Plaidoyer pour la statistique linguistique

Étienne Brunet

Plaidoyer pour la statistique linguistique

Etienne Brunet

Charles Muller, le maître de la statistique linguistique, assistait au colloque de Strasbourg, à côté de son ami Paul Imbs. On lui prêterait à tort, en cette occasion, un rôle de promoteur de ce qui allait devenir la lexicométrie. Sa conversion n'intervint qu'un peu plus tard, sur le chemin de Besançon, auprès de Quemada, Evrard et Moreau. Mais en 1957, Pierre Guiraud, qui allait publier l'année suivante « Problèmes et méthodes de la statistique linguistique » se trouvait aussi à Strasbourg parmi les intervenants. Certes la statistique ne fut guère évoquée dans les débats (on parlait plutôt de relevés et de dénombrements) et Guiraud choisit de parler d'un sujet un peu moins provocateur : l'argot. Mais les participants au colloque avaient pour la plupart une idée plus confuse encore de l'informatique (le mot n'existait pas alors), que certains confondaient alors – ce n'était pas le cas de Quemada ou de Wagner – avec la mécanographie.

Mais dans les années qui ont suivi le Colloque de Strasbourg, la statistique, présumée un peu plus accessible que l'informatique aux esprits littéraires, bénéficia soudain d'un essor remarquable auquel contribuèrent le Centre d'étude du vocabulaire français de Besançon, l'entreprise du « français fondamental », les travaux d'Evrard en Belgique et du Père Busa en Italie, les publications de Guiraud et, bien sûr, la thèse et le manuel de Muller.

Cela est si vrai que la première publication du TLF a été de nature statistique. C'est en 1971, l'année même où le premier tome du Trésor allait être livré aux lecteurs, que paraissent, avec la signature de Robert Martin, les quatre volumes du *Dictionnaire des fréquences*.

Cinquante ans après on doit reconnaître que la veine statistique dans la mine de Nancy n'a pas été exploitée autant qu'on pouvait l'espérer. Le filon qui devait conduire au trésor n'a pas suscité les vocations attendues, malgré les efforts des directeurs du TLF. Inversement, l'informatique, un peu timide les premières années, a pris un essor spectaculaire dans l'entreprise nancéenne, quand un informaticien de grand talent, Jacques Dendien, a rejoint l'INaLF.

Il n'y a pas concurrence entre l'informatique et la statistique. Les deux sont associées dans les moteurs de recherche, le data mining et la plupart des industries de la langue. Et si l'ATILF rend des services inégalés en matière documentaire, les informations statistiques qu'il distribue sont d'un grand intérêt et donnent à tout le moins les données de base pour les études quantitatives. On se propose d'en donner quelques exemples tirés de la BHVF, du TLFI et de *Frantext*, et peut-être aussi quelques leçons.

Mais avant de procéder à la défense et illustration de la statistique linguistique, il convient, puisque le présent colloque est une rencontre où le témoin va du passé à l'avenir et du bilan au projet, de s'interroger sur les raisons qui expliquent le piétinement d'une discipline scientifique qui s'est pourtant imposée dans la plupart des sciences sociales : sociologie, psychologie, géographie humaine, économie, sciences politiques. Les compteux de mot, pour reprendre une expression qui a cours au Québec, n'ont pas bonne presse, ni chez les linguistes, ni chez les littéraires. Les premiers s'appuient sur des exempliers et trouvent la garantie dans l'attestation d'un fait de langue et non dans sa fréquence. Les seconds se fondent sur les sources, les lectures, la culture et cherchent la garantie dans l'accord avec les jugements d'autrui. En réalité, les réticences s'adressent moins à l'informatique qu'à la statistique. Les vertus domestiques de l'ordinateur ont fini par être reconnues : rares sont les critiques ou écrivains qui écrivent à la main ou qui utilisent encore une machine à écrire. Et depuis qu'*Internet* a gagné en puissance, en extension et en rapidité, les mêmes littéraires sont sensibles aux facilités documentaires que permet le réseau mondial. Ce n'est plus seulement la référence d'un livre qu'on trouve dans le catalogue monstrueux de Google qui tend à se donner l'image de la bibliothèque universelle imaginée par Borges. C'est aussi à l'information primaire, au contenu d'un article ou d'un document, qu'*Internet* donne un accès immédiat. Plus besoin d'attendre, plus besoin de commander, ni même de payer. Certes les moteurs de recherche ne sont pas des critiques avertis : ils mêlent l'ivraie au bon grain. Mais l'indice de notoriété qu'ils utilisent est assez souvent suffisant pour faire apparaître le document pertinent parmi d'autres qui le sont moins. Les détracteurs littéraires de l'informatique et d'*Internet* sont ainsi devenus moins virulents. Ils acceptent les services auxiliaires et ancillaires de la machine, qui réunit les matériaux sur la table de travail. Ils se réservent la part noble de la sélection, de l'exploitation et de l'interprétation. C'est là ce qui compte.

Et pour cela, croient-ils, nul besoin de ceux qui comptent. On ne compte, disent-ils, que ce qui est quantifiable : les prix et les produits, les carottes et les avions. On ne compte pas les idées… Pourtant la démocratie s'exerce en comptant les votes, les opinions, les hommes. Et le jugement littéraire lui-même n'est pas tout à fait dépourvu de compteurs inconscients : beaucoup des jugements qu'on croit qualitatifs sont inspirés par une statistique implicite qui n'avoue pas son nom et qui autorise l'emploi des mots typique, spécifique, caractéristique, si fréquents sous la plume de la critique lorsqu'elle analyse un auteur, un genre ou une époque. Le mot fréquent lui-même relève de cette approche, comme aussi rare, original, banal, courant, cliché, surprise, rupture. Les littéraires parlent d'horizon d'attente quand les statisticiens parlent d'espérance mathématique. L'espérance des uns et l'attente des autres, ce n'est qu'une prévision fondée sur les observations répétées que la conscience enregistre.

Cette réticence des milieux littéraires est d'autant plus regrettable que les données textuelles ont des propriétés très avantageuses que Guiraud avait soulignées, dès 1959, deux ans après le colloque de Strasbourg : « La linguistique est la science statistique type ; les statisticiens le savent bien ; les linguistes l'ignorent encore.[1] » Ces données sont d'abord bon marché ; il suffit d'un traitement de texte pour les enregistrer, ou d'un scanner, ou tout simplement d'une liaison à *Internet* où les textes foisonnent. Elles sont abondantes ; or la statistique se plaît dans les grands nombres et ses conclusions sont moins sûres quand les observations sont en nombre limité parce qu'elles sont chères, ce qui est le cas des enquêtes et des sondages, ou que la nature des choses le veut ainsi, comme il arrive en médecine. Les données textuelles sont faciles à contrôler, à reproduire, à transmettre, à corriger, et se prêtent volontiers à la répétition et aux variations des expériences. Elles sont exemptes de filtrage sélectif et subjectif, et la conscience du chercheur intervient peu à l'entrée. Enfin et surtout les textes sont immédiatement interprétables, sans le truchement des machines. La lecture et la connaissance du texte sont des garants contre les calculs aberrants et les manœuvres avortées. La conscience a au moins une idée de l'ordre de grandeur des résultats attendus et peut les rejeter s'ils sont exagérément erronés ou délibérément triviaux, alors que d'autres disciplines travaillent

[1] Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, D. Reidel Publishing Company, Dordrecht-holland, 1559, p. 15.

sans vision directe et se trouvent liées aux instruments, sans possibilité de récuser leur témoignage.

Ce qui est un avantage est aussi un inconvénient. Puisqu'en matière littéraire ou linguistique, on peut se débrouiller sans appareillage, pourquoi s'encombrer de méthodes indirectes et grossières qui obscurcissent l'évidence ou brutalisent les nuances. Pour beaucoup d'esprits, le refus d'accueillir les méthodes quantitatives est une manière de réserver un espace de liberté pour la conscience individuelle. Qu'il y ait au moins un domaine préservé, une réserve naturelle comme il y des parcs du même nom pour la sauvegarde des paysages, un refuge contre l'invasion technologique, où puissent vivre et subsister les disciplines menacées : l'art, le langage, la philosophie, la religion, la musique, la cuisine…

Là s'arrête notre plaidoirie à charge et à décharge. La défense et illustration de la statistique ne passe pas par des discours mais par des exemples. Nous nous proposons d'en donner quelques-uns.

Avec près de 4 000 textes de la littérature nationale, engrangés méthodiquement depuis trente ans, *Frantext* n'a guère d'équivalent dans les autres langues, ni pour l'étendue, ni pour l'homogénéité des données, ni même pour leur accessibilité. La recherche de contextes est de loin la fonction la plus utile et la plus sollicitée de *Frantext*. Certains utilisateurs souhaiteraient certes que la transmission des textes s'ajoute à celle des contextes. Mais ce serait violer les prescriptions du copyright qui n'autorisent qu'un maximum de 300 caractères s'il s'agit d'un texte sous ayant droit. En revanche on n'a pas limité le nombre des contextes restitués pour un mot donné (ou une liste de mots). Les sorties peuvent donc se développer à loisir et cela est favorable à la statistique qui n'aime pas être à l'étroit.

Car l'utilité de *Frantext* ne se réduit pas aux seules opérations documentaires, si sophistiquées soient-elles. Les fonctions statistiques qu'on y trouve ne le cèdent en rien pour la puissance et la portée, et leur exploitation intensive et systématique permet d'atteindre des résultats dont la conscience linguistique, réduite à ses seuls moyens, serait incapable. Ce domaine est toutefois plus technique et moins familier aux populations littéraires qui constituent la clientèle privilégiée de *Frantext*. Et pour ne pas trop effrayer les néophytes, les fonctions statistiques offertes par le logiciel *Stella* ont une simplicité voulue, qui s'arrête aux pourcentages. Cela est suffisant pour donner une idée de la distribution

d'une forme parmi les époques, les écrivains, les textes ou les genres, et ce qui vaut pour une forme peut s'étendre à une constellation lexicale constituée librement autour d'un thème ou d'une construction syntaxique.

Mais trop de simplicité peut conduire à l'erreur d'interprétation et il est dangereux d'accorder une confiance illimitée aux effectifs non pondérés ou aux méthodes trop frustes de pondération que sont les pourcentages et les fréquences relatives. Pour tirer pleinement profit du gisement, il a paru utile d'installer à la sortie de *Frantext* une unité de transformation, qui puisse assurer le traitement quantitatif des matériaux. C'est l'objet du présent logiciel dont le nom THIEF indique assez sa relation de dépendance parasitaire à l'égard de *Frantext* et dont l'écran d'accueil (figure 1) propose deux choix principaux, selon qu'on souhaite être en liaison directe ou différée avec *Frantext*. Dans le premier cas on activera les fonctions rangées verticalement sur la droite de l'écran. Dans le second, on ne s'intéressera qu'aux fonctions disponibles horizontalement au haut de l'écran.

L'exploitation statistique off line

Comme on ne manipule ici que des nombres, en dehors de tout contexte, l'embarras du copyright n'est plus à craindre. Et le traité signé à Genève le 20 décembre 1996 exclut les banques de données des mesures de protection qui entourent la propriété intellectuelle et artistique. Il a donc été possible de livrer au public un dictionnaire de fréquences, qui est issu de *Frantext* par les voies autorisées et qu'on pourrait remettre à jour par les mêmes moyens. L'avantage recherché n'est pas seulement de faire l'économie des liaisons télématiques car, une fois réglé l'abonnement annuel et forfaitaire, le coût pour l'usager est le même, quels que soient le nombre et la longueur des séances de consultation. Ce qu'on recherche surtout, c'est le gain de temps et le confort de l'utilisation que permet tout traitement local. La base locale est accessible aux boutons rangés au haut de l'écran (figure 1). Elle rend compte de l'usage littéraire, les textes dits « techniques » ayant été volontairement écartés. Si elle ne livre pas l'exhaustivité du corpus actuellement disponible, qui s'accroît chaque année, elle n'en représente pas moins l'essentiel des données de *Frantext*, soit 2 376 textes et 117 millions d'occurrences, c'est-à-dire le contenu littéraire de *Frantext* en 1995.

L ’exploitation statistique de Frantext

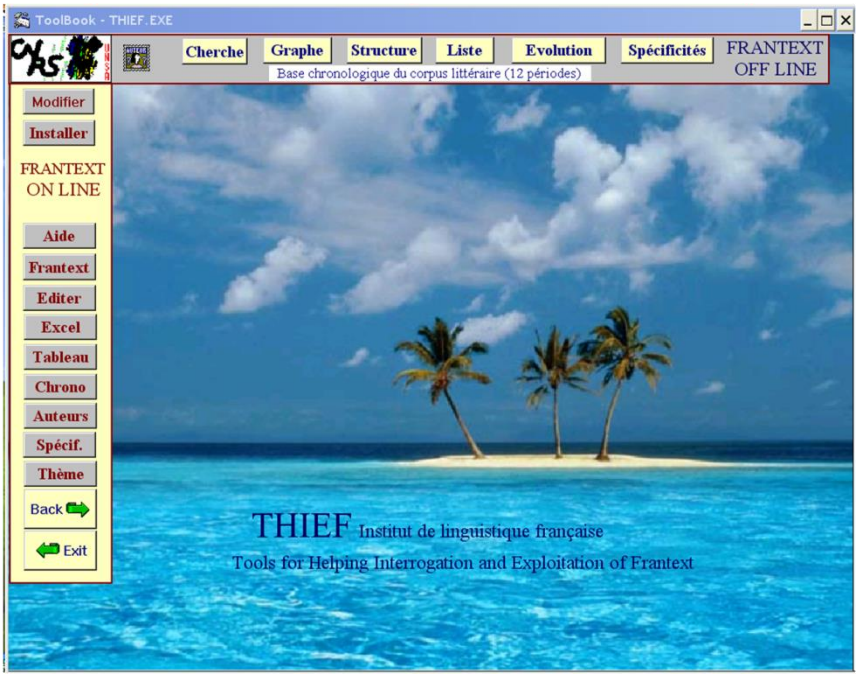


Figure 1. Le logiciel THIEF. Écran d’accueil de la version Windows

	Nb,mots	Nb,Formes	prob,p	prob,q	époque
1	1 719 178	67 014	0,014625	0,985375	1550
2	8 346 862	101 892	0,071006	0,928994	1630
3	6 087 533	69 612	0,051786	0,948214	1692
4	9 380 093	77 841	0,079796	0,920204	1735
5	11 946 384	99 028	0,101627	0,898373	1780
6	11 124 272	98 90	0,094633	0,905367	1820
7	16 184 517	124 845	0,13768	0,86232	1855
8	13 780 168	116 085	0,117227	0,882773	1885
9	8 695 375	98 488	0,073971	0,926029	1910
10	11 361 661	109 218	0,096653	0,903347	1928
11	10 083 262	106 498	0,085777	0,914223	1942
12	8 842 284	112 367	0,07522	0,92478	1960
TOTAL	117 551 589	393 848			

Tableau 2. Limites des 12 tranches

Cette base locale comprend le relevé de toutes les formes du corpus littéraire et des sous-fréquences de chacune dans 12 tranches

chronologiques distinguées du XVI^e^ siècle à nos jours. Les limites des tranches n'ont pu être établies sur un pied d'égalité, car les textes dépouillés sont très inégalement répartis selon les siècles. Afin d'équilibrer la taille des sous-ensembles, l'empan chronologique a été élargi là où les textes étaient rares, c'est-à-dire au XVI^e^ siècle, et resserré là où ils abondaient, aux XIX^e^ et XX^e^ siècles. La première tranche s'étend ainsi sur un siècle (on l'a représentée par son année médiane : 1550) tandis que les plus proches ne recouvrent guère que deux décennies. Voir tableau 2.

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

Ajoutons qu'une autre base a été extraite des données de *Frantext*, pour rendre compte non plus de la chronologie, mais de l'écriture propre aux écrivains. On a choisi ceux qui étaient les mieux représentés dans *Frantext*, soit 70 écrivains de Rabelais à Gracq, et pour chacun d'eux une fonction puissante de *Stella* a permis de recueillir le dictionnaire des fréquences observées dans son œuvre. Nous ne montrerons ni l'une, ni l'autre de ces deux bases, qui ont été largement exploitées et qui exploitent les fonctions statistiques de notre logiciel HYPERBASE, mais non les fonctions documentaires puisque seules les données numériques y sont accessibles.

Il semble opportun de montrer plutôt comment on pourrait les reconstruire, à partir du corpus enrichi de *Frantext*, ou, mieux encore,

---

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

comment on peut exploiter *Frantext*, en direct, sans s'enfermer dans un corpus figé.

L'exploitation statistique de *Frantext* on line

Les fonctions statistiques apparaissent clairement dans les propositions de *Frantext*, à côté des fonctions purement documentaires. On les a mises en relief dans la figure 3 :

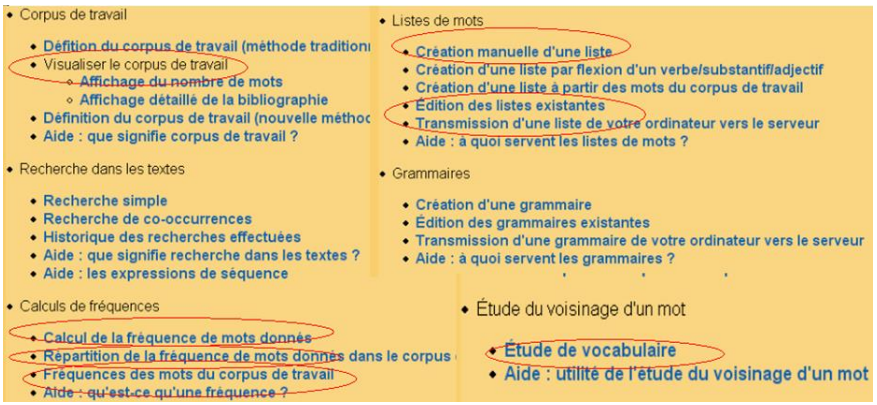


Figure 3 : Les fonctions de *Frantext*

1. L'évolution d'un mot ou d'un groupe de mots

La fonction la plus simple est celle qui rend compte de l'évolution d'un mot (ou d'une liste de mots). Elle suppose comme toute interrogation adressée à *Frantext* que l'on ait d'abord choisi un corpus de travail. On le choisira généralement large, pour permettre au temps de faire son effet. Le bouton à solliciter est le deuxième du groupe « Calculs des fréquences » et s'intitule « Répartition de la fréquence… ». Bien sûr il incite à sélectionner un mot ou une liste mais il permet aussi deux options : l'une s'engage du côté des auteurs, l'autre du côté de la chronologie. C'est ce dernier choix qu'il faut faire en précisant en outre la périodicité que l'on veut, par exemple 20 ou 30 ans.

Si le mot est rare, ou les tranches courtes, il peut se faire que l'effectif étant nul dans une tranche particulière, il n'en soit pas fait mention dans les résultats. Or cette absence peut constituer un déficit intéressant. Le calcul prend donc pour base uniquement les périodes qui contiennent au moins une fois le mot considéré. Les tranches chronologiques qui ont complètement ignoré le mot en question sont

tenues à l'écart du vote et considérées comme abstentionnistes. Ce cas ne se produit guère, si le mot est fréquent ou s'il s'agit d'une liste de mots. Et cela peut se justifier si le mot est de création récente et qu'un déficit dans les siècles anciens est une évidence trop triviale pour être soulignée.

Quand le fichier des résultats a été sauvegardé[3], on peut abandonner *Frantext* et le navigateur et solliciter le bouton *Chrono* du menu principal de THIEF. Le résultat final est un histogramme qui apparaît d'abord sur l'écran puis est transmis à l'imprimante, le cas échéant (figure 4).

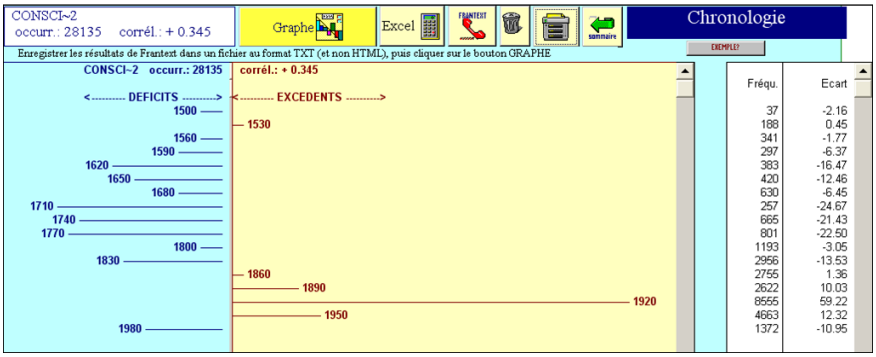


Figure 4. Évolution du mot CONSCIENCE dans le corpus entier de *Frantext* 2008

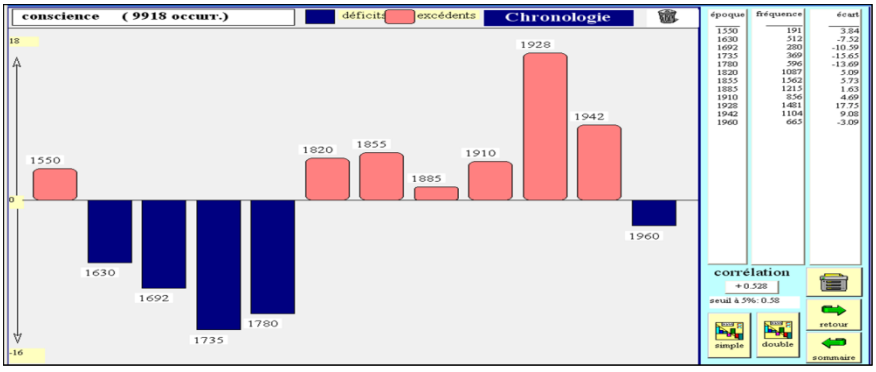


Figure 5. Évolution du mot CONSCIENCE dans le corpus littéraire de *Frantext* 1995

[3] Prendre garde à enregistrer les résultats en mode texte. Malheureusement certains navigateurs ignorent dans ce format les tabulations du fichier original, si bien que les nombres étant collés les uns aux autres la lecture en devient impossible. Le navigateur *Firefox* ne présente pas cet inconvénient, non plus que *Camino* dans le monde Mac.

Si l'on préfère une autre présentation, solliciter le bouton « Excel », en précisant que les données sont dans le fichier *chrono.xls*. On peut aussi solliciter la base locale (off line) qui donne un résultat équivalent où la faveur du mot CONSCIENCE doit certes beaucoup à Freud, mais aussi au mouvement spiritualiste qui l'a précédé et dont Bergson est la figure de proue (figure 5). Mais les choix ne sont plus modifiables, ni dans la composition du corpus ni dans la périodicité.

2. La répartition d'un mot ou d'un groupe de mots chez les écrivains

Il s'agit là de la deuxième option du programme « Répartition » de *Frantext*. La procédure d'interrogation et de sauvegarde est la même. Mais le résultat est sensiblement plus long, ayant autant de lignes que d'écrivains recensés pour le mot en question, même si ce mot ne figure qu'une fois dans son œuvre. On en donne ci-dessous un extrait, relatif au mot AME (figure 6).

Fréquence absolue totale : *59303*
Fréquence relative maximale : *4977* chez *RODENBACH Georges*

Diagramme des fréquences relatives
Échelle : un astérisque représente une fréquence relative de *200 millionième(s)

		freq. abs.	freq. rel.	
1	RODENBACH Georges	108	4977	
	*************************			
2	SAMAIN Albert	84	3807	
	********************			
3	LESPINASSE Julie de	556	3374	******************
4	DESBORDES-VALMORE Marceline	304	3261	******************
5	NOAILLES Anna de	43	2953	***************
6	POULET Georges	19	2533	*************
7	GUÉRIN Charles	53	2458	*************
8	GUÉRIN Maurice de	509	2407	*************
9	GUÉRIN Eugénie de	586	2119	**********
10	MONOD Henri	70	1959	**********
11	BRIZEUX Auguste	48	1927	**********
12	DIERX Léon	80	1817	**********
13	MAETERLINCK Maurice	259	1762	*********
14	MAINE DE BIRAN	608	1715	*********
15	JOUBERT Joseph	173	1639	*********
16	LAFORGUE Jules	40	1576	********
17	MORÉAS Jean	83	1564	********
18	SULLY PRUDHOMME Armand	85	1536	********

Figure 6. Le résultat proposé par *Frantext* pour le mot AME

Le résultat final est un histogramme qui apparaît d'abord sur l'écran puis est transmis à l'imprimante, si on le désire. Noter que la place est

trop exiguë pour représenter la série dans son intégralité. Dans certains cas, c'est plus d'un millier d'auteurs qui sollicitent chacun un « bâton » de l'histogramme. Le programme permet de ne retenir que les plus importants, c'est à dire les mieux représentés dans le corpus. Le seuil proposé pour l'étendue est de 200 000 occurrences, mais l'utilisateur peut déplacer ce seuil vers le haut ou le bas.

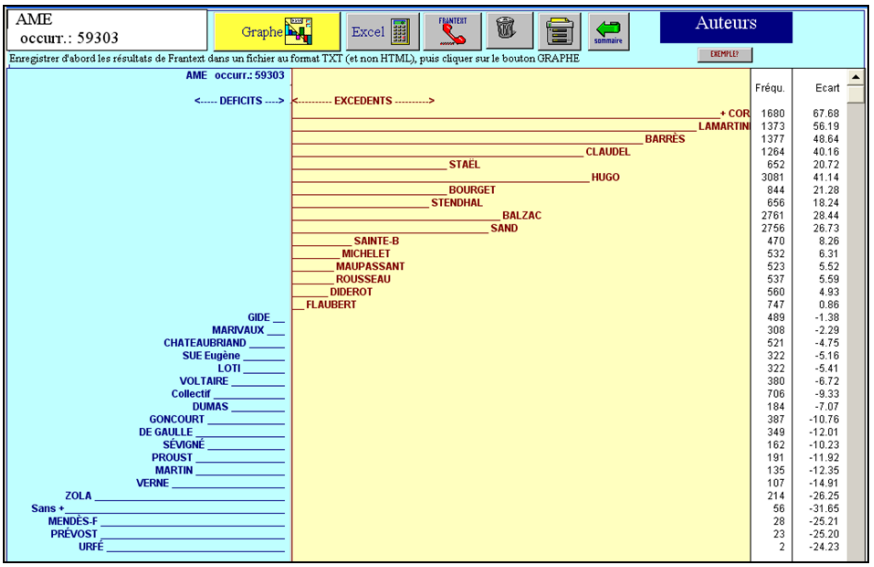


Figure 7. Le graphique du mot AME dans *THIEF*

Le classement est fondé sur les fréquences relatives. La décroissance est régulière. Le même classement est repris dans l'histogramme proposé par THIEF (graphique 7). Mais il est contesté par l'écart réduit, qui suit la tendance générale mais s'en écarte dans les détails. Si on teste le supplément d'âme, l'écart réduit donnerait l'avantage à Hugo sur Mme de Staël et à Balzac sur Stendhal. Les fréquences relatives ont l'inconvénient d'ignorer la taille des observations et d'établir l'égalité entre 20/100 et 20 000/100 000, ce qui est fort différent en terme de probabilité. Là-dessus on renvoie le lecteur à l'avertissement de Charles Muller qui dès 1968 opposait l'impasse et le bon chemin[4]. Comment retrouver le bon

[4] « Cette présentation par pourcentages a une petite allure statistique, mais il faut résister à la tentation de l'appliquer, car c'est une impasse qui n'aboutit à aucune conclusion. La bonne façon n'est d'ailleurs pas plus compliquée […] Une simple règle de trois conduit à une valeur théorique, à partir de laquelle on obtiendra un écart absolu avec la valeur observée ou réelle. », *Initiation à la statistique linguistique*, 1968, p 44.

chemin ? Prenons la ligne relative à Albert Béguin qui indique une fréquence absolue de 330 et une fréquence relative de 1 488. Sachant que la fréquence relative a été obtenue en divisant la fréquence du mot par la taille du texte (résultat multiplié par 1 000 000), on retrouve la taille du texte en faisant l'inverse, soit : (330 / 1 488) × 1 000 000 = 221 774. On procède ainsi pour toutes les lignes, en faisant le total pour la fréquence du mot et la taille du corpus entier, soit : fréquence de âme : 59 303, taille du corpus : 202 633 773. On peut alors calculer la fréquence théorique par la règle de trois : $59\,303 \times (221\,774/202\,633\,773) = 64{,}90$

puis l'écart absolu : $330 - 64{,}90 = 265{,}10$

et enfin l'écart réduit (formule simple) : $z = 265{,}10 / \sqrt{64{,}90} = 32{,}90$

3. Constitution et traitement des tableaux

Cette fonction est plus riche et plus puissante que les précédentes. Mais elle requiert une consultation plus longue de *Frantext*. Car il faut de façon répétitive interroger *Frantext* en lui proposant la même liste de mots mais en variant la composition du corpus. Le choix de la liste peut être fait en conversationnel et se maintient tant que dure la communication. Mais on peut l'établir dans un fichier local qu'on transmet à *Frantext* au moment de la consultation. Quand on a épuisé la liste des auteurs (ou des textes, des genres ou des époques), on dispose d'autant de fichiers, qu'on va réunir dans un tableau (tableau 8), chaque fichier servant à remplir une colonne.

GRAPHIQUE CLIC sur un mot ou un texte. — Lecture du tableau — Tableau (EXEMPLE) — ARBRE — FACTOR — ECART FREQU. — EXCEL — Sommaire

EFFACER un mot CLIC+MAJ

	ALEM	BALZ	CLAU	DIDE	FLAU	GIDE	GRAC	HUGO	MARI	PROU	RIMB	SART	STEN	VERN	VIGN	VOLT	ZOLA
ils	524	9565	2027	3517	5261	2920	969	8080	1725	2634	66	2295	1337	2939	928	555410178,	60519
je	495	46084	18600	17683	32501	34026	5406	42216	21346	15127	424	8155	9929	2484	4902	1047014094,	283942
leur	682	8806	1383	2174	3037	2153	1101	5988	1071	2552	66	776	1120	1968	905	3825	5207, 42814
lui	1255	22796	5535	5464	7208	6862	2482	13995	5075	7095	55	1747	5876	2792	1650	683018962,	115679
m'	109	10405	3745	4032	8720	9060	1445	8260	5361	4347	72	1593	2063	369	1135	2440	2954, 66110
me	134	15891	5559	6084	10964	12458	2439	11024	8535	7312	126	2428	4160	566	1878	3049	3938, 96545
moi	123	9352	6347	3476	6148	5048	1395	9824	4017	3734	61	1557	2077	508	1190	1951	3587, 60395
nous	1661	17876	10025	5299	8766	9493	2404	16140	4814	6676	188	2021	2092	2762	1341	4991	5901, 102450
on	1104	13579	4812	9129	10679	8525	4304	26133	5064	6820	138	2145	4154	4117	1972	1176014579,	129014
se	1184	33073	5691	6941	11991	8963	4894	21127	2917	6051	142	3268	6443	7271	2213	703927885,	157093
soi	8	302	184	179	288	406	132	361	41	232	0	50	124	23	61	121	165, 2677
t'	47	2226	1307	343	3503	1209	271	3285	487	147	15	600	187	53	206	474	1265, 15625
te	72	3615	1946	385	3363	1410	353	3620	772	251	34	797	380	94	217	500	1842, 19651
toi	52	1963	2377	256	2655	1101	262	3550	351	134	23	547	203	69	184	559	1319, 15605
tu	119	8918	6002	1365	9080	3657	1284	9311	1750	730	87	2955	726	322	601	1398	6365, 54670
vous	488	49764	12438	16415	16322	8443	1402	37687	19743	6183	89	3214	5346	3034	2774	1074112975,	207058
	8057	254215	87978	82742	140486	115734	30543	220601	83069	70025	1586	34148	46217	29371	22157	71702131216,	1429847

Tableau 8. Un tableau de contingence

À l'intersection d'une ligne et d'une colonne, on lit alors la fréquence du mot *x* dans le corpus *y*. Dès lors toutes les manipulations qu'on peut faire avec un tableau de contingence sont possibles. Il suffit de

s'appuyer sur le total général et les totaux marginaux de ligne et de colonne pour produire fréquences théoriques et écarts réduits, et conséquemment histogrammes, analyses arborées et analyses factorielles (figure 9). La page Tableau de THIEF offre tous ces outils traditionnels du traitement statistique.

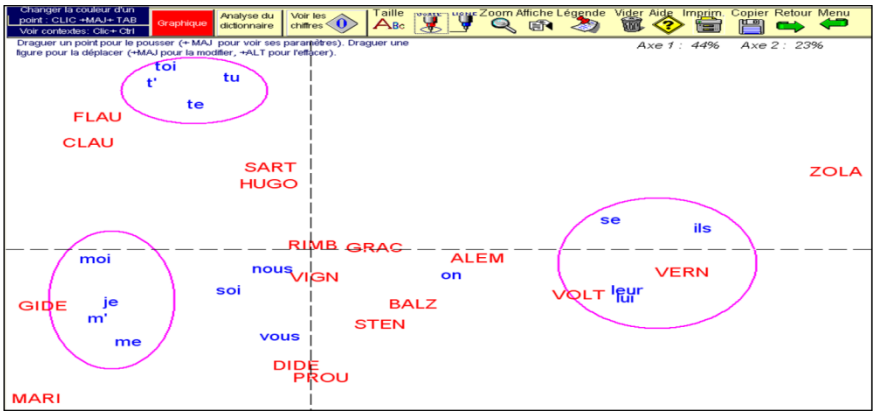


Figure 9. Analyse factorielle du tableau 8

4. Le vocabulaire spécifique d'un texte

Cette proposition de *Frantext* est puissante et précieuse, puisqu'elle permet en quelques secondes de connaître et d'enregistrer la liste alphabétique de tous les mots d'un corpus, si gros soit-il, chaque forme étant suivie de sa fréquence. On appelle cela un dictionnaire de fréquences. Les fréquences sont absolues et livrées brutes. Mais il est facile d'en tirer parti, si on dispose d'un élément de comparaison. THIEF propose un tel modèle, qui est lui-même issu de *Frantext* et qui regroupe les textes littéraires, soit 117 millions de mots (cela correspond à la base locale évoquée plus haut). Prenons pour exemple le Roman de Gracq, *Un Balcon en forêt*, actuellement au programme de l'agrégation. Si un agrégatif se trouve dans la salle, il sera peut-être intéressé par la figure 10, qui dresse la liste des spécificités de ce texte, en ordre hiérarchique et alphabétique. Il n'y a pas lieu d'admirer la position avantageuse des noms propres (Grange, Meuse, Mona) liée à des privilèges évidents. Mais que le premier terme significatif soit le mot «forêt » n'est pas dû au hasard. À lire la liste de ces descripteurs quantitatifs, on devine le sujet et l'atmosphère, comme s'il s'agissait d'une fiche signalétique.

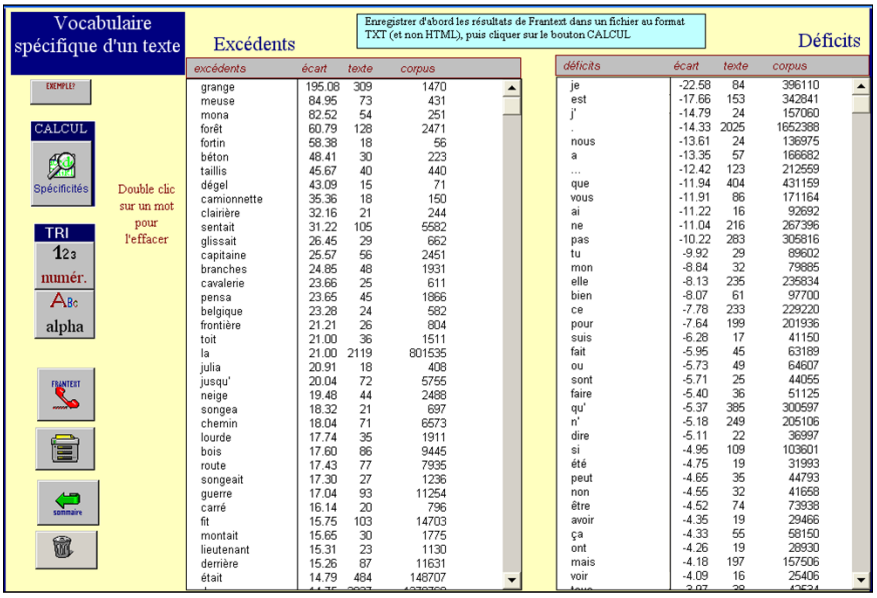


Figure 10. Le vocabulaire spécifique d'*Un balcon en forêt*

5. Étude du voisinage d'un mot

Cette fonction de *Frantext* ressemble à la précédente et livre une liste semblable dotée de fréquences brutes. Mais les mots retenus ne sont pas ceux d'un texte suivi, mais d'un agrégat de contextes, réuni autour d'un mot-pôle (ou d'une liste de mots). Il appartient au chercheur ici comme partout de définir le corpus, puis de proposer un mot (ou une liste de mots) que le logiciel *Stella* doit rechercher dans le corpus. *Stella* ne montre pas le monceau de contextes qu'il a trouvés mais en restitue le contenu sous la forme d'un dictionnaire de fréquences. On s'attend que ces fréquences soient liées par quelque raison, thématique ou syntaxique, voire métrique, au mot-pôle, mais la fréquence d'un mot dépend aussi de lui-même[5], quel que soit le contexte. Quel que soit le mot-pôle, il est probable que les mots HOMME, JOUR et FEMME[6], que l'on voit presque à

[5] Pierre Guiraud avait imaginé que la fréquence était une propriété attachée à chaque mot de la langue. On lui a objecté que cette propriété était variable et dépendait de la situation de discours. Il n'en reste pas moins que la disponibilité n'est pas la même pour tous et qu'il est rare qu'un mot soit premier dans un discours et dernier dans un autre. On voit souvent les mêmes au premier rang.

[6] Ce sont les trois substantifs que *Frantext* place en tête dans le corpus littéraire.

chaque page d'un texte en français, vont se presser au premier rang, que le mot proposé soit TENDRESSE ou ARTICHAUT.

Base textuelle FRANTEXT

Étude du vocabulaire dans le voisinage des occurrences d'un mot
Paramètres de l'étude
Nombre d'occurrences du pivot : *1836*
Taille en occurrences des voisinages explorés : *70073*
Nombre de graphies trouvées dans ces voisinages : *10282*

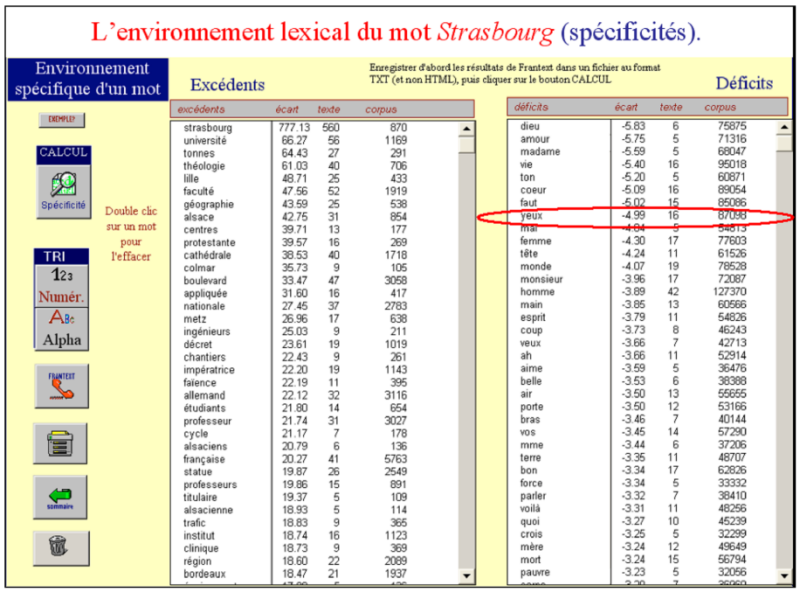
Résultats

1	%me
1	%xbre
1	%xiie
3	'
3	**
1	*à
7	*abbé
1	*achard
1	*achenheim
1	*adam
1	*adolf
1	*adolphe
5	*afrique
1	*agentur
1	*agrippina
6	*aix
1	*ajaccio
1	*alba
2	*albert
1	*alemaigne
2	*alexandre
1	*alfred
5	*alger

Voisinage du mot Strasbourg

1	wissembourg
2	witemberg
1	wittenberg
1	worms
2	xie
1	xiv
1	xix
1	xl
1	xvie
2	xviie
3	xviii
1	xviiie
1	xxe
1	xxix
1	xxx
1	xxxvii
1	yacht
16	yeux
2	yo
1	york
1	yougoslavie
1	yseut
1	yves
2	zapfenschuppe
1	zarathoustra
1	zélés
2	ziegler
1	zigzaguaient
8	zone
1	zoologie
1	zouaves
1	zum

Tableau 11. Le voisinage du mot STRASBOURG. Résultats bruts.



excédents	écart	texte	corpus
strasbourg	777.13	560	870
université	66.27	56	1169
tonnes	64.43	27	291
théologie	61.03	40	706
lille	48.71	25	433
faculté	47.56	52	1919
géographie	43.59	25	538
alsace	42.75	31	854
centres	39.71	13	177
protestante	39.57	16	269
cathédrale	38.53	40	1718
colmar	35.73	9	105
boulevard	33.47	47	3058
appliquée	31.60	16	417
nationale	27.45	37	2783
metz	26.96	17	638
ingénieurs	25.03	9	211
décret	23.61	19	1019
chantiers	22.43	9	261
impératrice	22.20	19	1143
faïence	22.19	11	395
allemand	22.12	32	3116
étudiants	21.80	14	654
professeur	21.74	31	3027
cycle	21.17	7	178
alsaciens	20.79	6	136
française	20.27	41	5763
statue	19.87	26	2549
professeurs	19.86	15	891
titulaire	19.37	5	109
alsacienne	18.93	5	114
trafic	18.83	9	365
institut	18.74	16	1123
clinique	18.73	9	369
région	18.60	22	2089
bordeaux	18.47	21	1937

déficits	écart	texte	corpus
dieu	-5.83	6	75875
amour	-5.75	5	71316
madame	-5.59	5	68047
vie	-5.40	16	95018
ton	-5.20	5	60871
coeur	-5.09	16	89054
faut	-5.02	15	85086
yeux	-4.99	16	87059
mal	-4.94	9	54513
femme	-4.30	17	77603
tête	-4.24	11	61526
monde	-4.07	19	78528
monsieur	-3.96	17	72087
homme	-3.89	42	127370
main	-3.85	13	60566
esprit	-3.79	11	54826
coup	-3.73	8	46243
veux	-3.66	7	42713
ah	-3.66	11	52914
aime	-3.59	5	36476
belle	-3.53	6	38388
air	-3.50	13	56655
porte	-3.50	12	53166
bras	-3.46	7	40144
vos	-3.45	14	57290
mme	-3.44	6	37206
terre	-3.35	11	48707
bon	-3.34	17	62826
force	-3.34	5	33332
parler	-3.32	7	38410
voilà	-3.31	11	48256
quoi	-3.27	10	45239
crois	-3.25	5	32299
mère	-3.24	12	49649
mort	-3.24	15	56794
pauvre	-3.23	5	32056

Figure 12. Le voisinage de STRASBOURG. Calcul des spécificités

6. Recherche de contextes

C'est là la voie royale de *Frantext*, nullement cataloguée comme statistique. Elle permet pourtant une démarche quantitative lorsqu'on assemble dans un fichier les contextes qu'elle restitue. Un tel fichier a nécessairement été créé dans la fonction précédente, qui explore le voisinage d'un mot. On fait donc ici un pas en arrière en demandant à la machine locale de refaire ce qu'avait fait le serveur. Il en ressort une liste de spécificités comme précédemment. Mais cette fois, comme on dispose du texte, les investigations peuvent aller plus loin. Au lieu de calculer seulement la liaison de chaque mot avec le pôle unique, on peut étudier les relations que chacun des mots entretient avec tous les autres. On rejoint là la démarche très générale où l'étude systématique des cooccurrences aide à trouver des solutions pour des problèmes divers : le data mining, les candidats-termes, la traduction assistée, le résumé automatique, le traitement des profils…

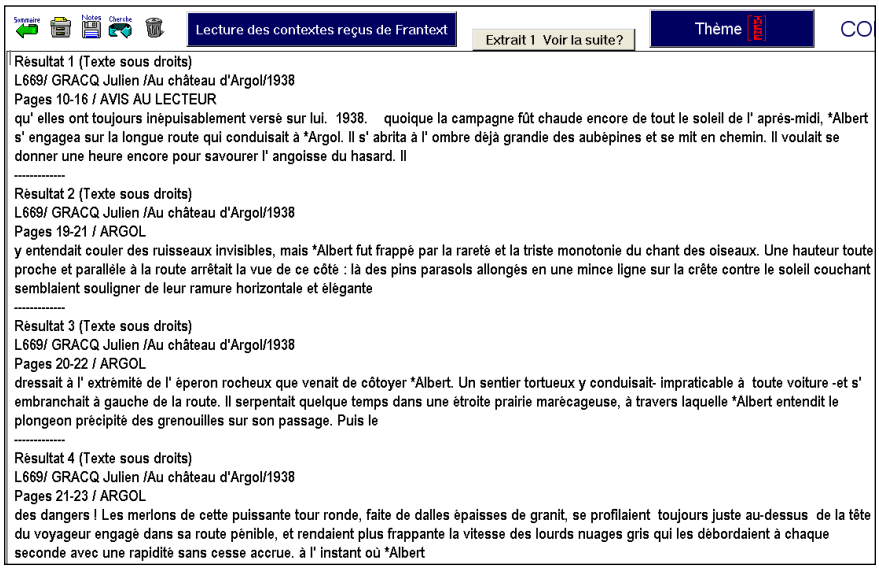


Tableau 13. Les contextes du mot ROUTE dans l'œuvre de Gracq (extrait)

Quand le traitement des cooccurrences est achevé, plusieurs outils permettent d'appréhender le tableau triangulaire où chaque mot est relié à chacun des autres par un indice de proximité, à l'image des atlas où la distance de ville à ville se lit dans un triangle analogue. Les ressources habituelles s'appliquent ici : histogrammes, analyses arborée et

factorielle. Nous y ajouterons un graphe où la ROUTE est un carrefour où se rejoignent les compagnons de ROUTE ; les arcs sont en rouge quand il s'agit d'une liaison avec le pôle, en bleu lorsqu'il s'agit d'une ligne intérieure ou transversale. La force du trait (gras, maigre ou pointillé) indique la force de la liaison. Les liaisons de la ROUTE[7] sont bien celles qu'on attendait chez un écrivain-arpenteur qui a beaucoup marché et circulé dans la campagne française et qui est l'un des premiers paysagistes de notre littérature.

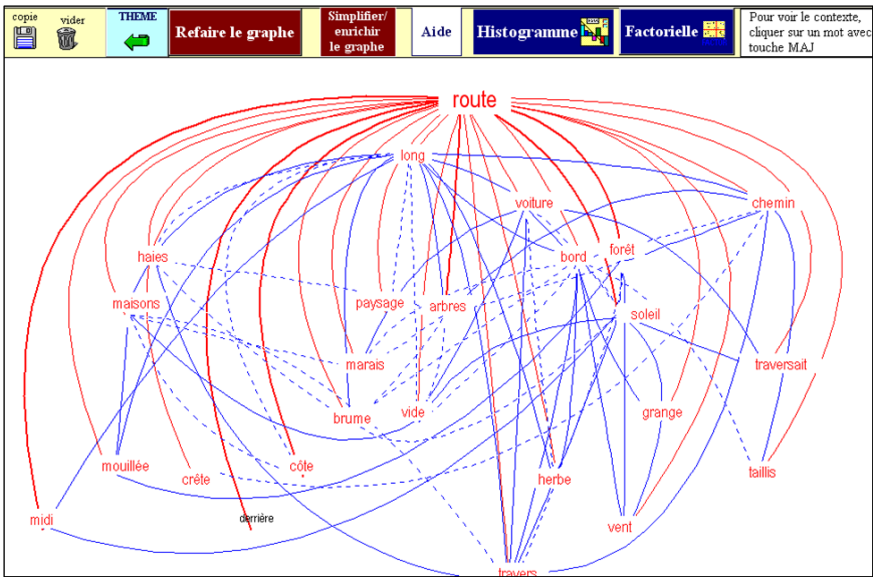


Figure 14. Le graphe de la ROUTE

Conclusion

1- Le graphe qu'on vient de montrer rend possible le retour au texte. Il suffit de cliquer sur un mot pour voir apparaître les contextes où on l'emploie. Un tel retour est malheureusement impossible pour la plupart des fonctions statistiques disponibles dans *Frantext*, qui ne transmettent que des nombres. Si *Frantext* délivrait aussi les textes, les ateliers spécialisés de statistique auraient un matériau plus fruste mais plus riche. Or là où le copyright n'est pas un droit opposable, cette libre transmission

[7] Un de ce que Gracq qualifie de « fragment » porte ce nom. Et son dernier livre a pour titre « Carnets du grand chemin ».

du trésor serait très souhaitable. Au reste il y a des précédents, dont ont bénéficié l'ARTFL de Chicago, la Bibliothèque Nationale et quelques privilégiés, dont je suis.

2- Dans son état actuel les fonctions statistiques de *Frantext* ne s'appliquent qu'aux formes graphiques On souhaiterait qu'elles s'étendent aux lemmes, puisqu'aussi bien les textes ont été, dans leur majorité, étiquetés et lemmatisés.

3- On aimerait qu'un véritable atelier statistique soit ouvert à Nancy et que le serveur puisse offrir des traitements statistiques évolués, et pas seulement des fréquences et des pourcentages. Un tel atelier irait au bout du traitement, y compris jusqu'aux graphes comme ceux que délivre le Crisco de Caen ou le CNRTL de Toulouse. La situation actuelle, où les produits échangés sont demi-finis, engendre trop de contraintes et de corrections, dès que change le format des données.

4- On réclame enfin que dans le domaine des données chiffrées le copyright des éditeurs soit abandonné et que l'abonnement ne soit plus nécessaire, puisqu'à aucun moment le texte n'est communiqué.

5- Ce sont là des projets à court terme dans le développement des bases constituées à Nancy. Ils n'excluent pas des projets plus ambitieux que je laisse aux plus jeunes le soin d'évoquer. Je me suis laissé dire qu'un projet se préparait qui porterait sur l'étiquetage non plus seulement grammatical, mais sémantique des textes. Voilà une recherche d'avenir, à côté de laquelle ce que je viens de raconter n'est qu'un jeu d'enfant.