Changements et variations quantifiables dans la langue française

Etienne Brunet

Changements et variations quantifiables dans la langue française

Étienne BRUNET
Institut de linguistique française (CNRS)

Un linguiste a l'oreille aux aguets. S'il se promène dans la campagne angevine qui nous entoure, il peut encore surprendre des tournures anciennes et patoisantes, même *icitte* (les québécois se retrouvent chez eux), même *anuit*, c'est à dire aujourd'hui. Au besoin Pierre Réseau , qui est comme moi un régional de l'étape, pourra lui servir d'interprète. Mais le linguiste entendra plus souvent, ici comme ailleurs, des mots nouveaux, que la mode et l'actualité introduisent comme une marée lente dans l'usage et que les dictionnaires finissent par officialiser si leur emploi perdure. La nouvelle édition du Larousse vient d'accepter, parmi les entrants, la *séniorie* belge (une seigneurie résidentielle pour personnes âgées) et la proposition québécoise *clavarder* (bavarder à l'aide du clavier) qui vaut assurément mieux que le *chat* anglais. Quel dommage que le Robert ait préféré dans sa nouvelle édition le *mail* anglais au *courriel* adopté avec raison par le Québec! Ces changements linguistiques sont pour une part inconscients. Emporté par le courant, on perd de vue la rive où les siècles déposent leurs alluvions et où les mots roulés par le flot finissent par s'échouer. L'attention se porte plus volontiers aux apports, aux affluents du fleuve, et souvent à l'invasion redoutée des anglicismes. La conscience est moins sensible aux pertes lexicales et par exemple les dictionnaires restent évasifs sur leurs abandons, auxquels ils doivent consentir pour ne pas gonfler démesurément le volume. S'il s'agit de rendre compte des mouvements du lexique, et de tenir le registre des acquis et des pertes, on pourrait prolonger l'étude faite en 1960 par Dubois pour la première moitié du 20e siècle[1]. Mais pour cette enquête, nul n'est mieux placé que Jean Pruvost, qui doit prendre la parole dans un instant.

[1] J. Dubois, L. Guilbert, H. Mitterand, J. Pignon, "Le mouvement général du vocabulaire français de 1906 à 1960", le *Français moderne,* avril et juillet 1960, pp. 86-107 et 196-211.

Je placerai plutôt mon observatoire au milieu des échanges linguistiques, au niveau des textes ou des communications. De telles données sont nécessairement brutes et partielles, au lieu que celles qu'on extrait des dictionnaires sont épurées et complétées par le choix, subjectif mais collectif, des lexicographes. Même en traitant d'immenses corpus, on n'atteint jamais l'exhaustivité du lexique: il y aura toujours des recoins inexplorés, des régions inaccessibles, des limbes vagues où flottent les mots qui attendent le baptême. Et inversement les textes produisent des mots avortés, vaines tentatives que la langue abandonne sitôt la naissance. Le vocabulaire qu'on atteint à travers un corpus, n'est qu'un échantillon de la nomenclature du dictionnaire, lequel à son tour n'épouse le mouvement du lexique que d'une façon asymptotique, dans une poursuite fuyante qui s'épuise à suivre, plus mollement encore, la trajectoire incertaine de la langue.

Toutefois quand les instruments de mesure sont installés au coeur des textes, l'observateur a dans son champ de vision des objets qui ne se limitent pas à la composante lexicale de la langue. Les phénomènes relevés peuvent appartenir à la syntaxe, ou à la sémantique, à la situation de communication, voire à l'évolution des sensibilités et des mentalités, et en fin de compte au mouvement de l'histoire. C'est à de telles observations que nous nous emploierons au cours de cette étude.

La base *Frantext*

On ne fera pas la revue détaillée des données ni des instruments. Il nous suffira de renvoyer le lecteur à Charles Muller pour l'exposé des méthodes statistiques[2]. Les données sont celles auxquelles a accès la communauté scientifique depuis que *Frantext* est accessible sur Internet. Rappelons que celle base textuelle a servi de réservoir inépuisable aux rédacteurs du *Trésor de la Langue Française* et que, depuis l'achèvement de ce dictionnaire, elle alimente en exemples ou attestations les recherches qui se poursuivent à tous les coins du globe sur la langue et la littérature françaises[3]. Chacune des 500 000 graphies, rencontrées tout au long de 3000 textes dépouillés, est instantanément restituée dans son contexte. Certes les ressources textuelles virtuellement disponibles sur le Web sont d'une autre ampleur, même si l'on se contente de la part exiguë réservée au français (moins de 5%). Mais les milliards de mots que les moteurs de recherche explorent hâtivement ne sont qu'un amas énorme mais informe, où l'image de la langue française apparaît nécessairement déformée. Les 200 millions de mots rassemblés dans Frantext ne doivent rien au hasard: le catalogue des textes recensés est le fruit d'une réflexion collective, qui vaut aux

[2] Charles Muller, *Initiation à la statistique linguistique*, Larousse, 1968, ouvrage réédité, en deux volumes, chez Hachette, puis chez Champion (collection Unichamp).

[3] Frantext a servi de modèle à des entreprises similaires nées dans les pays francophones, en Belgique (Beltext), en Suisse (Suistext), au Québec (Québétext).

données une valeur ajoutée, faite de représentativité, d'homogénéité et de disponibilité. Bien peu de langues au monde jouissent d'un tel outil.

La base *Chrono* extraite de *Frantext*

Cependant pour donner plus d'homogénéité encore au corpus, nous avons préféré écarter les textes techniques et réduire le champ d'observation à l'espace littéraire, soit 117 millions de mots, répartis dans 2000 textes environ. Cette première base extraite de Frantext porte de nom de CHRONO. C'est dire qu'elle est orientée vers l'étude diachronique, le corpus étant divisé en 12 tranches, échelonnées de 1500 à 2000. L'espace entre les tranches est inégal, qu'il s'agisse du nombre d'années ou du nombre de mots. Mais des calculs de pondération permettent la comparaison et la mise en relief des changements intervenus en cinq siècles de production littéraire.

L'accroissement lexical

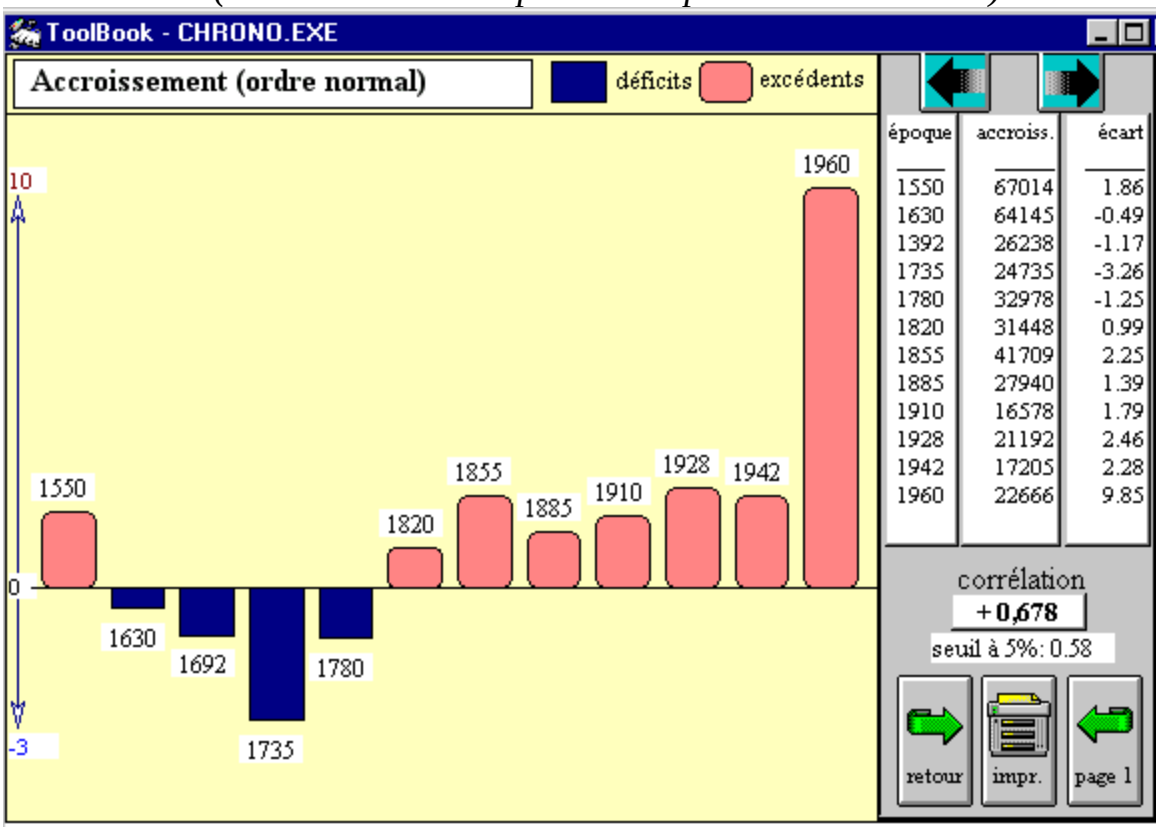
Personne ne s'étonnera que cette production aille croissant. Nous voulons parler ici non du volume des textes, mais de la richesse du vocabulaire, dont rien ne semble devoir arrêter l'inflation ou la créativité. Nous ne voulons pas répéter ici ce que nous avons expliqué jadis dans notre *Vocabulaire français*[4], et plus récemment dans la *Nouvelle histoire de la langue française*[5]. Observons seulement à l'aide du graphique ci-dessous que les échanges linguistiques obéissent aux mêmes forces que les échanges monétaires: la monnaie des mots s'use, des besoins nouveaux apparaissent, surtout en matière de terminologie, et la faveur de la mode va à ce qui est neuf. La planche à mots est donc sollicitée comme la planche à billets. La mesure peut être statique, si l'on compte dans chaque tranche l'étendue du vocabulaire (ou nombre de mots différents qu'on y trouve). Mais elle peut être aussi dynamique, si l'on recense à chaque étape du parcours le nombre de mots nouveaux, non encore rencontrés dans les tranches précédentes. Bien sûr le stock disponible n'étant pas inépuisable, l'accroissement lexical devrait s'épuiser au fur et à mesure que s'entassent les textes dépouillés. La collecte est en effet à la baisse (voir colonne "accroiss." du tableau). Mais ce mouvement est freiné par une force contraire, qui pousse au renouvellement accéléré et que l'histogramme reproduit de façon claire. La diagonale régulière qui monte de gauche à droite rend compte de l'intensité du flux des mots nouveaux, très supérieur au débit attendu. Le fleuve, faisant fi des rives et des digues, s'élargit. Précisons qu'il s'agit ici des graphies, ce qui explique l'anomalie constatée dans les premières tranches. On est tenté d'y voir l'invention hardie dont font preuve les écrivains de la Renaissance, Rabelais à leur tête, et à

[4] Etienne Brunet, *Le Vocabulaire français de 1789 à nos jours*, 3 tomes, Slatkine-Champion, 1981

[5] Jacques Chaurand (sous la direction de), *Nouvelle histoire de la langue française*, Seuil, 1999.

laquelle Malherbe aurait mis un terme pour canaliser la langue dans un classicisme plus sobre et plus sévère. En réalité la variété de la langue au XVIe siècle est aussi orthographique et cela perturbe la comparaison: quand l'évêque apparaît dans les textes de cette époque, il peut revêtir jusqu'à neuf habits différents, suivant la distribution des accents, alors que l'orthographe stabilisée lui impose le même uniforme sacerdotal. Si les données avaient été lemmatisées, comme l'on fait pour les dictionnaires, on peut parier que le XVIe siècle aurait perdu une part de son originalité lexicale.

Figure 1. L'accroissement du vocabulaire, au cours de cinq siècles (les tranches sont représentées par l'année médiane)

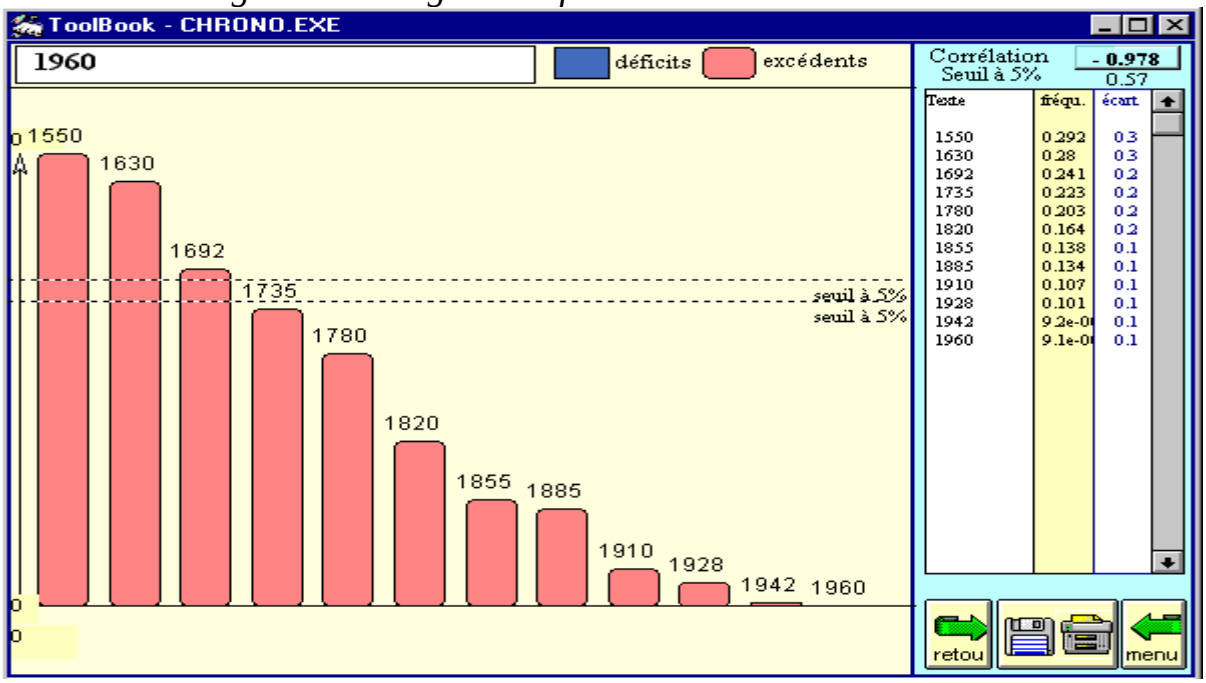


La mesure de l'accroissement devrait être complétée par celle des allégements, chaque tranche se distinguant ce celle qui précède non seulement par ce qu'elle gagne, mais aussi par ce qu'elle perd. Pour ce faire, il suffit de lire le corpus à rebours, en commençant par la dernière tranche. Le calcul a été fait et confirme les résultats acquis dans la perspective chronologique. Pour mettre en relief le processus évolutif, procédons à une méthode voisine qui calcule la distance lexicale qui s'établit entre chaque tranche et toutes les autres. La démarche la plus simple consiste à considérer pour chaque couple de textes (ou ici de tranches) l'effectif des mots communs aux deux textes, et l'effectif de ceux qui sont privatifs et n'appartiennent qu'à l'un des deux. La formule qui met en rapport ces deux quantités donne un résultat plus proche de 0, quand la distance est faible, et de 2 quand l'hétérogénéité s'accentue. La seconde méthode fait intervenir la fréquence respective d'un même mot dans les deux textes, une

répartition équilibrée tendant à rapprocher les textes tandis qu'un partage léonin contribue à les éloigner. Une fois réalisés les calculs - qui doivent être réitérés pour chaque mot dans chaque couple de textes - on obtient un tableau triangulaire, qui ressemble à ceux que fournissent les atlas et où l'on peut apprendre, au croisement d'une ligne et d'une colonne, quelle distance kilométrique ou lexicométrique sépare une époque - ou une ville - d'une autre.

Partons par exemple de la dernière tranche représentée dans l'histogramme ci-dessous (figure 2). La distance est faible lorsque l'époque est voisine (tranches 1942, 1928, et 1910) et s'accroît régulièrement quand l'éloignement dans le temps se fait sentir.

Figure 2. Histogramme focalisant la dernière tranche



La distance lexicale

L'influence du temps est encore plus sensible dans la figure 3, qui rend compte de toutes les tranches à la fois et leur assigne la place exacte que la chronologie leur accorde par ailleurs. Rappelons que l'ordinateur est tenu dans l'ignorance de l'ordre chronologique et que le classement auquel il parvient dérive de l'emplacement des 117 millions d'occurrences analysées. L'analyse arborée qui est ici employée dessine le chemin qui relie les tranches entre elles et qui suit docilement la série chronologique. L'analyse factorielle, représentée dans la figure 4, est fondée sur les mêmes données et met en relief pareillement la suite ordonnée des époques selon un parcours courbe, qui relie dans l'ordre les maillons de la chaîne chronologique.

L'évolution du lexique ne laisse donc place à aucun doute, ni dans l'analyse délivrée par la machine, ni dans la conscience des usagers. L'analyse quantitative fournit toutefois sur le rythme de l'évolution une information que l'intuition seule ne remarquerait peut-être pas. Dans les deux graphiques les

Figure 3. Analyse arborée de la distance lexicale

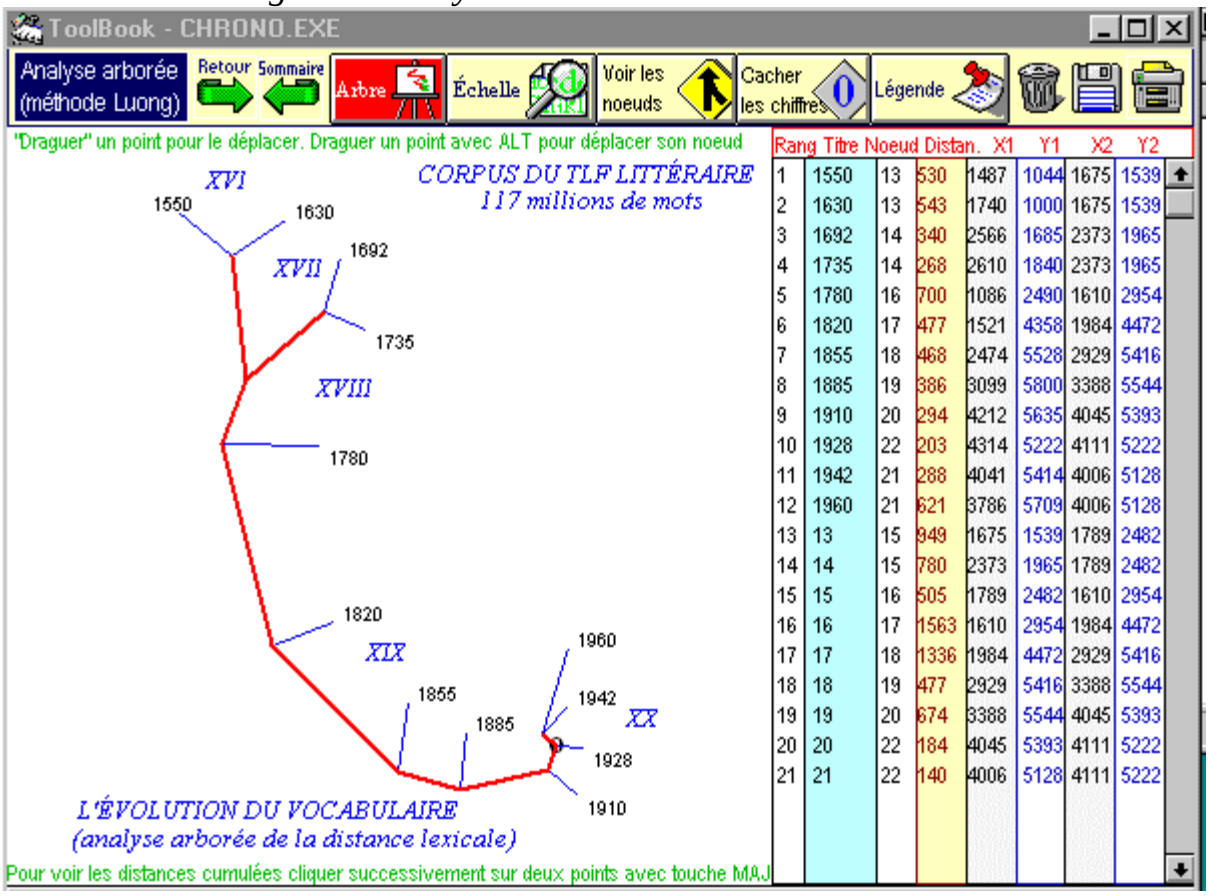
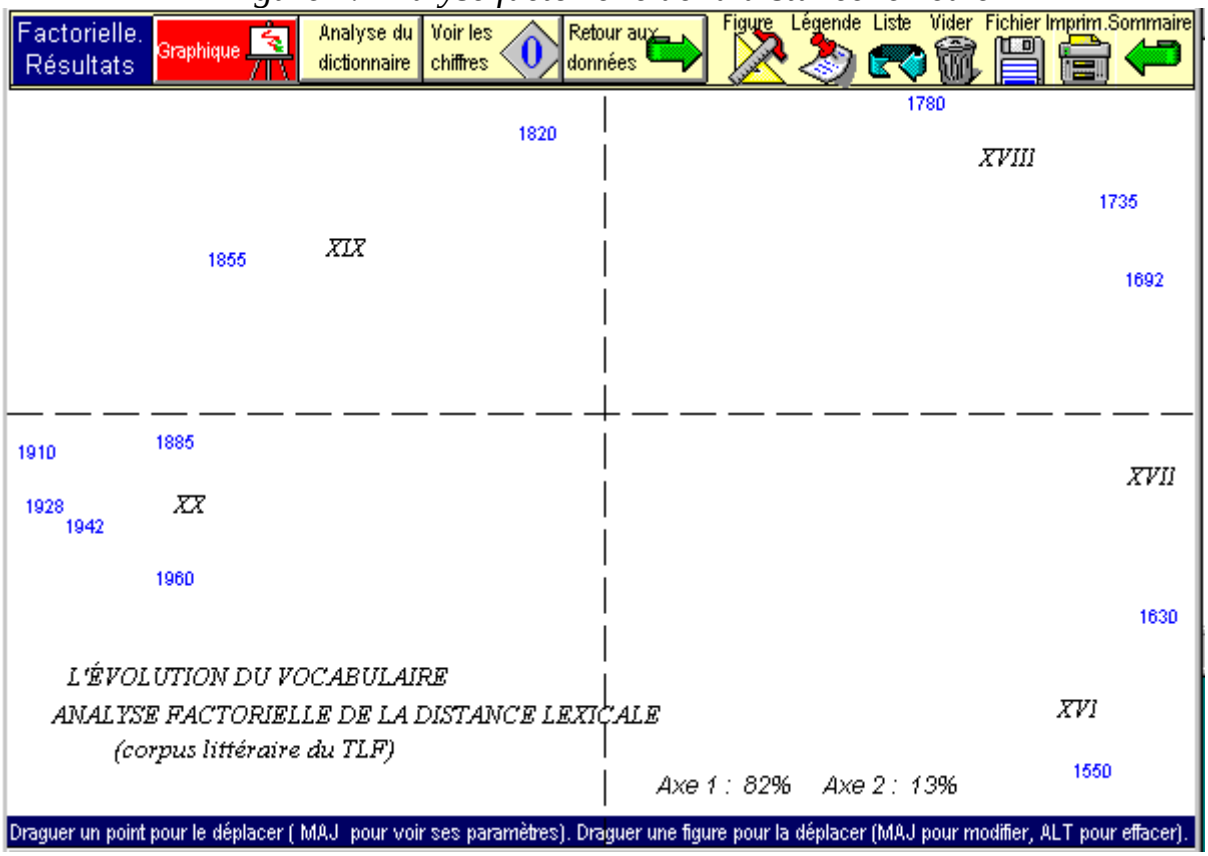
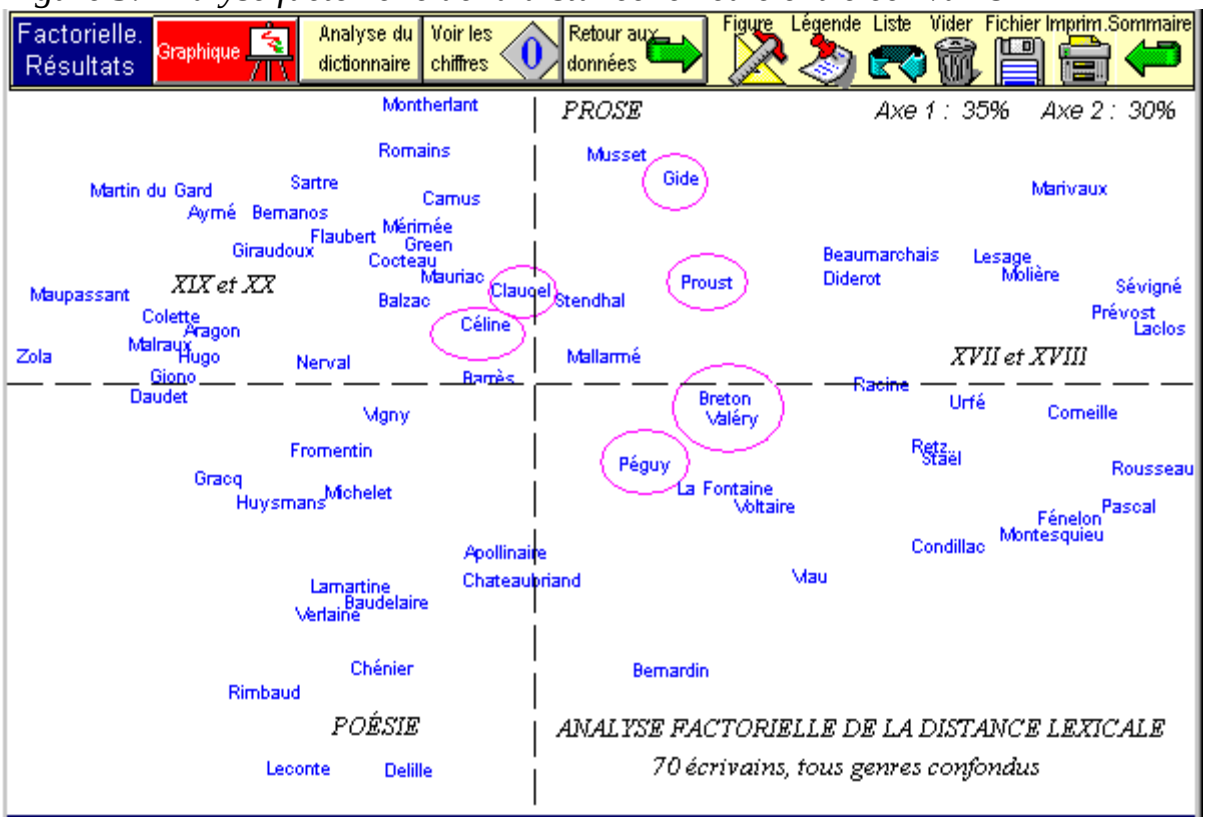


Figure 4. Analyse factorielle de la distance lexicale



distances se raccourcissent comme si le lexique tendait à se stabiliser. Certes l'orthographe s'est en effet figée dans la première moitié du XIXe siècle et les variantes d'un même mot ne peuvent plus contribuer à l'enrichissement fictif du vocabulaire. Mais ce ralentissement de la créativité, qui dépasse les questions d'orthographe, ne laisse pas de surprendre quand on est habitué à constater dans d'autres domaines l'accélération de l'histoire. Tout se passe comme si les écrivains se souciaient davantage d'exploiter le terrain conquis plutôt que de l'étendre. L'explication réside peut-être, au moins partiellement, dans la composition du corpus que l'on voulait représentatif du français littéraire et de la langue soutenue. L'évolution semblerait sans doute plus rapide si l'on avait fait appel au salon des refusés. Mais on peut penser aussi qu'un certain conservatisme préside au destin de la langue et qu'il est dû peut-être à la frilosité jalouse des usagers plutôt qu'aux directives officielles, comme le montre l'échec relatif de la dernière réforme de l'orthographe.

Figure 5. Analyse factorielle de la distance lexicale entre écrivains



La base *Écrivains*

La composition du corpus du TLF n'a pas eu pour souci premier de réaliser un échantillonnage raisonné des tranches chronologiques, dont nous avons posé les jalons a posteriori, non sans quelque arbitraire. Le choix initial s'est exercé, non sur les époques, mais sur les textes et leurs auteurs, en s'entourant de quelques garanties d'impartialité - un peu illusoires - pour retenir les meilleurs de notre littérature. Quoi qu'il en soit de ce choix, il peut donner lieu à une seconde base où seront mis en parallèle les écrivains les mieux représentés. Comme ces écrivains se répartissent tout au long de la chaîne chronologique,

c'est là un moyen, indirect, de mesurer les variations du temps. La même mesure de distance lexicale peut être appliquée aux 70 écrivains retenus. Le résultat est dans l'analyse factorielle de la figure 5. L'interprétation de tels graphes est rendue aisée dès qu'on a compris qu'il s'agit approximativement d'une carte géographique[6]. Les écrivains qui sont voisins sur la carte partagent un lexique commun et sans doute aussi quelques autres propriétés. Or les regroupements ne se font pas par affinités mais par périodes, les écrivains du XVIe et du XVIIe siècles s'orientent à droite, tandis que le XIXe et l'époque contemporaine investissent la moitié gauche du graphique. Il serait piquant de comparer la sagacité de l'ordinateur à celle d'un lecteur moyennement cultivé qui aurait à répartir les 70 écrivains sur l'échiquier. Il n'est pas certain que la machine aurait le dessous.

Les genres

L'ordinateur à qui la biographie des écrivains est tenue cachée ignore également tout des genres littéraires. Il en repère pourtant la trace dans le vocabulaire, puisqu'il range de préférence les textes poétiques dans la moitié basse, en réservant à la prose la moitié supérieure. Si la décantation n'est pas aussi claire qu'on le souhaiterait c'est que beaucoup d'écrivains, comme Hugo, se sont illustrés dans plusieurs genres. Les solidarités sont plus fortes quand les écrivains partagent tout ensemble le même genre et une époque voisine (comme c'est le cas pour le théâtre de Molière, Marivaux et Beaumarchais, situés à proximité dans le quadrant supérieur droit). Un écrivain n'échappe-t-il donc pas aux deux mâchoires de la tenaille, au temps et au genre? Certains trouvent le moyen d'affirmer leur indépendance à cet égard. On les retrouve indécis dans le marais central du graphique: si Claudel, Péguy, Breton, Valéry et Céline sont là, c'est parce que les valeurs et les mots qu'il chérissent sont éloignés des choix de l'école réaliste et naturaliste qui fait la loi à gauche, derrière Flaubert, Maupassant et Zola. En réalité c'est bien la chronologie qui explique ce reflux, mais une chronologie non linéaire, faite de ruptures. Et la coupure est nette à la fin du XIXe, quand divers mouvements spiritualistes rejettent la tutelle du positivisme et du naturalisme.

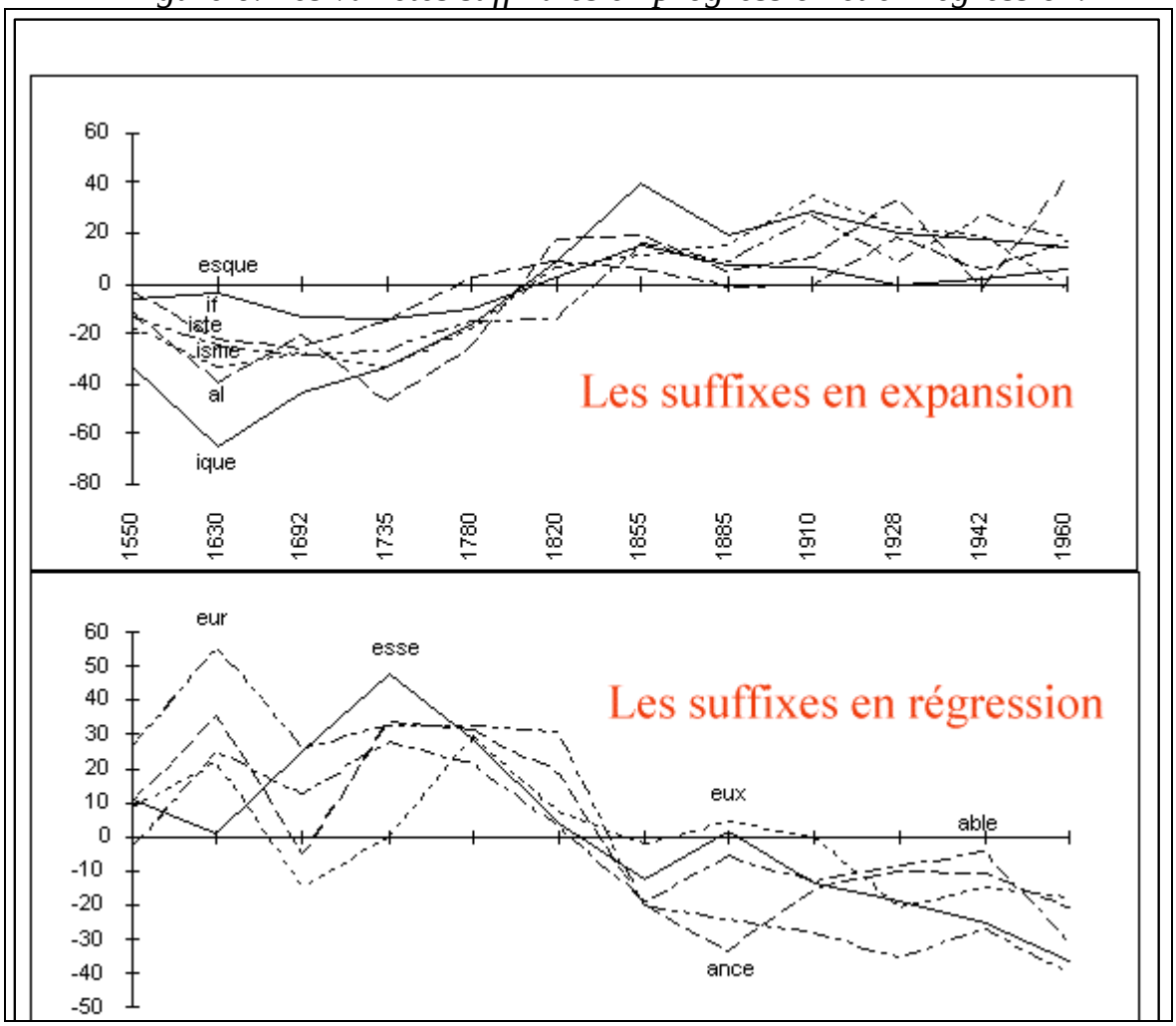
Les suffixes

Reste à analyser les mouvements internes qui animent et grossissent la masse lexicale. La créativité lexicale se manifeste surtout par la suffixation. Il est rare qu'un mot naisse de parents inconnus. Le plus souvent un radical préexistant s'affuble d'un ou plusieurs affixes, généralement un suffixe, pour constituer une nouvelle unité de discours. Or le processus suit des modes changeantes. Parmi une cinquantaine de variétés analysées, retenons celles dont l'emploi est croissant (on les a superposées dans la moitié supérieure de la figure 6) et celles qui sont en régression (partie basse de la même figure). Les variétés

[6] La même analyse factorielle, appliquée à un tableau des distances entre villes d'un même pays, restitue la carte géographique, un peu déformée, de ce pays. La déformation vient de que les distances sont plus courtes en plaine qu'en région montagneuse, ce qui, par exemple, a pour effet de déporter Paris vers le Nord.

suffixales en *esque*[7], *if*, *al* et surtout *iste* et *isme* sont en expansion, tandis que refluent les noms en *-eur* (du type *valeur*, *chaleur*, *longueur*), en *-esse* (agent du type *comtesse* ou qualité du type *tendresse)*, en *-eux* (comme *heureux* ou *honteux*) ou en *-able*.

Figure 6. Les variétés suffixales en progression et en régression.



En réalité on devrait faire une distinction entre la fécondité active de certains moules et la descendance plus ou moins nombreuse de certains autres dont la fertilité est épuisée. Ainsi les mots en *-tude* comme *habitude* peuvent se maintenir fermement dans l'usage même si l'on ne crée plus guère de mots nouveaux pourvus de ce suffixe (il y a pourtant, par analogie, *foultitude*).

Les emprunts

La place nous manque pour caractériser davantage les espèces auxquelles profite l'expansion lexicale et celles aussi qu'elle ignore. La suffixation qui alimente l'essentiel des créations verbales contribue à l'embonpoint croissant du mot. L'inflation prend alors la forme de l'enflure. L'emprunt aux langues

[7] Le célèbre *abracadabrantesque* est venu naguère sur les lèvres du Président de la République non point à cause d'un précédent rimbaldien, mais parce que le suffixe *-esque* était disponible et qu'il s'est présenté le premier à l'esprit dans une occasion où il importait d'amplifier l'invraisemblance.

étrangères pose également à la langue des problèmes de digestion. Ils sont résolus dans notre corpus grâce à la prudence parcimonieuse des écrivains, encore soucieux de la pureté de la langue, et le nombre des anglicismes que nous avons relevés reste bien modique (quelques milliers sur plus de 100 millions de mots). Il est vrai qu'ils sont concentrés dans les deux tranches les plus proches du temps présent. Les digues de protection, qui ont rompu dans la publicité, les médias et la chanson, tiennent encore dans le domaine littéraire. Au niveau infralexical, d'autres observations auraient leur intérêt, comme l'embarras qui frappe l'écrivain devant le tréma (faut-il le mettre sur le *u* ou le *e* de aiguë?) et qui aboutit souvent à l'abstention, ou comme la désaffection qui entoure de plus en plus le circonflexe (cela donne raison à la dernière réforme de l'orthographe, qui le rend facultatif en certains cas).

La corrélation chronologique

Les calculs qui portent sur la richesse ou la distance lexicale aboutissent à des résultats sûrs quant à l'existence de l'évolution mais ils restent peu explicites dès qu'on veut préciser le détail de cette évolution. D'autres outils s'offrent pour l'examen individuel des mots, dont le coefficient de corrélation chronologique. Ce test statistique permet d'isoler les mots qui sont le plus sensibles à l'emprise du temps soit à la hausse, soit à la baisse. Parfois l'explication est triviale: les objets qui sont apparus et se sont développés au cours du temps entraînent dans leur sillage les mots qui les désignent, comme l'*avion*, le *train*, la *cigarette*, le *cinéma*. Il s'agit moins ici de faits de langage que de société. La disparition complémentaire de certains traits attachés aux époques révolues est moins sensible à la conscience et c'est là, dans cette moitié négative, que le coefficient de corrélation apporte les enseignements les plus utiles. En s'en tenant aux seuls substantifs, on voit qu'un certain langage noble et poétique, en usage dans le vers classique, est tombé en désuétude. Si l'humour ou la parodie ne se mêlent pas à ses intentions, un écrivain moderne trouvera difficilement l'emploi de certains mots au charme suranné que signale le coefficient: *alarmes, courroux, muses, joug, vaisseaux, ornement, serment, oracles, temples, enfers, hymen, langueur*. De même les termes qui désignent les sentiments, les valeurs morales ou les facultés mentales, et dont le roman psychologique a fait grand usage, se raréfient au cours des siècles. Certes l'ostracisme dont ils ont été victimes au temps du naturalisme explique en partie cette désaffection, mais le déclin se prolonge bien au-delà de Zola, comme si les valeurs morales, trop liées à la religion et à l'ordre établi, avaient suivi le sort du roi. Voir tableau 7.

Tableau 7. Liste partielle des substantifs en régression

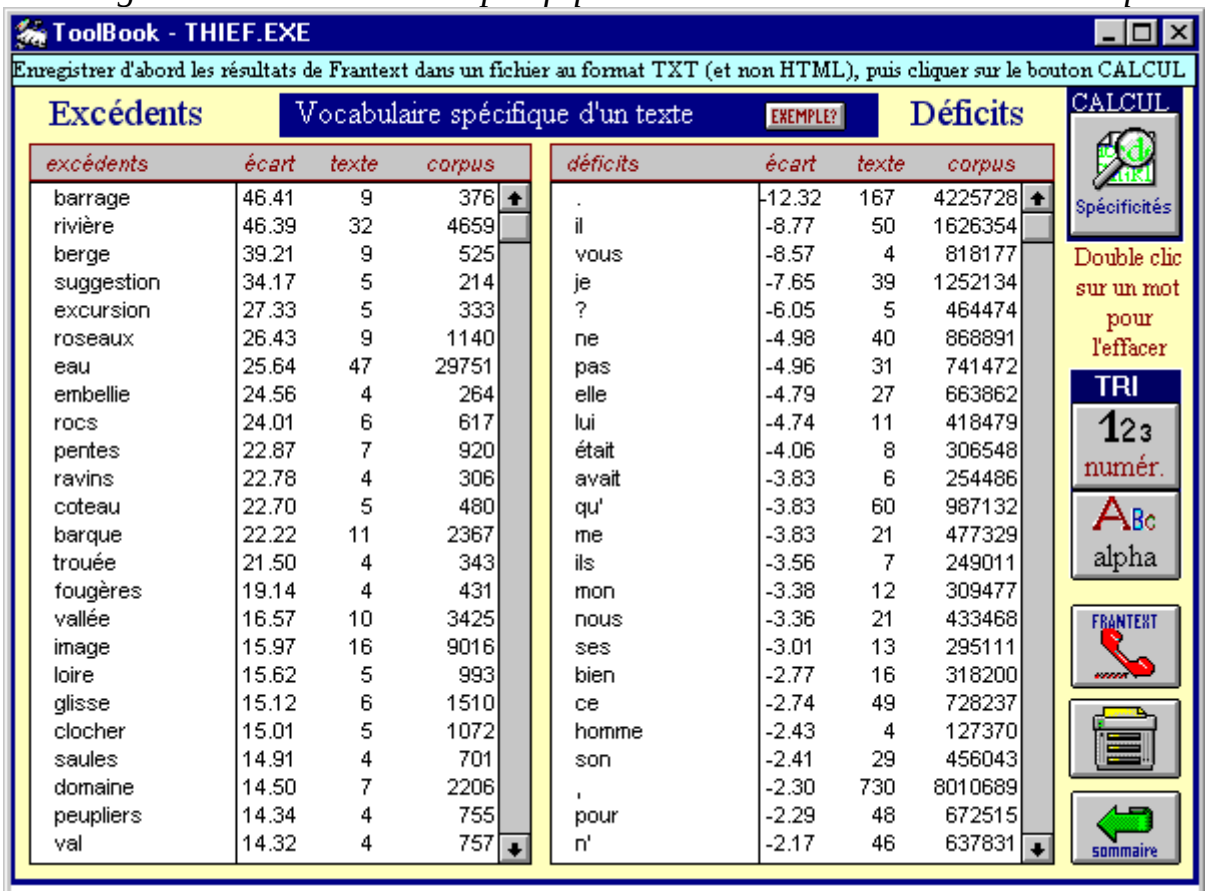
honneur	honneurs	frayeur	froideur	coeur	coeurs
malheur	querelle	repentir	offense	affront	vengeance
injures	ambition	fureurs	punition	désespoir	audace
gloire	bonté	prudence	crainte	secours	danger
esprit	âme	raison	conseil	finesse	douceur
amour	amours	passion	jalousie	douleurs	dispute

Si l'on s'adresse à Frantext et qu'on l'interroge sur un mot ou une famille de mots, une fonction spéciale du logiciel permet de saisir les variations du succès ou de l'oubli qui frappe au cours du temps l'objet isolé. Par exemple la Loire qui coule à nos pieds se répand à travers les Lettres françaises, avec des crues et des décrues, et une progression d'ensemble qui se manifeste surtout dans la dernière tranche et à laquelle Gracq n'est pas étranger.

Les spécificités

Le second outil est le plus classique dans le monde de la statistique linguistique. C'est le calcul des spécificités. Comme Gracq est notre voisin immédiat, c'est à lui que nous emprunterons l'illustration de la figure 8. Le texte analysé prend pour objet une humble rivière, l'Èvre, qui embrasse les lieux de ce colloque, de Beaupreau à Liré, et que Gracq évoque dans les *Eaux étroites*. La liste fournie par le calcul dans la colonne de gauche est suffisamment évocatrice pour rendre inutiles les commentaires. Les mots qui dans ce texte ont une fréquence inhabituelle sont en relation directe avec l'eau et la promenade en barque. À travers ces mots-clés, c'est presque un résumé que propose la machine.

Figure 8. Le vocabulaire spécifique des Eaux étroites *de Julien Gracq*



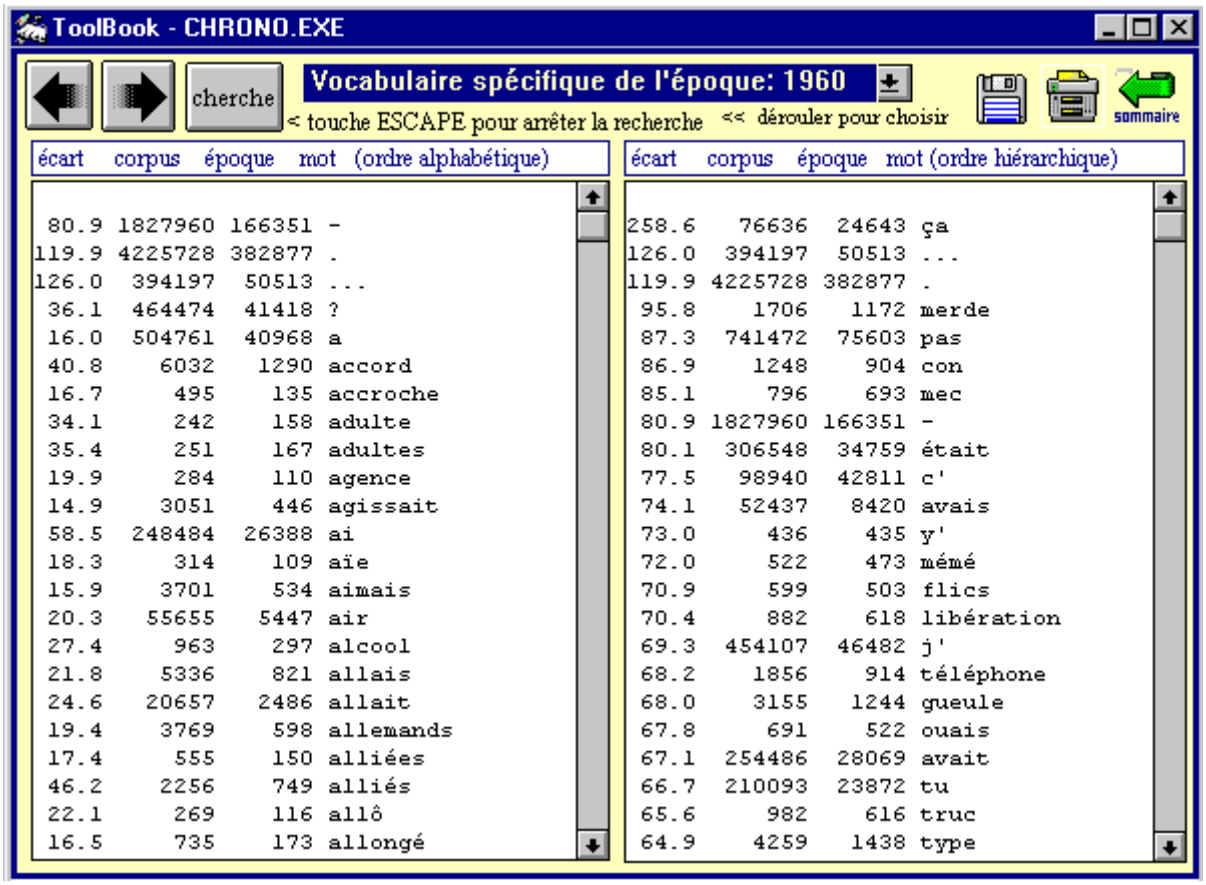
excédents	écart	texte	corpus
barrage	46.41	9	376
rivière	46.39	32	4659
berge	39.21	9	525
suggestion	34.17	5	214
excursion	27.33	5	333
roseaux	26.43	9	1140
eau	25.64	47	29751
embellie	24.56	4	264
rocs	24.01	6	617
pentes	22.87	7	920
ravins	22.78	4	306
coteau	22.70	5	480
barque	22.22	11	2367
trouée	21.50	4	343
fougères	19.14	4	431
vallée	16.57	10	3425
image	15.97	16	9016
loire	15.62	5	993
glisse	15.12	6	1510
clocher	15.01	5	1072
saules	14.91	4	701
domaine	14.50	7	2206
peupliers	14.34	4	755
val	14.32	4	757

déficits	écart	texte	corpus
.	-12.32	167	4225728
il	-8.77	50	1626354
vous	-8.57	4	818177
je	-7.65	39	1252134
?	-6.05	5	464474
ne	-4.98	40	868891
pas	-4.96	31	741472
elle	-4.79	27	663862
lui	-4.74	11	418479
était	-4.06	8	306548
avait	-3.83	6	254486
qu'	-3.83	60	987132
me	-3.83	21	477329
ils	-3.56	7	249011
mon	-3.38	12	309477
nous	-3.36	21	433468
ses	-3.01	13	295111
bien	-2.77	16	318200
ce	-2.74	49	728237
homme	-2.43	4	127370
son	-2.41	29	456043
,	-2.30	730	8010689
pour	-2.29	48	672515
n'	-2.17	46	637831

Le même procédé peut s'appliquer à un ensemble de textes et par exemple à ceux qui sont regroupés dans la dernière tranche. On obtient alors les traits significatifs qui s'attachent à l'époque contemporaine et qui permettent une extrapolation point trop imprudente sur l'avenir de la langue et de la littérature

françaises. L'image que la figure 9 en donne n'est rien moins que flatteuse. Dans la galerie de portraits qu'on pourrait dresser en enfilade pour reproduire la généalogie du français, le dernier rejeton apparaît comme un avorton indigne de la lignée. Un vocabulaire familier, voire ordurier, émerge de la liste comme une écume sale. La liste est fort longue et nous n'en donnons qu'un extrait mais cela suffit pour constater la vulgarité des propos (*con, mec, flics, mémé, gueule truc, type*, etc..).

Figure 9. Les spécificités de la dernière tranche
(ordre alphabétique à gauche, ordre hiérarchique à droite)



Le premier mot-plein de la liste est le mot de Cambronne qui est un vieux mot, apprécié par Rabelais, et dont la distribution dans les 12 tranches est la suivante:

1550	1630	1692	1735	1780	1820	1855	1885	1910	1928	1942	1960	total
39	26	0	1	2	6	58	108	38	50	312	1334	1974

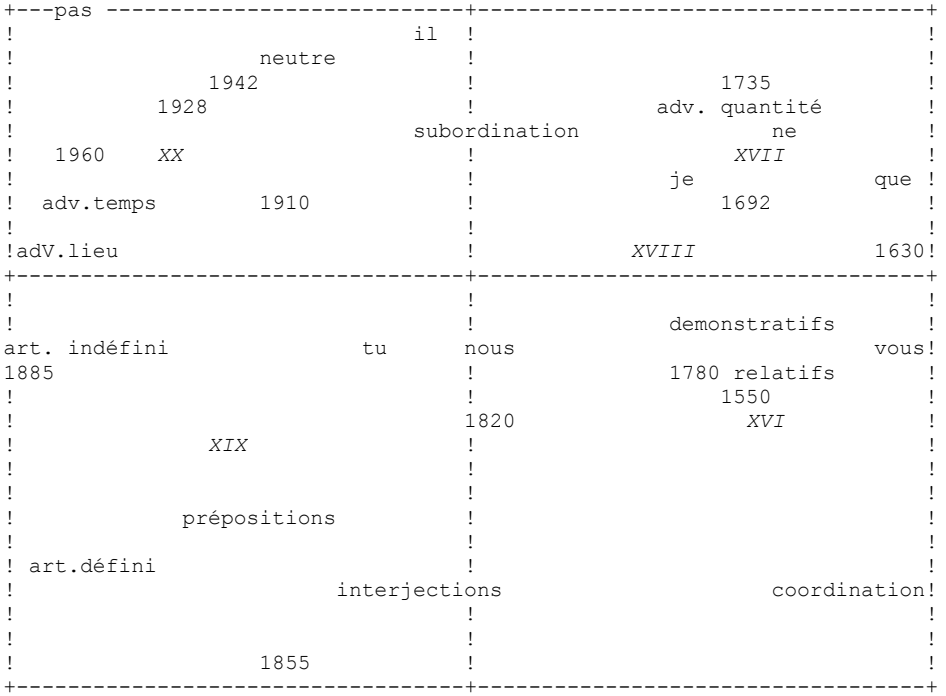
Nul besoin de pondération: en accaparant les deux tiers des occurrences l'époque moderne affirme son goût. Cela est dû à Céline (148 occurrences) et à Sartre (61), mais aussi au Queneau de Zazie dans le métro et à bien d'autres écrivains, moins soucieux que leurs aînés de beau langage. Le roman policier et le roman populaire ne sont plus victimes d'ostracisme et l'on y accueille Simonin (*Touchez pas au grisbi*), Japrisot, Vautrin, Chabrol et jusqu'aux *Valseuses* de B. Blier. Reste à savoir si ce ton populaire tient à l'évolution de la langue ou à la composition du corpus. La seconde hypothèse est la plus prudente. Il se trouve en effet que le corpus a été constitué et les auteurs choisis pour fournir des attestations à un dictionnaire de la langue soutenue - où les expressions

argotiques avaient peu de place. Or après l'achèvement du TLF, les données ont continué d'affluer, surtout aux deux bouts de la chaîne: au XVIe siècle et à l'époque récente, mais avec un objectif différent, qui n'était plus associé à la rédaction d'un dictionnaire. Plus besoin d'un filtre réducteur ni de passeport. Le succès de librairie devenait le seul critère. Cela certes n'offre pas nécessairement la garantie de la qualité et de la pérennité. Mais la représentativité est mieux assurée, donnant une image plus fidèle de la production et du goût du public.

Les mots grammaticaux

Dans la trop courte liste de la figure 9, des détails apparaissent qui ne relèvent pas du seul lexique. Des mots comme *ça* et *c', j'* et *tu, était* et *avais*, donnent des indications sur la syntaxe et les caractéristiques stylistiques de l'usage contemporain qui est favorable au neutre, au tutoiement, à l'emploi narratif de l'imparfait. L'étude systématique des mots grammaticaux permet une approche, certes indirecte, mais révélatrice de la tectonique du discours. Au total l'analyse factorielle réalisée à partir des classes ainsi constituées (figure 10) 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 XX^{e} siècle.

Figure 10. Analyse factorielle des mots grammaticaux.



Les verbes

Malheureusement, en l'absence de codage grammatical[8], il n'est guère possible d'isoler les catégories ouvertes dans Frantext. Difficile de rendre un compte exact des substantifs, des adjectifs, de la plupart des adverbes, et des verbes. Ces derniers cependant permettent un repérage partiel. Une fonction spéciale de FRANTEXT 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.

Elle est aussi moins variée dans l'expression des modes et des temps. Le simple indice du circonflexe suffit pour illustrer le déclin du passé simple et du subjonctif imparfait, au moins dans les formes où cette graphie apparaît. On peut entreprendre l'étude plus systématique des temps et des modes en éliminant toutefois les formes qui prêtent 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 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 11 et 12.

Les temps verbaux

La figure 11 est réservée aux temps que le temps n'attaque pas. Il 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

[8] Nous n'ignorons pas qu'une grande partie de FRANTEXT a été catégorisée et lemmatisée. Mais l'accès au codes n'est pas encore autorisé pour les calculs de fréquence.

présent, et gagne des parts de marché au XIXe siècle, à partir de Balzac et de Flaubert.

Figure 11. Temps et modes en progression

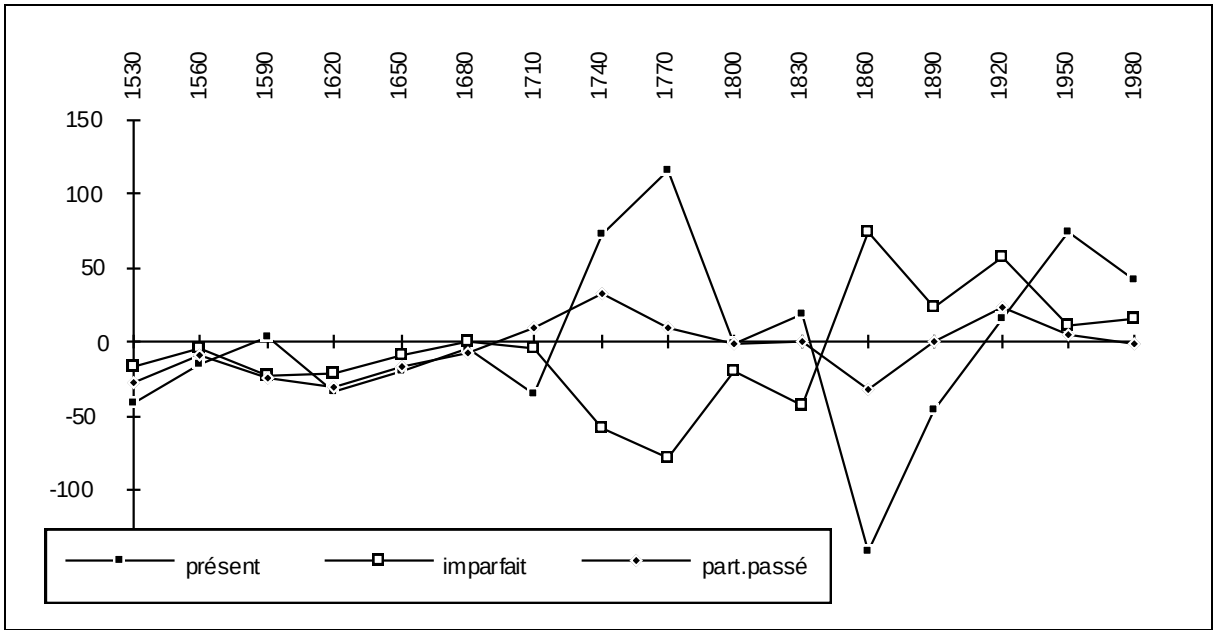
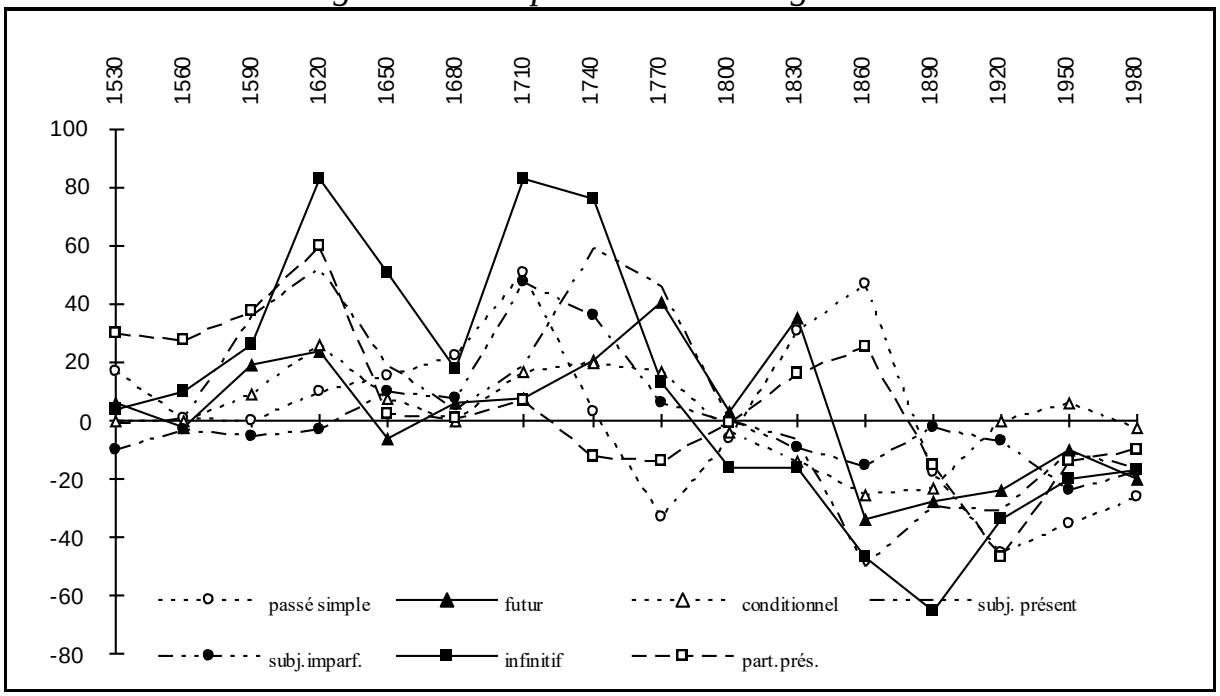


Figure 12. Temps et modes en régression



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 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.

La base *Francil*

Pour en avoir une perception nette, on croit devoir abandonner Frantext et consulter une autre base que nous avons appelée Francil et qui gagne dans l'espace ce qu'elle perd dans le temps. La dimension diachronique y est neutralisée, les textes choisis appartenant tous ou presque à notre temps. Mais l'extension y est recherchée dans l'espace francophone puisqu'à quelques témoins de la littérature française, nous avons ajouté des textes québécois, suisses, belges, maghrébins et africains. Et surtout les usages du français n'y sont pas limités à la littérature, l'oral, la presse et la technique entrant aussi dans la composition d'un corpus hétérogène qui mêle volontairement les lieux, les genres et les auteurs, afin d'observer quelle influence se révèle la plus forte dans les analyses.

Cette base, grosse de quatre millions d'occurrences, gagne aussi en précision ce qu'elle perd en étendue. Chaque forme a été analysée par le logiciel Cordial, et pourvu de codes grammaticaux, sémantiques et fonctionnels. À l'issue de ce premier traitement, cinq indexations successives ont été mises en oeuvre par notre logiciel Hyperbase, pour rendre compte successivement des graphies, des lemmes, des fonctions et des codes grammaticaux, des structures syntaxiques et enfin des champs sémantiques.

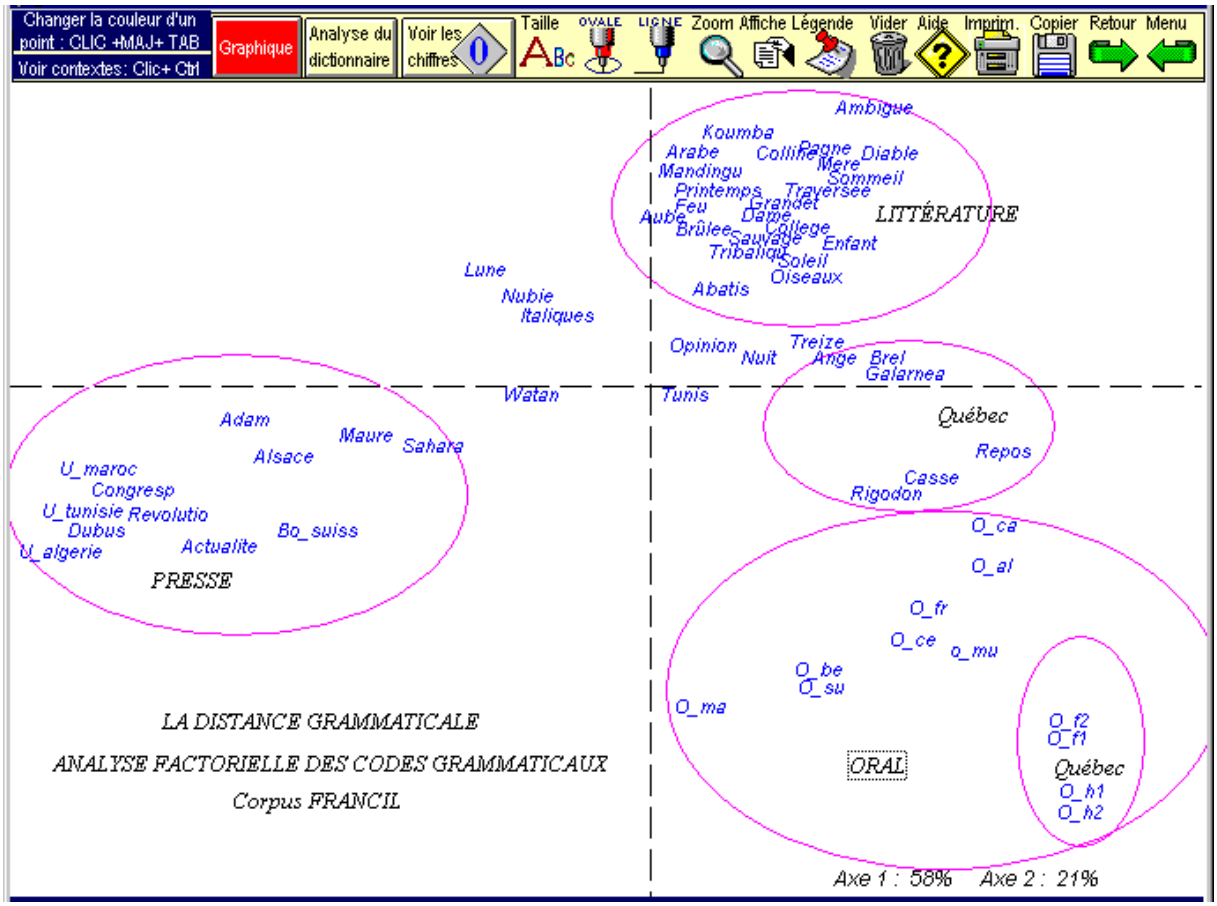
La distance grammaticale

L'analyse factorielle représentée ci-dessous (figure 13) rend compte des quatre millions de codes répertoriées dans le corpus lemmatisé. Ces codes sont très précis et distinguent les homographes. Ainsi le substantif *savons* reçoit l'étiquette *Nc_mp_N* (= nom commun, masculin pluriel, groupe apposition) tandis que le verbe *savons* est représenté par le code *VmiP1pV* (= verbe principal, indicatif présent, première personne du pluriel, base de proposition). Ces codes dont la variété est considérable (plusieurs milliers) sont soumis aux mêmes calculs de distance qu'on a expliqués précédemment et dont la représentation a recours à l'analyse factorielle, comme dans le graphique 4.

L'interprétation est aisée puisque trois groupes fort distincts se partagent l'espace: à l'extrême droite les textes transcrits de l'oral se regroupent sans qu'un seul manque à l'appel (on les reconnaît à leur initiale O_). À l'opposé un groupe compact se forme autour des textes techniques et de la presse. La tension la plus forte est donc celle qui oppose l'oral et la presse et 51% de la variance se résume

à ce duel. La littérature se répartit entre les deux et prend position en haut du graphique, le second facteur (31%) opposant le littéraire et l'utilitaire. Mais l'unanimité ne règne pas parmi les textes littéraires. Certains se rapprochent des textes techniques ou journalistiques, c'est le cas des récits de voyage ou d'aventure (par exemple *Un été au Sahara*, de Fromentin, *ou De la terre à la lune*, de J. Verne). D'autres lorgnent du côté de l'oral, comme *Voyage au bout de la nuit* et *Rigodon* de Céline, les chansons de Brel, ou les romans québécois (*Ange*, *Galarneau*, *Repos* et *Cassé*).

Figure 13. Corpus lemmatisé FRANCIL. La distance grammaticale.



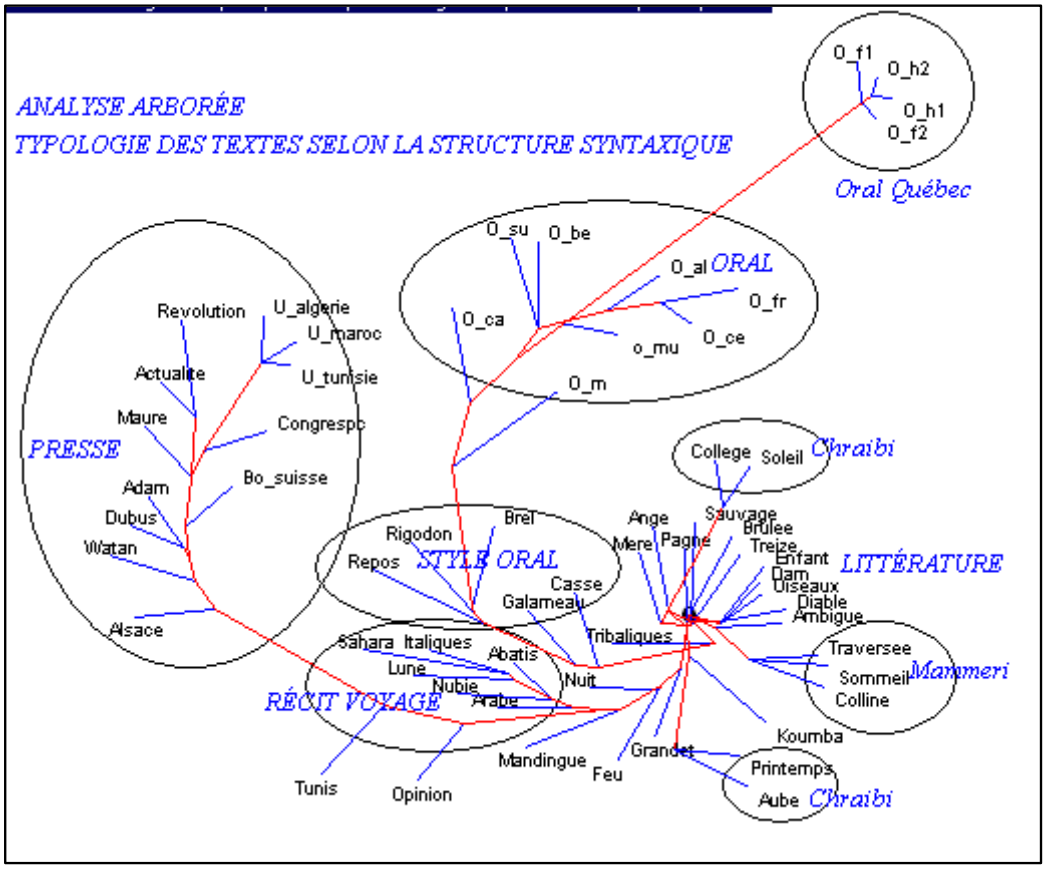
La distance lexicale

Le fait le plus remarquable est que les lignes de force sont les mêmes lorsque la distance est établie à partir des graphies, ou à partir des lemmes (comme ces deux graphiques sont quasi superposables au présent graphique, il n'est pas très utile de les représenter). Rien n'impose pourtant a priori ce parallélisme. Dans le cas des lemmes, les occurrences de *savons*, si c'est un verbe, sont portées au compte du verbes *savoir*. Dans le cas des codes, le même lot d'occurrences sera totalisé avec ceux de *pouvons*, *marchons*, *chantons*. Il n'y a rien de commun dans ces deux principes de regroupement. Rien de commun non plus dans l'objet mesuré: la composante sémantique dans le lemme, et la composante grammaticale dans le code.

La distance syntaxique

Ajoutons, pour faire bonne mesure, que la surdétermination du langage fait apparaître la même aimantation quand la combinatoire syntaxique entre en scène. On peut réduire tout segment du discours (entre deux ponctuations) à une séquence de codes simples symbolisant les parties du discours (ainsi la suite *dnav* représente la structure syntaxique *déterminant+nom+adjectif+verbe*). Ce traitement a été appliqué à l'ensemble du corpus et soumis à une nouvelle indexation et de nouveau à un calcul de distance. Ici seul compte l'ordre des éléments, ce qui ne dépend que faiblement de la proportion dans l'urne des éléments isolés. Et pourtant là encore l'image obtenue (figure 14) montre les mêmes alliances et les mêmes oppositions qu'on a observées dans la figure 13. On y voit deux routes diverger dont l'une conduit à l'oral, et l'autre à la presse. À l'embranchement la littérature hésite, certains textes s'orientant vers l'oral (ce sont les mêmes que dans le graphique 13), d'autres vers les journaux et les essais (là aussi on reconnaît les mêmes éléments).

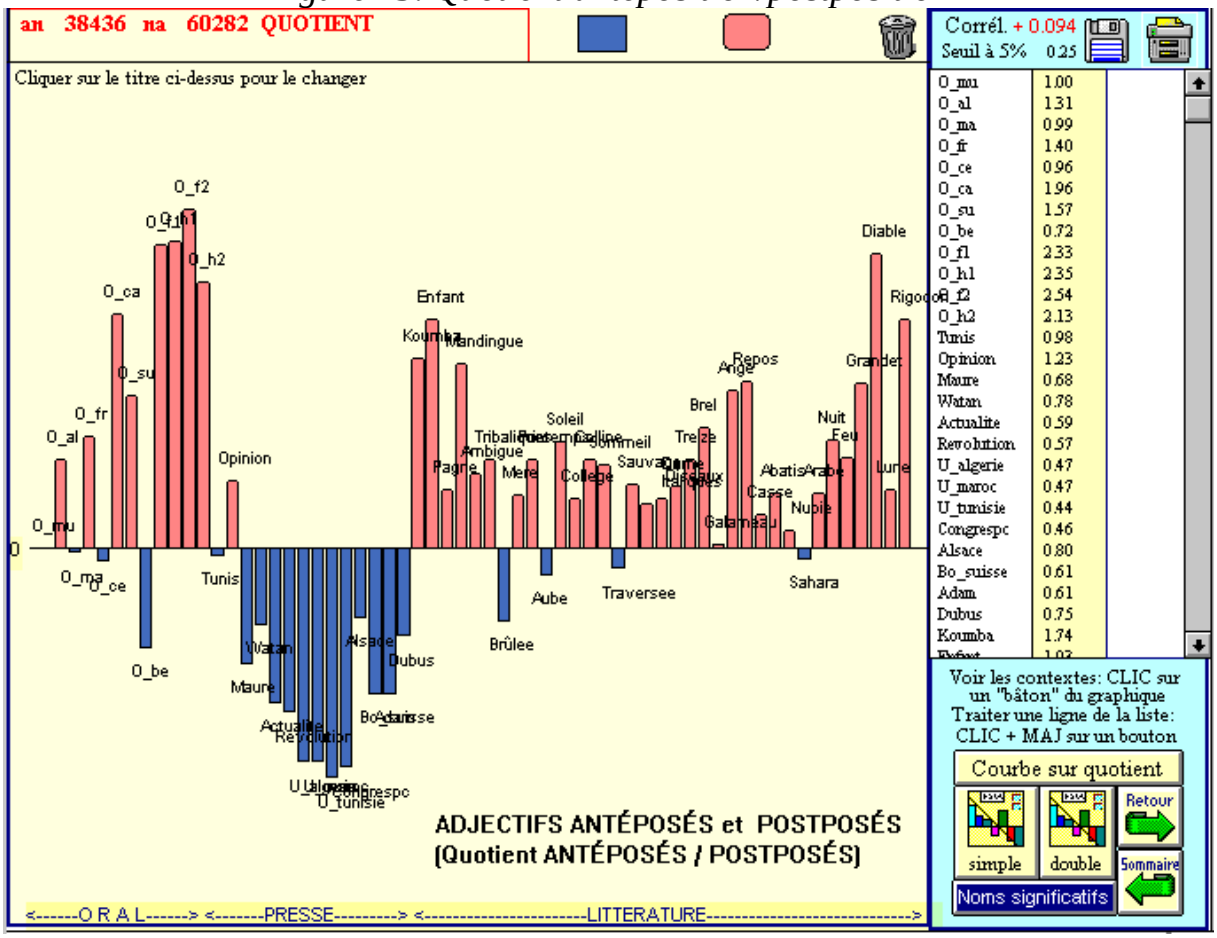
Figure 14 Typologie des textes selon la structure syntaxique



La combinatoire syntaxique est ici extrêmement riche et beaucoup des séquences, quand elles sont longues, n'apparaissent qu'une fois. On peut s'intéresser à des structures partielles à deux ou trois éléments, afin d'obtenir des

effectifs plus importants. Isolons le couple adjectif-substantif qui peut se présenter sous deux formes selon que l'adjectif est antéposé ou postposé. Le rapport entre ces deux séquences est établi dans l'histogramme 15. La décantation des genres ou domaines est particulièrement nette: l'oral et la littérature sont plus favorables à l'antéposition, la presse à la postposition.

Figure 15. Quotient antéposition/postposition

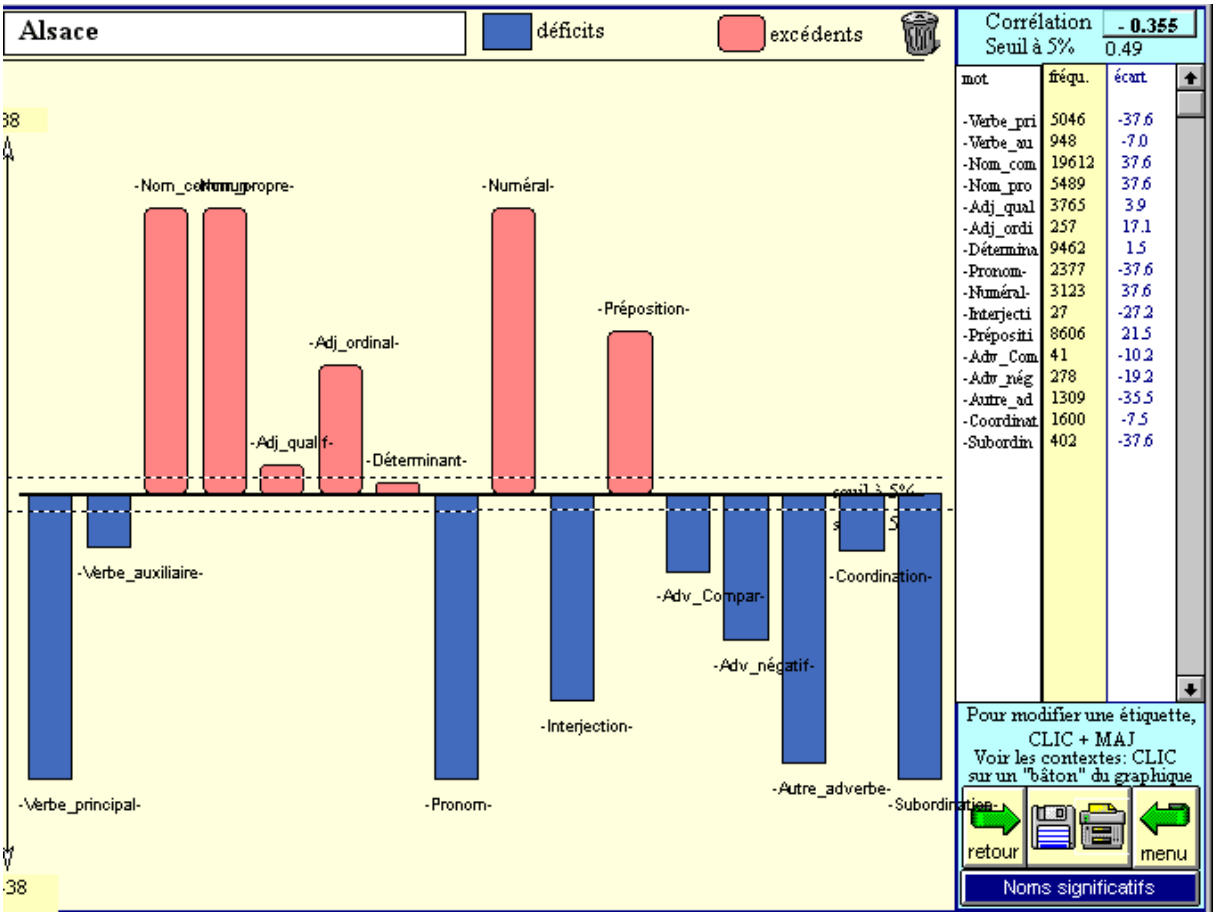


Les parties du discours

Constituons enfin un tableau de contingence où l'on détaillera dans chaque texte la distribution des parties du discours. Un tel tableau se prête à l'analyse, soit qu'on focalise son attention sur une catégorie, soit qu'on dresse le profil grammatical d'un texte (figure 16), soit qu'on soumette l'ensemble des données à l'analyse factorielle (figure 17). Le graphique 16 montre excellemment que le discours de l'information (il s'agit ici d'un numéro complet des *Dernières nouvelles d'Alsace*) fait surtout appel aux classes nominales: les excédents sont considérables tant pour le nom propre que pour le nom commun. Les déterminants et les prépositions ont partie liée avec le substantif et sont pareillement excédentaires. L'adjectif suit le même mouvement, surtout lorsqu'il donne une information chiffrée, ordinale ou numérale. Inversement parmi des catégories déficitaires figurent le verbe et ses acolytes, à savoir les pronoms, les

adverbes et les conjonctions. Tous les journaux et essais font sensiblement le même choix, tandis que les textes oraux font le choix inverse, comme le montre l'analyse factorielle dans la figure 17, qui n'est pas sans rappeler celle de la figure 10. Mais la portée du présent graphique est plus grande, non seulement parce que toutes les catégories, fermées ou ouvertes,- et donc tous les mots - sont pris en compte, mais parce que les données sont ici désambiguïsées et lemmatisées et qu'on ne confond pas un *le* article (codé comme déterminant) et un *le* troisième personne (codé comme pronom).

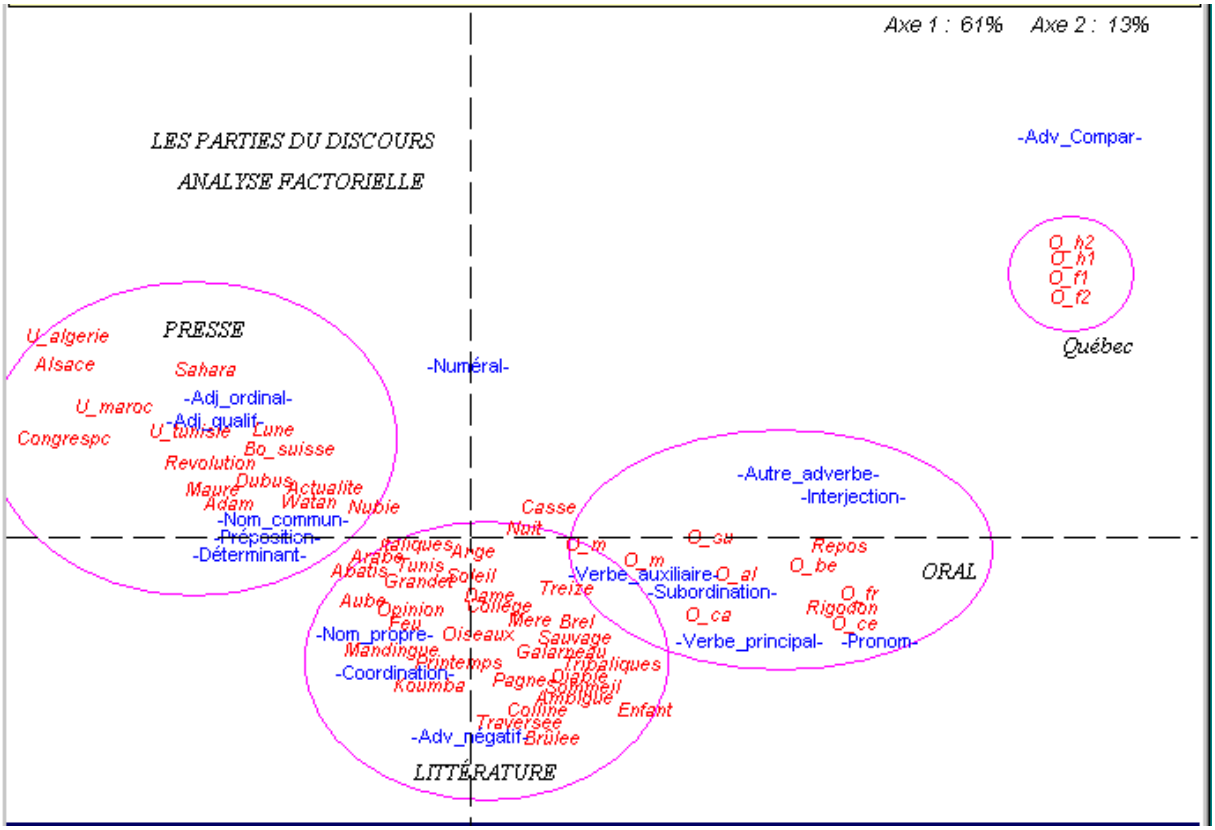
Figure 16. Les parties du discours dans le journal Dernières nouvelles d'Alsace



Le rapprochement avec la figure 13 s'impose aussi, puisqu'il s'agit du même objet d'étude (les codes grammaticaux) et du même corpus. Mais la méthode est différente: le graphique 13 rendait compte de la distance grammaticale, obtenue à partir de milliers de codes différents, dont la multiplicité ne pouvait être explicitée dans le détail. Il montrait les mêmes alliances et les mêmes antagonismes, mais sans donner le moyen de les expliquer. Ici les codes ont été réduits à 16 variétés, qui trouvent aisément place sur le graphique et en facilitent l'interprétation, puisque les lignes et les colonnes du tableau sont représentées conjointement. Non seulement le regroupement étroit des textes oraux prouve la force du lien générique qui les unit, mais aussi la présence constante, dans les mêmes parages, des catégories liées au verbe met en relief l'une des

caractéristiques les plus fortes du style oral: l'abondance des verbes, des pronoms, des adverbes et des conjonctions de subordination. La conjonction *que* à elle seule est un indice d'oralité, comme inversement la préposition *de* est un indice de technicité. Le simple rapport *que/de* pourrait servir de critère efficace pour déterminer auquel des deux genres considérés un texte appartient.

Figure 17. Analyse factorielle des parties du discours



La littérature

Mais notre corpus ne se limite pas à ces deux genres antagonistes. Que devient le troisième: la littérature? Observons que dans toutes les analyses qui précèdent, le genre littéraire n'impose jamais sa loi. Il a certes voix au chapitre, mais son influence ne s'exerce qu'en second lieu, sur le second facteur, quand les acteurs principaux, oral et presse, ont fait leur choix. On a souvent dit qu'il y avait deux français: l'écrit et l'oral. On observe plutôt un jeu à trois, l'écrit se partageant en deux groupes, selon qu'il est littéraire ou utilitaire. Le français utilitaire, qui poursuit un but de communication et d'information, a des caractéristiques propres qui l'éloignent de plus en plus de la tradition littéraire. Il est vrai que les langues techniques du passé, la langue religieuse par exemple, ou la langue juridique, n'entrent pas ici dans nos calculs et qu'elles abusaient aussi des classes nominales. Mais il s'agissait de langues spécialisées que l'honnête homme au siècle classique se gardait bien d'employer. Or cette forme du français, autrefois cantonnée dans des jargons spécialisés, s'est répandu dans l'usage courant, dans les médias, chaque fois que l'on transmet une information,

situation commune à la presse, à l'édition scientifique, à la littérature politique, économique ou technique. Le langage de l'information, comme les langages de programmation, tend à n'être plus qu'un jeu de substantifs emboîtés, la préposition *de* jouant le rôle des parenthèses dans les formules mathématiques. Le verbe dans un tel discours se trouve sans emploi. Lorsqu'il apparaît cependant, ce n'est plus qu'une copule logique, le signe de l'équivalence ou du transfert, simple liaison entre ce qui précède et ce qui suit. On imagine très bien qu'un journaliste, rendant compte de nos travaux et obéissant aux lois du genre (tous ne sont pas de cette espèce), puisse télégraphier à son journal un message aimable, du genre: " L'avenir de la langue française, thème du Colloque de Liré "À la croisée des mots", a fait l'objet de remarquables communications lors de la séance de cet après-midi, au château de la Turmelière." Rien de parodique dans ce propos. On pourrait aisément allonger le chapelet des substantifs, qui remplissent ici tout l'espace. Un seul verbe, maigrelet, dans la phrase, *"a fait l'objet"*, dont on pourrait même faire l'économie si l'on adoptait un style encore plus télégraphique, en disant seulement *"objet"*. Si l'on relit la première page de présentation du présent colloque, on y trouve une trentaine de mots dont aucun n'est un verbe. Certes il s'agit de titres, où il faut bien indiquer les thèmes, les lieux et les dates, en étant précis et concis. Mais cette tendance à la nominalisation du langage peut être à terme un danger, et je n'ose relire mon propre texte, craignant qu'il ne relève du même genre insipide, abstrait, impersonnel, intemporel, sans rien qui évoque les personnes, les temps, les modalités, bref sans verbe. Imaginons le pire: que le français l'ait emporté sur l'anglais et qu'il ait gagné le marché des échanges linguistiques. C'est ce français-là que le monde entier aurait parlé: un français affadi, sans conjugaisons, sans difficultés, sans nuances. Un "petit français", comme on disait jadis "petit nègre".

Mais nous n'en sommes pas là et il y a des recours. Certes, pour sauver la langue, il ne faut pas trop compter sur les linguistes, du moins si j'en juge par cette affiche que j'ai photographiée dans les rues de Nice et qui combine curieusement les adjectifs et les noms: "centre linguistique d'apprentissage accéléré". Les rébus de ce type seraient plus clairs et plus engageants s'ils utilisaient un verbe, comme on faisait jadis à l'ouverture d'un chapitre: "où l'on apprend les langues par une méthode accélérée". Je ferais davantage confiance aux journalistes ou aux publicitaires, malgré ce que j'ai pu dire de la presse. On a vu en librairie il y a dix ans un titre proposé par l'un d'entre eux: "Ne dites pas à ma mère que je suis dans la publicité: elle me croit pianiste dans un bordel". Pourrait-on dire la même chose en supprimant les verbes et les personnes?

Heureusement la tendance à l'abstraction et à la nominalisation se heurte à deux résistances, qui viennent de deux horizons que l'on croyait opposés. L'oral et l'écrit littéraire, ne sont plus face à face, mais de plus en plus côte à côte. On

vient de le voir dans les analyses qui précèdent: si les récits de voyage, écrits en France il y a un siècle, sont tournés vers la catégorie nominale, qui convient à la description pittoresque, la littérature la plus jeune, notamment celle que produit l'Afrique et le Québec, tend à emprunter les ingrédients stylistiques du langage parlé. En France aussi, même si, sur cinq siècles de littérature, le progrès du substantif paraît assuré, la progression n'est pas linéaire et il y a bien des méandres de Rabelais à Malherbe, de Chateaubriant à Zola et de Proust à Céline. C'est précisément Céline, qu'on aperçoit en filigrane dans l'image de la dernière tranche de Frantext. Certes le mot de Cambronne s'y trouve valorisé plus que de raison. Mais c'est un mot de la langue française, que les écrivains les plus gourmés emploient au moins une fois dans leur oeuvre, pour montrer qu'ils ne sont pas bégueules et qu'ils connaissent le français. On peut considérer que c'est de bon augure pour la langue française et que cela...porte bonheur[9].

[9] Même Proust évoque le terme à mots couverts, à propos de Zola, "l'Homère de la vidange": " Il grandit tout ce qu'il touche. Vous me direz qu'il ne touche précisément qu'à ce qui ... porte bonheur", *Le côté de Guermantes*, ancienne édition de la Pléiade, p.499.