Que disent les tables ? Que disent les chiffres ?

Étienne Brunet

Que disent les tables ?
Que disent les chiffres ?

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

Les tableaux de la statistique sont-ils les tables de la loi ou des tables tournantes ? Disent-ils quelque chose ? Et si oui, doit-on le croire ? Quel crédit accorder au médium – au statisticien – qui parle en chiffres comme le médium de Hugo dans les réunions spiritistes de Guernesey parlait en vers ? Dans nos sociétés modernes les batailles de chiffres se sont substituées aux guerres de religion et aux conflits idéologiques. Mais on voit souvent les mêmes chiffres servir successivement dans les deux camps, comme les mercenaires des combats d'antan, au point que Platon, s'il écrivait le Gorgias à notre époque, orienterait vers les chiffres les cours de rhétorique et de sophistique et ferait de Calliclès un statisticien. Les chiffres paraissent avoir sur les mots une supériorité de droit divin, car ils donnent lieu à des effectifs dits absolus. Or cet absolutisme, qui en impose à la pensée, recouvre très souvent des opérations relatives et contingentes, qui n'ont rien de nécessaire. Comme la nature reproduit rarement le même objet, le temps le même événement, le discours les mêmes mots et les mots la même signification, les faits dont on fait un effectif ne sont pas rigoureusement identiques. Ils ne le deviennent que par abstraction, par neutralisation des conditions particulières de leur réalisation. Et ils ne se prêtent à la comparaison que si les conditions sont supposées ne pas varier et si l'unité de mesure reste constante. Ce qui se produit rarement. On est ainsi amené à biaiser, à relativiser l'absolu, en pondérant, et, par exemple, en exploitant les données économiques « corrigées des variations saisonnières », ou en corrigeant les chiffres des sondages pour tenir compte des sondés et des sondeurs.

– I –

1 – Ces difficultés, inhérentes à toutes les sciences humaines, obèrent les travaux de statistique linguistique. Faut-il une fois de plus les passer en revue ? Ce qui manque à cette discipline c'est d'abord une **unité de mesure**. Que d'incertitudes dans la délimitation du mot, même si l'on adopte la plus simpliste : la formule graphique qui définit le mot comme étant l'espace imprimé entre deux blancs ou deux séparateurs. Car beaucoup des signes de ponctuation ne sont pas des séparateurs univoques, et c'est le cas du point, de l'apostrophe et surtout du trait d'union dont le nom même récuse cette fonction de séparation. Les séparateurs de la typographie ne sont pas d'une plus grande sûreté lorsqu'on s'appuie sur eux pour définir des unités de discours plus larges, comme la phrase ou le paragraphe. Quel statut donner aux ponctuations affectives (interrogation, exclamation, suspension) dont la portée varie selon le contexte ? Et que faire des signes intermédiaires (point-virgule et deux points) ? Au reste ces jalons, déjà ambigus par eux-mêmes, ont par surcroît une autorité incertaine, car dans bien des cas ils appartiennent à l'éditeur, non à l'auteur. Et que dire de la lemmatisation, qui prétend normaliser dans certaines langues à flexion la multiplicité foisonnante des formes et qui n'a jamais pu se soumettre elle-même à la standardisation. Faute de disposer d'un mètre-étalon, la statistique linguistique parcourt les textes comme les arpenteurs de jadis qui mesuraient les champs en comptant leurs pas.

2 – Ce qui a longtemps manqué à cette discipline, ce sont les **termes de référence**. La statistique est comparative dans son essence même, comme le sont les sciences expérimentales. La mesure même implique une comparaison, à tout le moins avec le mètre-étalon. Or que peut-on comparer ? Osera-t-on comparer deux langues différentes, deux époques éloignées, deux genres opposés, deux auteurs étrangers l'un à l'autre ou deux textes qui diffèrent entre eux par un ou plusieurs de ces paramètres ? Fort heureusement cette situation périlleuse a été épargnée aux premiers représentants de la discipline, précisément parce que le domaine étant vierge aucune étude préalable ne leur offrait de points de repère pour s'orienter et comparer. Et les pionniers n'ont souvent pu que rendre compte de leurs observations, sans pouvoir généraliser ni conclure. Davantage d'éléments se trouvent disponibles aujourd'hui. Mais à l'embarras du choix s'ajoute une aporie paralysante : avant de risquer une comparaison avec les relevés d'un autre, on ne peut pas ne pas se demander non seulement s'ils sont honnêtes et fiables, mais s'ils sont compatibles avec ceux qu'on veut rapprocher d'eux et si les principes de

sélection et de traitement sont les mêmes. À partir de mesures établies de part et d'autre avec une bonne foi égale mais avec des options différentes, la statistique peut conduire à des conclusions imprudentes ou impudentes.

3 – L'**objectif** même de la statistique linguistique ou lexicale n'a pas été clairement défini et le fait qu'on hésite sur l'adjectif montre qu'on n'a pas toujours su préciser la portée et les limites de la discipline. Ce qui est lexical est en effet plus facile à comptabiliser mais dans les mots se cachent des réalités morphologiques, syntaxiques, rythmiques, thématiques. Bien loin d'être une unité élémentaire, le mot est une réalité complexe dont les composants infra-lexicaux (suffixes, préfixes, flexions verbales, phonèmes et morphèmes de toutes sortes) permettent l'approche de réalités plus larges qui appartiennent au supra-lexical. Le mot d'autre part n'est pas une unité indépendante. L'essentiel d'un discours est dans l'enchaînement des mots. En coupant la chaîne, la statistique ne détruit-elle pas son objet, comme cela arrive dans la physique corpusculaire où la lumière jetée sur les éléments perturbe leurs mouvements qu'on veut précisément étudier ?

4 – Ces apories n'ont pourtant pas empêché qu'on sollicite la statistique pour résoudre les **problèmes** les plus épineux, comme celui de la datation des textes ou de leur attribution à tel ou tel auteur. On a demandé ingénument à l'ordinateur de fournir les réponses scientifiques, puisqu'on supposait la machine dotée d'une mémoire infaillible, d'une attention imperturbable et d'une impartialité irréprochable. En somme on attendait des appareils ce qu'on obtient du carbone 14 pour la datation d'un gisement préhistorique, ou de l'empreinte digitale pour l'attribution d'un meurtre. Mais où se trouve l'empreinte digitale d'un auteur ? Où dénicher, dans le lacis des mots, la signature, invisible mais indélébile et irrécusable, de l'écrivain ? Plus généralement la problématique de la discipline oscille entre deux tendances : il y a la démarche classique – et les problèmes d'attribution et de datation sont de ce type – qui consiste à émettre d'abord une hypothèse et à la soumettre ensuite à l'épreuve des faits. Le danger est ici celui de l'échec ou plus exactement du refus de l'échec. Il y a le risque de l'acharnement statistique, quand les preuves ne donnent pas le résultat escompté et qu'on veut à toute force ne pas perdre le bénéfice de son effort[1]. La démarche inverse est modeste mais naïve

1. Le risque majeur est le silence – parfois inconscient – qui entoure les essais avortés et les résultats provisoirement (?) défavorables à l'hypothèse. On pense ici à la pratique assez répandue des analyses factorielles, qu'on recommence en écartant, comme « éléments supplémentaires », les lignes ou les colonnes qui influent trop ou mal sur le

qui consiste à n'avoir pas d'idée préconçue – donc pas d'hypothèse – et à observer les écarts de toutes sortes que délivre le calcul. Or dans le domaine du discours il y a surdétermination. Rien n'y est aléatoire. Partout des oppositions, des clivages et des écarts. Mais les causes pour un même effet peuvent être multiples : si j'observe un écart pour la forme AIMAIT, est-ce dû au temps, au mode, à la personne, au nombre, aux sonorités du mot ou à son contenu sémantique ? On n'a guère avancé quand on a trouvé un écart. Reste à démêler la cause, c'est-à-dire à chercher ce qu'on vient de trouver. Et la situation n'est pas plus favorable que celle de la démarche classique qui tend à trouver et à prouver ce qu'elle cherche.

5 – Ce qui manque à la statistique c'est aussi un consensus sur les **méthodes**. On s'est d'abord contenté de produire des relevés bruts et, lorsqu'une pondération s'avérait nécessaire, de les transformer en simples pourcentages. Mais cette démarche un peu courte n'a pas satisfait le statisticien, qui aspire à entrer dans l'univers probabiliste. Cette aspiration est-elle légitime ? Pour qu'elle le soit pleinement il faudrait que le choix des mots dans un texte ait quelque rapport avec le schéma d'urne. Or chacun l'admet, même les promoteurs de la méthode probabiliste, un écrivain ne puise pas ses mots dans une urne et – sauf par expérimentation surréaliste – il ne les épingle pas au hasard dans un dictionnaire. Il y a donc généralement de grands écarts entre les faits observés et le modèle probabiliste, au point que l'exception devient la règle. Ce qui a caché quelque temps cette distorsion, c'est l'échelle modeste des premiers résultats. Quand les effectifs sont faibles, l'hypothèse nulle est plus difficile à rejeter et les écarts dits significatifs sont moins nombreux à franchir le seuil, ce qui donne l'illusion d'un schéma d'urne. Mais dès que se fait sentir la loi des grands nombres, l'univers qu'on décrit est manifestement orienté, les mots se font des signes de connivence qui ne doivent rien au hasard. Cette constatation, qui rassure le chercheur littéraire sur l'intelligibilité irréductible des textes, inquiète le statisticien et l'invite à se mettre en quête d'un autre modèle. Hélas, rien n'a été trouvé jusqu'ici qui puisse constituer une alternative au schéma classique. On a vu apparaître un foisonnement d'indices de toutes sortes, de quotients divers, de formules ingénieuses, par quoi on a voulu mesurer la richesse lexicale, ou le dosage des catégories grammaticales, ou la coloration stylistique d'un texte. Mais

---

résultat. Quelle foi accorder à un sondage ou à une élection si on récuse *a posteriori* les réponses ou les votes les plus prévisibles ou les plus atypiques ?

toutes ces tentatives manquent de généralité et de cohérence et font regretter la perfection théorique que le schéma d'urne contenait dans ses flancs.

6 – Le procès de la statistique n'est pas nouveau. Bien des pièces de ce procès se trouvent dans les archives même de l'ALLC, et ce dossier a été ouvert notamment aux colloques de Pise et de Louvain et dans ce pays, à Waterloo et à Victoria. Mais ce n'est pas de ce procès que la statistique a le plus à craindre (car il y a quelque complicité parmi les avocats, même quand ils se combattent). Plus redoutable est l'**accueil** qui a été fait dans la communauté littéraire et linguiste aux travaux de la discipline. En dehors du cercle restreint des spécialistes et de leur approbation jalouse ou intéressée, il faut reconnaître que les observateurs littéraires voient les débats statistiques parfois d'un œil amusé, plus souvent d'un œil indifférent et dans certains cas d'un œil courroucé et indigné. Sur cet accueil du public je renvoie à l'analyse plaisante quoique peu complaisante que Charles Muller a faite à Nice à l'ouverture du Colloque ALLC de 1985[2].

7 – En réalité cet accueil réservé prend souvent l'aspect d'une dérobade. Car les **résultats** statistiques sont plus souvent ignorés que condamnés. *Graecum est non legitur*, disent les esprits littéraires devant une formule mathématique, même enfantine. À plus forte raison renâclent-ils devant ces listes dont on ne voit jamais en même temps la tête et la queue, devant ces tableaux de nombres accueillants comme des buissons d'ajoncs et ces monceaux d'index et de concordances qu'on voit à l'abandon sur les quais de la recherche. On doit avouer que ces filets pleins de « résultats » n'ont souvent rien d'engageant. Et il arrive que leur auteur ne soit guère engagé, amassant des matériaux qu'il livre à l'état brut ou à peine transformés. Il est vrai que dans le passé l'effort était fort long pour obtenir un résultat et ce qui eût dû être l'étape initiale devenait le terme de l'épreuve. Mais les résultats qui ne sont pas interprétés par celui qui les a obtenus, les listes qui ne sont pas analysées, les concordances qui ne sont pas exploitées, tout cela appartient-il à la statistique linguistique ? N'est-ce pas plutôt seulement une prestation de service documentaire ?

---

2. La conclusion de Charles Muller, le pionnier de la discipline, n'invitait pas à l'optimisme : « Confiance excessive d'un côté, méfiance abusive de l'autre... La cause est bien compromise ! », *Méthodes quantitatives et informatiques dans l'étude des textes*, Slatkine-Champion, 1986, p. 11 (1986c).

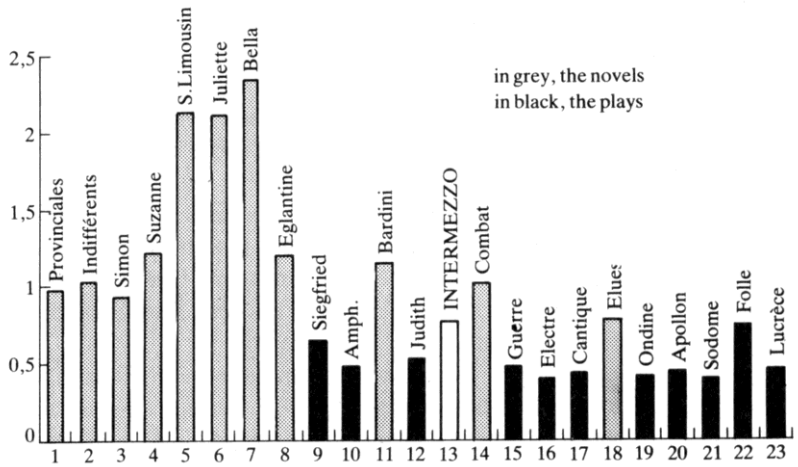
– II –

Mais à l'heure actuelle la collecte des données n'a plus ce caractère épuisant, au moins dans le domaine français. Parce que Chicago n'est pas loin qui abrite un centre de distribution de la base de données *Frantext*, on doit savoir ici ce que représente ce riche trésor, qui est gros de 160 millions de mots et dont la cohérence est aussi précieuse que son abondance. Au reste le créateur de cette base, Jacques Dendien, est dans nos rangs et il est mieux qualifié que moi, pour expliquer les fonctions de son système, sinon pour en vanter les mérites. Bornons-nous à dire que de tout point de l'hexagone national, et même de tout point du globe, y compris de Toronto, il est possible, au bout d'une ligne télématique jetée par-dessus les océans, de pêcher n'importe quel mot, ou expression ou association, au sein des 3000 textes complets qui sont disponibles dans la base, et que cette pêche miraculeuse s'accomplit en quelque secondes à toute heure du jour et de la nuit. Maintenant que le chercheur peut s'approvisionner lui-même, on peut espérer moins de gaspillage, plus de cohérence et d'engagement et une meilleure qualité des résultats.

Cette qualité tient d'abord au calibrage, c'est-à-dire à la possibilité de constituer des faits qui partagent certaines propriétés communes, afin de comparer ce qui est comparable. La condition première qui commande tout le reste est l'homogénéité des données et la constance des traitements qu'elles ont reçus. La base *Frantext* donne cette garantie, car tous les textes qu'elle contient ont été saisis et traités selon des normes qui n'ont pas varié d'un iota depuis vingt ans. Mais la cohérence du corpus qu'on sélectionne dans cet ensemble relève de l'initiative du chercheur. Elle peut se justifier par des critères d'auteur, de genre ou d'époque.

1 – La sélection la plus naturelle et la plus fréquente est celle qui se fonde sur l'auteur et elle vise à produire des **monographies**, comme celles que nous avons réalisées sur Giraudoux, Proust, Zola et Hugo. En de tels cas on a tendance à s'enfermer dans l'univers propre à l'écrivain choisi et à considérer l'ensemble de ce corpus comme étant la « norme » pour les textes qui le constituent. Ces textes s'ordonnent alors selon l'évolution de l'auteur, à moins que la chronologie ne soit perturbée par l'influence prépondérante du genre littéraire. Divers points de vue peuvent être alors explorés qui envisagent tour à tour les faits de structure lexicale, ceux de la segmentation, de la syntaxe, ou de la sémantique. Et parallèlement les outils utilisés peuvent varier des plus classiques (courbes établies sur l'écart réduit) aux plus complexes (analyses factorielles).

Afin de donner une idée des résultats qu'on obtient ainsi, nous passerons rapidement en revue quelques-uns de nos ouvrages[3] en variant les auteurs, les méthodes et les objets de recherche.

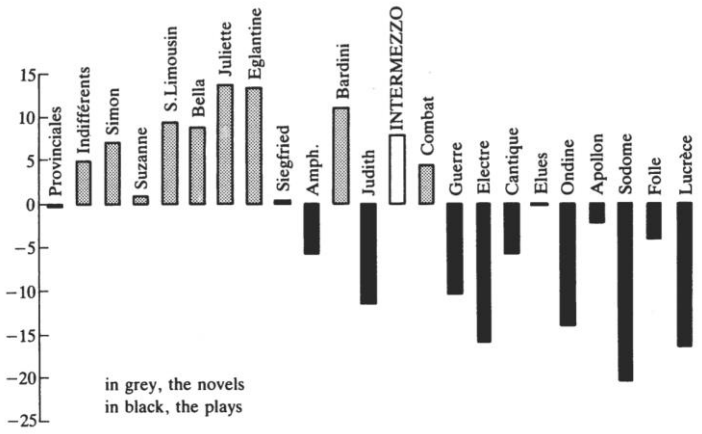


Graphique 1. La structure lexicale chez Giraudoux

Le premier graphique 1 aborde une question qu'on a sans doute trop souvent traitée et qui a trait à la structure lexicale. Depuis la loi de Zipf que n'a-t-on pas écrit sur le rapport mathématique complexe qui lie le nombre d'occurrences à celui des vocables et gouverne la distribution des classes de fréquence. Que de débats et de propositions sur ce rapport qui est apparu longtemps comme l'équivalent moderne de la pierre philosophale et du nombre d'or. L'intérêt littéraire ne va guère au delà de l'appréciation de la richesse (ou de la variété) lexicale. Dans le cas de Giraudoux, ce rapport est favorable au roman (en gris sur le graphique) et défavorable au théâtre, et dans le théâtre les comédies modernes sont mieux traitées que les sujets antiques et tragiques.

Le graphique 2 offre exactement la même stratification des genres littéraires, quoiqu'il analyse un objet tout différent : la longueur moyenne du mot. Le mot est plus volumineux dans le roman, et plus court dans les pièces antiques que dans les pièces modernes, comme si les deux faits étaient liés. La sécheresse de la tragédie refuse tout à la fois l'inflation numérique du vocabulaire et l'enflure adipeuse des mots.

3. On nous pardonnera le choix de ces exemples en considérant que ces ouvrages (aux éditions Slatkine-Champion, Paris-Genève), que l'auteur a nécessairement sous la main, sont par contre assez peu disponibles pour le lecteur, vu leur prix exorbitant.

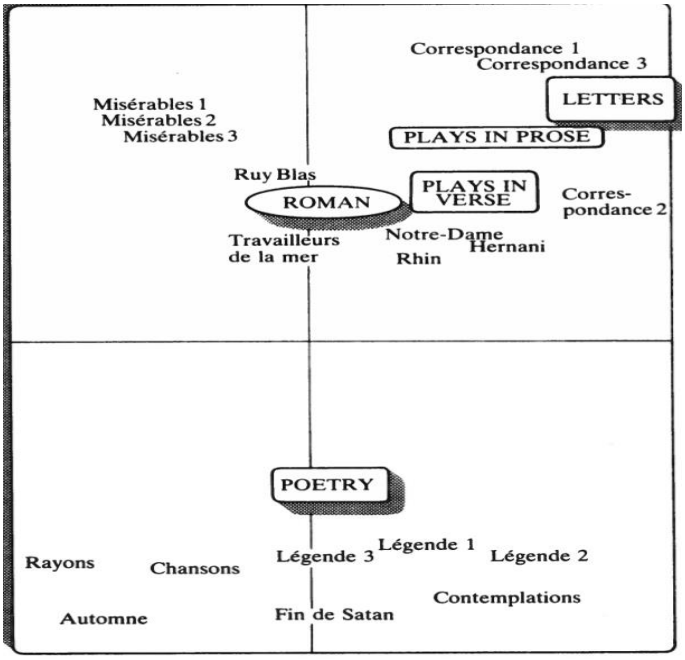


Graphique 2. La longueur moyenne du mot chez Giraudoux

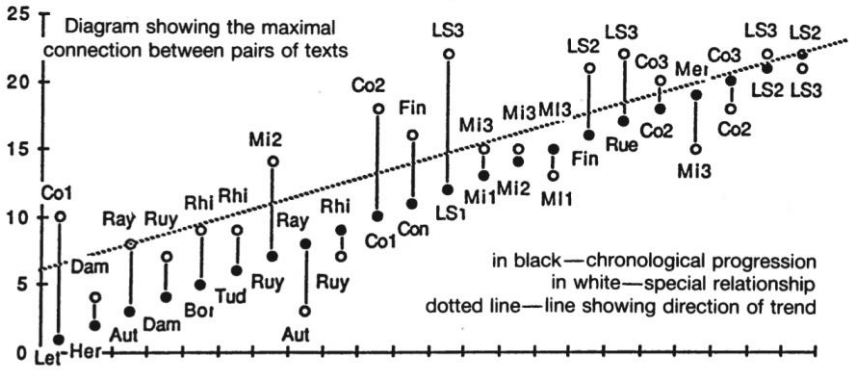
Le corpus de Hugo est animé, comme celui de Giraudoux, par l'opposition des genres. C'est ce que montre l'analyse factorielle de la figure 3, qui rend compte de la connexion lexicale, c'est-à-dire de la distance entre le vocabulaire de deux textes. On n'effraiera pas le lecteur avec la description des calculs infinis qui sont nécessaires pour établir cette distance par une mesure globale, sachant qu'il faut considérer tous les mots des deux textes considérés et la fréquence de chacun, réelle et théorique, dans chaque texte, et que ce calcul complexe doit être renouvelé pour chaque paire, soit 231 fois pour 22 textes. Le résultat de l'analyse est pourtant d'une parfaite limpidité : en bas se groupent tous les recueils poétiques tandis que le roman, la correspondance et le théâtre prennent place dans la moitié supérieure, sans trop se mélanger les uns aux autres (mis à part le théâtre en vers que sa division interne désoriente).

Cette loi des genres ne laisse pas de surprendre chez l'auteur de la *Préface de Cromwell*, qui affirmait vouloir abattre ces barrières trop contraignantes. Et n'y a-t-il donc aucune évolution chez cet écrivain tout à la fois précoce et tardif, dont la production s'étend sur soixante ans ?

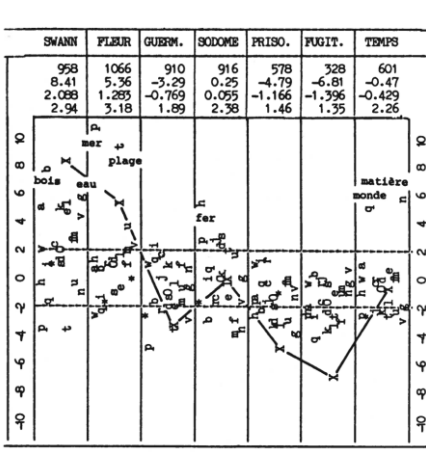
La même mesure de la connexion lexicale laisse apparaître en filigrane la chronologie dans le graphique 4, qui pour chaque texte indique son associé le plus proche. Certes tous les couples respectent la loi du milieu social (et appartiennent au même genre). Mais cela n'empêche pas la chronologie d'assortir les couples, ce qui produit la diagonale du graphique.



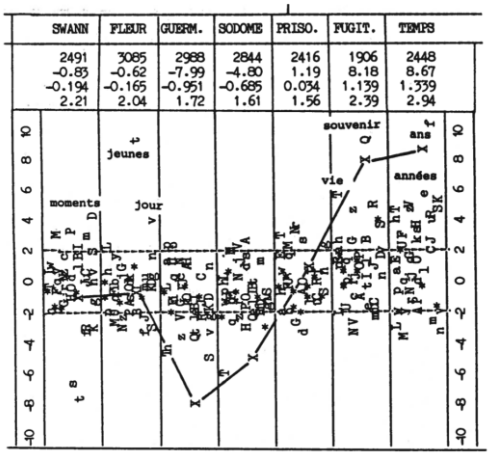
Graphique 3. Analyse factorielle de la connexion lexicale chez Hugo



Graphique 4. L'influence du temps dans le vocabulaire de Hugo

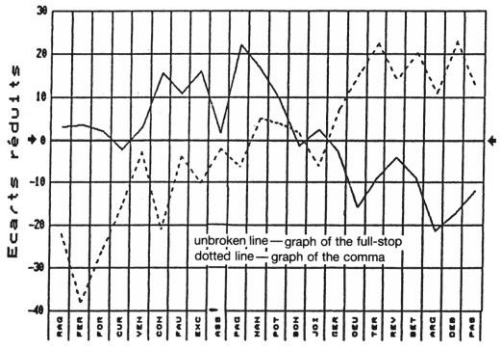


Graphique 5. La nature chez Proust



Graphique 6. Le temps chez Proust

La chronologie règne en principe sans partage dans les corpus où les oppositions de genre n'interviennent pas, comme c'est le cas de *La recherche du temps perdu* et des *Rougon-Macquart*. On en donnera deux illustrations, précisément empruntées à Proust et Zola. La première envisage le **contenu** lexical dans les sept livres de *La recherche*, et plus précisément deux champs sémantiques : le thème de la nature et le thème, proustien entre tous, du temps. On a mis en parallèle les courbes 5 et 6 parce qu'elles sont en opposition : alors que la nature s'estompe progressivement au cours de la rédaction, le temps se fait de plus en plus obsédant, avec une tonalité plus sombre qui se penche plutôt sur les ANNEES que sur les JOURS, sur la VIE que sur les MOMENTS, sur les SOUVENIRS que sur la JEUNESSE.



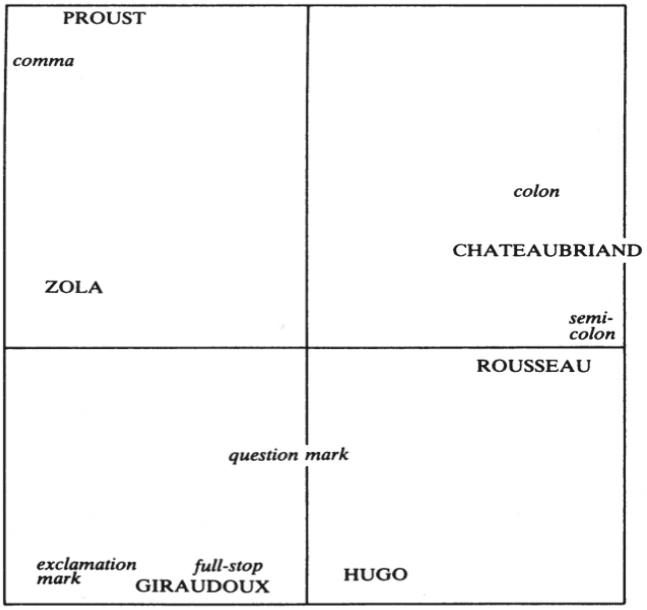
Graphique 7. Évolution de la ponctuation de Zola dans les *Rougon-Macquart* (en traits pleins, courbe du point ; en pointillés, courbe de la virgule)

La seconde illustration est relative à la ponctuation des *Rougon-Macquart* qui est fiable, car fidèle aux volontés de Zola. Il y a là 132 114 points et 340 479 virgules, dont la répartition dans les vingt années du cycle suit deux mouvements contrariés. La courbe du point montre des excédents dans la première moitié et des déficits dans la seconde, alors que la virgule poursuit une progression continue. La phrase de Zola que structure la ponctuation change donc au cours de la rédaction, et gagne en ampleur[4].

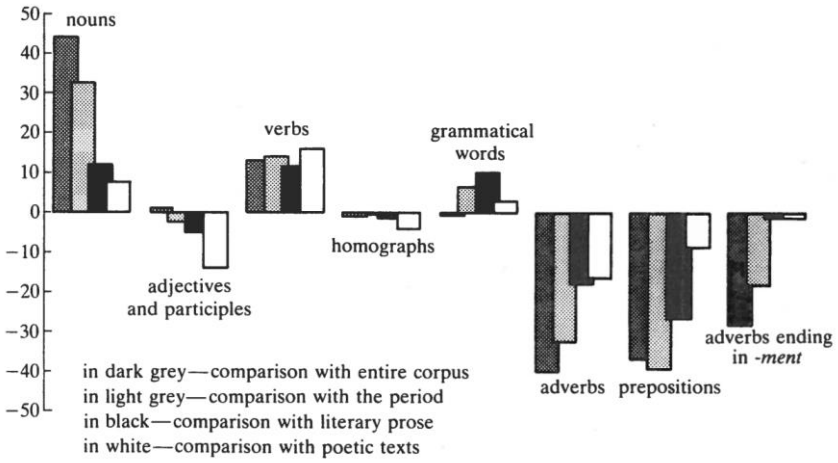
2 – Les monographies peuvent être reliées entre elles, si elles sont issues du même gisement de données. Ainsi le graphique 8 réunit six auteurs dont les données ont subi un traitement homologue. On voit qu'en deux siècles de littérature le système des ponctuations a nettement évolué, puisque Rousseau et Chateaubriand sont les seuls à cultiver les signes intermédiaires (deux points et point-virgule), tandis que Hugo et Giraudoux multiplient les points (et aussi les signes affectifs) et Proust la virgule. Ce même Proust, comme on s'y attendait, bat le record de la longueur des phrases (31 mots en moyenne), devant Rousseau (27) et Chateaubriand (22). Hugo et Zola préfèrent les phrases courtes (14 mots en moyenne).

Mais tant que les monographies d'auteurs restent en nombre limité, la comparaison de chacune peut se faire avec l'ensemble du corpus XIX^{e}-XX^{e} siècles. On dispose ainsi d'une toile de fond commune, sur laquelle se détache la spécificité de chaque écrivain. Au besoin on peut extraire de ce grand corpus un sous-corpus mieux adapté au rôle de modèle qu'on veut lui prêter, si on sélectionne les dates et les genres appropriés.

4. NDÉ : voir tome III, chapitre 9, « La phrase de Zola » (1985d).



Graphique 8. Analyse factorielle des ponctuations de six écrivains



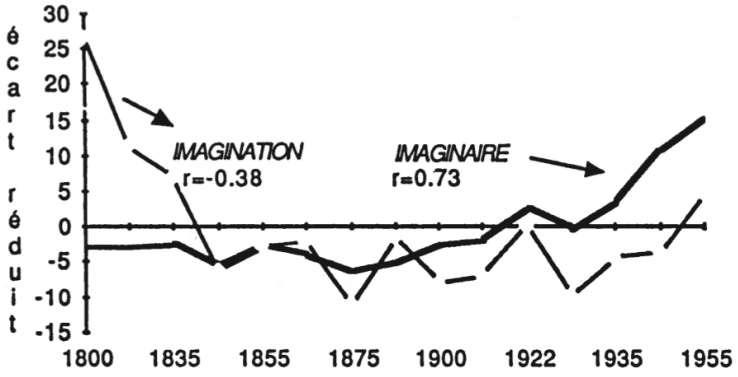
Graphique 9. Les catégories grammaticales chez Hugo. Comparaison externe

L'exemple des catégories grammaticales (graphique 9) chez Hugo est rassurant à ce propos. Que la confrontation s'opère avec le corpus entier ou celui du temps de Hugo, qu'elle mette en jeu seulement la

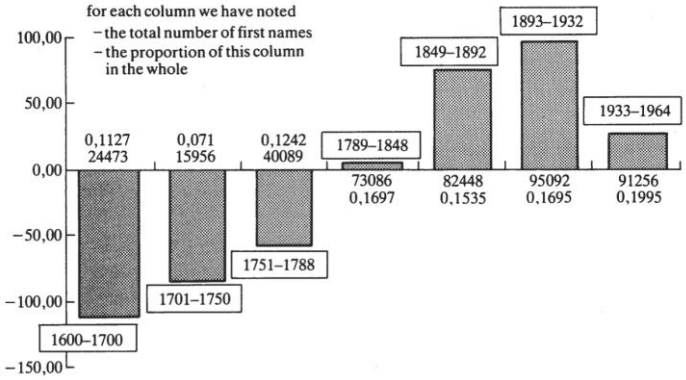
poésie ou seulement la prose littéraire de l'époque, les conclusions ne varient guère : Hugo choisit parmi les parties du discours les plus pleines, c'est-à-dire les substantifs et dans une moindre mesure les verbes et il prise peu les prépositions et les adverbes. Les choix de Proust ou de Zola ne sont pas les mêmes.

Mais c'est surtout en matière de thématique que cette comparaison extérieure s'avère utile. Le calcul de l'écart réduit (ou du Chi2) pour tous les mots d'un écrivain, suivi d'un tri, fait apparaître deux lots de mots également intéressants : ceux que cet écrivain chérit plus que d'autres (plus que les autres mots et plus que ne le font les autres écrivains), et ceux qu'il évite. Ainsi les listes significatives de Proust et de Zola sont l'inverse l'une de l'autre. Fidèle à sa théorie naturaliste, Zola montre le milieu, le corps, les choses, alors que Proust montre les sentiments et les réalités psychologiques. Chez Hugo les tests statistiques mettent en lumière... la lumière et plus encore l'ombre. C'est le mot OMBRE qui apparaît en tête de liste parmi les substantifs, comme le mot SOMBRE parmi les adjectifs, tandis que la rime SOMBRE-OMBRE atteint la fréquence la plus élevée. Le sentiment du lecteur est là-dessus en accord avec la statistique, mais il ne l'est pas toujours, les listes réservant des surprises. Ainsi le vocabulaire de la sensation et du sentiment est déficitaire chez Chateaubriand, qui est l'auteur de *René* et d'*Atala*, et qui pourtant éprouve une retenue classique devant les mots PASSION, AFFECTION, IMPRESSION, ATTENTION, SENSATION, EMOTION, PLAISIR, EXPRESSION.

3 – Comme il y a des monographies d'auteurs, il peut exister aussi des monographies de mots, ou plus largement de familles de mots et de thèmes. Au lieu de s'intéresser à tous les mots d'un auteur, on envisage cette fois un seul mot (ou un groupe restreint de mots) à travers tous les auteurs de la base. Ce point de vue est plus souvent partagé par les linguistes et les historiens, quand les monographies d'auteurs correspondent mieux à la démarche des littéraires. À titre d'exemple on présente ci-dessous la courbe du mot IMAGINATION, qui est descendante depuis 1789. Est-ce à dire que l'imagination se tarit de nos jours ? Ce serait imprudent de l'affirmer dans le siècle de l'image. De fait l'IMAGINAIRE vient suppléer aux défaillances de l'IMAGINATION (les deux mouvements sont contraires dans la figure 10). Et le verbe IMAGINER vient en renfort, comme le mot IMAGE lui-même et dans ces deux cas la progression est soulignée par un coefficient de corrélation fortement positif (respectivement r=+0,93 et r=+0,70).



Graphique 10. Les mots IMAGINATION et IMAGINAIRE.



Graphique 11. Les prénoms de 1600 à nos jours

Graphique 12. Le couple ADAM et ÈVE

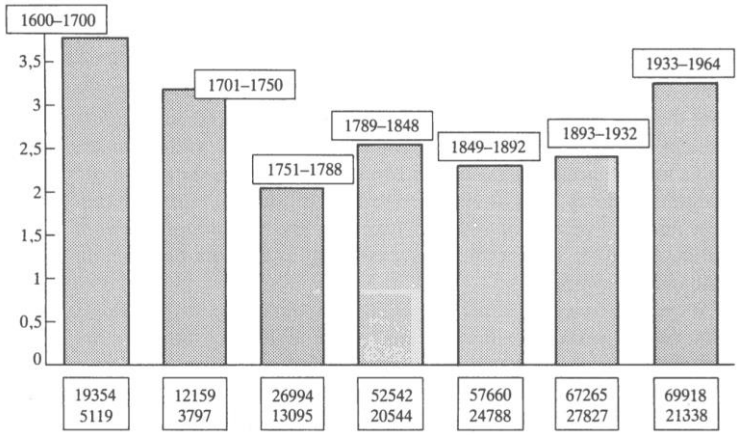
En réalité l'intérêt de telles études transversales est plus vif quand on parcourt le corpus à la recherche non pas d'un mot mais d'une série de mots. Le cas du mot IMAGINATION vient de nous prouver que les conclusions

sont plus nuancées quand on consulte les autres membres de la famille lexicale. On pourrait élargir le cercle et admettre dans la série les mots ESPRIT, AME, CŒUR, et quelques autres. On verrait alors apparaître des phénomènes de substitution, soit que des termes à la mode comme CONSCIENCE ou PSYCHOLOGIE captent l'héritage des mots qui ont cessé de plaire (comme le mot AME), soit que des déplacements s'opèrent parmi les catégories grammaticales, l'adjectif prenant la place du substantif pour les facultés mentales, sans doute parce qu'on n'ose plus se les représenter comme des substances.

Plus largement on peut tailler dans *Frantext* (grâce à la commande LISTEMOTS) des séries aussi riches qu'on le souhaite et délimiter de larges champs sémantiques, comme le bestiaire dans la littérature, ou le système de la parenté, ou la représentation du corps, de l'espace ou du temps. Nous choisirons un exemple curieux qui fait rarement l'objet de recherches statistiques et qui met en jeu les noms propres, ou plus précisément les prénoms du calendrier romain, soit plus de 300 unités. Et comme le XVII^{e} et le XVIII^{e} siècles sont désormais disponibles, sans poser dans ce cas-ci des problèmes d'orthographe ou de lemmatisation, on a représenté dans le graphique 11 quatre siècles de littérature. On ne se cache pas que le matériau brut est composite et que ces prénoms sont souvent des noms ou des lieux et qu'ils désignent des êtres fictifs ou réels, anciens ou modernes. Il n'en reste pas moins que leur effectif (près d'un demi-million d'occurrences) se distribue de façon fort claire et montre une progression continue de 1600 à 1900. Sans doute l'époque moderne, depuis la Révolution et le romantisme, donne-t-elle davantage de prix à l'individu et à la spécificité des êtres et des lieux dont le nom propre est le symbole.

On a considéré le lot des prénoms en bloc. Mais cette approche globale peut être affinée. On peut considérer chaque ligne du tableau (chaque prénom de la liste) ou chaque colonne (l'une des sept périodes du corpus) et établir des profils particuliers pour les uns et les autres ou étudier l'ensemble du tableau par une analyse factorielle. Un seul exemple suffira qui est le premier chronologiquement puisqu'il concerne le couple ADAM et ÈVE (mais ce ne sont pas les premiers dans l'ordre des fréquences, la palme revenant plutôt aux martyrs, soit respectivement JEAN, LOUIS, JACQUES, CHARLES, MARIE, PIERRE, HENRI, ANTOINE, PAUL, PHILIPPE). Le couple est réhabilité à partir du XIX^{e} siècle, avec un avantage plus marqué pour ÈVE. En réalité ÈVE profite d'un mouvement d'ensemble qui au cours du XIX^{e} siècle atténue le déséquilibre des sexes. Le rapport entre prénoms masculins et prénoms féminins, qui est de 4 au XVII^{e}

siècle, est corrigé au cours du XVIII[e] et du XIX[e], pour atteindre une valeur voisine de 2. Mais cette promotion de la femme semble précaire puisque le rapport remonte à 3 dans la tranche la plus récente, et au total MARIE est le seul prénom féminin à figurer dans la distribution des dix premiers rôles.



Graphique 13. Rapport prénoms masculins / prénoms féminins

4 – Peut-on aller plus loin encore et étudier le corpus *Frantext* dans son ensemble, en s'intéressant à tous les mots, à tous les textes, tous les auteurs, tous les genres, toutes les époques ? Cette ambition serait trop grande pour être réellement compatible avec les contraintes de la télématique. Elle se heurterait en outre à l'absence de lemmatisation dans les données disponibles. Cette étude globale ne peut guère être envisagée que par un traitement spécial, de type *batch*, et hors ligne. Au reste nous avons réalisé cette entreprise en 1981, alors que les données ne recouvraient que les deux derniers siècles. Mais elle pourrait être reprise avantageusement si les données de *Frantext* se trouvaient disponibles sur un support optique qui permettrait le transfert des textes et s'ouvrirait aux traitements locaux. Un grand projet de CD-ROM qui va dans ce sens est présentement en cours d'élaboration, sous la responsabilité de Jacques Dendien.

– III –

On peut estimer en effet que les temps ont changé en matière de recherche documentaire et statistique. Combien d'index et de concordances n'a-t-on pas réalisés dans le passé, qui n'ont guère servi qu'à leur auteur. Et encore certains fabricants ont livré leur œuvre vierge, sans la moindre exploitation. Mais ces produits étaient longs et coûteux à

réaliser, leur diffusion était lente et rare, et leur utilisation malaisée. Quand l'information se trouve sur un support non accessible à l'ordinateur, comme le papier ou la microfiche, elle est difficile d'emploi, chaque fois qu'on a à trancher, à ajouter, bref à sélectionner – et ces manipulations sont de pratique courante s'il s'agit d'exploitation statistique. De plus les options prises au moment de la réalisation sont contraignantes et notamment les tris, les rebuts, les regroupements, les modalités de lemmatisation, la longueur du contexte... Au moins la puissance, la rapidité, l'étendue et la souplesse de la base de données *Frantext* représentaient sous ce rapport des progrès décisifs.

Ce à quoi doit tendre l'informatique documentaire et statistique, c'est à plus de souplesse encore, par la rupture des chaînes télématiques. Ce que demande actuellement la communauté scientifique, c'est un **accès facile, immédiat, et libre**, non pas aux sous-produits figés que sont les index, les concordances et les dictionnaires de fréquences, mais au texte même. L'utilisateur veut bien être aidé par un logiciel d'interrogation, mais il veut choisir ses questions, ses textes, et la présentation des résultats. Il souhaite aussi avoir l'illusion de partir seul et le premier à la découverte, sans avoir à suivre les pas d'un devancier et à citer les pages d'une publication préalable.

Quelle réponse donner à cette attente ? Sans doute le **CD-ROM**. C'est dans cette direction que nous avons travaillé depuis un an, en imaginant ce que pourrait être un tel CD-ROM. À l'occasion du Bicentenaire, le Centre Georges Pompidou avait demandé à l'Institut national de la langue française de lui fournir une base de données sur la Révolution de 1789, que le public puisse interroger. Un prototype a donc été expérimenté qui assure certaines des fonctionnalités essentielles de *Frantext* et qu'on a délibérément orienté non pas seulement vers la recherche documentaire mais aussi vers l'exploitation statistique. Et ce sont les deux branches qui s'ouvrent à l'utilisateur dès la présentation du menu principal, représenté dans la figure 14. À droite on s'engage dans la recherche documentaire; à gauche on s'oriente vers les traitements statistiques.

Figure 14. HYPERBASE. Menu principal

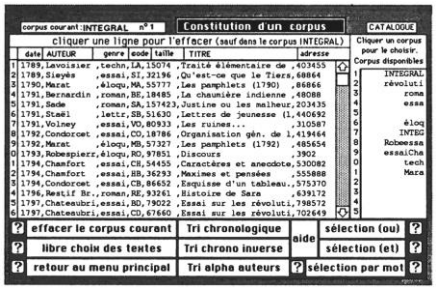


Figure 15. Choix du corpus

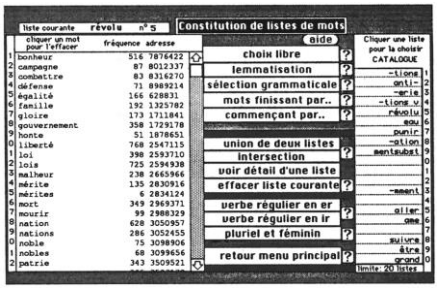


Figure 16. Choix des mots

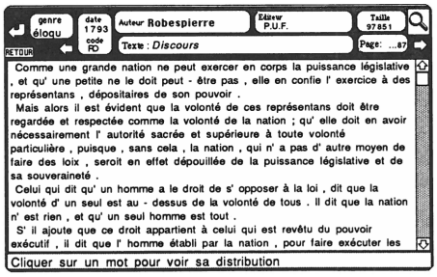


Figure 17. Une fiche-page

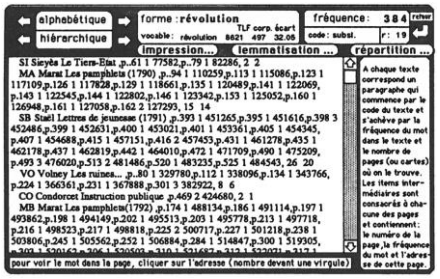


Figure 18. Une fiche-mot

Mais avant de choisir une de ces deux directions (d'ailleurs non exclusives), le chercheur est invité à fixer ses choix. Il délimite lui-même le sous-ensemble de textes qu'il veut consulter (par un programme CHOIX DU CORPUS, qui multiplie les critères de sélection : genres, auteur, date, titre, contenu et leur applique les opérateurs booléens. Voir

figure 15). Il établit aussi librement la liste des mots qui l'intéressent (programme CHOIX DES MOTS, qui propose des listes automatiques fondées sur le suffixe, le préfixe ou la lemmatisation, et qui autorise les ajouts, les suppressions, les croisements. Voir figure 16).

L'utilisateur peut vouloir d'abord procéder à des opérations de contrôle et examiner les textes, page après page (programme VOIR UN TEXTE, graphique 17), ou consulter le fichier des mots, par ordre alphabétique ou par fréquences décroissantes (programme VOIR DES MOTS, graphique 18) ou passer instantanément, par un simple « clic », des textes aux mots et des mots aux textes, selon les méthodes de l'hypertexte. S'il désire vérifier le contenu du corpus courant et de la liste courante – qui sont variables et se modifient à volonté – un aperçu rapide fait défiler les fiches bibliographiques des textes du corpus ou les fiches signalétiques des mots de la liste en cours.

Si le chercheur veut s'engager dans la recherche **documentaire**, il choisit dans les boutons de droite les actions à mener, qui peuvent conduire à un index (programme INDEX), à une concordance (programme CONCORDANCE, figure 19), à une liste de contextes-phrases (voir le programme CONTEXTE, graphique 20), ou à une recherche de cooccurrence (programme COOCCURRENCE). Liberté lui est donnée pour chacun de ces traitements de fournir soit un mot, soit un vocable (dont les différentes formes seront automatiquement fournies), soit une chaîne de mots (une expression), soit une liste de mots.

Figure 19. Concordance

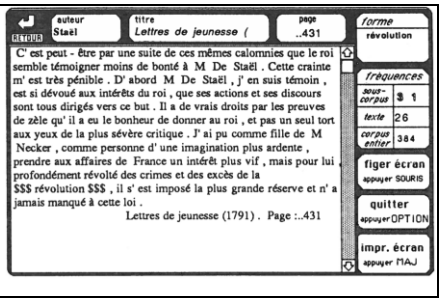


Figure 20. Contexte

Le logiciel HYPERBASE mène aussi du côté des méthodes **quantitatives**. Il établit des tableaux de contingence (programme FREQUENCES, graphique 21), et il dresse des courbes qui comparent la distribution de deux mots dans les différents textes, ou le profil de deux textes à travers une série de mots (programme GRAPHIQUE, figure 22). On dispose d'une interface pour l'exploitation de ces tableaux par les

méthodes d'analyse multi-dimensionnelle (programme ANALYSE FACTORIELLE). Enfin divers modules sont proposés qui font apparaître la distribution des classes de fréquences, dans chaque texte et dans l'ensemble (boutons DISTRIBUTION et LOI DE ZIPF) ou qui détaillent le vocabulaire spécifique d'un texte ou les propriétés spécifiques d'un mot (bouton SPECIFICITES. Voir graphique 23).

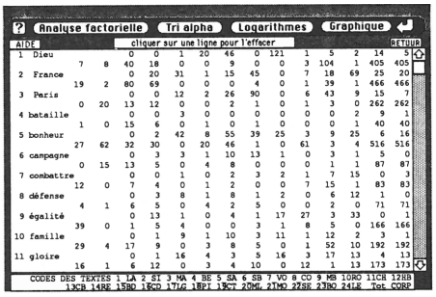


Figure 21. Fréquences

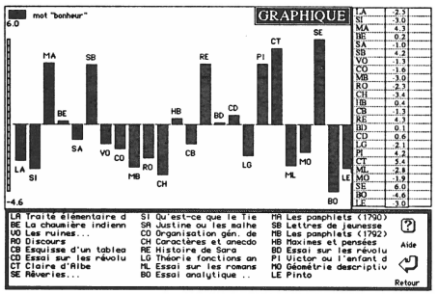


Figure 22. Graphique

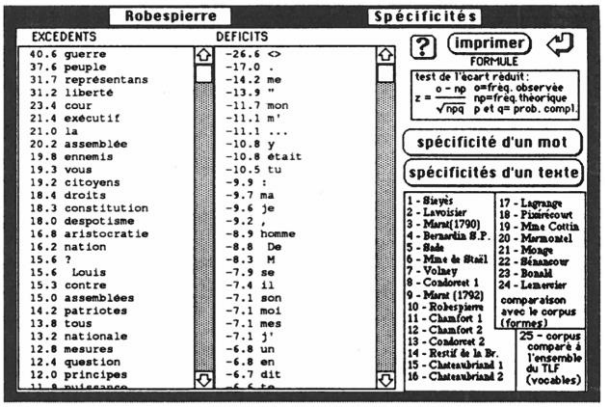


Figure 23. Le vocabulaire spécifique

Enfin les **résultats** apparaissent sous deux formes qui ne sont pas toujours identiques : à l'**écran**, au moment même où ils sont obtenus, et dans un **fichier** où ils sont enregistrés, avant d'être contrôlés et imprimés. Les résultats sont soumis à des opérations de tri, et notamment les concordances, qu'on offre sous diverses présentations : ordre chronologique, ordre chronologique inverse, ordre alphabétique des auteurs, tri du contexte gauche ou tri du contexte droit. Un éditeur (avec module d'impression) est fourni qui permet les retouches et la mise en page. Notons que les résultats eux-mêmes ne constituent pas un terminus, et que les méthodes de l'hypertexte leur sont applicables : en particulier il

suffit de désigner une ligne de la concordance obtenue pour faire apparaître aussitôt la page correspondante du texte en question.

Précisons que le **moteur de recherche** mis en œuvre tient compte des contraintes liées au **CD-ROM**, et en particulier de la relative lenteur des accès-disque de ce support. Pour toute exploration, il suffit de deux déplacements de la tête de lecture. Les accès eux-mêmes ont été optimisés (à partir de la fréquence présumée des appels).

La rapidité nécessaire n'a pas conduit aux sacrifices habituellement consentis par les produits documentaires, qui s'intéressent rarement aux mots-outils. Ces mots ont au contraire beaucoup de prix dans une exploitation de type linguistique, qui vise à l'**exhaustivité**. Le fichier inverse contient tous les mots, et incorpore aussi les marqueurs du récit (en particulier les signes de ponctuation).

Tout le logiciel a été écrit en gardant le souci de la **convivialité**. Même s'il s'adresse à un public averti, on n'a pas cru devoir lui donner un visage sévère. Au contraire on a tenté d'exploiter les ressources **graphiques** de l'écran (on montre par exemple le portrait des écrivains ou le fac-similé de l'édition originale), et l'on a utilisé pleinement la facilité opératoire des menus déroulants ou des boutons de commandes et la spontanéité de l'utilisation que permet la « souris ». Le son même n'a pas été banni. Et une **aide** a été prévue pour chacune des fonctionnalités, au moment même où celle-ci est proposée : il suffit de solliciter le point d'interrogation qui accompagne chaque bouton.

Ce souci de convivialité et de souplesse explique le choix d'un **langage objet** pour l'ensemble de ce logiciel écrit en *Hypertalk* et complété par des commandes et fonctions externes. Comme il s'agit d'un langage interprété[5] le logiciel est ouvert et admet toutes sortes de compléments.

Dans son état actuel, le prototype exploite une base de 24 textes complets de l'époque révolutionnaire, de 1789 à 1800. Robespierre y voisine avec Sade, Siéyès avec Marat, Mme de Staël avec Chateaubriand, et Chamfort avec Condorcet. Si ce corpus fait la part belle aux essayistes (parmi lesquels Volney, Marmontel, Bonald), il n'exclut pas les romanciers (par exemple Restif de la Bretonne, Madame Cottin ou

5. En réalité de larges parts du traitement vont être compilées, afin de gagner en rapidité et en sûreté. On ne désespère pas de compiler l'ensemble quand le langage *Hypertalk*, actuellement trop jeune, sera stabilisé et qu'un compilateur efficace sera disponible pour ce langage.

Bernardin de Saint Pierre), ni les auteurs dramatiques (Pixérécourt, Lemercier), ni les savants (Lavoisier, Monge, Lagrange). Au total c'est un ensemble de 1 300 000 mots qui se trouve disponible.

La version exposée – qui offre la couleur sur MAC II – fonctionne à partir d'un disque dur (de 40 Mo). La base a été installée également sur un disque compact WORM, afin d'expérimenter la technologie du laser. Mais la version CD-ROM n'est pas prévue dans l'immédiat, parce que le langage *Hypertalk* ne semble pas stabilisé et que de grandes améliorations sont attendues dans les prochains mois, du côté de la rapidité et de la confidentialité[6]. Le logiciel HYPERBASE ayant fait l'objet d'un contrat de développement (société APPLE), on peut espérer profiter parmi les premiers des améliorations prochaines du langage.

Bien entendu on a pensé à traiter d'autres données, par exemple les textes du Moyen Âge, ou la poésie au XIX^{e}. On a aussi projeté des versions dépourvues de données mais munies de programmes de préparation. Enfin pour les textes soumis aux droits de copyright, une version du logiciel a été étudiée qui escamoterait le texte intégral pour ne délivrer que des contextes. En somme les versions envisagées s'orientent tour à tour sur les deux versants où se sont engagés jusqu'ici les producteurs de bases de données : le premier est plus touffu, plus passionnant peut-être, parce que la découverte y est plus sensible, c'est celui du texte intégral. Le second est balisé, canalisé. Les voies y ont été tracées à l'avance et l'on y circule plus vite. C'est le versant des bases structurées.

Quant à la statistique lexicale, qui est indifférente au copyright, elle va d'un versant à l'autre du traitement documentaire. Virtuelle, discrète et toujours disponible, elle est intégrée aux données mais elle se cache derrière le texte et n'apparaît que sur ordre exprès, quand l'utilisateur demande aussi des comptes et des calculs. En rendant ces modestes services, elle peut espérer qu'on lui rende un jour justice.

---

6. Nous avons connu quatre versions de ce langage en un an. L'exemple de *Supercard* montre que la gestion des fenêtres multiples, des grands écrans, de la couleur, n'est pas impossible et qu'on peut attendre dans un avenir proche les avantages de la compilation.