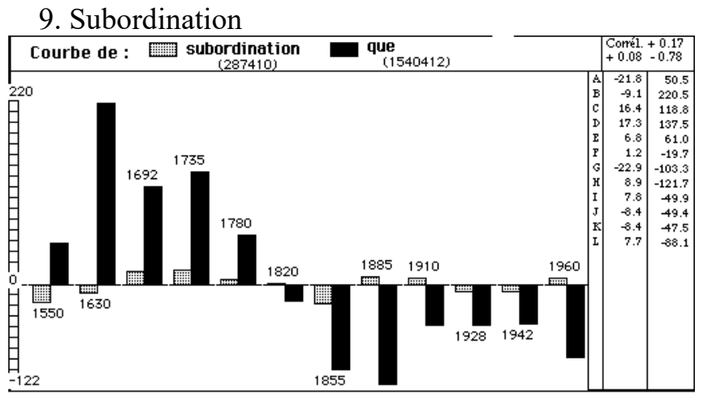
7. Démonstratifs



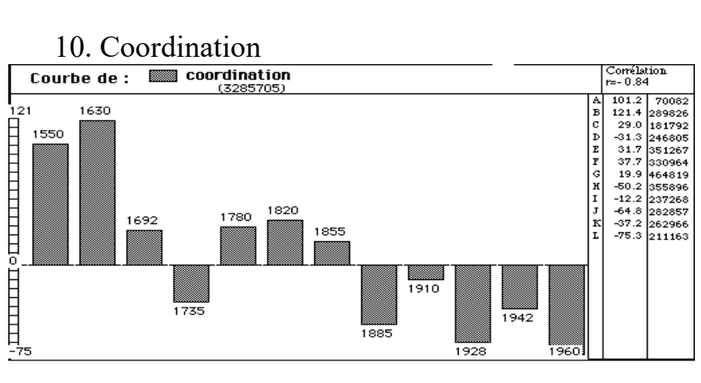
8. Relatifs



9. Subordination



10. Coordination



Avant de passer aux catégories qui déclinent, le cas des négations servira de transition (graphique 36.5). Car il réunit deux particules *ne* et *pas* qui sont en principe solidaires dans le discours et qui se croisent dans le graphique en suivant deux destins diamétralement opposés. La première subit une forte régression (r=-0,66), quand la seconde bénéficie d'une promotion remarquable (r=+0,91). On avait observé depuis longtemps la raréfaction de la négation *ne* dans le discours oral et même sa disparition presque complète dans la conversation des québécois. On voit que la même tendance se rencontre aussi dans l'écrit.

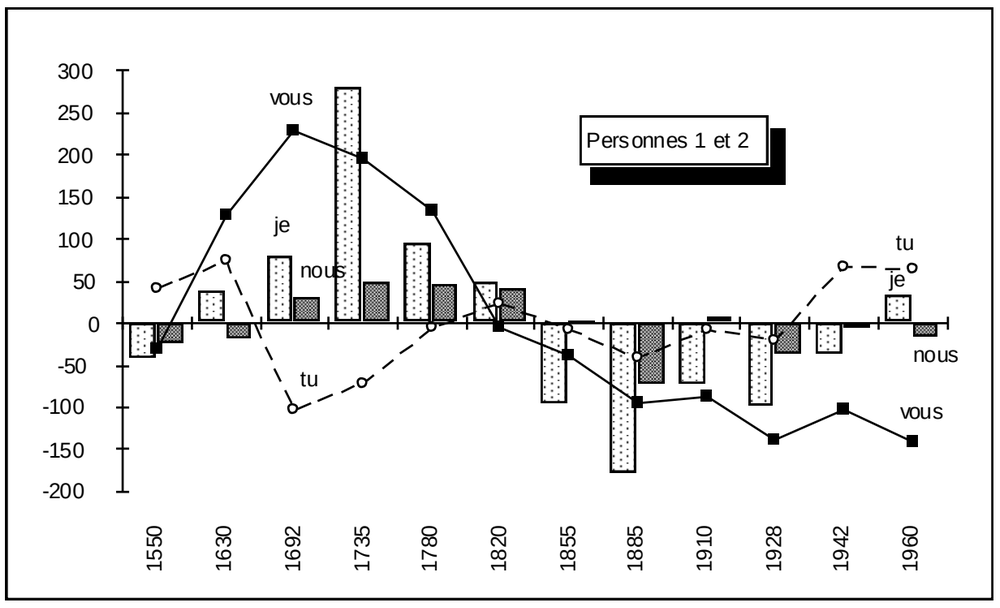
Il y a une certaine cohérence dans la figure 37 qui recense les classes en difficulté. Qu'il s'agisse des subordonnants (r= -0,78 pour le principal d'entre eux: *que*), des coordinations (-0,84) ou des relatifs (-0,91), il s'agit toujours des articulations du discours et de la gestion des propositions dans la phrase ou du rapport des phrases entre elles. L'économie de la phrase tend donc à se transformer et à se simplifier et cela n'est pas sans rapport avec le temps et le mode des verbes, et notamment avec le déclin du subjonctif. Les constructions architecturales dont s'honorait la rhétorique du Grand Siècle semblent passées de mode. L'heure est à la phrase plate, sans étage, où les circonstances d'espace ou de temps ont recours à la classe nominale et sont ajoutées librement, sous le toit unique, comme les communs des fermes basses de Bretagne.

Les démonstratifs - dont la régression apparaît dans la figure 37.7 - relèvent-ils de la même logique? Partiellement. Ceux qui figurent ici appartiennent à deux lots, celui de l'adjectif (*cet cette ces*) et celui du pronom (*celui, celle, celles, ceux*). Cette deuxième espèce est le plus souvent associée aux relatifs et partage leur sort. Quant à l'adjectif les pertes subies sont compensées par les gains de *ce*, qu'on a fait figurer ailleurs - un peu illégitimement - parmi les neutres (fig.34).

Reste une catégorie, qui est la plus sensible aux conditions de l'énoncé et à laquelle nous avons réservé le graphique 38. On a souvent remarqué que de tous les mots auxquels s'applique la statistique les pronoms personnels sont les plus violents à réagir au milieu qui les entoure. Les écarts, stylistiques ou thématiques, qu'on observe parmi eux ont tendance à se porter aux extrêmes (un regard attentif jeté sur l'échelle du graphique 38 montre un espace énorme - de -100 à +300 - dévolu ici à l'écart réduit) et ils convient parfois de les faire taire si l'on veut entendre d'autres voix plus discrètes. Au reste ils s'accordent rarement entre eux, comme le montrent les courbes divergentes de *vous* et de *tu* (derrière ces têtes de liste il faut comprendre toutes les formes de la seconde personne, possessifs inclus, de même que toutes les formes de la première sont représentées par *je* et *nous*). Le progrès de *tu* au détriment du *vous* est l'indice d'un changement de ton en littérature, et d'une plus grande familiarité. *Tu* est le seul élément en progrès - d'ailleurs faiblement. Tous les autres voient leur emploi baisser, même le moi que les

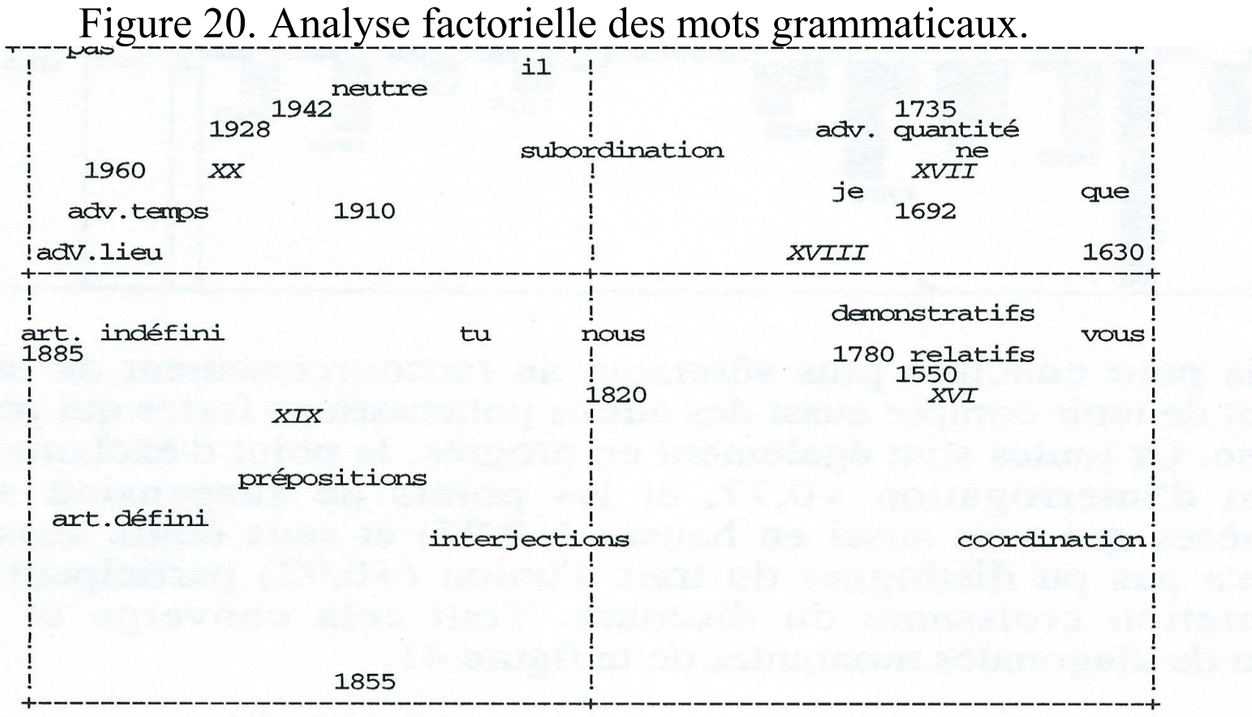
classiques disaient haïssable et dont la faveur se maintient du XVII^e^ siècle au XVII^e^ et ne survit pas au romantisme.

Figure 38. Personnels et possessifs



Au total l'analyse factorielle réalisée à partir des mots de relation (figure 39) met 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 et c'est là que s'oriente le XXe siècle.

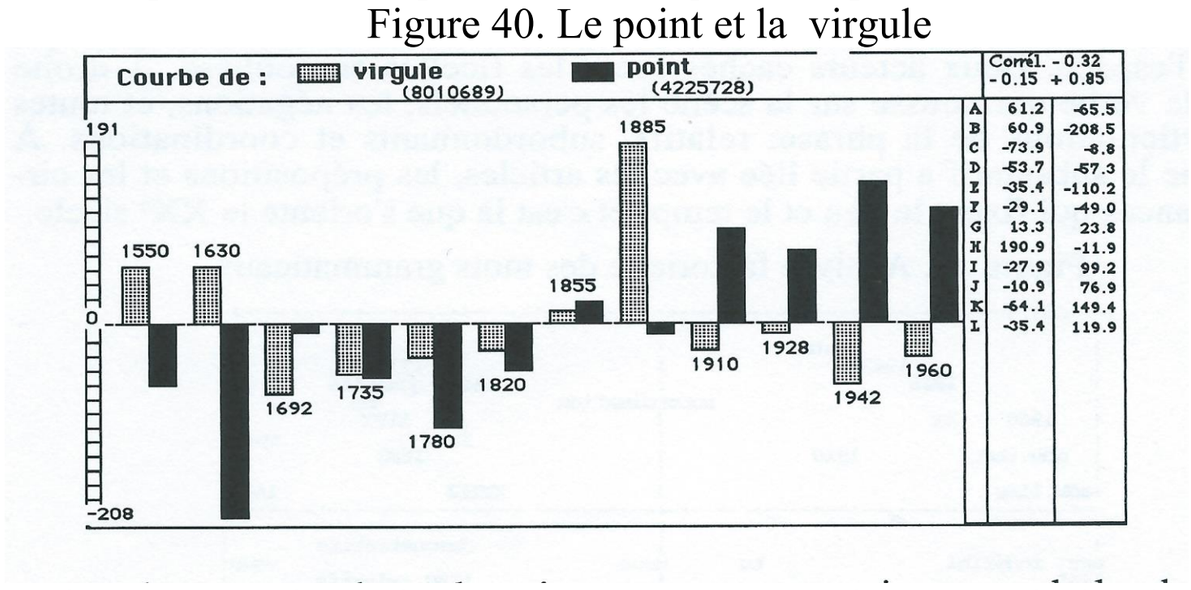
Figure 20. Analyse factorielle des mots grammaticaux.



Le rythme du discours. La ponctuation

La ponctuation est liée à la structure du discours. Si la phrase est complexe, elle n'aura pas la même segmentation qu'une phrase simple, et le dosage des virgules et des points sera différent. Comme on vient de constater que les constructions phrastiques avaient tendance à se simplifier, on peut prévoir qu'un raccourcissement de la phrase accompagne ce mouvement. C'est ce que l'on vérifie en effet quand on compte les points. Avec 4 millions de points et 8 millions de virgules, la figure 40 évolue là où se plaît la statistique, c'est-à-dire dans les grands nombres. Au surplus la série compte peu d'éléments différents et il n'y a pas d'homographie à redouter. Mais les fonctions d'un même signe peuvent être diverses (comme celles du point) ou variables, et même diverses et variables. Ainsi la valeur du point-virgule s'est affaiblie au cours du temps: de signe fort qu'il était à l'origine, au XVIIe siècle, il est devenu la marque d'une pause moyenne dans l'énoncé. Mais la difficulté la plus grande vient de ce qu'on ne sait pas toujours si les signes qu'on lit et qu'on compte sont ceux de l'auteur ou de l'éditeur. Quand une édition d'un texte ancien est modernisée, l'éditeur est amené à normaliser l'orthographe et plus encore la ponctuation, afin de ne pas heurter les habitudes du lecteur. Lorsqu'il s'agit d'un texte contemporain, le rôle de l'éditeur ne consiste guère qu'à surveiller la ponctuation et à ajouter la pagination. Cependant les résultats montre une évolution si nette qu'on est en droit de négliger ces incertitudes. La courbe du point est particulièrement claire: avec un coefficient de +0,85 elle rend compte à elle seule d'une progression régulière. Si l'on définit grossièrement la phrase comme l'espace textuel compris entre deux points, on voit que cet espace se réduit.

Figure 40. Le point et la virgule



Mais pour conclure plus sûrement au raccourcissement de la phrase, il convient de tenir compte aussi des autres ponctuations fortes qui peuvent finir la phrase. Or toutes sont également en progrès: le point d'exclamation +0,66, le point d'interrogation +0,77, et les points de suspension +0,91. Les parenthèses qui sont aussi en hausse (+0,25) et sans doute aussi les tirets qu'on n'a pas pu distinguer du trait d'union (+0,92) participent aussi à la segmentation croissante du discours. Tout cela converge et produit le

faisceau de diagonales montantes de la figure 41. Mais il est une autre segmentation, plus courte, un autre rythme, interne à la phrase, qu'on peut traduire aussi en oscillations. La virgule permet de le mesurer. Dans les différentes études que nous avons menées jusqu'ici (sur Proust, Giraudoux, Rousseau, Hugo ou Zola), nous avons généralement observé l'indépendance de la virgule qui ne cherche ni à suivre docilement le point, ni à lui offrir un contrepoint, si l'on peut dire, systématique. Ce signe (qui a lui seul représente la moitié de l'effectif des ponctuations) est relativement stable et le coefficient (-0,15) est trop faible pour être significatif. Il en va de même pour les deux points, dont le signe est certes négatif mais la baisse peu accentuée (-0,30). Une ponctuation pourtant donne des signes manifestes d'épuisement: le point-virgule. Elle n'est pas la plus ancienne du système (le point, la virgule, le point d'interrogation et les deux points l'ont précédée). Elle n'est pas non plus la plus jeune (les points d'exclamation et de suspension sont apparus après). Mais sa place a toujours été flottante et sa survie n'est pas assurée. Une enquête récente menée parmi les écrivains montrent qu'ils ne savent guère lui trouver un emploi et que l'oubli dont le point-virgule est victime est conscient.

Figure 41. La croissance des ponctuations fortes

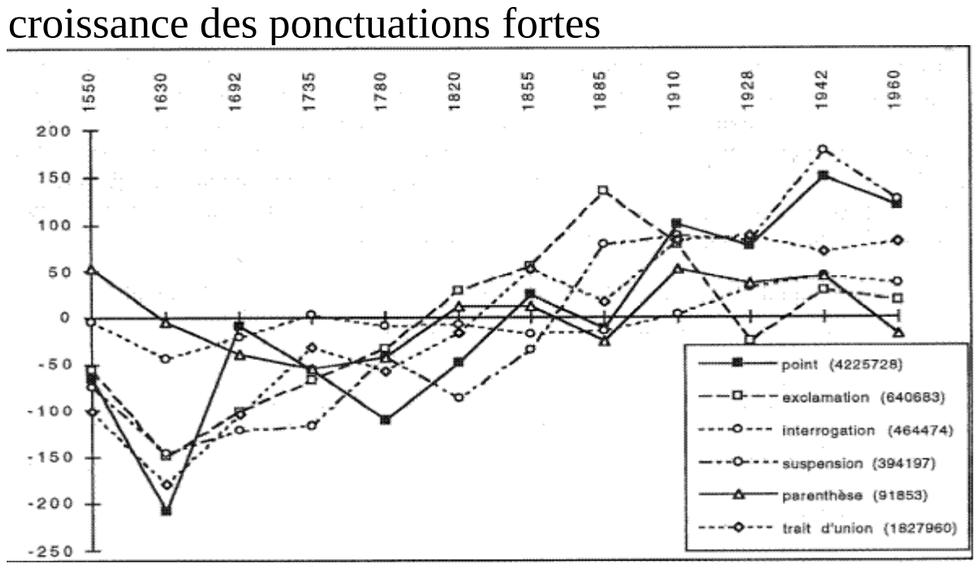
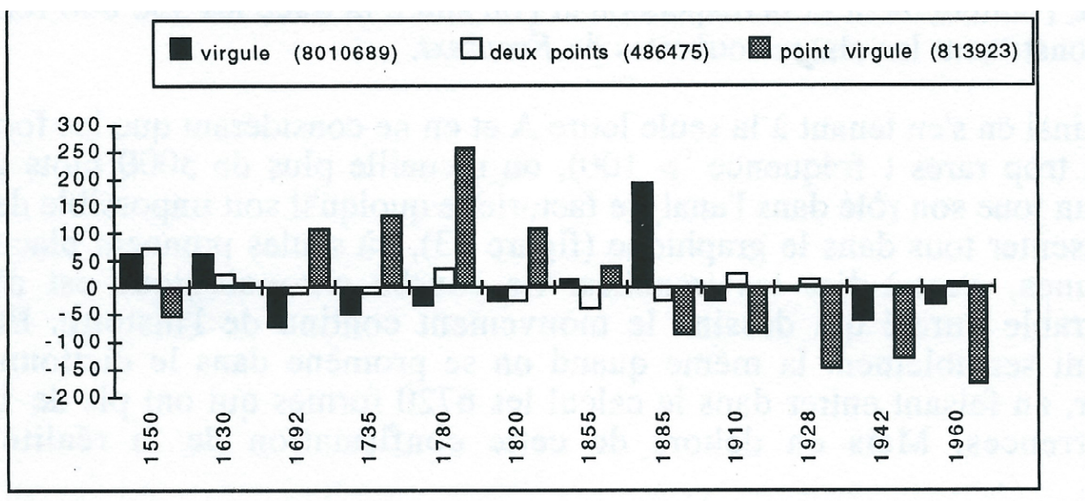


Figure 42. Les ponctuations faibles



Le contenu

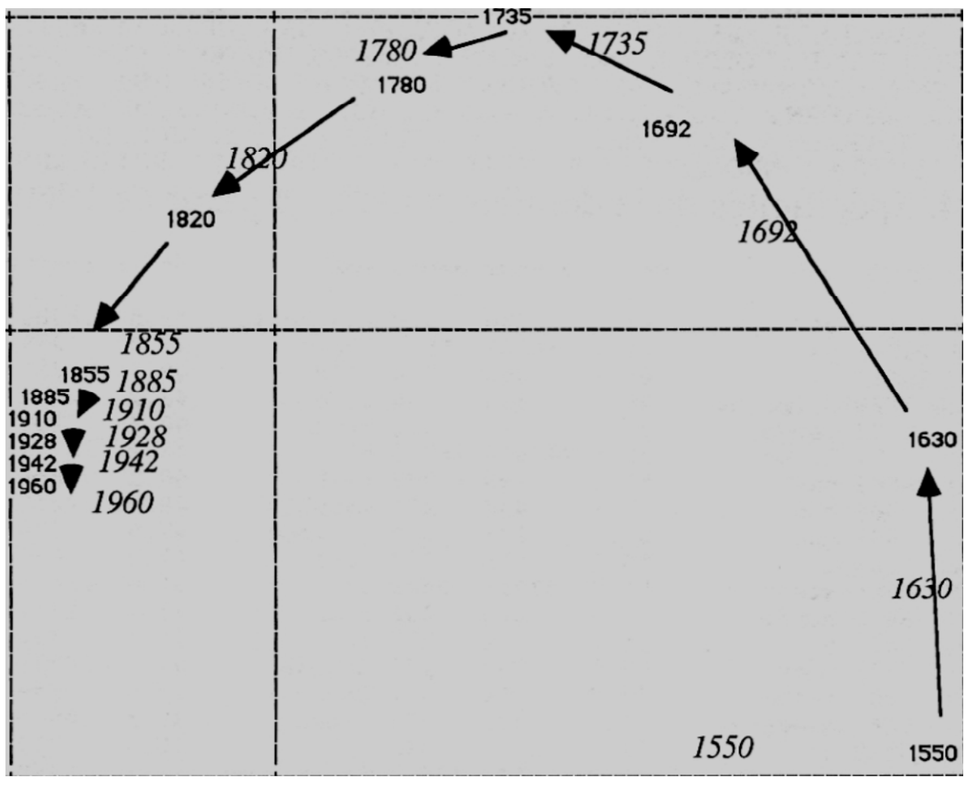
La statistique peut nous mener plus loin, du côté du contenu lexical, et considérer les mots individuellement et non plus dans des classes anonymes où leur chair disparaît. La classe des mots grammaticaux une fois constituée, avec deux ou trois centaines d'éléments, rien n'empêchait de rendre la liberté à chacun d'eux et de réaliser une étude d'ensemble, sans groupement d'aucune sorte, *que* voisinant avec *le* et *de* avec *et.* Cette expérience a été tentée dans une analyse globale, dont on fera grâce au lecteur parce que le vote des individus confirme en tous points celui des catégories. Mais on risque l'émiettement et la dispersion si l'on suit à la trace les 500 000 formes qui constituent les classes ouvertes de *Frantext.*

Ainsi en s'en tenant à la seule lettre A et en ne considérant que les formes point trop rares (fréquence > 100), on recueille plus de 3000 mots dont chacun joue son rôle dans l'analyse factorielle quoiqu'il soit impossible de les représenter tous dans le graphique (figure 43), où seules prennent place les colonnes, c'est-à-dire les tranches. La courbe chronologique est d'une admirable pureté qui dessine le mouvement continu de l'histoire. Et l'on obtient sensiblement la même quand on se promène dans le dictionnaire entier, en faisant entrer dans le calcul les 6720 formes qui ont plus de 1000 occurrences. Mais en dehors de cette confirmation de la réalité du mouvement, les éléments ont disparu qui permettraient de caractériser cette évolution. Il est donc nécessaire de constituer là aussi des regroupements, afin d'isoler par exemple un champ sémantique, ou un thème littéraire. Mais ce faisant, on poursuit un objectif détourné qui est moins l'histoire de la langue que celle de la littérature, même s'il est vrai qu'on ne peut guère atteindre la langue qu'à travers ses réalisations, littéraires ou non.

Figure 43. Analyse factorielle du contenu lexical

-en italique un échantillon de 3000 formes (lettre A, fréquence > 100)

-en romain tous les mots de fréquence supérieure à 1000 (6720



Le deuxième moyen fourni par la statistique sert moins à focaliser l'attention, comme un téléobjectif, sur un détail révélateur qu'à construire une perspective, selon l'optique du grand angle. Comme pour tous les effets d'optique, cet accessoire (il s'agit du coefficient de corrélation chronologique) ne va pas sans déformation : il exagère les mouvements réguliers qui suivent une pente rectiligne, à la baisse ou à la hausse, mais il est impuissant à filtrer les mouvements désordonnés ou contradictoires, qu'on rencontre aussi dans les faits de langage, par exemple une hausse suivie d'une décrue, ce qui produit une courbe en cloche, ou l'inverse qui prend la forme d'une cuvette. En de tels cas le coefficient de Bravais-Pearson est désorienté et ne pipe mot.

On ne le regrettera pas trop, ayant déjà une matière trop abondante. Pour en rendre la lecture possible, on a dû se limiter aux substantifs et aux adjectifs et ne retenir que les fréquences hautes (supérieures à 10000). Les adjectifs sont rares dans ces listes (tableau 39). Ceux qu'on voit dans celle des gains désignent une entité souvent politique (*gauche, droite, français, libre*), précisent une couleur, qui peut d'ailleurs être politique (*noir, rouge, blanche*) ou situent un objet ou un évènement (*première, dernière*). L'adjectif qui grandit le plus dans l'usage est le mot *petit* sous toutes ses formes (*petit, petite, petits*) et l'opposition est forte avec la liste des pertes où figurent *grand, bons* et *belles*.

Les choix des substantifs sont plus clairs encore. Le coefficient souligne bien sûr la perte d'influence des grands de l'Ancien Régime (*seigneur, prince, roy*) et le déclin des préoccupations religieuses (*dieu, dieux*). Mais il accuse surtout la chute des valeurs, morales ou sociales (*honneur, vertu, justice, devoir, digne*) et des aspirations à la *puissance*, à la *gloire*, à la *beauté*. La question du bien et du mal n'a plus la même acuité (*mal, bien, malheur, bons, mauvais, heureux*), non plus que celle de la raison (esprit, raison, âme) et du coeur (*amour, amitié, cœur, amant, passion, mariage, crainte, pitié*). Si le cours des valeurs et des sentiments baisse, c'est au profit d'un regard jeté sur le monde, sur les réalités concrètes du milieu physique (*route, rue, place, porte, table, lit, fenêtre*), sur l'épaisseur palpable du corps (*corps, tête, lèvres, nez, voix, silence, regard*) et sur la présence tangible du temps (*midi, nuit, journées, ans, années*).

Tableau 44. Substantifs et adjectifs en progression et en régression (f> 10000)

En progression				En régression			
coeff.	*fréqu.*	*mot*	*coeff. fréqu. mot*	*coeff.*	*fréqu.*	*mot*	*coeff. fréqu. mot*
+0.94	82613	fois	+0.80 10987 mesure	-0.92	10092	bons	-0.75 20158 nouvelle
+0.93	13102	midi	+0.80 59123 petit	-0.92	26050	honneur	-0.75 13174 passion
+0.92	47595	nuit	+0.80 10798 train	-0.90	15509	moyen	-0.75 22925 prince
+0.90	17855	route	+0.79 55655 air	-0.89	89598	toutes	-0.75 35812 raison
+0.89	13934	années	+0.79 23695 enfants	-0.88	34920	lieu	-0.74 17035 amitié
+0.89	11782	marche	+0.79 24965 nouveau	-0.86	11004	digne	-0.74 14057 devoir
+0.88	11303	fenêtre	+0.79 12240 salle	-0.85	17088	fortune	-0.74 90085 grand
+0.88	11973	libre	+0.79 10459 voiture	-0.85	11403	pitié	-0.74 11088 roy
+0.88	23124	lit	+0.78 11702 journée	-0.85	11142	repos	-0.73 19182 garde
+0.88	34707	place	+0.78 15611 noir	-0.85	16286	seigneur	-0.73 25908 heureux
+0.87	12429	droite	+0.78 53166 porte	-0.85	16450	vertu	-0.73 12033 personnes
+0.87	61526	tête	+0.78 25365 rue	-0.82	71316	amour	-0.73 10053 puissance
+0.86	16710	lumière	+0.78 50550 voix	-0.81	10137	armes	-0.72 11977 belles
+0.86	25515	silence	+0.77 23682 bout	-0.81	17240	gloire	-0.72 318200 bien
+0.85	14932	français	+0.77 12951 froid	-0.80	10957	conseil	-0.72 75875 dieu
+0.85	13072	gauche	+0.77 49649 mère	-0.80	16902	douleur	-0.71 14826 ame
+0.84	15918	question	+0.77 42792 petite	-0.79	38388	belle	-0.71 15541 beauté
+0.83	21444	regard	+0.77 21852 petits	-0.79	54826	esprit	-0.71 14573 dieux
+0.83	13086	rouge	+0.76 11923 garçon	-0.78	89054	coeur	-0.71 40027 nom
+0.83	22228	suite	+0.75 14639 dernière	-0.78	15868	nouvelles	
+0.82	46243	coup	+0.75 13077 pièce	-0.77	10840	crainte	
+0.82	28017	première	+0.74 43389 ans	-0.77	15590	malheur	
+0.82	17750	travail	+0.74 10326 blanche	-0.76	14754	mauvais	
+0.81	10718	coin	+0.74 34575 fond	-0.76	138131	point	
+0.81	11620	lèvres	+0.74 10536 nez	-0.75	10483	amant	
+0.81	21438	table	+0.73 13083 chef	-0.75	14022	justice	
+0.80	12179	bord	+0.73 10187 hôtel	-0.75	13086	mariage	
+0.80	34620	côté	+0.73 23318 mots				

Sans s'attarder aux changements technologiques qui expliquent la promotion du train, de la voiture et de l'hôtel, ou aux modifications de la famille bourgeoise qui bénéficient aux enfants, au garçon, et à la mère, on notera la banalisation croissante des mots courts, presque vides, qui servent à tous les usages et entrent dans beaucoup d'expressions figées. À travers le progrès de *bout, bord, fond, coin, côté, coup, fois, suite, pièce, air,* on devine comment un substantif peut se vider à la longue de sa substance et comment s'usent les mots. Ils servent plus souvent dans les échanges, mais leur charge utile diminue.

On pourrait s'arrêter là, sur la frontière qui sépare et unit la linguistique et la littérature. Si la sémantique appartient à la première, la thématique relève de la seconde. La statistique peut inspirer bien d'autres commentaires, oiseux ou pertinents, non seulement sur l'évolution des mots, mais sur l'histoire des mentalités, les variations du goût, la réception des œuvres. Elle réserve ces commentaires à une éventuelle histoire de la littérature, s'il advenait qu'une telle entreprise eût besoin de son concours.

Avant d'achever cette exploration du contenu lexical, il convient d'essayer une troisième méthode apte à évaluer le débit et les variations du flux verbal. Cette méthode, très classique, consiste à calculer les spécificités pour chacune des tranches considérées. On obtient une série de portraits, chacun mettant en relief les mots, les tournures ou les thèmes caractéristiques de l'époque. La place nous manque pour disposer cette galerie en enfilade, et nous nous contenterons du dernier

portrait de la série, en espérant que l'image un peu vulgaire qu'il donne de la langue contemporaine n'est due qu'aux aléas de la composition du corpus. Il se trouve en effet que le corpus qui devait servir de fondement au TLF a été arrêté dans les années 60. Depuis lors les ajouts de textes, aux deux bouts de la chronologie, n'ont plus eu comme but exclusif de fournir des exemples nouveaux et distingués aux rédacteurs du dictionnaire mais plutôt de constituer une base de données représentative du français à travers les siècles.

Et c'est pourquoi la dernière tranche bénéficie d'une ouverture plus large au français non conventionnel, et aux expériences hardies où la littérature s'est engagée depuis Céline. Le vocabulaire familier, voire ordurier, qui émerge comme une écume sale de la liste 45, n'est pas dû seulement à Céline ou au Queneau de *Zazie dans le métro*, mais à bien d'autres écrivains, moins soucieux que leurs aînés de beau langage. Le roman policier et le roman populaire n'y sont plus victimes d'ostracisme et l'on y accueille Simonin *(Touchez pas au grisbi)*, Japrisot, Forlani, Vautrin, Chabrol et jusqu'aux *Valseuses* de B. Blier. On ne s'étonnera pas dès lors que le mot de Cambronne figure en tête de liste.

Tableau 45.Spécificités de la dernière tranche (1960)

écart	corpus	époqu.	mot	écart	corpus	époqu.	mot	écart	corpus	époqu.	mot
426.3	98940	42811	c'	62.6	530	420	manivelle	49.8	36008	5203	jusqu'
258.6	76636	24643	ça	61.5	36167	5807	étais	49.3	959	475	cinéma
126.0	394197	50513	...	60.1	459	374	flic	49.3	466	316	mômes
119.9	4225728	382877	.(point)	59.1	436	358	fric	49.2	540	342	marre
95.8	1706	1172	merde	58.7	669	451	trucs	48.3	768	411	métro
89.4	699	676	hippo	58.5	248484	26388	ai	48.2	32780	4769	as
87.3	741472	75603	pas	58.1	298	287	rock	48.2	859	437	p'
86.9	1248	904	con	57.8	438	352	bagnole	48.1	2255	772	vacances
85.1	796	693	mec	57.0	275	270	blouson	47.9	378	274	z'
80.9	1827960	166351	-	56.5	436	344	pote	47.8	680	380	boulot
80.1	306548	34759	était	56.2	63217	8482	t'	47.8	21510	3467	derrière
74.1	52437	8420	avais	55.7	440	341	cons	47.4	1737	652	cigarette
73.0	436	435	y'	55.6	301	277	ze	47.1	331	251	super
72.0	522	473	mémé	55.3	292	271	dingue	47.1	10798	2102	train
70.9	599	503	flics	55.0	364	304	tronche	47.0	936	450	radio
70.4	882	618	libération	54.7	347	295	samba	46.5	50645	6569	va
69.3	454107	46482	j'	54.5	2127	823	cul	46.4	210	193	combattante
68.2	1856	914	téléphone	54.2	714	436	britanniques	46.3	780	400	putain
68.0	3155	1244	gueule	54.1	354	295	mecs	46.2	2256	749	alliés
67.8	691	522	ouais	52.0	222	221	artisse	46.1	222	198	piges
67.1	254486	28069	avait	51.1	1377	604	britannique	46.0	211	192	prof
66.7	210093	23872	tu	50.5	1109	527	foutre	45.9	614	346	photo
65.6	982	616	truc	50.4	651	388	immeuble	45.8	15768	2702	vite
64.9	4259	1438	type	50.3	1144	535	copains	45.6	721	377	copain
64.8	344	343	télé	49.8	1973	732	comité				

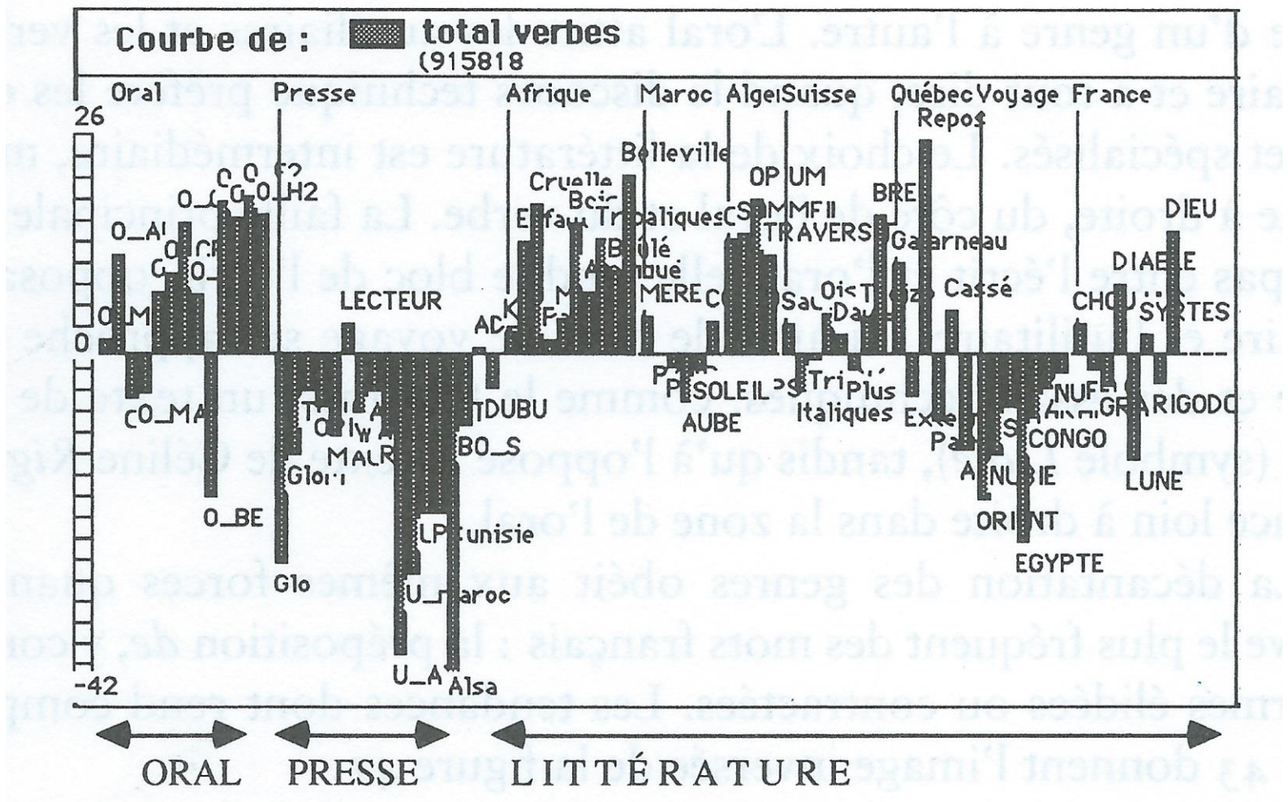
Il est toujours prudent de retenir d'abord les explications les plus simples, la plus triviale étant d'attribuer à la composition du corpus les particularités observées, ce qu'on vient de faire. Mais il y aurait quelque lâcheté méthodologique à s'en tenir là. Le tableau 45 est si provocant qu'il exige quelque approfondissement. Il évoque immédiatement les

caractéristiques du style parlé. Et pourtant les textes dont il rend compte appartiennent principalement au genre romanesque, rarement au théâtre ou à la correspondance, dont le statut se rapproche de l'oral, et rien n'autorise à penser que la part du dialogue est plus grande dans le roman contemporain que chez Balzac, Flaubert ou Zola. Pour comprendre ce qui se passe, il devient nécessaire de comparer, chiffres à l'appui, l'écrit et l'oral, ces deux variétés de français qu'on a l'habitude d'opposer et qui semblent se rejoindre dans la production littéraire de notre temps. C'est qu'en réalité une troisième variété, qu'on peut appeler utilitaire, technique ou communicationnelle, s'installe dans les usages, en s'éloignant de l'oral expressif comme de l'écrit littéraire. Dans un champ de forces devenu triangulaire, s'ouvre le jeu mouvant des oppositions et des alliances, et il peut se produire que l'oral et l'écrit ne soient plus face à face, mais côte à côte

Pour étudier ce jeu, *Frantext* n'est pas d'un grand secours. Certes des textes techniques ont été incorporés dans la base, et la distance du littéraire à l'utilitaire pourrait y être mesurée mais l'oral manque absolument. On a donc eu recours à une base spécifique, que nous avons constituée dans le cadre du projet FRANCIL de l'AUPELF-UREF et avec le concours des équipes qui étudient le français dans l'ensemble de la francophonie. Les variables diachroniques ont été cette fois neutralisées au profit de la variabilité géographique et sociolinguistique. Au lieu d'isoler la littérature à travers les siècles, on a cherché à isoler le français contemporain à travers l'espace et dans ses emplois multiples, où la littérature a sa place à côté de l'oral et des journaux. La base ainsi constituée n'a ni l'ampleur, ni l'homogénéité de *Frantext.* Avec plus de 4 millions d'occurrences, elle donne cependant une image de ce qui se dit ou s'écrit en français dans les pays francophones, y compris l'Afrique noire et le Maghreb. Il a fallu harmoniser les conventions de transcription, qui différaient d'un laboratoire à l'autre, et en particulier adopter un système de ponctuation qui traduise les pauses et la segmentation du discours oral.

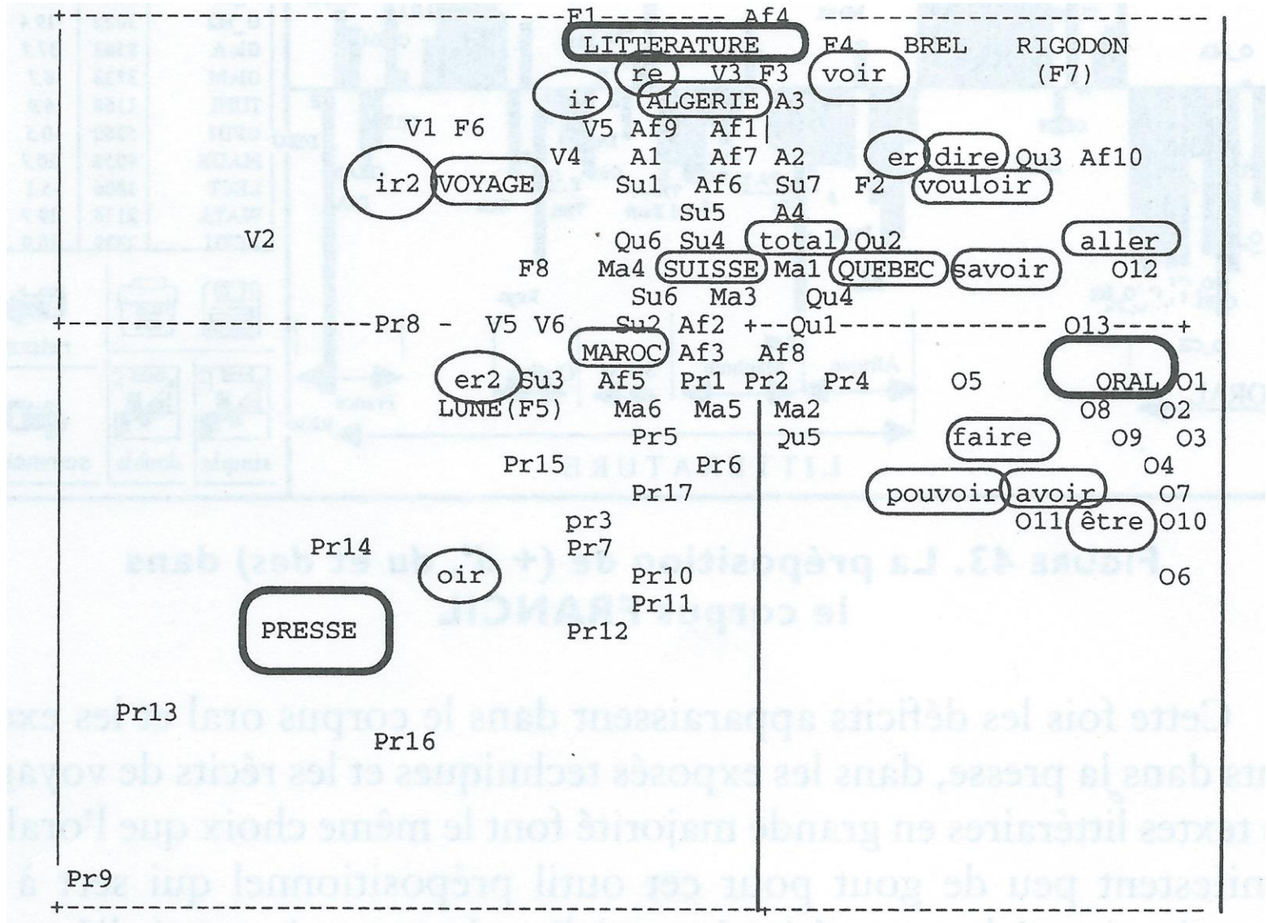
L'étude est en cours et nous n'en donnerons que certains résultats fragmentaires. Reprenons par exemple l'étude du verbe. On a coutume de dire que c'est là un des critères qui distinguent le mieux l'oral de l'écrit. Encore faut-il préciser de quel écrit il s'agit. C'est l'écrit de la communication technique qui refuse l'emploi du verbe ou du moins le réduit au rôle de simple copule, comme le signe de l'égalité dans une formule mathématique. L'utilisation littéraire du verbe est beaucoup plus riche. Même si les temps et les modes ont perdu de leur complexité, comme nous l'avons vu, le verbe se maintient, au moins dans ses temps simples, aussi bien dans le discours oral que littéraire.

Figure 46. Les verbes dans la base *FRANCIL*



La figure 46, qui porte sur la totalité des verbes rencontrés dans la base étudiée (soit près de 1 million d'occurrences), montre que l'oral (à gauche) et la littérature (à droite) se rangent dans la moitié supérieure du graphique où le verbe concentre ses excédents. En revanche, les textes qui représentent la presse et le discours technique (au centre gauche) révèlent un sous-emploi manifeste de la catégorie verbale. Une exception est à signaler qui concerne le courrier des lecteurs dans un journal populaire d'Alger et qui précisément se rapproche de l'usage oral et spontané de la langue. Inversement un texte d'un type particulier fait défection dans la zone orale. Le code O_BE recouvre des interviews d'hommes politiques belges, dont le discours, oral ou écrit, obéit aux mêmes contraintes de la propagande électorale. Pareillement dans la zone littéraire, le récit de voyage fait bande à part et se situe dans les déficits, comme si le style descriptif recourait aux mêmes catégories que le style des journaux et des essais, la description comme l'information recourant principalement aux catégories nominales.

Figure 47. Analyse factorielle des verbes dans la base FRANCIL



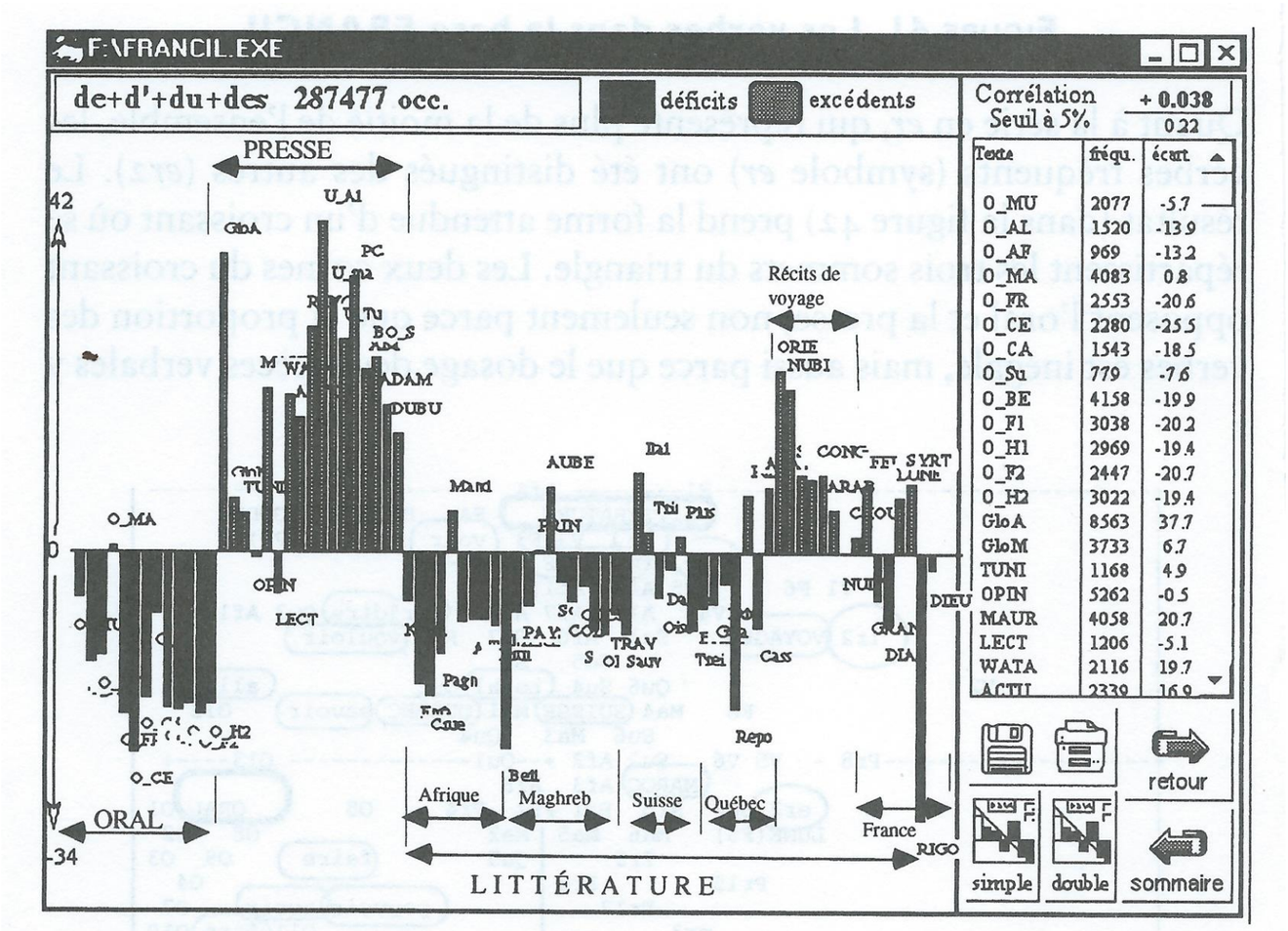
L'étude des verbes (soit un mot du corpus sur quatre) peut donner lieu à une analyse factorielle, où l'on distinguera les séries fermées : verbes *en re, oir* et *ir (ant*), et les séries ouvertes : en *er* et *ir (issant*). Certains verbes qui ont une fréquence exceptionnelle et des emplois modaux ont été traités individuellement, à savoir *être, avoir, pouvoir, voir, savoir, faire, dire* et *aller*. Les verbes en *ir* ont été séparés, suivant le groupe auquel ils appartiennent (*ir* pour le 3e groupe, *ir2* pour le 2e).

Quant à la série en *er*, qui représente plus de la moitié de l'ensemble, les verbes fréquents (symbole *er*) ont été distingués des autres (*er2*). Le résultat (dans la figure 47) prend la forme attendue d'un croissant où se répartissent les trois sommets du triangle. Les deux cornes du croissant opposent l'oral et la presse, non seulement parce que la proportion des verbes est inégale, mais aussi parce que le dosage des espèces verbales y diffère d'un genre à l'autre. L'oral attire les auxiliaires et les verbes à tout faire et à tout dire, quand le discours technique préfère les outils rares et spécialisés. Le choix de la littérature est intermédiaire, mais il penche à droite, du côté de l'oral et du verbe. La faille principale n'est donc pas entre l'écrit et l'oral; elle fend le bloc de l'écrit, opposant le littéraire et l'utilitaire. Ici aussi le récit de voyage se rapproche de la presse et des essais techniques, comme le fait aussi un texte de Jules

Verne (symbole *Lune*), tandis qu'à l'opposé le texte de Céline *Rigodon* s'avance loin à droite dans la zone de l'oral.

La décantation des genres obéit aux mêmes forces quand on observe le plus fréquent des mots français: la préposition *de*, y compris les formes élidées ou contractées. Les tendances dont rend compte la figure 48 donnent l'image inversée de la figure 46. Cette fois les déficits apparaissent dans le corpus oral et les excédents dans la presse, dans les exposés techniques et les récits de voyage. Les textes littéraires en grande majorité font le même choix que l'oral et manifestent peu de goût pour cet outil prépositionnel qui sert à la détermination, à la caractérisation et à l'analyse et qui a partie liée avec les substantifs, spécialement lorsqu'ils sont abstraits. Si l'on compare cette distribution à celle de l'article *la* qui accompagne tant de suffixes abstraits (en *tion*, en *ie*, en *té*, etc.), le parallélisme est presque parfait. Pour 79 paires d'observation, le coefficient de corrélation (r = 0,8o) est très élevé.

Figure 48. La préposition *de* (+ *d', du* et *des*) dans le corpus *Francil.*

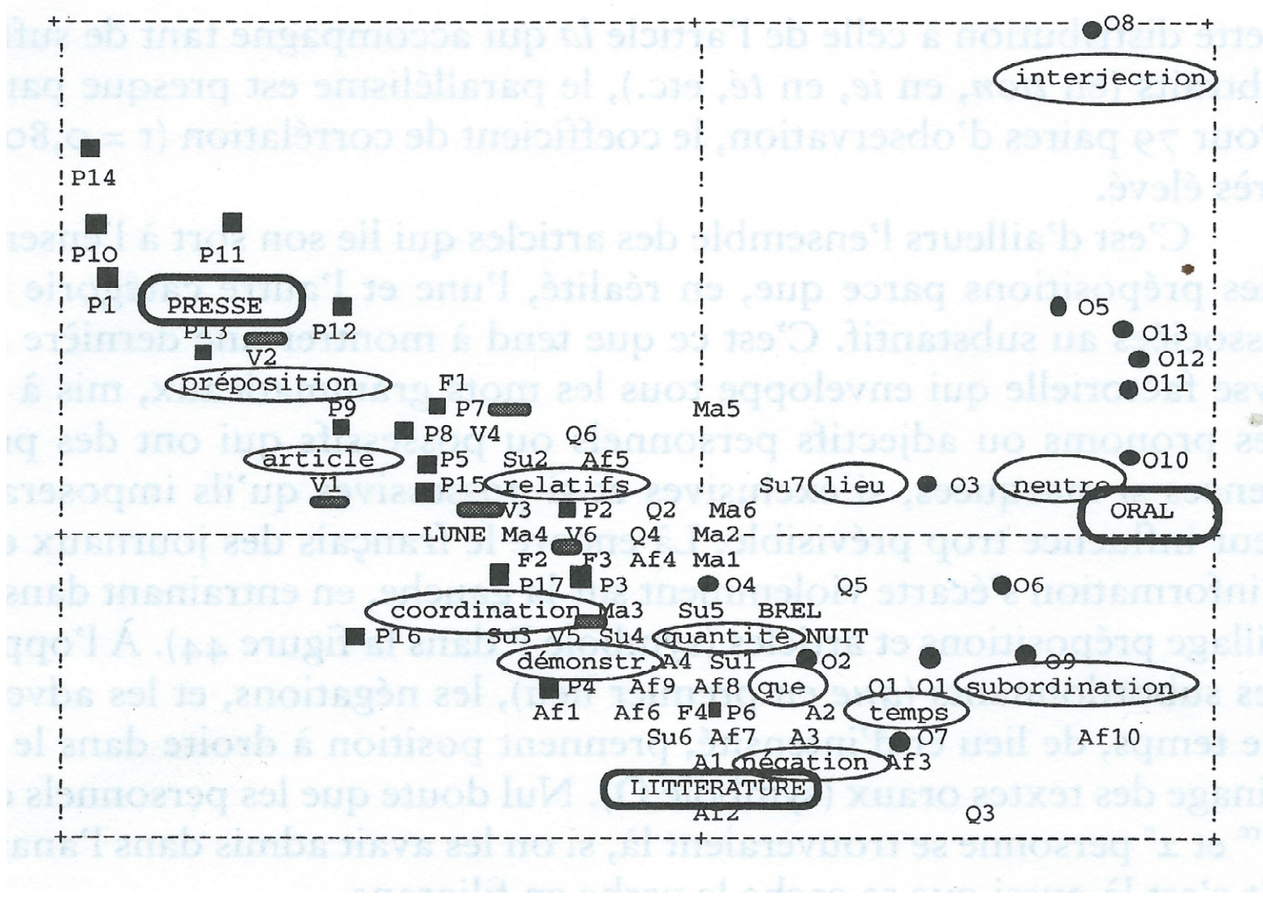


C'est d'ailleurs l'ensemble des articles qui lie son sort à l'ensemble des prépositions parce que, en réalité, l'une et l'autre catégories sont associées au substantif. C'est ce que tend à montrer une dernière analyse factorielle qui enveloppe tous les mots grammaticaux, mis à part les pronoms ou adjectifs personnels ou possessifs qui ont des préférences si marquées, si exclusives et si possessives qu'ils imposeraient leur

influence trop prévisible. Là encore le français des journaux et de l'information s'écarte violemment sur la gauche, en entraînant dans son sillage prépositions et articles (symbole P dans la figure 44). À l'opposé, les subordonnants (*que* en premier lieu), les négations, et les adverbes de temps, de lieu et d'intensité, prennent position à droite dans le voisinage des textes oraux (symbole O). Nul doute que les personnels de la première et deuxième personne se trouveraient là, si on les avait admis dans l'analyse. Et c'est là aussi que se cache le verbe en filigrane.

Or la littérature occupe l'espace intermédiaire. Comme précédemment, les récits de voyage (symbole V) se mêlent à la prose journalistique, et c'est le cas aussi du roman de Jules Verne *De la terre à la lune*. Mais nombreux sont les textes littéraires, surtout ceux qui viennent d'Afrique, d'Algérie ou du Québec (et aussi un texte de Céline et les chansons de Brel), à franchir la ligne médiane et à se confondre avec les représentants de l'oral. On voit bien que les traits syntaxiques dont les mots grammaticaux portent témoignage sont structurés dans les deux cas par l'opposition classes nominales-classes verbales. On observe certes une tendance qui donne de plus en plus l'avantage au substantif dans l'usage contemporain, mais il ne faut pas accuser la littérature, qui résiste autant qu'elle peut à ce déséquilibre des parties du discours, encore moins le français tel qu'on le parle dans la conversation de tous les jours.

Figure 48. Analyse factorielle des mots de relation dans le corpus FRANCIL



La substantification croissante du français vient du langage qu'on utilise quand on transmet une information, situation commune à la presse, à l'édition scientifique, à la littérature politique, économique ou technique et à beaucoup de médias. Le langage de l'information, comme les langages de programmation, tend à n'être plus qu'un jeu de variables emboîtées (de substantifs), la préposition *de* jouant le rôle qu'ont les parenthèses dans les formules mathématiques. Peu de place pour le verbe ou l'action, la principale opération étant celle de l'équivalence et du transfert, ce qui peut se faire avec les verbes *get* ou *put*, mais aussi bien avec le signe de l'égalité. Pas de personnes, pas de formulation expressive, pas de modalités, pas de temps, pas de modes. Au moment où l'on admet, avec Searle et la pragmatique, que toute parole est une action, le discours paradoxalement efface les traces perlocutoires et se fige dans le présent intemporel et impersonnel, comme les livres de recettes culinaires ou le mode d'emploi des appareils domestiques, où tous les verbes sont à l'infinitif.

L'abus des constructions nominales peut obscurcir le message en prétendant le réduire à l'essentiel. Combien de titres sibyllins dans la presse qui mettent en jeu la combinatoire étriquée des constructions nominales et n'aboutissent qu'à un rébus opaque. On peut voir des affiches où noms et adjectifs sont distribués presque au hasard, comme les *beaux yeux*, *l'amour* et la *marquise* de Monsieur Jourdain. On verrait mieux la fonction et l'utilité d'un *Centre linguistique d'apprentissage accéléré* (cette affiche est bien réelle), si l'on précisait en sous-titre, comme au commencement des chapitres dans les livres anciens: *"où l'on apprend les langues par une méthode accélérée"*. Les publicitaires les mieux avertis flairent l'impasse et l'on a vu récemment un livre à succès prendre pour titre une longue phrase où la surprise syntaxique s'ajoute au paradoxe sémantique: *"Ne dites pas à ma mère que je suis dans la publicité: elle me croit gardien de nuit dans un bordel"*. Pourrait-on dire la même chose en supprimant les verbes et les personnes?

Osera-t-on, après Hugo, parodier la bible? Le verbe c'est la chair, c'est le goût, c'est la variété des nuances et des modalités. Les chiffres bien entendu n'autorisent pas ce genre de plaidoyer. Mais ils inquiètent et rassurent à la fois. Ce qu'on peut craindre, c'est que l'évolution s'engage plus avant dans la voie dangereuse qu'on a constatée sur plusieurs siècles. Mais on peut penser aussi que l'évolution n'est pas nécessairement linéaire; de Rabelais à Malherbe et de Voltaire à Céline le chemin va en serpentant. Au reste l'usage utilitaire et, pensons-nous, un peu dégradé du français, a la vie courte des choses destinées à la poubelle, comme le papier biodégradable qui le véhicule. Le français

spontané, le français parlé, qui est fait de rupture, de surprise et d'expressivité offre plus de garanties, d'autant qu'une alliance avec le français littéraire tend à s'affirmer de nos jours. Sur d'autres terrains on voit aussi artisans et artistes s'opposer aux industriels.

Ce qui peut rassurer enfin sur l'avenir du français, c'est précisément sa défaite face à l'anglais. Si le français avait gagné le marché des échanges linguistiques, il aurait perdu ses difficultés, ses conjugaisons, ses subtilités et ses nuances. Le monde aurait parlé petit nègre, petit français. Faut-il le regretter?