Les emprunts dans le vocabulaire français. Inventaire et analyse

Etienne Brunet

Les emprunts dans le vocabulaire français
Inventaire et analyse

Etienne Brunet, BCL(CNRS), Université de Nice Côte d'Azur

Il est malaisé de faire un compte précis du vocabulaire français. L'approche la plus naturelle consiste à ouvrir le dictionnaire et à faire le relevé des entrées. Mais quel dictionnaire choisir ? Et, pour un même ouvrage, quelle édition ? La nomenclature est variable selon la chronologie, selon le public visé, selon que l'intention est encyclopédique, historique, normative ou simplement utilitaire. Même les ouvrages les plus sérieux sont soumis à des impératifs extrascientifiques et à la nécessité commerciale d'économiser la place. Quand un dictionnaire procède à une mise à jour, il met en avant les entrées nouvelles mais reste généralement discret sur les mots qu'il a fallu sacrifier, comme étant vieillis. Sans cet équilibre entre les pertes et les gains, le volume imprimé grossirait indéfiniment[1]. Mais il y a quelque arbitraire dans la condamnation des mots. S'il est facile de repérer les mots nouveaux dans la rue ou dans les journaux – car leur apparition inattendue force l'attention –, il est plus difficile de capter le silence dont s'entourent les mots absents que l'usage abandonne. Il y a une asymétrie entre les naissances lexicales et les décès. La naissance des mots est ponctuelle, datée, repérable et inscrite – même si c'est à retardement – dans les registres, alors que la mort des mots est lente, insensible et ne donne lieu à aucun avis de décès. Les mots que la langue et les dictionnaires déposent comme des alluvions sur la rive du temps sont-ils vraiment morts à jamais ? Pour peu qu'on les rencontre dans un texte

1. Une étude ancienne a été faite à partir des gains et des pertes observés dans les éditions successives du *Petit Larousse* : J. Dubois, L. Guilbert, H. Mitterand, J. Pignon, « Le mouvement général du vocabulaire français de 1906 à 1960 d'après un dictionnaire d'usage », *Le Français moderne*, 1960, avril p. 86-107 et juillet p. 196-211.

ancien ou sous la plume d'un lettré, ils parlent encore à voix basse, et sous la cendre survit une lueur de sens.

1. La *Base Historique du Vocabulaire Français* (BHVF)

Sensibles à cette difficulté de repérage, les observatoires de la langue pointent leur objectif sur l'avance de la tête de la comète, sans trop s'attacher à la queue qui se disperse. Ainsi en est-il de l'entreprise à laquelle Bernard Quemada a attaché son nom et qui depuis cinquante ans amasse les « matériaux pour l'histoire du vocabulaire ».

Dans les quelque 50 volumes publiés de la collection « Datations et Documents lexicographiques », ce qu'on enregistre ce sont les premières attestations relevées dans les textes ou les dictionnaires. C'est le catalogue des naissances. Celui des disparitions fait défaut, même si les mêmes méthodes auraient pu pareillement le constituer : il suffirait de remonter la chronologie en parcourant les ouvrages des plus récents aux plus anciens et en notant à chaque étape les mots non encore rencontrés, au moment de leur dernière attestation[2].

Tableau 1. Effectif des datations relevées dans la BHVF (le premier chiffre correspond à l'année, le second à la décennie)

1400	0	11	1550	71	394	1700	32	190	1850	171	1665
1410	1	23	1560	17	240	1710	33	239	1860	117	2123
1420	0	17	1570	9	279	1720	40	342	1870	98	1513
1430	1	12	1580	31	269	1730	21	268	1880	118	1805
1440	6	19	1590	13	310	1740	50	471	1890	325	3250
1450	4	16	1600	12	604	1750	98	633	1900	421	3924
1460	1	21	1610	71	431	1760	64	741	1910	311	2459
1470	22	47	1620	20	380	1770	90	831	1920	272	3135
1480	4	22	1630	15	266	1780	50	1079	1930	434	3124
1490	20	53	1640	56	296	1790	555	2557	1940	41	1311
1500	8	72	1650	45	312	1800	229	1723	1950	282	1946
1510	4	158	1660	33	283	1810	102	1425	1960	264	2053
1520	30	125	1670	29	397	1820	95	1718	1970	206	2195
1530	15	397	1680	37	228	1830	297	2504	1980	116	806
1540	33	528	1690	86	281	1840	197	2312		5822	54824

2. Cette navigation à rebours peut s'appliquer aussi à un corpus textuel. L'accroissement inverse, ainsi mesuré à partir du dernier texte de la série, signale les ruptures ou plus exactement les clôtures, c'est-à-dire les cycles qui se ferment, alors que la perspective chronologique normale met en relief les orientations nouvelles, les cycles qui s'ouvrent. À l'aller les ruptures dénotent le renouvellement de l'inspiration, au retour elles marquent son épuisement. Cette double exploration est réalisée par notre logiciel HYPERBASE.

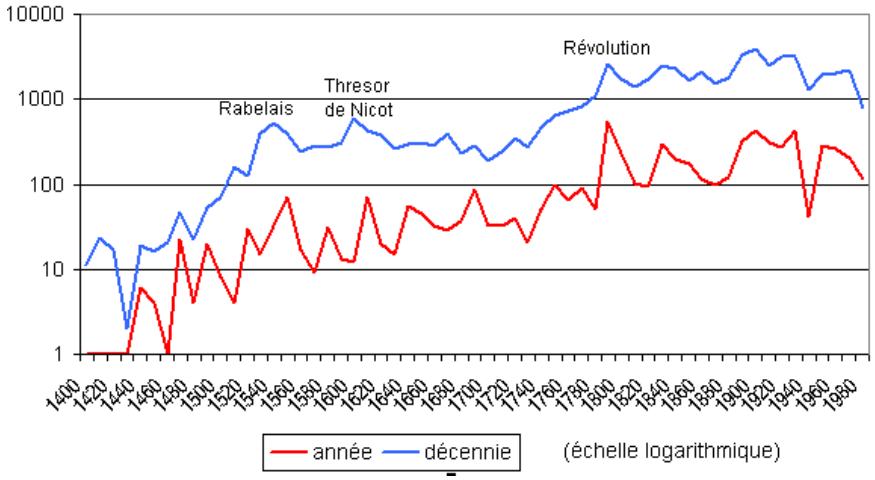


Figure 2. Effectif des datations lexicales relevées dans la BHVF de 1400 à 1989

Cette patiente enquête, poursuivie parmi 240 ouvrages lexicologiques, dont 42 dictionnaires, mais aussi au sein même des textes, au hasard des lectures, a été rendue disponible sur *Internet*, ce qui nous a permis de construire le tableau 1, en interrogeant la base BHVL autant de fois que nécessaire, année après année. Pour rendre les résultats lisibles, on a procédé par décennie, en commençant au quinzième siècle. Car les décennies antérieures n'offrent que de maigres effectifs, peu exploitables[3]. Il est en effet difficile de préciser les datations, à l'année près, lorsqu'il s'agit du Moyen Âge. Au reste les documents sont de plus en plus rares à mesure que l'on remonte le temps, ce qui donne à la courbe une pente abrupte, que l'échelle logarithmique a adoucie dans la figure 2. Les datations se comptent par unités au Moyen Âge pour une même décennie, par dizaines au XV^{e} siècle, par centaines à partir de Rabelais et par milliers après la Révolution. Au total c'est plus de 50 000 mots recensés. À chacun une carte d'identité est délivrée au moment où il entre dans les textes du patrimoine et, pour la plupart et pour quelque

3. On ne relève que 3 exemples avant 1100, après quoi les effectifs s'établissent comme suit :

1100	5	**1110**	3	**1120**	3	**1130**	7	**1140**	2	**1150**	10	**1160**	19	**1170**	10	**1180**	12	**1190**	2
1200	4	**1210**	5	**1220**	0	**1230**	7	**1240**	10	**1250**	10	**1260**	41	**1270**	10	**1280**	17	**1290**	13
1300	36	**1310**	156	**1320**	13	**1330**	6	**1340**	13	**1350**	11	**1360**	7	**1370**	38	**1380**	12	**1390**	6

temps, dans le dictionnaire. On a la trace contrôlable de leur passage, la preuve de leur existence à un moment de l'histoire. Mais leur survie n'est pas garantie et leur descendance n'est pas évoquée. Au reste ce contrôle douanier, fait *a posteriori*, est sujet aux lacunes et aux révisions. Bien des mots ont passé la frontière sans se faire remarquer, et ceux qu'on a épinglés n'en sont peut-être pas à leur première tentative. Pour certains d'entre eux la base signale leurs datations successives, la première attestation étant toujours considérée comme provisoire, en attendant un éventuel document antérieur. Enfin il ne faut pas se cacher que l'exploitation numérique, nécessairement neutre et impartiale, est établie sur des données filtrées par une conscience humaine, le nombre et l'expérience des experts ne garantissant ni l'exhaustivité ni l'objectivité. En outre les aléas de la conservation des textes et l'imprécision qui enveloppe certains de ceux qu'on trouve imposent au recensement des limites évidentes.

Il n'en reste pas moins que les accidents de la courbe reflètent les événements de l'histoire. On notera en particulier la brusque élévation du niveau au XVI^{e} siècle, dès que Rabelais entre en scène, et l'accès de fièvre qui correspond à la prise en compte du *Thresor* de Nicot. Le niveau reste stationnaire au cours du XVII^{e} et du XVIII^{e}, non pas parce que les textes feraient défaut, mais parce que une phase de stabilisation succède à l'explosion de la Renaissance. La Révolution française est explosive aussi, même dans le domaine de la langue. On légifère sur tout, sur les choses et sur les mots. On donne des noms aux mesures, aux divisions du temps et à celles de l'espace. Surtout la liberté, nouvellement conquise, s'exerce sans retenue et sans censure, sans attendre Hugo pour mettre « un bonnet rouge au dictionnaire ». La créativité lexicale se nourrit de passion, de discours improvisés, de pamphlets enflammés et le vocabulaire de l'invective atteint des sommets dans les feuilles du *Père Duchesne* (par exemple AFFAMEUR, ACCAPAREUR, AGITATEUR, ALARMISTE, ANTIFRANÇAIS, ANTIREVOLUTIONNAIRE, ARCHINOIR, etc.). Mais les périodes de guerre ne sont pas nécessairement favorables à l'expansion du vocabulaire, car la liberté d'expression et la création littéraire peuvent être bridées soit par la censure de l'occupant, soit par la pression contraignante de l'union nationale, et l'on observe un fléchissement de la courbe quand la France est en guerre (1870, 1914, et 1940).

On n'ose trop approfondir les raisons qui expliquent le tassement observé à la fin de la série. Peut-être hésite-t-on à cataloguer un néologisme qui n'a pas fait ses preuves, faute de temps et de recul. Et c'est le cas lorsque l'invention verbale est cultivée pour elle-même, sans

souci de propagation ni de pérennité, comme il arrive aux romans de San Antonio. Mais on craint surtout qu'il s'agisse d'un artefact, dû pour l'essentiel au tarissement des sources. Les dictionnaires consultés sont pour un tiers antérieurs à 1970 et pour les trois quarts antérieurs à 1980. Leur témoignage est donc muet pour les dernières tranches.

Cette distorsion n'existe pas si l'on s'en tient à un dictionnaire unique, assez représentatif de la pensée majoritaire des experts. C'est le privilège, pour l'anglais, de l'*Oxford English Dictionary*. En exploitant la version 2 publiée sur cédérom, on obtient une courbe qui n'est pas sans analogie avec celle du français. Nous n'avons procédé que par sondage, en ne retenant qu'une année sur 10. Car le parallélisme observé dans la figure 2 entre l'échantillon et la population autorisait ce raccourci. On observe pareillement une poussée de la sève au temps de la Renaissance, dont le couronnement coïncide avec le temps de Shakespeare, puis une retombée aux XVII^{e} et XVIII^{e} siècles, et de nouveau une puissante progression au XIX^{e} siècle, ce qui semble correspondre à l'hégémonie de l'Angleterre sur les mers et le monde. La différence essentielle concerne l'époque antérieure à la Renaissance : la conquête des normands a visiblement bouleversé le vocabulaire local en introduisant des milliers de mots français, au point de menacer même l'existence de l'anglais (les rois Plantagenets parlaient le français à la Cour), menace que connut aussi le français, face à l'italien, au temps des reines Médicis et de Mazarin.

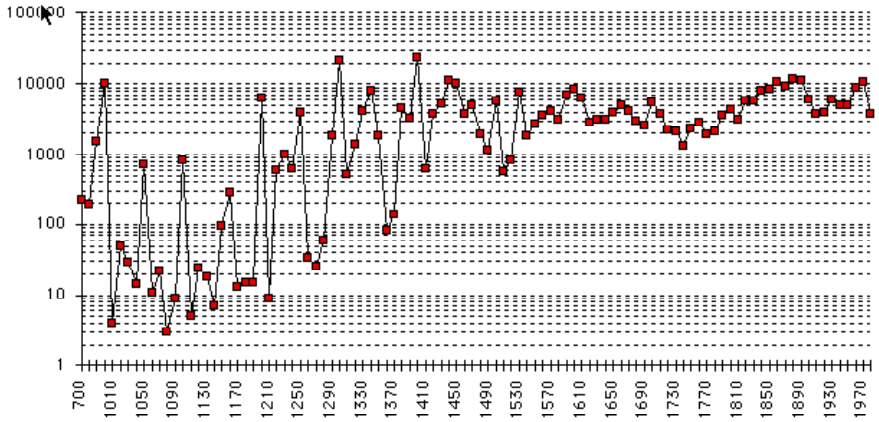


Figure 3. Les datations de mots dans l'*Oxford English Dictionary* (échelle logarithmique)

2. Le *Petit Robert* (sur cédérom)

Comme le *Petit Robert* est également disponible sur cédérom dans une édition récente (1996), et que l'interrogation peut, comme dans l'OED, s'appliquer à la datation (parmi d'autres critères : entrée, étymologie et citation, etc., voir figure 4), la tendance observée dans la BHVF (graphique 2) se trouve confirmée. La figure 5 en rend compte très simplement, en portant en ordonnée l'effectif absolu de chaque siècle. Comme précédemment c'est le XIX^{e} qui est le plus productif (avec 11 738 unités, soit plus du quart du total), devant le XX^{e}. Et de la même façon le XVI^{e} l'emporte sur le XVII^{e} et le XVIII^{e}.

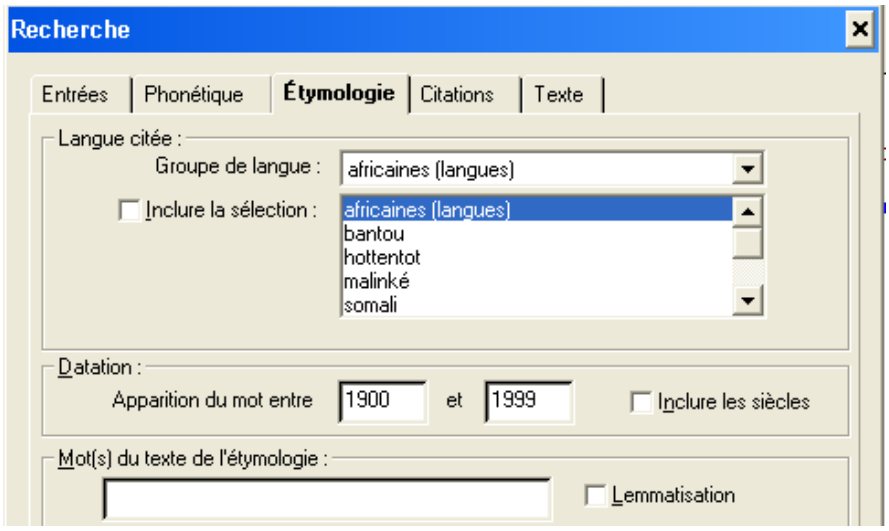


Figure 4. L'interrogation du *Petit Robert* sur cédérom

Le recours à une source unique évite certes les contradictions, mais ne résout pas toutes les incertitudes. Cependant, s'agissant de datation, la part de l'interprétation reste faible. Car les documents modernes ont une référence précise et la date n'est incertaine que pour quelques textes plus anciens. Il n'en va pas de même lorsqu'il s'agit de préciser la provenance des emprunts. Plus que les marchandises, les mots circulent d'un pays à l'autre, et d'une époque à l'autre, en se prêtant tour à tour aux coutumes locales, si bien qu'il est difficile de suivre ce mouvement brownien sous les déguisements et les truchements successifs. Du persan au français (par exemple le mot AZUR) un mot peut faire un long détour, par l'arabe, l'espagnol, l'italien, sans qu'on sache exactement les étapes du circuit. Sans doute la connaissance des évènements historiques et des moyens de

communication aide à fixer l'orientation des flux, le mouvement des hommes, des capitaux, des marchandises, des idées et des mots. L'invasion des peuplades germaniques, la conquête des vikings et des normands, les croisades, les guerres d'Italie, la découverte de l'Amérique créent des remous dans la circulation des monnaies et des langues, et dans les ports et les foires les mots s'échangent dans un sens, puis dans l'autre, et souvent dans les deux sens, comme le remarque avec humour Henriette Walter : « En argot l'entremetteur est appelé MAQUEREAU , mot qui vient selon toute vraisemblance du néerlandais MAKELAER « courtier ».[…] Mais, d'autre part, on apprend que le mot néerlandais MACKEREL, qui désigne le poisson, viendrait lui-même du français MAQUEREAU, ce qui signifie qu'en matière d'étymologie on n'est jamais sûr de rien »[4].

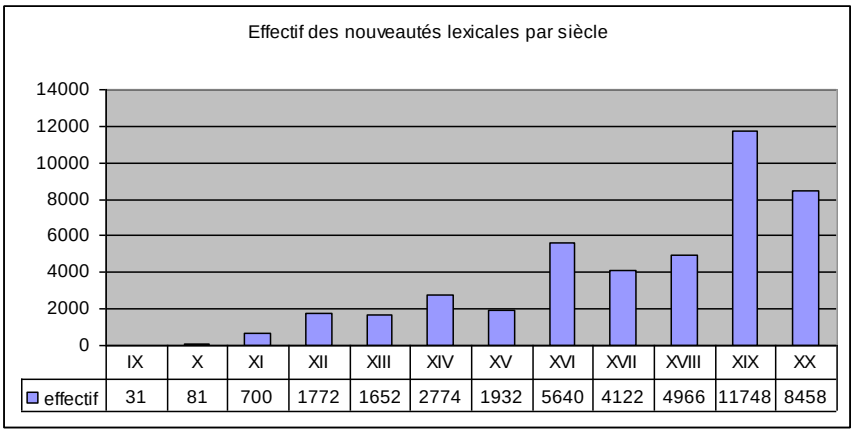


Figure 5. Le registre des naissances verbales, cataloguées dans le *Petit Robert*

Cela est particulièrement vrai pour l'arabe qui n'a guère eu de contact direct avec le français, mis à part un début de conquête vite stoppé en 732 à Poitiers et l'entreprise tardive de la colonisation au XIX^{e} siècle. Le rayonnement de la civilisation arabe durant le Moyen Âge a pourtant laissé des traces profondes dans notre vocabulaire, en

4. Henriette Walter, *L'aventure des mots français venus d'ailleurs*, Livre de poche, 1999, p. 123. Parmi les nombreux ouvrages que Henriette Walter a consacrés à l'origine du vocabulaire français, signalons surtout le *Dictionnaire des mots d'origine étrangère*, Larousse, 1991. Précisons que l'enquête menée par elle en 1991 a été remise sur le chantier et les résultats réajustés en 1999. Cependant notre propre enquête peut n'être pas inutile, car elle s'appuie sur des ressources linguistiques qui n'étaient pas précédemment disponibles et elle tire profit des progrès accomplis dans les méthodes statistiques et les outils informatiques.

empruntant divers canaux : par l'Espagne, par Venise et l'Italie, ou par les ports du midi. Si tous les dictionnaires sont d'accord pour reconnaître la source arabe du mot CALIBRE, les uns invoquent le truchement de l'espagnol CALIBRE ou de l'italien CALIBRO (Henriette Walter penche pour cette hypothèse), quand d'autres récusent l'un et l'autre. Voici ce qu'en dit le *Trésor de la langue française* :

Tableau 6. L'étymologie du mot CALIBRE dans le TLF

Étymol. et Hist. 1. 1478 (*Statuts des Platriers, ap. Ouin-Lacroix. Hist. des anciennes corp. de Rouen*, 717 d'apr. DELB. *Notes*) ; av. 1571 « capacité d'une chose par rapport au volume qui doit la remplir (ici en parlant d'un canon) » (Carloix, VII, 7 ds LITTRÉ) ; 1636 « volume d'un projectile ; d'un objet cylindrique ou sphérique » (Le P. MONET, *Invantaire des deus langues fr. et lat*., Genève) ; d'où fig. 1548 « importance » (N. DU FAIL, *Cont. d'Eutr*., XIX ds GDF. *Compl*.) ; 1611 *n'être pas du même calibre* « différer de sentiments et d'opinions » (COTGR., *s.v. qualibre*) ; **2.** 1690 « instrument servant à vérifier le calibre d'une arme » (FUR.) ; **3.** 1694 technol. « (dans la fabrication de divers objets d'art, d'industrie) modèle sur lequel sont tracés les contours, les dimensions de l'objet à fabriquer » (CORNEILLE). Empr. à l'ar. , (LOK., no 1030 ; *FEW* t. 19, p. 82b ; BL.-W.5) « moule où l'on verse les métaux » (IX^{e} s., Aboûl'Atâhiya ds LAMMENS, p. 70) ; « forme de cordonnier » (début XII^{e} s., Harîrî ; 1505, P. de Alcala ds DOZY t. 2, p. 391a) ; « forme de marbre servant de support pour un turban » (début XV^{e} s. ds SACY, *Chrestomathie ar.*, Paris, t. 1, 1826, pp. 235-236). L'ar. est lui-même empr. au gr., « forme en bois pour fabriquer des chaussures » (composé de « bois » et de « pied », v. CHANTRAINE, s.v. ; v. aussi LAMMENS, pp. 70-71 ; DOZY, *loc. cit.*). L'hyp. d'un intermédiaire ital. *calibro* (KOHLM., p. 35 ; *DG* ; *EWFS2* ; *Webster's* ; DAUZAT 1973) fait difficulté du point de vue chronol., ce mot n'étant pas attesté av. le XVII^{e} s. (Galilée ds BATT.) ; il en va de même pour l'esp. *calibre* (RUPP., p. 286 ; *REW3*, no 4663a) attesté seulement dep. 1583 (sous la forme *calibio*, Escalante ; *calibre* en 1594, B. de Mendoza d'apr. COR.). Selon BL.-W.5 ; COR. ; *FEW*, *loc. cit.* ; HOPE, p. 330, l'ital. et l'esp. seraient empr. au français.

Comme la dispute, à coup d'attestations, se répète d'un bout à l'autre du dictionnaire, en invoquant tour à tour le témoignage de Littré, de Wartburg et de Dauzat, il nous a paru sage de fonder notre enquête sur deux dictionnaires récents (qui d'ailleurs concordent pour le mot CALIBRE), le TLF et le *Petit Robert*. Le premier offrait l'avantage d'être interrogeable sur le réseau *Internet*, et le second d'être disponible sur un support numérique.

Tableau 7. L'origine des mots français dans le *Petit Robert*

Afrique	25
bantou	8
hottentot	1
malinké	1
somali	1
soudanais	2
swahili	2
wolof	
allemand	520
alémanique	18
bas allemand	11
haut allemand	39
alsacien	11
frison	3
amérindien	45
algonquin	29
araucan	1
arawak	3
aztèque	12
caraïbe	23
guarani	11
huron	2
inuit	9
iroquois	2
nahuatl	2
quechua	16
taino	1
tupi	36
anglais	2337
anglais ancien	26
américain	274
arabe	416
arabe Espagne	2
Maghreb	35
Asie	
chinois	35
thaï	5
tibéto-birman	9
vietnamien	4

coréen	1
basque	7
celtique	17
breton	35
gaélique	19
gallois	5
gaulois	161
Inde (dravidien)	19
malayalam	4
tamoul	12
égyptien	6
espagnol	
castillan	427
esp. d'Amérique	32
étrusque	4
finno-ougrien	5
hongrois	19
lapon	2
fr. argot et verlan	182
fr. ancien	1039
fr. dialectal	281
fr. régional	
franco-prov.	34
langues d'oïl	156
occitan	513
fr. extérieur	
Antilles	39
Belgique	25
Canada	41
Suisse	13
germanique	
francique	378
germanique	184
saxon	2
grec	
ancien.	3770
byzantin	15
ecclésiastique	9

moderne	9
hébreu	63
indo-européen	
(langues isolées)	
ligure	1
lituanien	1
tsigane	6
indo-iranien	
hindi	59
indo-iranien	10
sanskrit	46
iranien	
avestique	1
iranien	4
persan	79
italien	1150
corse	6
ital. régional	53
japonais	63
latin	10578
bas latin	1218
biblique et chr.	434
cabalistique	15
faux latin	3
Polynésie	
indonésien	8
javanais	5
khmer	2
malais	58
malgache	13
polynésien	10
tahitien	3
néerlandais	250
flamand	22
hollandais	12
nordique	26
danois	7
islandais	3
norvégien	20

scandinave	57
suédois	25
Océanie	
Australie	9
mélanésien	1
portugais	
Amérique	8
Portugal	113
roman	
catalan	15
roman	36
romanche	2
roumain	2
sarde	1
sémitique	
amharique	1
araméen	4
assyrien	3
berbère	4
punique	1
sémitique	6
syriaque	4
slave	
bulgare	2
polonais	16
russe	104
serbo-croate	4
slave	10
tchèque	8
turc	
kirghiz	1
mandchou	1
mongol	5
ouzbèke	1
tatar	4
toungouze	3

Commençons par le *Petit Robert* en utilisant la grille de la figure 4 (*langue citée* dans la rubrique *étymologie*). Les divisions du tableau obtenu (tableau 7) sont fort détaillées, même si nous n'avons pas distingué certaines rubriques à l'intérieur de l'occitan ou des langues d'oïl[5]. On prendra garde que les titres retenus ont une extension variable. Par exemple au sens large l'origine germanique inclut le francique ; si

5. L'occitan enveloppe la Gascogne (19 emprunts), le Béarn (6), l'Auvergne (2), le Languedoc (10), la Provence (452). Quelques provinces sont aussi précisées pour les langues d'oïl : Normandie 86, Picardie 66, Poitou 8.

cependant deux effectifs distincts sont proposés, c'est parce que dans certains cas, la variété spécifique ne pouvant être précisée, on a recours au terme générique. Les effectifs sont donc en principe cumulables : ceux qu'on a mis en corps gras n'incluent jamais ceux des sous catégories en corps maigre. On s'abstiendra pourtant de considérer ces relevés comme indépendants et superposables : car les filiations complexes, où l'on relève des transitions, donnent lieu parfois à plusieurs mentions. Ainsi le mot AZUR figure parmi les mots persans, mais aussi parmi les mots arabes. Les pays d'accueil ou de transit – l'Italie joue souvent ce rôle – sont souvent pris en compte au même titre que le pays d'origine. Cela gonfle un peu les effectifs, ce qui rend difficile la comparaison avec d'autres relevés qui excluent la double appartenance.

Le classement des langues vivantes qu'on peut extraire de l'enquête de Henriette Walter[6] (1 anglais, 2 italien, 3 arabe, 4 allemand, 5 espagnol, 6 néerlandais) se retrouve pourtant dans le tableau 7, à une différence près, relative à l'arabe, qui recule de deux places. Mais l'apport de l'arabe est ancien et, mis à part quelques souvenirs récents, souvent populaires, que la colonisation a rapportés d'Afrique du Nord, l'arabe est aujourd'hui moins sollicité que les langues européennes, lorsque, sortant du vocabulaire courant, auquel se borne le relevé de H. Walter, on a besoin d'emprunter. À plus forte raison les relevés sont-ils voisins, quand il s'agit du vieux fonds celtique et germanique du français, lequel est moins un emprunt qu'un héritage : on a respectivement 550 et 564 pour le germanique, francique compris, et 160 et 161 pour le gaulois. Ces langues non écrites n'ayant pas survécu, la source est épuisée et le stock ne varie plus. Le latin a servi plus abondamment encore à constituer le socle du français, mais il accompagne beaucoup plus longtemps le destin du français, particulièrement à la Renaissance.

Or l'intérêt paraît devoir s'attacher moins au dosage absolu des emprunts linguistiques qu'aux fluctuations des échanges au cours du temps. Il convient donc de croiser l'espace et le temps, en procédant siècle après siècle, pays après pays[7].

---

6. H. Walter, *L'aventure des mots français venus d'ailleurs*, Livre de Poche, 1999, p. 20. Une autre raison explique que les effectifs soient moindres dans cet ouvrage, car ils concernent « 4200 mots courants », notion qui recouvre imparfaitement la nomenclature d'un dictionnaire de langue.

7. Chaque élément du tableau résulte de deux interrogations combinées, la première portant sur la date précise de l'entrée, quand elle est mentionnée (par exemple un millésime compris entre 1500 et 1599), la seconde sur le siècle en chiffres romains (par

Tableau 8. L'origine des mots dans le *Petit Robert*, par langue et par siècle

	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
latin	51	253	542	1784	1305	1605	922	1892	704	878	1197	459
grec	1	14	40	233	224	232	140	815	285	444	1048	459
hébreu	0	3	2	10	5	1	3	7	5	7	11	8
arabe	1	1	2	19	35	34	32	64	58	40	85	50
anglais	1	4	16	50	41	60	41	74	105	202	663	1094
italien	0	2	8	6	46	76	112	360	224	150	149	67
espagnol	0	2	2	4	9	15	16	88	103	72	88	47
allemand	0	0	3	16	4	9	11	36	36	68	202	151
néerlandais	0	0	1	18	33	26	29	29	51	42	25	12
portugais	0	0	0	0	0	0	3	26	39	16	16	10
russe	0	0	0	0	0	0	2	6	6	15	27	41
japonais	0	0	0	0	0	0	0	2	0	2	25	30
persan	0	0	2	3	4	8	8	8	22	5	16	2
chinois	0	0	0	2	0	1	0	1	4	5	9	10
dialectal	1	0	2	20	20	16	15	35	31	35	99	32
occitan	1	1	3	34	38	60	41	119	78	60	85	29
fr. prov.	0	0	0	0	1	1	4	8	1	6	13	1
oïl	0	0	1	17	17	14	12	24	31	20	32	4
gaulois	0	0	19	51	31	18	10	25	10	6	11	2
francique	1	6	55	153	54	46	27	39	26	11	20	1
germanique	0	1	23	47	23	18	14	22	24	15	16	3

Le tableau 8 reproduit les effectifs absolus, ce qui donne au latin une écrasante domination, même si cette domination s'atténue au fil du temps puisque le grec arrive exactement au même niveau dans la dernière tranche (459 exemples dans les deux langues) et que l'anglais en revendique le double (1094 unités). Mais un tableau de ce type se prête mieux à l'analyse quand les nombres absolus sont traités par les calculs classiques de la statistique et convertis en écarts réduits. Comparons en effet les deux représentations que les données du latin (première ligne du tableau) peuvent générer : la première (figure 9) donne l'impression que le latin contribue de façon presque constante à l'enrichissement du français, bien qu'un fléchissement soit visible aux deux bouts de la

exemple XVI), le critère relatif à la langue empruntée étant fixé de la même façon dans les deux cas. Les datations précises de la première espèce sont évidemment plus rares quand l'époque est plus reculée. En de rares occasions il y a redondance, voire contradiction apparente, certaines entrées plus développées comportant plusieurs dates. C'est le moment de rappeler que le *Petit Robert*, comme le TLF et tous les dictionnaires antérieurs, n'est devenu qu'*a posteriori* une base de données. Les rubriques qu'on y a distinguées après coup sont des chaînes de caractères approximatives dont le format n'a pas la rigueur et la constance espérées. Voir là-dessus la thèse de Chantal-Édith Masson : *Le traitement des substantifs dans le Robert-CD-Rom : modélisation, formalisation et proposition méthodologique en vue de son informatisation*, Université de Sherbrooke, 2001.

chaîne. La raison de cette baisse est aisément explicable dans les premiers siècles, où les sources manquent. Mais pour le XX[e] il faut invoquer un autre motif qui ne peut être que la concurrence des autres langues.

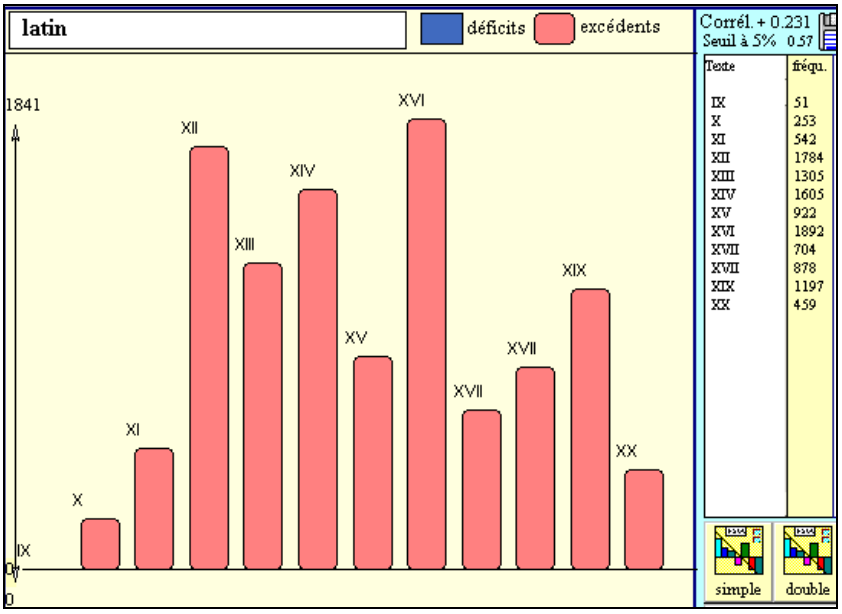


Figure 9. L'évolution du latin. Effectifs absolus

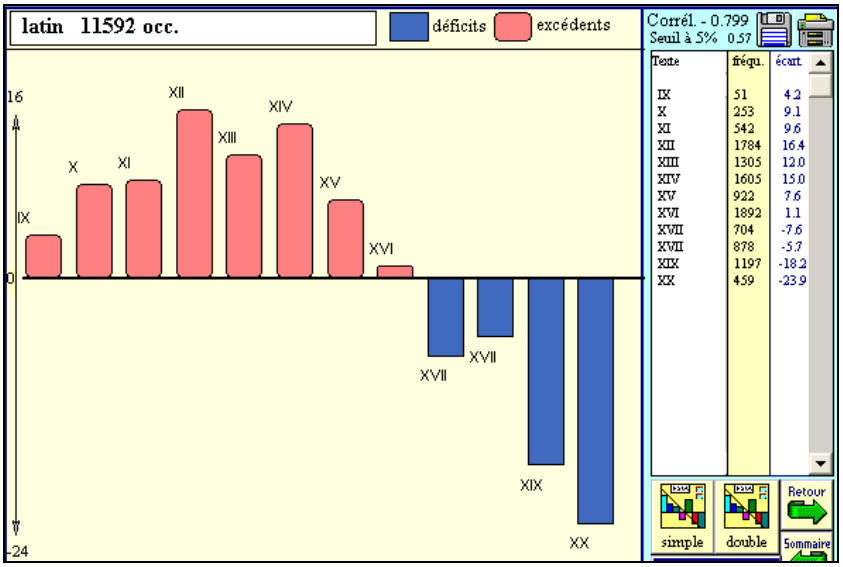


Figure 10. L'évolution du latin. Écarts réduits

Or ces autres langues interviennent dans le calcul quand la statistique s'empare des données. Les écarts réduits (figure 10) manifestent alors beaucoup plus clairement que la part du latin, d'abord prépondérante, est de plus en plus contestée dans le vocabulaire français. Le sommet, rencontré précédemment au XVI^{e} siècle, est atteint dès le XII^{e} et le déficit se creuse à partir du XVII^{e}. La veine semble donc s'épuiser, comme si la plupart des racines latines avaient déjà été mises à contribution et que la combinatoire des radicaux et des affixes avait été trop largement exploitée. Lorsqu'un besoin de la terminologie se fait sentir, le grec se propose de plus en plus souvent à la place du latin, avec des racines et des préfixes moins usés.

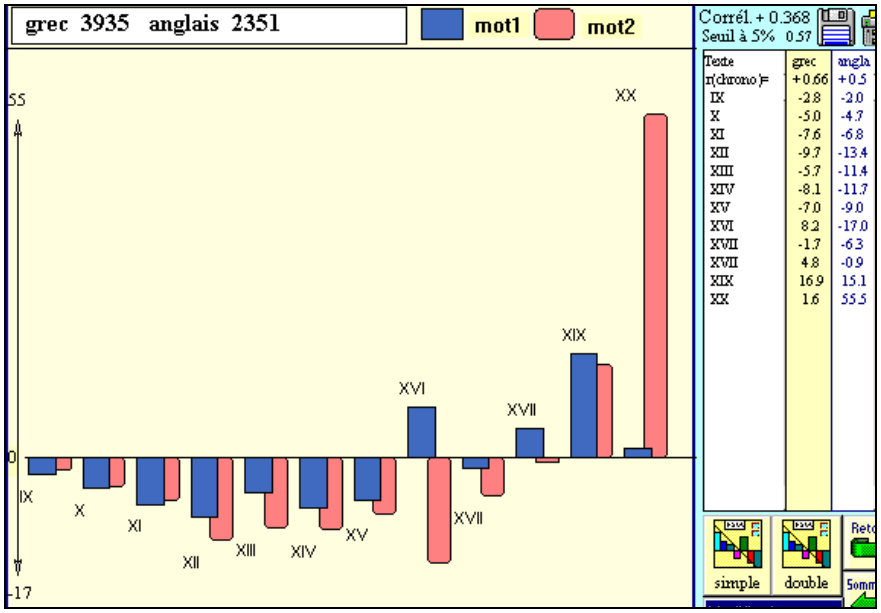


Figure 11. Courbes *comparées* des emprunts grecs et anglais

Mais la concurrence du grec n'est pas seule en cause. À mesure que les communications se développent, que les moyens d'information se répandent et que les relations commerciales et linguistiques s'intensifient entre les peuples et les langues, les mots passent par-dessus les frontières, souvent sans prendre la peine ou le temps de s'adapter au parler local, ni pour l'orthographe, ni pour la prononciation. L'anglais est communément désigné comme le principal responsable de ce qui apparaît à certains comme une dégradation plutôt qu'un enrichissement. La statistique se tient à l'écart de tout jugement de valeur sur ce point, mais elle aide à

prendre la mesure du phénomène. Avec un millier d'emprunts, l'anglais est effectivement le fournisseur le plus sollicité par la langue française au XX^{e} siècle et la promotion de l'anglais est récente et rapide, comme le montre la figure 11.

Mais la courbe juxtaposée du grec, dont le mouvement est grossièrement parallèle, permet d'apaiser les craintes que provoque le franglais. Avec un apport de 4000 éléments, le grec est plus productif au total, même si sa progression est plus ancienne et moins abrupte. D'autres sources étrangères, même moins abondantes, participent aussi à l'extension du lexique français (536 unités venant d'Allemagne, 97 de Russie, 59 du Japon et 32 de Chine, etc.). L'invasion de l'anglais n'a donc pas encore le caractère massif qu'on lui prête parfois. Son ampleur est très inférieure au raz-de-marée français qui submergea la langue anglaise au temps de la conquête normande. La rubrique « étymologie » de l'*Oxford English Dictionary* attribue en effet 37 022 mots anglais à la source française, certes derrière le latin (50 725 mots), mais loin devant les apports grecs (18 675) et germaniques (12 322)[8].

En réalité la mode actuelle de l'anglais fait songer à celle de l'italien à la fin du XVI^{e} siècle, sous le règne de Catherine de Médicis. Or les modes linguistiques changent et le profil d'une époque, par exemple celui de la Renaissance, représenté dans la figure 12 et orienté vers le sud, est fort opposé à celui du XX^{e}, que le nord polarise. L'italien domine la distribution au XVI^{e} siècle mais plus largement l'emprunt s'adresse alors à tout ce qui se rattache aux langues latines : l'espagnol, le portugais, le grec, l'occitan et le franco-provençal participent activement à la souscription. Le latin, pourvoyeur habituel, est encore sollicité mais il n'a plus le monopole. D'autres instantanés peuvent être reproduits, qui délivrent le détail d'une époque ou d'une langue, selon qu'on isole une colonne ou une ligne du tableau 8. Mais un tel tableau à double entrée se prête à une exploitation synthétique si l'on utilise les méthodes multidimensionnelles.

8. Voici le classement qu'on peut extraire de l'OED :

1	50 725	Latin	**7**	6 286	Dutch	**13**	2 824	Teutonic
2	37 022	French	**8**	5 795	Spanish	**14**	2 294	Provençal
3	18 675	Greek	**9**	4 430	Norse	**15**	2 120	Frisian
4	14 119	English	**10**	3 438	Swedish	**16**	1 889	Saxon
5	12 322	German	**11**	3 130	Portuguese	**17**	1 744	Anglo-French
6	7 893	Italian	**12**	3 046	Danish	**18**	1 480	Gothic

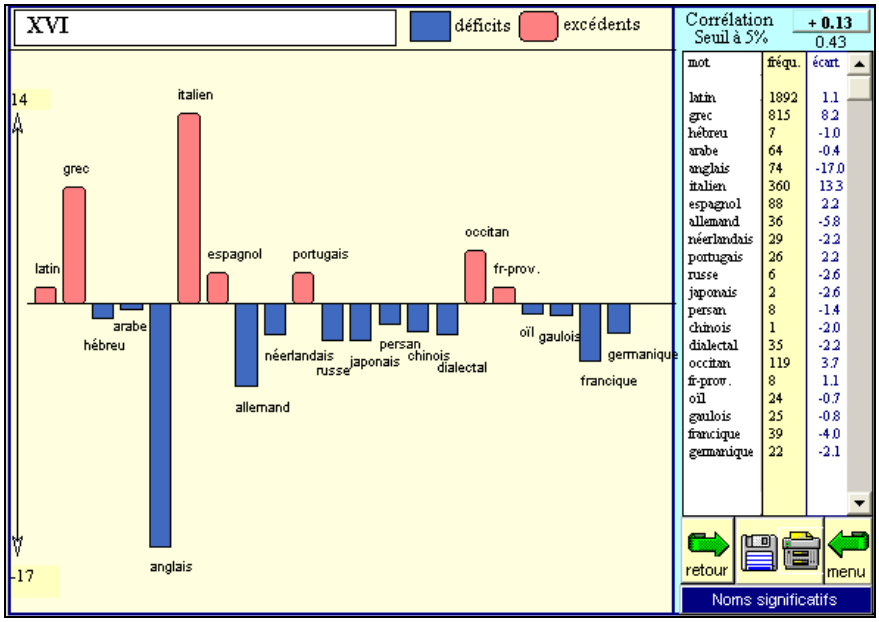
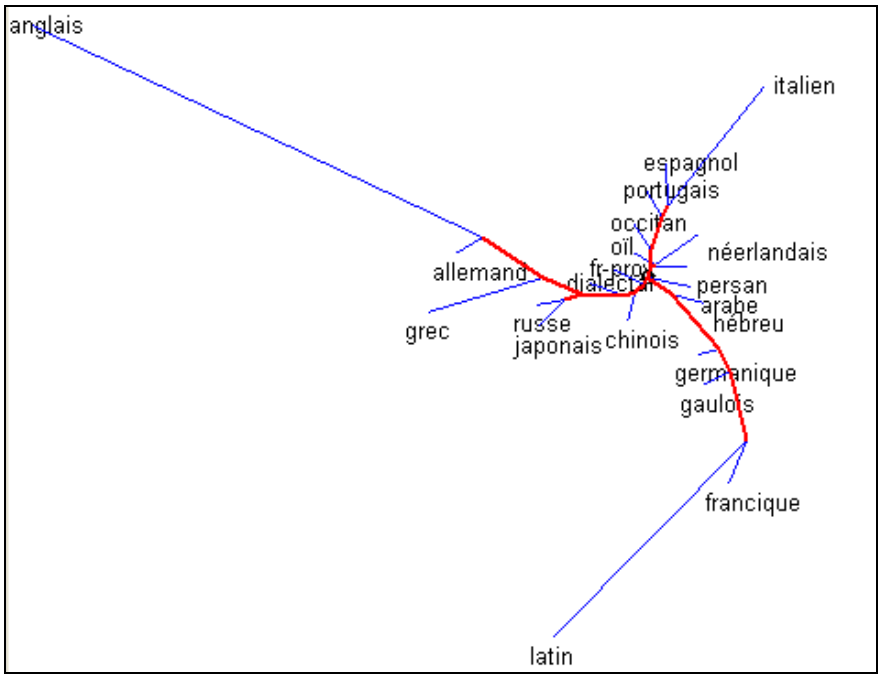


Figure 12. Distribution des emprunts au XVI^e^ siècle

La première approche – dite analyse arborée – classe selon leurs affinités les langues auxquelles le français doit une part de son vocabulaire. Entendons par affinités, non pas leur origine et leur parenté historiques, mais la ressemblance de leur distribution, c'est à dire un calcul de distance statistique. Trois pôles prennent position aux extrêmes : le latin, l'italien et l'anglais. Le latin (c'est la branche basse dans le graphique 13) mêle son vocabulaire à celui des gaulois, des Francs, et de certaines peuplades germaniques. La branche droite est dévolue à l'italien et aux parlers du sud qui lui sont associés : l'espagnol, le portugais, l'occitan, le franco-provençal. La troisième branche qui s'étire sur la gauche est dominée par l'anglais. C'est là que se rattachent les langues des grands pays du monde, auxquelles le français est confronté à l'époque contemporaine : allemand, russe, japonais, chinois. Le grec, qui répond abondamment aux besoins de la terminologie scientifique, se place aussi sur la branche de la modernité.

L'analyse factorielle, réalisée dans la figure 14, confirme cette division tripartite, de façon plus explicite. Comme elle rend compte en même temps des lignes et des colonnes, son pouvoir explicatif est supérieur, puisqu'elle établit une relation entre les unes et les autres, entre les langues et les époques. C'est le temps qui gouverne la distribution, de

la droite à la gauche : dans le quadrant inférieur droit sont concentrés tous les siècles antérieurs à la Renaissance, du IX^{e} au XIV^{e}. La gauche est accaparée par les temps modernes, XX^{e} et XIX^{e}, tandis que les siècles intermédiaires, du XVI^{e} au XVIII^{e}, occupent la zone médiane, tout en s'écartant vers le haut du graphique. Ce profil linéaire, en forme de croissant, est caractéristique des données sérielles ou chronologiques, qu'un même courant anime.

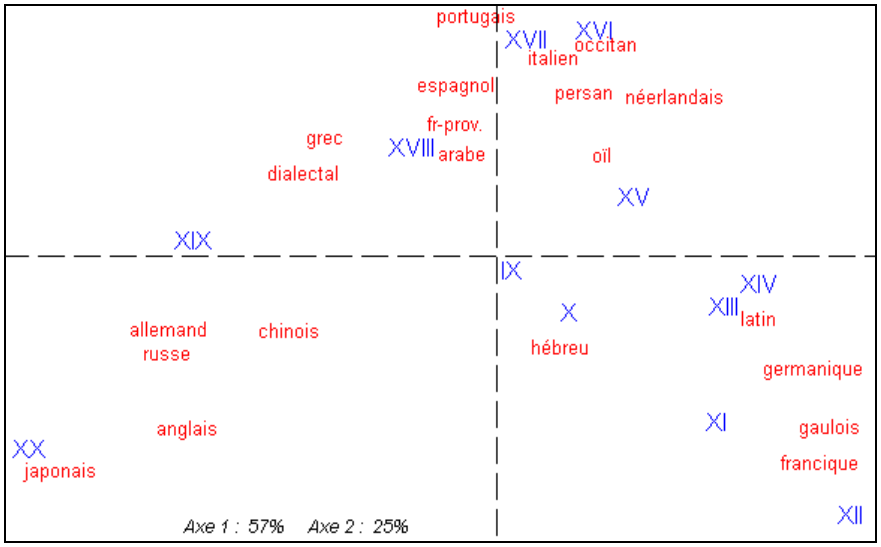


Graphique 13. Analyse arborée des emprunts relevés dans le *Petit Robert* (classification des langues sources)

Or sur cette chaîne du temps, dont les maillons sont régulièrement ordonnés d'un bord à l'autre, prennent place les trois groupes proposés précédemment par l'analyse arborée. Le gaulois, le francique et les parlers germaniques accompagnent le latin dans la zone consacrée aux origines. À l'opposé l'allemand, le russe, le japonais et le chinois font cercle autour de l'anglais et du XX^{e} siècle. Enfin les langues latines, italien en tête, ont la faveur de la Renaissance et des siècles classiques[9]. On prêtera une attention particulière aux éléments intermédiaires : par

9. Ces pavillons latins, surtout l'espagnol, couvrent souvent la marchandise arabe.

exemple le néerlandais et les langues d'oïl tendent à se rapprocher des origines (le XV[e] est à mi-chemin entre les deux premières divisions). Et de l'autre côté le grec est partagé entre l'époque classique et les temps modernes.



Graphique 14. Analyse factorielle des emprunts relevés. Temps et espace

3. Le *Trésor de la langue française* (TLFi)

Tout cela était attendu. La statistique n'intervient ici que pour confirmer les enseignements de l'histoire. La même enquête, conduite sur les données du TLF, aboutit aux mêmes conclusions, ce qui rend inutile un second développement. Reste à caractériser ces emprunts et à les rattacher à quelque champ du savoir. On s'attend à retrouver dans leur sémantisme les spécificités locales du pays prêteur : le vocabulaire de la vie rurale pour le gaulois, de la mer pour le néerlandais, du bois pour les parlers germaniques, de la montagne pour le franco-provençal, de la guerre et des arts pour l'italien, etc. Peut-on faire un relevé précis de ces emprunts, discipline par discipline ? Cette fois on se propose de croiser les langues, non plus avec les siècles, mais avec les disciplines[10].

10. Si les langues sont liées à la fois aux siècles et aux disciplines, il y a chance que les siècles et les disciplines ne soient pas indépendants. De fait les champs du savoir et de l'activité humaine sont soumis au temps : certains champs disciplinaires sont cultivés dès la plus haute antiquité, comme le droit ou la religion, quand d'autres sciences sont plus

Le tableau 15 reproduit la distribution observée dans le TLF. Le latin est dominant parmi les colonnes, et le droit parmi les lignes, et l'effectif le plus élevé est précisément à l'intersection du droit et du latin. La préférence que le droit donne au latin est si marquée (1262 emprunts au latin contre 52 aux autres langues) qu'on pourrait se dispenser de pondération. Ce sont cependant des données réduites qui sont fournies à l'analyse arborée (figure 16) et à l'analyse de correspondance (figure 17).

Tableau 15. Relevé des emprunts par langue et par discipline dans le TLF

	latin	**grec**	**arabe**	**ang.**	**ital.**	**esp.**	**all.**	**néerl.**	**prov.**
marine	132	5	3	32	45	10	4	41	35
agriculture	43	1	1	5	5	2	0	2	4
armée	41	3	1	3	10	1	0	4	2
industrie	43	3	0	14	6	3	0	5	3
médecine	730	108	4	27	19	7	6	2	1
droit	1262	5	3	11	24	2	3	1	1
commerce	82	1	4	14	11	0	0	2	1
politique	232	16	0	18	8	1	6	0	2
artisanat	189	4	7	21	28	11	3	6	14
religion	549	27	7	13	6	2	1	0	1
économie	395	64	1	27	9	2	5	0	3
musique	280	27	6	12	54	11	1	1	5
peinture	86	5	1	1	9	1	3	0	0
architecture	115	11	1	2	47	10	0	2	5
arts	127	7	0	6	31	0	3	2	1
linguistique	395	64	1	27	9	2	5	0	3
littérature	114	8	1	3	21	3	0	0	1
géographie	82	4	7	7	4	5	3	4	0
histoire	649	71	10	34	73	21	16	8	13
psychologie	216	21	0	18	5	0	8	0	0
astronomie	165	18	1	4	3	0	0	0	0
mathématique	257	19	0	5	1	1	3	0	0
physique	320	31	0	32	6	2	1	0	0
zoologie	280	45	2	10	15	10	3	6	3
botanique	663	57	14	15	20	32	9	9	14
mécanique	96	6	0	4	3	1	1	0	2
métallurgie	25	1	0	5	3	1	1	0	1
imprimerie	48	0	0	4	10	0	0	0	0

La première propose une typologie des disciplines, à partir des emprunts effectués parmi les langues[11]. Le bloc des arts, en haut du graphique, forme un groupe très homogène, qui s'oppose aux activités

---

récentes, comme la psychologie ou l'aéronautique.

11. La perspective est réversible : on aurait pu tout aussi bien mettre en relief la typologie des langues, à partir de leurs contributions aux différents champs du savoir. C'est le point de vue qui avait été adopté dans le graphique 13.

utilitaires situées à gauche et aux sciences constituées, installées dans la moitié basse. Ces dernières se différencient selon qu'il s'agit de sciences dures, de sciences humaines ou de sciences de la vie.

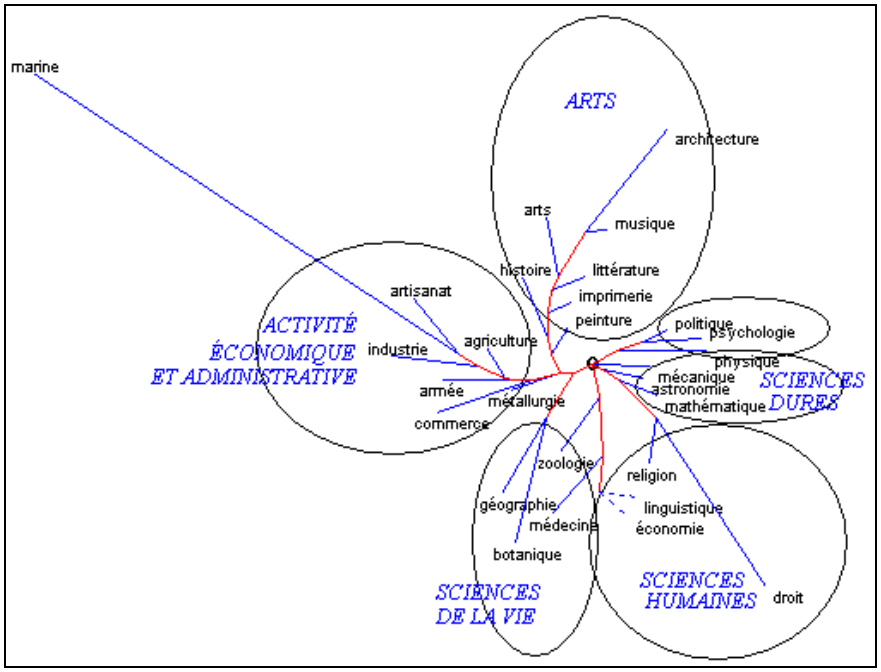


Figure 16. Représentation arborée des disciplines dans le TLF

La lisibilité de la figure 17 est meilleure encore. Car la représentation simultanée des langues et des disciplines facilite l'interprétation. Les diverses sciences, groupées dans le quadrant inférieur gauche, font surtout appel au grec et à l'anglais pour satisfaire leurs besoins de terminologie spécialisée. Le latin jouit quasiment d'un monopole pour tout ce qui touche au droit et à la religion (quadrant supérieur gauche). De l'autre côté de l'axe vertical, on aborde le royaume des arts, tous réunis autour de l'italien, de l'espagnol et de l'arabe. De la littérature à la musique, de la peinture à l'architecture, toutes les voies où s'engage la création artistique mènent vers le sud. Enfin la boucle se ferme avec les activités traditionnelles (commerce, agriculture, artisanat, industrie, armée, marine) auxquelles le pays s'est consacré depuis toujours et dont la terminologie puise dans le fonds ancien de la langue. Le provençal et le

néerlandais ont été alors les fournisseurs, à côté du latin et de l'héritage gaulois et germanique[12].

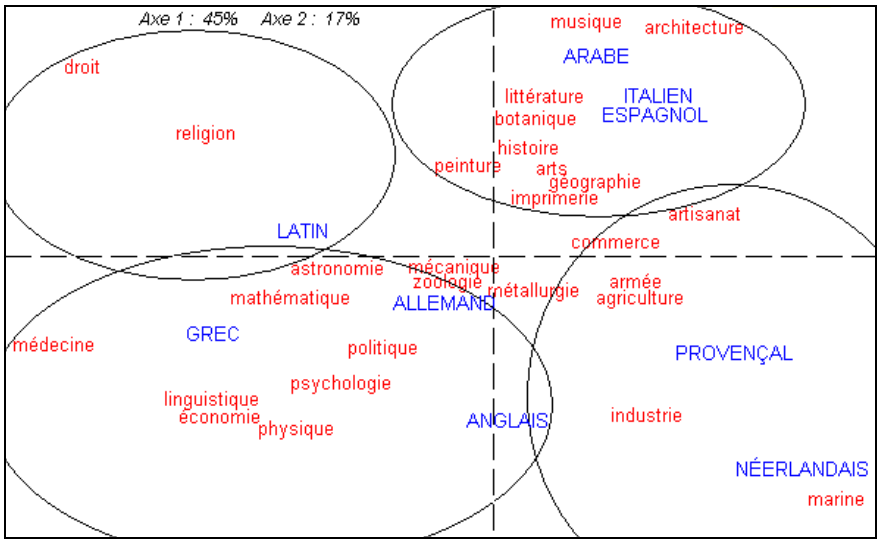


Figure 17. Analyse factorielle des langues et des disciplines dans le TLF

Parmi les nombreuses rubriques que le TLF permet d'exploiter, on peut encore isoler ou croiser le code grammatical, ce qui est réalisé dans la figure 18. C'est à l'évidence la catégorie du substantif qui est retenue de préférence, quand le français a recours à un emprunt, et le plus souvent l'emprunt se fait sans marque de genre ni de nombre. Comme les immigrés candidats à l'intégration, la plupart viennent sans famille et sans bagages. Mais le latin et, à un moindre degré, le grec ont des privilèges : les verbes et les adjectifs peuvent passer directement la frontière, et les substantifs féminins ou pluriels ne sont pas systématiquement dépouillés de leurs caractéristiques. Les autres langues n'ont pas d'autre choix que le substantif, mis à part le provençal et le néerlandais dont l'intégration est plus ancienne. Les langues latines sont naturellement plus accueillantes au féminin que les langues, comme l'anglais, que le genre indiffère.

Reste une question sensible : les immigrés lexicaux ont-ils un faciès ? Peut-on à première vue les reconnaître à leur timbre ou à leur coloration ? Ceux qu'irrite le franglais ont vite fait de dénoncer les

12. Les effectifs pour cet héritage étaient trop étroits pour rentrer dignement dans les calculs.

graphies OO, OA ou ING, et les intrus qui ne savent pas cacher leur K ou leur W. On peut leur objecter que ces caractéristiques s'effacent au bout de quelques générations et que l'orthographe tend à les fondre dans le moule français. Beaucoup des mots qui ont franchi la Manche au XIX^{e} ou avant ne sont plus reconnaissables, comme la REDINGOTE qui recouvre un *riding-coat*. Certains, que les Normands ont d'abord exportés, ont franchi deux fois la frontière, et reviennent au pays après un long séjour en Angleterre. Il n'en reste pas moins que la structure phonétique des langues empruntées laisse des traces que l'analyse statistique peut déceler, même si la conscience linguistique n'y est guère sensible. Ainsi les mots d'origine anglaise ou germanique préfèrent les sourdes T et P, quand la latinité est plus accueillante aux sonores D et B. Les voyelles accentuées (surtout E et E) sont propres au fonds ancien de la langue et les emprunts récents n'y ont guère recours. L'analyse factorielle réalisée sur la composition graphémique des emprunts livrent d'autres informations, dont certaines sont triviales : le y est propre au grec comme son nom l'indique, et la graphie x dénonce le latin. Mais la combinaison des lettres a plus d'intérêt que le recensement des lettres isolées. Et c'est ce dont rend compte le graphique 19.

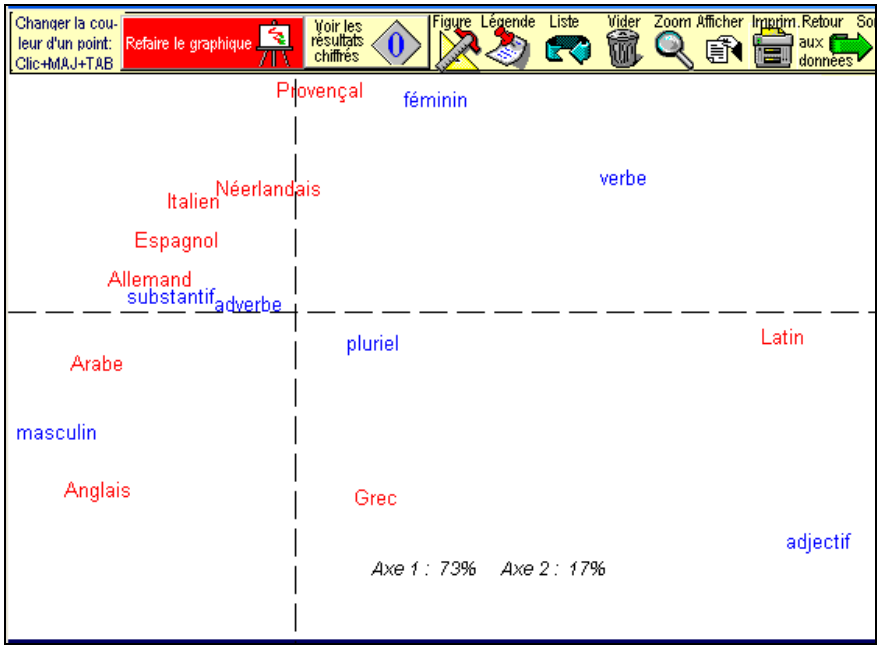


Figure 18. Analyse factorielle des codes grammaticaux dans les emprunts.

Pour ce faire, on a interrogé de nouveau le TLF mais cette fois, au lieu de se contenter de noter les effectifs, on a enregistré la liste de tous les mots qui répondaient au critère retenu : ici la nationalité des emprunts. Cependant les données obtenues pouvaient difficilement être assimilées à un texte suivi. C'étaient des listes, chaque mot étant un hapax. Dès lors les calculs de distance, fondées sur la répétition des mots, n'étaient plus applicables. Le découpage en n-grammes de quatre lettres a permis de sortir de l'impasse et d'obtenir des fréquences pour les séquences retenues. En déplaçant la fenêtre progressivement d'une lettre à la suivante, le mot DECOUPAGE aurait ainsi généré six segments élémentaires : DECO, ECOU, COUP, OUPA, UPAG et PAGE. Ces segments, pouvant être communs à d'autres mots, donnent lieu à une indexation et à différents calculs statistiques, comme celui des spécificités. L'anglais est friand des combinaisons TING, PPER, AKER, UNCH ; le grec des séquences IQUE, OÏDE, POLY, LOGI et de toutes celles qui contiennent CH, PH ou TH, par exemple PHOR, GRAPH, CHRO ; l'allemand est senti confusément dans les mots qui contiennent EISS, HUSS ou HALL et l'arabe se devine derrière une initiale en AR ou dans les combinaisons ABOU, HAMM, CHOU. Quant aux langues latines elles partagent quelques séquences communes : ESQU, ILLE, ELLE, NADE, ASSE, AILL, ETTE.

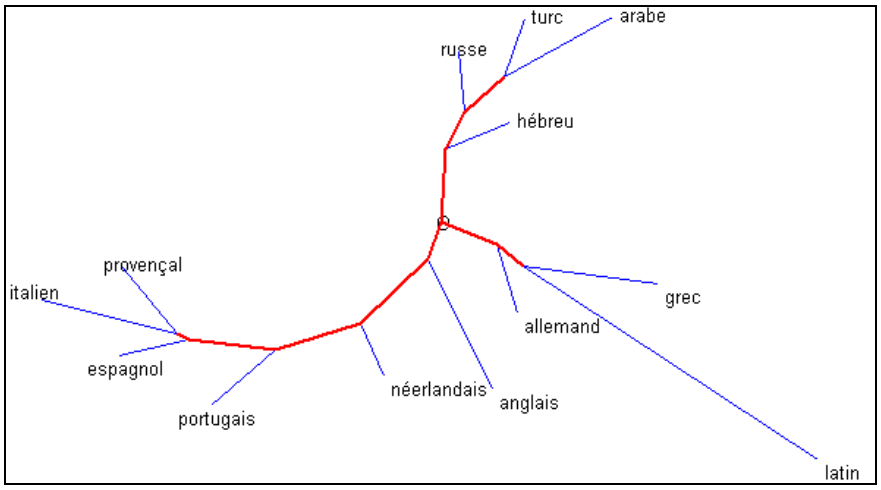


Figure 19. Analyse arborée de la distance lexicale observée dans les emprunts

Le calcul de distance est réalisé sur la totalité des n-grammes recensés dans le corpus. Le principe qui établit cette distance est fondé sur le coefficient de Jaccard, c'est-à-dire un rapport entre les n-grammes communs et les n-grammes exclusifs. Pour deux langues dont on mesure

la distance, les mots communs tendent au rapprochement, les mots privatifs à l'éloignement. Quand toutes les combinaisons des langues deux à deux ont été épuisées, on obtient un tableau triangulaire analogue à celui que fournissent les cartes géographiques, pour les distances kilométriques de ville à ville.

L'analyse arborée (figure 19), exploitant au mieux un tel tableau, en constitue un graphe, qui équivaut à une sorte de carte où la parenté explique les regroupements et les oppositions. Les langues romanes (italien, espagnol, portugais et provençal) se retrouvent entre elles à la gauche du graphique. Le grec et le latin font cause commune à droite, tandis que les langues germaniques se portent au centre. Restent les inclassables (hébreu, arabe, turc et russe), qui sont reléguées dans un *no man's land* en haut de la figure. Le *melting pot* n'est donc pas allé au bout du processus d'intégration et quelques traits physiques indélébiles dénoncent encore les origines.

*
* *

Quand on se fonde sur les dictionnaires, les emprunts ont un caractère officiel. Les mots sont alors naturalisés en bonne et due forme. Mais avant d'apparaître dans les nomenclatures, les mots de l'étranger apparaissent dans la rue ou dans les textes, souvent en contrebande. Il faut des années avant que le filtrage douanier des lexicographes sépare les importations clandestines et les emprunts régularisés, les modes éphémères et les usages qui se maintiennent. Il arrive aussi que le passage de la frontière linguistique soit momentané et provisoire. Beaucoup de mots étrangers, surtout parmi les noms propres, viennent en touristes dans les textes ou la conversation. Ils restent anglais ou russes ou espagnols et arborent souvent un signe distinctif, italique ou guillemets. Le dosage de l'apport étranger dans les textes exige donc des méthodes plus souples et plus sophistiquées, qui ne soient plus seulement des tests de présence/absence, mais des relevés sensibles à la fréquence. Qu'un mot anglais sans équivalent français soit introduit fugitivement dans une brochure technique n'a pas le même impact que l'adoption massive et brutale d'un anglicisme à la mode dans les médias. La base textuelle *Frantext*, qui a servi à la rédaction du TLF, permet de telles investigations. Celle que nous avons entreprise dans la *Nouvelle histoire de la langue française*[13] ne contredit nullement la présente enquête.

13. Sous la direction de Jacques Chaurand, Le Seuil, 1999, p. 673-727 (1999a).

Certes le progrès de l'anglais est fort perceptible dans la production littéraire du XX^e^ siècle, dans les textes aussi bien que dans la nomenclature des dictionnaires. Mais les effectifs sont d'une modicité rassurante : les formes en –ING n'y totalisent que 2400 occurrences dans un corpus littéraire de 120 millions de mots. Et le recensement total, qui cumule les mots anglais de passage et les transfuges incrustés, dépasse à peine 11 000 occurrences. Un millier d'emprunts anglais dans le dictionnaire, dont la plupart assimilés depuis longtemps (soit une entrée sur 50), dix milliers dans les textes littéraires (soit une occurrence sur 10 000), ce n'est pas un raz-de-marée. Il est vrai que le goût des écrivains, soucieux de maintenir la pureté de la langue, est sévère, surtout dans le passé, à l'endroit des emprunts. Les textes techniques n'offrent pas la même garantie, et, s'agissant de langues de spécialité, le mieux à faire, dans ce cas, est de passer le relais à Philippe Thoiron[14].

14. NDÉ : Rappelons que la présente communication s'inscrit dans un hommage au terminologue Philippe Thoiron.