

SATOR et la statistique

Étienne Brunet

▶ **To cite this version:**

Étienne Brunet. SATOR et la statistique. Etrange topos étranger, M. Vernet, 2006, Kingston, Canada. pp.14-37. hal-01572192

HAL Id: hal-01572192
https://hal.science/hal-01572192

Submitted on 5 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Étienne BRUNET
Institut de Linguistique française (CNRS)

Sator et la statistique

La base SATOR est fondée sur la notion de récurrence. Le *topos* qu'on recherche et qu'on reconnaît est un schéma narratif commun à plusieurs textes et repéré et répété de l'un à l'autre. La démarche lexicométrique est de la même façon centrée sur la répétition. Qu'il s'agisse de mots, de codes grammaticaux, de structures sémantiques, ou de segments infra ou supra lexicaux, il s'agit toujours de réunir les occurrences d'un même phénomène et de manipuler des fréquences. Lorsqu'on s'avance dans les champs thématiques, ce qui conduit du côté du topos, la statistique peut craindre que le sol ne se dérobe sous ses pieds. Le terrain sémantique est en effet plus mouvant que les autres. Ainsi quand l'ordinateur rencontre le mot *marche,* il peut – c'est le plus facile - le mettre en relation avec tous les autres emplois du même mot dans le corpus considéré. Si l'on dispose d'un logiciel de lemmatisation, le pluriel entrera aussi dans les décomptes. L'analyse de la phrase permettra aussi de fixer le code grammatical et de déterminer si l'on a affaire à un verbe ou à un substantif ce qui permet non seulement de rendre plus purs et plus sûrs les effectifs observés, grâce à la désambiguïsation des homographes, mais aussi d'étendre le champ de la statistique au domaine grammatical et syntaxique, et par exemple d'opposer les parties du discours les unes aux autres. Le codage automatique peut-il aller plus loin et coller des balises sémantiques aux mots rencontrés ? Certains logiciels comme *Cordial* ont osé s'engager dans cette voie. Car même pour un but limité à la simple correction orthographique d'un texte, il est parfois utile de fixer le sens des homographes et des mots polysémiques, et de distinguer *voler* (dérober) de *voler* (dans les airs). Mais l'ambiguïté en matière sémantique est beaucoup plus redoutable que celle qu'on peut rencontrer au niveau morpho-syntaxique. Il y a loin de la *marche* de l'escalier à la *marche* en montagne et plus encore à la *marche* du temps, même si l'image subsiste en filigrane dans les emplois les plus abstraits. Au moment où l'occurrence rencontrée doit être rangée dans les tiroirs des concepts, choisira-t-on celui qui recueille les éléments d'architecture, ou celui qu'on voue aux mouvements du corps et quel sort réserver aux emplois métaphoriques où la *marche* n'est plus qu'un déroulement temporel ? Et que faire du mot *nature* ? Mettra-t-on dans le même panier le sentiment de la nature et la nature du sentiment ?

Le recours à un thesaurus extérieur est donc plein de pièges, lorsqu'on veut rendre compte de la composante sémantique des mots. On arrive cependant à caractériser grossièrement les sujets que traite le texte analysé, même si les étiquettes proposées sont souvent mal choisies : ainsi *Cordial* rendant compte de l*'Évangile* y reconnaît certes un ouvrage de religion, mais aussi de chirurgie (à cause des guérisons miraculeuses) et de linguistique (à cause de la Sainte Écriture). On obtient un résultat peut-être meilleur en se contentant des spécificités, calculées sur les graphies, en dehors de tout codage (si l'on choisit Frantext comme référence externe, l'*Évangile* est alors caractérisé par les mots *Jésus, adeptes, disciples, pharisiens, Galilée, Pilate, Synagogue, royaume, Pâque, sabbat, fils, Seigneur, serviteur, messie, parabole, etc…)*.

Mais quelles que soient la méthode et la référence, dans le meilleur des cas on circonscrit un thème, jamais un topos. Le topos est une configuration complexe, structurée qui ne se réduit pas aux éléments qu'il met en jeu, qu'il s'agisse de mots, de thèmes, de concepts, de types ou de motifs. Et ici je renvoie à un article de Michèle Weil qui définit le topos en le distinguant de ce qu'il n'est pas[1]. Ces distinctions ont certainement fait l'objet de longues discussions au sein de l'entreprise Sator et j'ai personnellement entendu il y a dix ans, à l'École Normale Supérieure, un exposé qui précisait clairement les objectifs et les méthodes que les fondateurs avaient en vue à l'origine de ce projet.

La mise en œuvre d'un tel projet exclut que les données fournies à la machine puissent être des données brutes qui seraient enregistrées automatiquement - ce qui est le cas des corpus en texte intégral. Un filtre humain est placé à l'entrée et de la qualité de sa lecture dépend le traitement ultérieur. La subjectivité est redoutable dans une telle entreprise, car selon l'humeur de chacun l'observation et l'interprétation des faits narratifs peuvent varier et l'on peut craindre que les résultats à la sortie soient déterminés en partie par les choix exercés à l'entrée. Mais il en est ainsi de beaucoup de recherches menées dans les sciences sociales. L'entreprise de la BDHL (Banque de données d'histoire littéraire) qui ressemble au présent projet et qui a été menée à Paris par l'équipe de Henri Béhar et Michel Bernard a dû surmonter le même obstacle: pour définir les thèmes, c'est une conscience humaine qui proposait les mots-clés. De même si l'on s'intéresse aux images, la machine aura bien du mal à distinguer les emplois métaphoriques de ceux qui ne le sont pas et le chercheur aura non seulement à proposer une liste de mots éventuellement porteurs d'une image, mais aussi à trier dans les résultats ceux qui contiennent une métaphore. Encore peut-on parfois faire appel à la statistique pour faire émerger les thèmes ou les images. Mais il s'agit ici de structures profondes qui appartiennent à la narratologie et qui échappent complètement à « 1'intelligence » des ordinateurs. On approuvera donc que les responsables du projet prennent tant de précautions pour réunir les données et les soumettre à une critique collective sans cesse renouvelée, même si cette démarche laisse une impression d'inachèvement et de perpétuelle remise en question. Les "catégories" sous lesquelles sont rangées les

[1] Michèle Weil, « Un thésaurus informatisé pour la topique romanesque. Le projet de SATOR », in *Banques de données et hypertextes pour l'étude du roman*, PUF, 1997, pp. 44-62

faits sont parfois incertaines: leurs limites, tantôt trop étroites, tantôt trop larges, n'empêchent ni les lacunes ni les recouvrements et leur formulation est parfois peu claire. Les auteurs du projet reconnaissent qu'elles sont provisoires et qu'elles doivent être soumises à l'épreuve des utilisateurs. L'utilité du produit, dans son état actuel, est expérimentée dans des classes du secondaire et on envisage de garder la trace des interrogations pour étudier le cheminement pédagogique le plus efficace.

Si un tel espion est déjà opérationnel, il a probablement dénoncé l'effraction dont Satorbase a été victime ces derniers temps. J'avoue être l'auteur du forfait, accompli toutefois sans intention de nuire. La porte était ouverte, même aux simples curieux, sans qu'il soit nécessaire de décliner son identité et de fournir le mot de passe d'un abonné patenté. Le logiciel d'ailleurs invite l'utilisateur à s'introduire dans les chambres et à explorer la liste des topoi ou des catégories, ce que j'ai fait lors d'une première visite. Pour payer mon droit d'entrée, j'ai accepté une tâche subalterne de balayeur en rétablissant les accents là où le codage minimal de Sator s'était contenté des majuscules non accentuées, c'est à dire dans la dénomination des topoi et des catégories. Cette liste, grosse d'un millier de mots différents, a été rapprochée de celle de Frantext pour souligner les spécificités du métalangage de Sator. Voici les éléments que le calcul met en relief : *narrateur, causer, viol, emprisonner, partenaire, substitution, maltraiter, révéler, provoquer, infidélité, duel, ruser, échouer, venger, jalousie, chantage, feindre, suicide, substituer, femme, amants, aider, amant, séduire, fuir, meurtre, inceste, amante, amour*, etc. L'aspect causatif de beaucoup de verbes de la série (*causer, provoquer, révéler, aider, échouer*) montre assez qu'il s'agit d'une action dont les effets sont liés aux causes. L'action elle-même paraît se limiter au jeu de l'amour (*amant, amour, séduire, femme*), de la trahison (*infidélité, jalousie, chantage, ruser, feindre, substituer*), de la violence et de la mort (*emprisonner, maltraiter, venger, duel, viol, inceste, suicide, meurtre*).

On peut estimer cependant que la liste des topoi constitue un abrégé de la base, dont l'effectif est trop restreint (moins de 4000 occurrences) pour offrir un fondement solide à la statistique. C'est la base elle-même qu'il faut soumettre aux calculs, à tout le moins ce qu'on nomme le contexte, c'est à dire l'énoncé qui rend compte d'une occurrence rencontrée. Rien ne permet d'extraire l'ensemble de ces contextes, mais rien ne l'interdit non plus. Afin de mettre à l'épreuve la puissance du logiciel d'interrogation et aussi la générosité confiante des créateurs j'ai utilisé une ruse pour obtenir le maximum de fiches. Il a suffi de se lancer dans la « recherche avancée » et de demander toutes les fiches qui contenaient une entrée commençant par la lettre L. Comme les énoncés sans substantif sont rares et que ceux-ci se passent difficilement d'un déterminant, c'était là le moyen indirect de recueillir l'essentiel de la base. La figure 1 montre comment s'est effectué le cambriolage et ce que notre logiciel (figure 2) a fait du butin (figure 3).

Figure 1. Extraction des données (2486 fiches obtenues sur 2660)

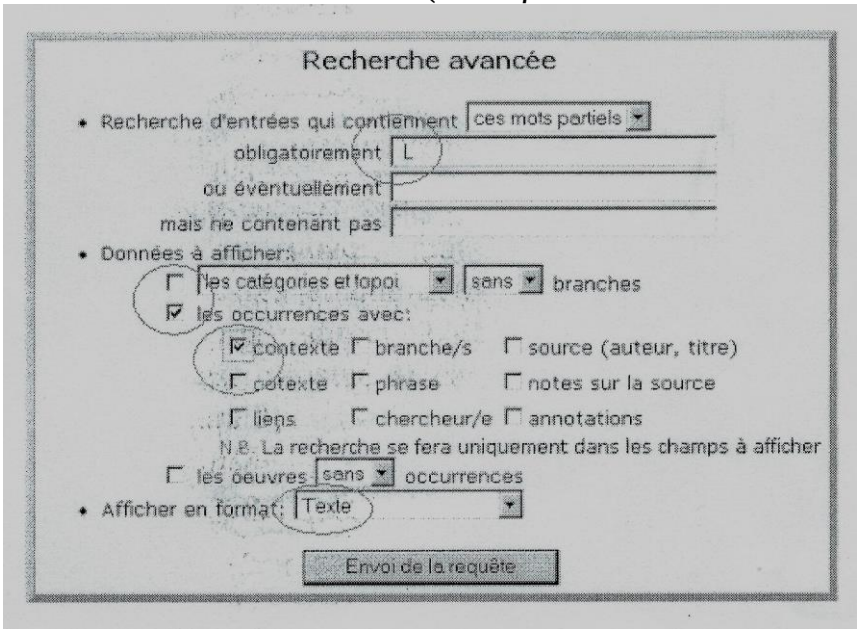
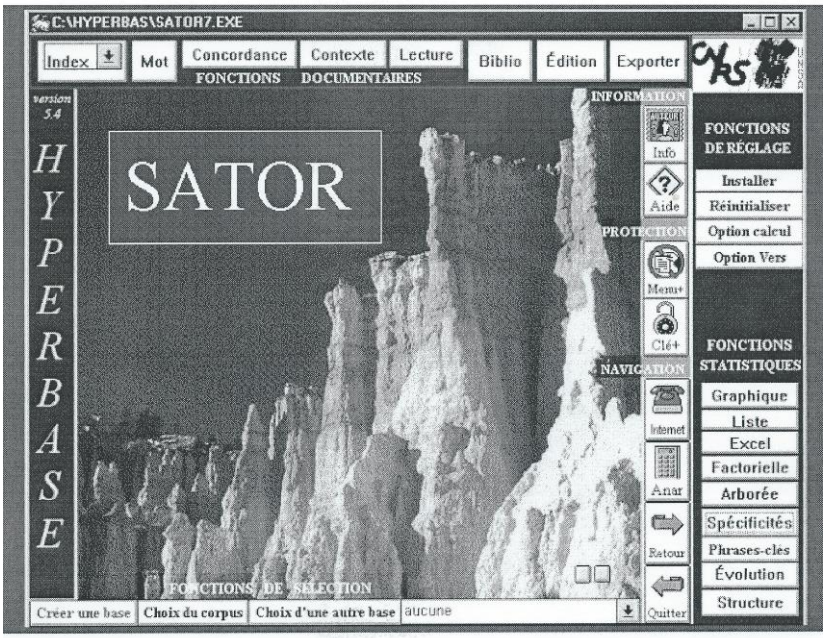
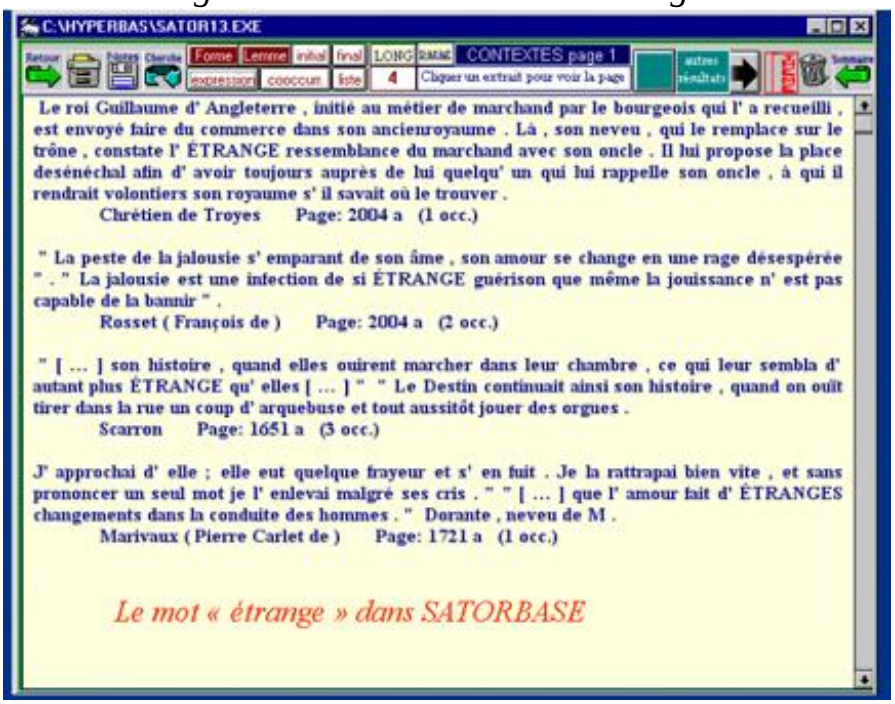


Figure 2 La base SATOR



Il a fallu faire le ménage et ranger dans des champs particuliers tout ce qui relève du cotexte, c'est à dire les précisions concernant le titre, l'auteur, la date et l'édition. La base a certes vu son volume croître d'un facteur 10 mais sans atteindre une dimension suffisante pour permettre à la loi des grands nombres de se manifester. On a certes trouvé quelques exemples des deux mots sur lesquels s'articule le présent colloque (les quatre contextes relevés pour *étrange* sont dans la figure 3), mais quelques unités ne suffisent pas. Et notre projet initial qui était d'étudier *l'étrange* et *l'étranger à* travers Satorbase ne pouvait être maintenu sans sacrifier le sujet ou la base. Ne pouvant choisir les deux à la fois, nous les convoquerons séparément, d'abord Satorbase, puis le couple *étrange/étranger,* sans appliquer le sujet à la base non plus que la base au sujet.

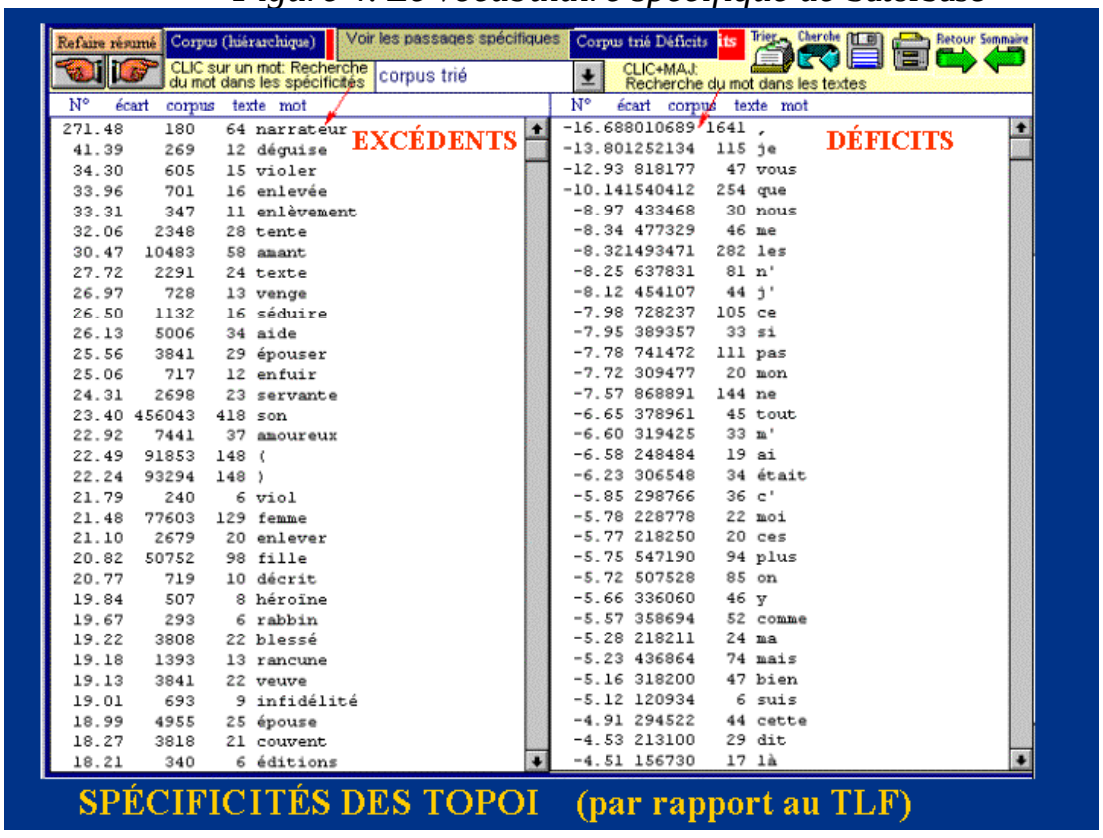
Figure 3. Contexte du mot « étrange »



I - SATORBASE

La figure 4 qui compare les mots relevés dans les contextes de Satorbase à ceux de Frantext donne à peu près les mêmes indications que l'étude des topoi proprement dits. C'est dire que les topoi ont été constitués en même temps que les fiches ou plus souvent en les résumant et en les regroupant, qu'ils partagent ainsi le même vocabulaire et qu'ils se prêtent aux mêmes commentaires, qu'il n'est pas utile de répéter.

Figure 4. Le vocabulaire spécifique de Satorbase

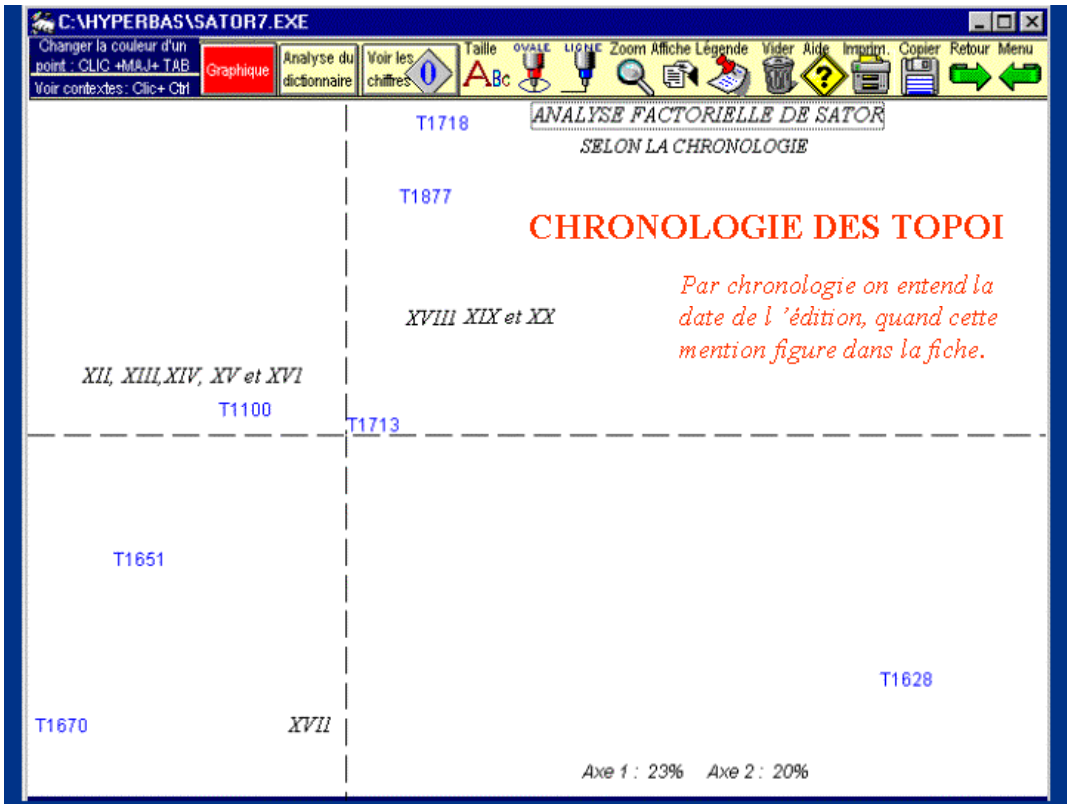


En dehors de cette comparaison externe avec *Frantext* pour référence, l'exploitation des fiches de *Satorbase* né peut aboutir à quelque résultat que si un critère interne permet de constituer des sous-ensembles qu'on opposera les uns aux autres, la statistique étant nécessairement une procédure comparative.

La chronologie.

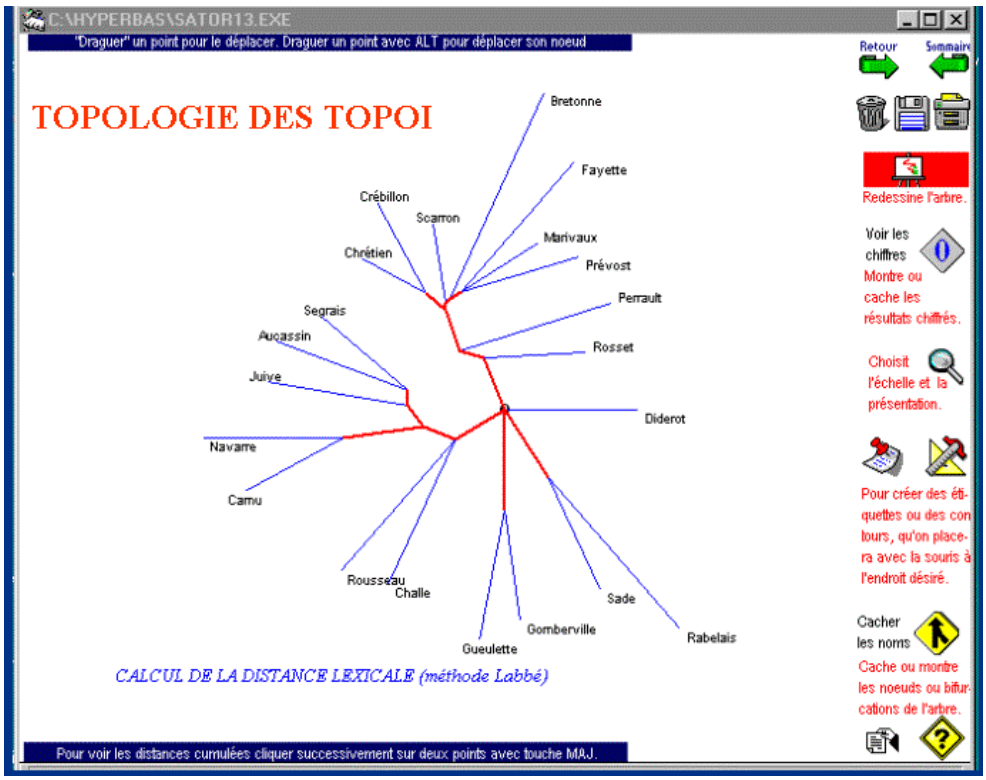
Le premier critère qui vient à l'esprit est la chronologie. Quand une base recouvre une période de plusieurs siècles, on peut s'attendre à une évolution. On peut par exemple imaginer que les héros de la chevalerie médiévale aient un comportement différent du paysan parvenu que Marivaux met en scène à la veille de la Révolution. Trouve-t-on la trace de tels changements dans les données? Pour cela on n'a d'autre ressource que de s'appuyer sur la date de l'édition de l'ouvrage exploré. Malheureusement cette mention de date ne s'applique pas toujours à l'édition originale, qui n'est pas toujours accessible. Toutes les références postérieures à la Révolution sont donc sujettes à caution et on les a regroupées dans la dernière tranche, laquelle est un fourre-tout sans grande signification. L'analyse factorielle de la figure 5 souffre certainement de cette distorsion, mais le mouvement chronologique y est toutefois sensible: les fiches qui se rattachent aux époques les plus lointaines se groupent à gauche (jalons 1100, 1651,1670), tandis que les textes les plus récents occupent la partie droite (jalons 1628, 1713, 1718 et 1877). Précisons qu'il s'agit d'une sorte de carte géographique exprimant les dis' tances qui s'établissent entre les tranches au seul vu de leur vocabulaire, sachant que deux tranches sont estimées plus proches quand elles partagent un lot plus important de mots communs.

Figure 5. La chronologie dans Satorbase



Un second essai peut être tenté en distribuant les fiches selon les écrivains dont elles rendent compte. Les mêmes données seront intégrées à une base nouvelle ordonnée autour des auteurs. Comme certains sont mal représentés, n'ayant qu'un nombre limité de mentions, on a pris le parti de retenir ceux qui obtenaient le plus de suffrages, soit une vingtaine d'écrivains (ou de titres lorsque l'auteur était anonyme). On procédera comme précédemment à un calcul de distance, mais pour exprimer ces distances on aura recours à l'analyse arborée. L'interprétation de la figure 6 doit toutefois veiller à ne considérer que la longueur du parcours pour aller d'un auteur à l'autre sans tenir compte des angles et des directions (la distance à vol d'oiseau n'est pas significative). Avouons que les regroupements manquent parfois de lisibilité. Certes on peut comprendre pourquoi Prévost, Marivaux et Mme de la Fayette se donnent la main, et pourquoi Rousseau et Challe se rapprochent. Mais la liaison établie entre Sade et Rabelais nous laisse perplexe, comme celle qui fait voisiner Crébillon et Chrétien de Troyes. Nous laissons aux spécialistes le soin d'expliquer la présence de certains couples (Marguerite de Navarre et Camus, Gueulette et Gomberville, Aucassin et la Juive) et le mouvement d'ensemble qui préside à la répartition des auteurs. Plusieurs facteurs, dont le genre et l'époque, semblent se mêler et se combattre pour justifier la position de chacun. Mais ce qui ajoute à la confusion, c'est le fait que n'est pas considéré le texte même de ces écrivains, mais la note d'une main étrangère à propos de ce texte.

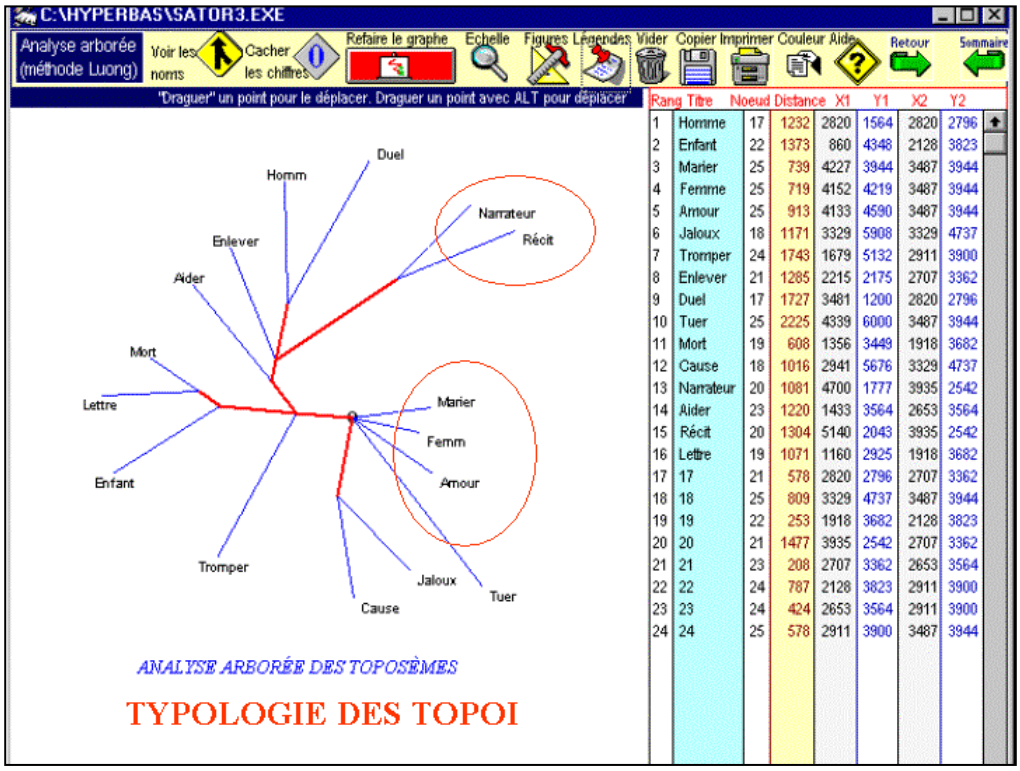
Figure 6. Le critère auteur



La typologie

Une dernière tentative sera-t-elle plus heureuse? Interrogeons la base sur des thèmes, ou plus exactement des mots, et enregistrons à chaque fois les résultats pour former un sous-ensemble d'un nouveau corpus rassemblant l'ensemble des interrogations successives. Il peut se faire qu'une même fiche réponde à deux interrogations différentes. Il y aura donc des doublons, qui ne sont pas gênants puisqu'ils contribuent à raccourcir les distances là où les fiches sont partagées entre plusieurs thèmes. Il y aura aussi des lacunes si une fiche ne contient aucun des termes demandés. Enfin la série des termes proposés pour la consultation de Satorbase comporte une part d'arbitraire, le choix ayant retenu les mots les plus fréquents parmi les verbes ou les substantifs. En réalité on espérait retrouver dans les résultats l'organisation des topoi en catégories et notamment la division majeure qui distingue le descriptif, le discursif et le narratif. Le lecteur pourra apprécier si l'attente est déçue ou satisfaite. La figure 7 qui est une nouvelle fois un calcul de distance lexicale montre bien que les topoi discursifs se distinguent des autres. La branche la plus excentrique de la figure tient à l'écart le *narrateur* et le *récit*. On voit aussi se former près de l'origine un noeud formé autour de la constellation *femme*, *amour*, *marier* et correspondant à la situation ordinaire. On comprend que le *duel* soit une affaire d'*homme* mais pourquoi le verbe *tuer* est-il séparé de la *mort*, pourquoi ne trouve-t-on pas réunis les couples attendus: *enfant* et *enlever*, *tromper* et *jaloux*?

Figure 7. La typologie



En conclusion, les résultats, plutôt médiocres, qu'on obtient de la statistique viennent d'une méprise initiale. *Satorbase* est une base structurée, non un corpus en texte intégral. Et sa conception l'oriente vers une exploitation documentaire plutôt que statistique. Plusieurs conditions n'ont pas été remplies dans cet essai prématuré qui n'a pas tenu compte des mentions *Provisoire*, *Proposé* ou *Accepté*, afin d'atteindre des effectifs moins réduits. Si l'on n'avait considéré que les fiches dûment ratifiées, certains défauts qui nuisent aux résultats auraient sans doute disparu: fiches incomplètes, dates manquantes, mélange d'éditions anciennes et de rééditions modernes, fiches sans texte et réduites à de simples références, enfin citations peu homogènes, certaines développées dans un état de langue ancien, incompatible avec le texte moderne qui l'entoure.

Pour ne pas rester tout à fait étranger au thème de ce colloque, force est de recourir aux sources primaires, c'est à dire aux textes, qu'ils soient ou non explorés par Satorbase. Et dans un premier temps j'utiliserai une base hypertextuelle que le ministère de l'Éducation nationale m'a demandé de constituer à l'intention des populations scolaires et universitaires. Grosse de trois millions d'occurrences, elle couvre une large période, qui va de Rabelais à Proust, et un large spectre parmi les genres littéraires, puisque le théâtre voisine avec la poésie et le roman avec les essais Ce sont précisément les *Essais* de Montaigne qui sont consultés dans le graphique 8, à la recherche de l'*étrange*. On a souvent dit la filiation qui lie à l'origine l'*étrange* et l'*étranger*. Il semble que le lien soit déjà rompu au temps de Montaigne, pour qui le mot *étrange* n'a plus rien d'exotique ou de merveilleux. La plupart du temps, le mot n'a plus qu'un sens logique dans les *Essais*: il s'applique à ce qui est contradictoire, illogique ou insensé. Et c'est encore plus vrai de Pascal qui dans les *Provinciales* utilise le mot comme une arme polémique, pour dénoncer la fausse charité, la théologie dévoyée, le langage égaré ou les maximes douteuses. Chez Pascal cet euphémisme ironique est toujours dirigé contre la partie adverse, il s'agit toujours d'un comportement ou d'un raisonnement humain, jamais de quelque spectacle merveilleux offert par la nature ou la magie.

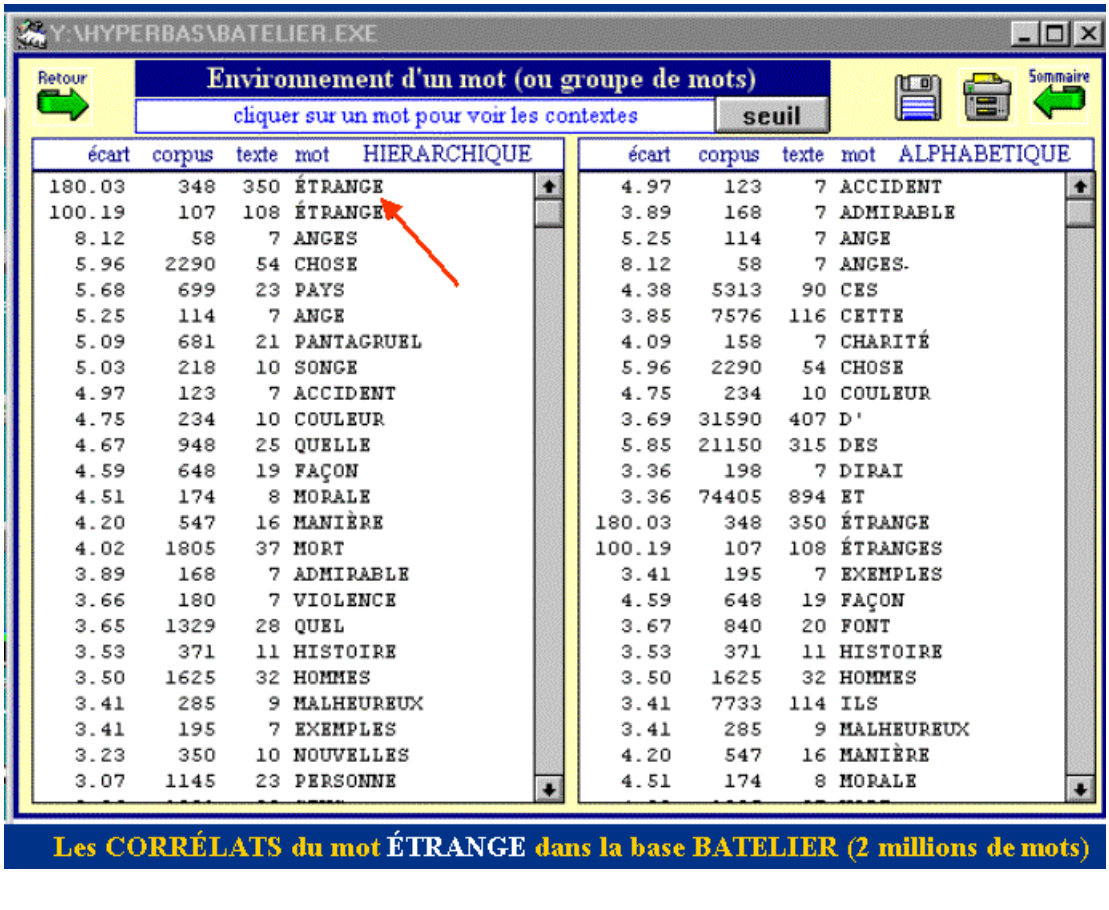
L'extrait très partiel de la figure 8, trop attaché à suivre la démarche des *Provinciales*, ne rend pas compte de la totalité des 350 emplois du mot. On aurait tort de croire que le sème poétique et mystérieux du mot ait complètement disparu au XVIIe siècle. On le retrouve toujours neuf et évocateur sous la plume des poètes, surtout des symbolistes, et dans le roman d'évasion: comme par exemple dans le *Grand Meaulnes*, dont un chapitre s'intitule *la fête étrange.* Plutôt que d'explorer l'une après l'autre les 350 lignes de la concordance, on a utilisé un calcul particulier des spécificités, voisin de celui de la figure 4. En rassemblant toutes les phrases où le mot *étrange* est relevé, on compare cet ensemble à la totalité de la base, pour mettre en évidence les mots qui se trouvent préférentiellement dans l'entourage de l'*étrange* et qu'on appelle des corrélats. Les raisons qui expliquent ce voisinage peuvent appartenir à la

versification (ce qui explique ici la place des *anges*) ou à la syntaxe (les substantifs sont évidemment plus nombreux à proximité de l'adjectif *étrange*).

Figure 8. Extrait de la concordance du mot étrange dans la base Batelier



Figure 9. Les corrélats du mot étrange



Mais l'aimantation des mots autour de l'*étrange* est surtout sémantique et thématique. Le sens premier s'impose encore lorsqu'il s'agit d'un *songe*, d'une *couleur*, d'une *chose*, d'une *personne*, d'un *accident*, d'une *histoire*. Le mot sert alors d'intensif ou de détersif pour éveiller l'attention et mettre l'accent sur le caractère surprenant de ce qu'on évoque. Il est alors souvent associé à l'expression de l'étonnement ou de l'admiration (*quelle, quel, admirable*), de la nouveauté (*nouvelles*) ou de l'excès (*violence, malheureux*). Mais la présence de quelques termes abstraits laisse à penser que le terme est souvent pris dans un sens ironique et polémique, comme on l'a vu chez Pascal.

Le même traitement appliqué au mot *étranger* dans la même base révèle un environnement lexical très différent (figure 10). Alors que *étrange* servait à caractériser, *étranger* sert à identifier et à désigner ce qui n'appartient pas à la communauté nationale. C'est le pôle opposé à soi, sous quelque forme qu'apparaisse l'affirmation de soi (*soi, nous, notre, nôtres, miennes, propres, français, maison, pays, nation*). L'opposition géographique est souvent de caractère politique (*affaires, sujets*), mais souvent aussi elle est linguistique (*langue, ouvrages*), et dans les deux cas un soupçon plutôt hostile accompagne l'étranger (*barbares, dangers*).

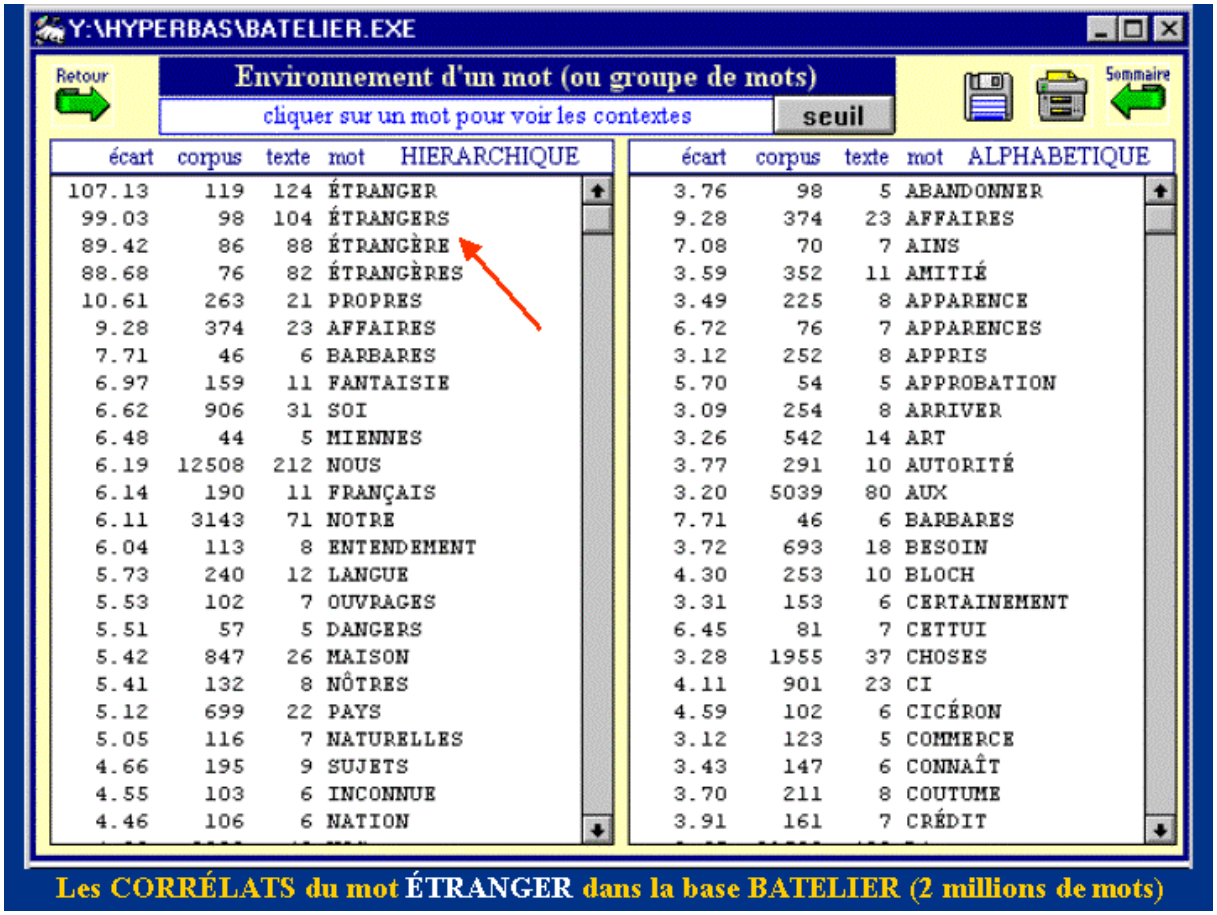
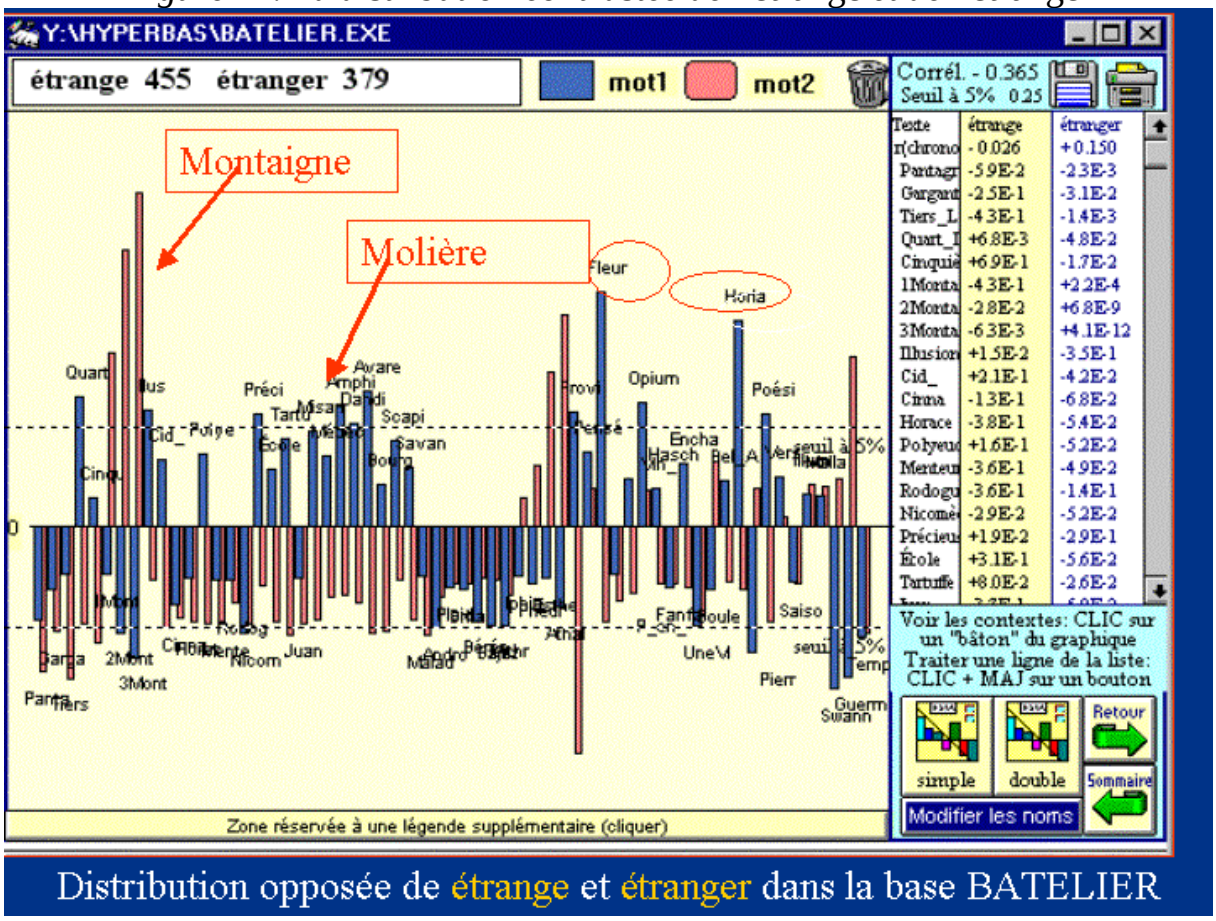


Figure 10. Les corrélats du mot étranger

Il apparaît donc que les deux mots vivent dans deux univers étrangers l'un à l'autre. Même s'ils se rattachent à la même étymologie, ils n'ont ni la même

famille, ni les mêmes fréquentations. On dirait même qu'ils évitent de se rencontrer, ce que tend à montrer la figure 11. Les deux termes ont une fréquence comparable (environ 400 occurrences dans la base *Batelier*), mais une répartition très différente dans les textes. Tantôt l'*étrange* l'emporte: chez les poètes symbolistes (Baudelaire, Rimbaud) ou les conteurs fantastiques (Rabelais, Maupassant), ou dans un emploi polémique, chez Pascal et Molière, tantôt c'est l'*étranger* qui prend le pas (notamment chez Montaigne et Proust). Le signe négatif du coefficient de corrélation (-0,37) rend compte de cette opposition, qui ne peut être attribuée au hasard (le seuil de 0,25 est largement dépassé).

*Figure 11. La distribution contrastée de l'*étrange *et de l'*étranger



La confirmation de cette opposition peut être obtenue, soit en limitant le corpus à des écrivains qui partagent la même époque et le même environnement culturel, soit en l'étendant à la dimension de *Frantext*.

Dans le premier cas on se contentera de trois écrivains qu'on a souvent rapprochés parce qu'ils ont en commun non seulement non seulement la même époque, mais aussi les mêmes préoccupations philosophiques et religieuses et la même manière d'aborder ces questions par le biais de l'essai (même si Rabelais donne à ses pensées une tournure burlesque et une trame narrative). L'analyse factorielle (figure 12) agglomère toutes les formes de l'*étranger* dans les trois livres de Montaigne, à l'extrême droite du graphique, tandis qu'à l'opposé, sur la gauche, Rabelais et Pascal se disputent l'*étrange*, dans sa variante fantastique chez Rabelais et argumentative chez Pascal, surtout dans les *Provinciales*.

Dans le second cas (figure 13) les auteurs disparaissent et se fondent dans les 12 tranches chronologiques, réparties de 1500 à nos jours. Les genres s'y mêlent aussi, même si on a écarté de *Frantext* les textes catalogués comme techniques. Il reste une masse imposante, de 117 millions de mots, qui ne réserve pas le même accueil aux notions d'*étrange* et d'*étranger*. L'*étranger* est accepté, ou du moins évoqué, dans les textes du XVIIIe siècle et au début du XIXe, à un moment où les écrivains commencent à voyager. L'*étrange* est plus en faveur dans les autres époques mais l'on devine qu'il ne s'agit pas du même emploi selon qu'on a affaire à la comédie de Molière ou à la description, fantastique ou pittoresque, qui a cours dans la deuxième moitié du XIXe siècle.

Figure 12. Rabelais, Montaigne et Pascal. Analyse factorielle

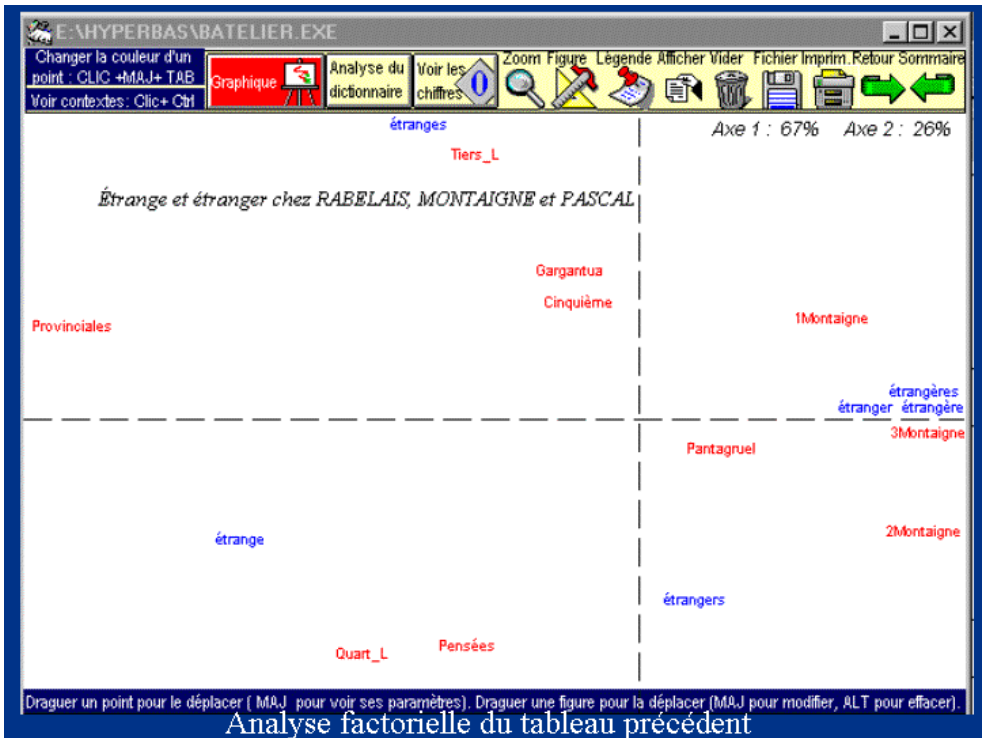
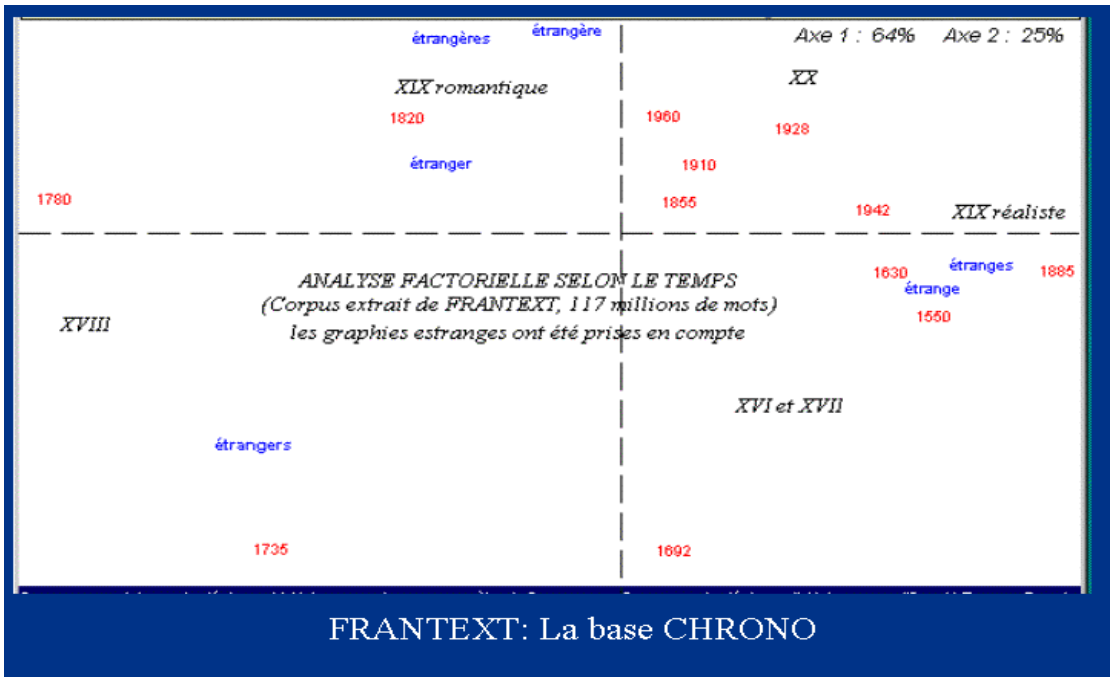
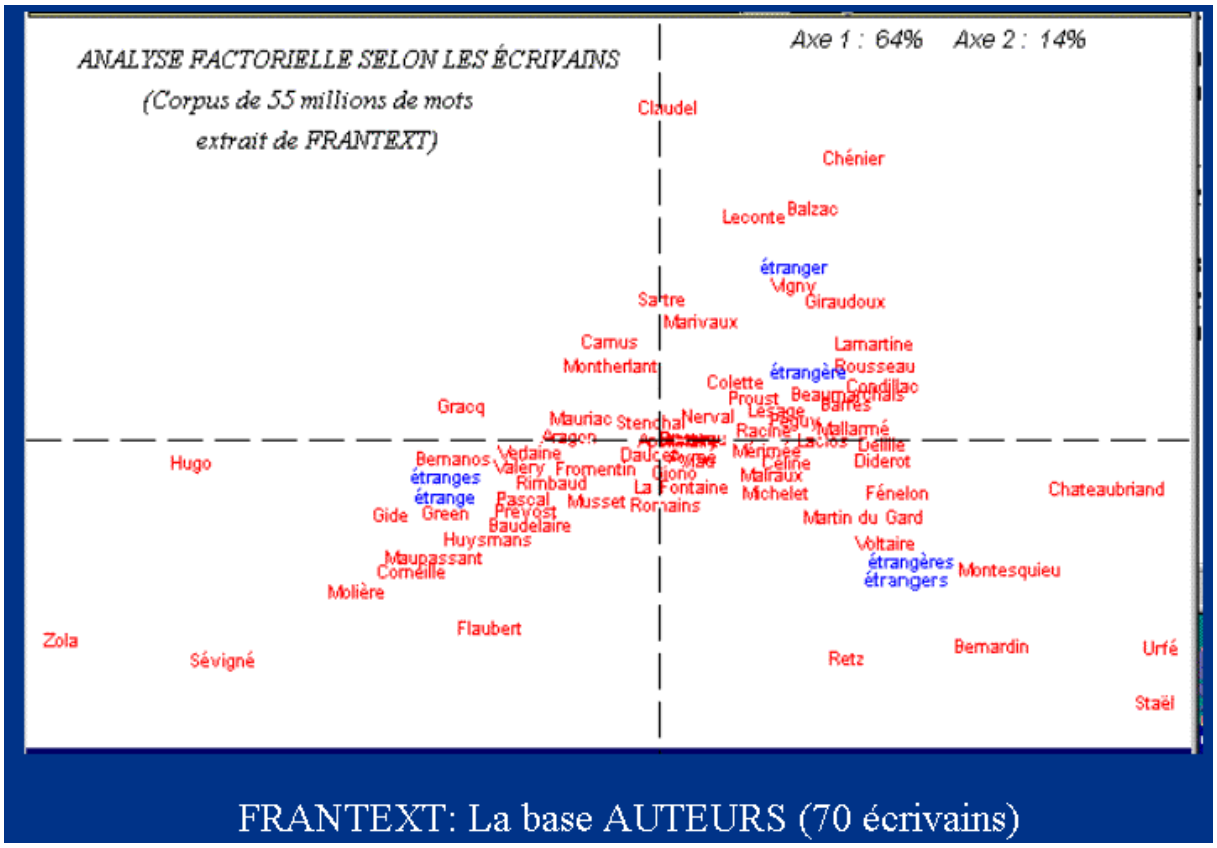


Figure 13. Le corpus littéraire et chronologique de Frantext (117 millions de mots)



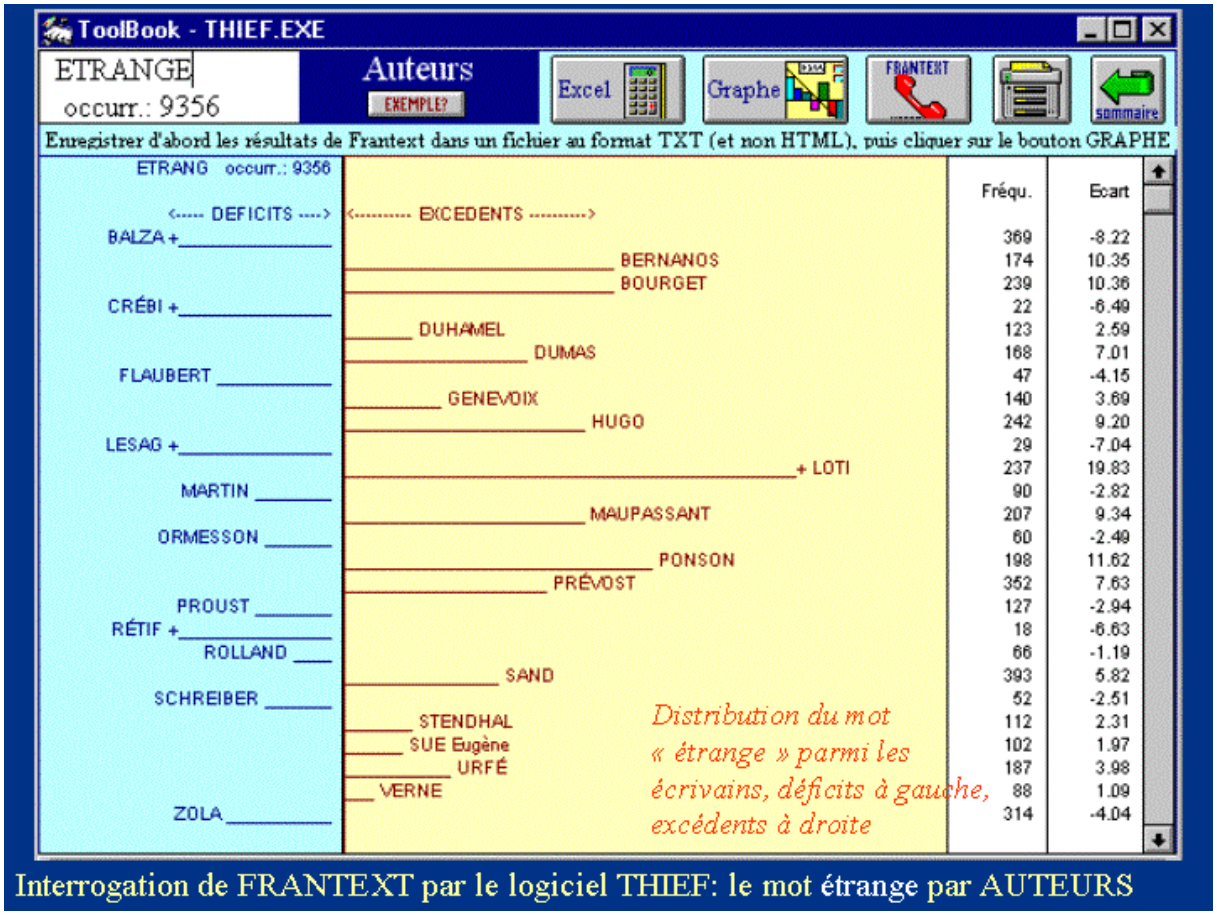
On peut mettre des noms pour voir à quelle influence particulière tient cette évolution. 70 écrivains (soit un corpus de 55 millions de mots) sont ainsi convoqués dans la figure 14 et sommés de choisir entre l'*étrange* et l'*étranger*. Les deux notions s'opposent de la même manière: à droite l'*étranger* attire les écrivains qui voyagent en dehors des frontières (Bernardin de Saint-Pierre, Chateaubriand, Lamartine) ou qui poursuivent une réflexion historique ou politique sur le sort des nations (Fénelon, Retz, Montesquieu, Voltaire, Diderot, Condillac, Rousseau). À droite les partisans de l'*étrange* viennent de plusieurs horizons: les uns sont sensibles à sa vertu descriptive et évocatrice, et l'on retrouve là les poètes (Hugo, Verlaine, Baudelaire, Rimbaud, Valéry, Aragon) et les romanciers réalistes (Flaubert, Maupassant, Huysmans et Zola); les autres utilisent sa force polémique de dérision dans le dialogue, la correspondance ou l'argumentation (Pascal, Corneille, Molière, Mme de Sévigné).

Figure 13. Analyse factorielle des écrivains



Il faut toutefois prendre garde que l'analyse factorielle de la figure 14 impose un choix draconien et que certains auteurs qui n'aiment ni l'un ni l'autre des deux termes doivent pourtant se ranger dans le camp dont ils se sentent le moins éloigné. Si l'on consulte librement les écrivains en dehors de toute alternative, on peut observer la tiédeur de certains suffrages. Ainsi parmi les romanciers réalistes, Maupassant est le seul à manifester franchement le goût de l'étrange et du fantastique. Flaubert et Zola sont plutôt réfractaires et on les observe parmi les déficits dans l'histogramme 15, qui isole le mot *étrange*.

*Figure 15. Les partisans déclarés de l'*étrange



Près de 10000 occurrences de l'*étrange*, toutes graphies confondues, ont été analysées dans le précédent histogramme. On les retrouve dans la figure 16, qui explore les 10000 phrases où l'*étrange* est apparu, en comptant tous les mots qui se rencontrent à proximité. Le procédé est le même qu'on a utilisé dans la figure 9[2], mais le champ d'observation est ici vingt fois plus vaste. Observons d'abord la présence insistante des formes verbales de la description[3] *paraissait, semblait, semblaient, paraître*, qui appartiennent en même temps, à cause de l'imparfait, au genre narratif (ou plus rarement poétique). La liste comporte des synonymes: *mystérieux, mystère, bizarre, inconnu*, et laisse apparaître une aimantation vers le *vide*, le *silence*, le *rêve* et la *folie*. La qualité de ce qui est étrange peut s'appliquer à des *choses* ou des personnes (*personnage, créature*) mais le plus souvent l'adjectif est là pour qualifier l'aspect inattendu d'un phénomène (*aspect, forme, façon, attitude*), les sens que l'étrange tient en éveil (*regard, yeux, bruits, paroles*) et l'impression que l'étrange exerce sur la sensibilité (*sensation, impression, expression*).

---

[2] *Frantext* ne fournit pas directement la liste pondérée des corrélats. Il délivre des fréquences brutes qu'il faut ensuite soumettre au calcul, pour mettre en évidence les corrélats réellement liés au mot choisi pour pôle. Notre logiciel *THIEF* réalise cette opération de filtrage, quand les données ont été extraites de *Frantext*, via Internet. Précisons que la même procédure est utilisée dans notre logiciel *Hyperbase*, quand les données sont indépendantes de *Frantext* et appartiennent à l'utilisateur.

[3] Ou plutôt de l'évocation. Les verbes *paraître* et *sembler* sont moins précis que suggestifs. Ils introduisent une image plutôt qu'un fait. Et ils s'accordent avec le flou mystérieux qui entoure l'*étrange*.

*Figure 16. Les corrélats de l'*étrange *dans* Frantext**.**

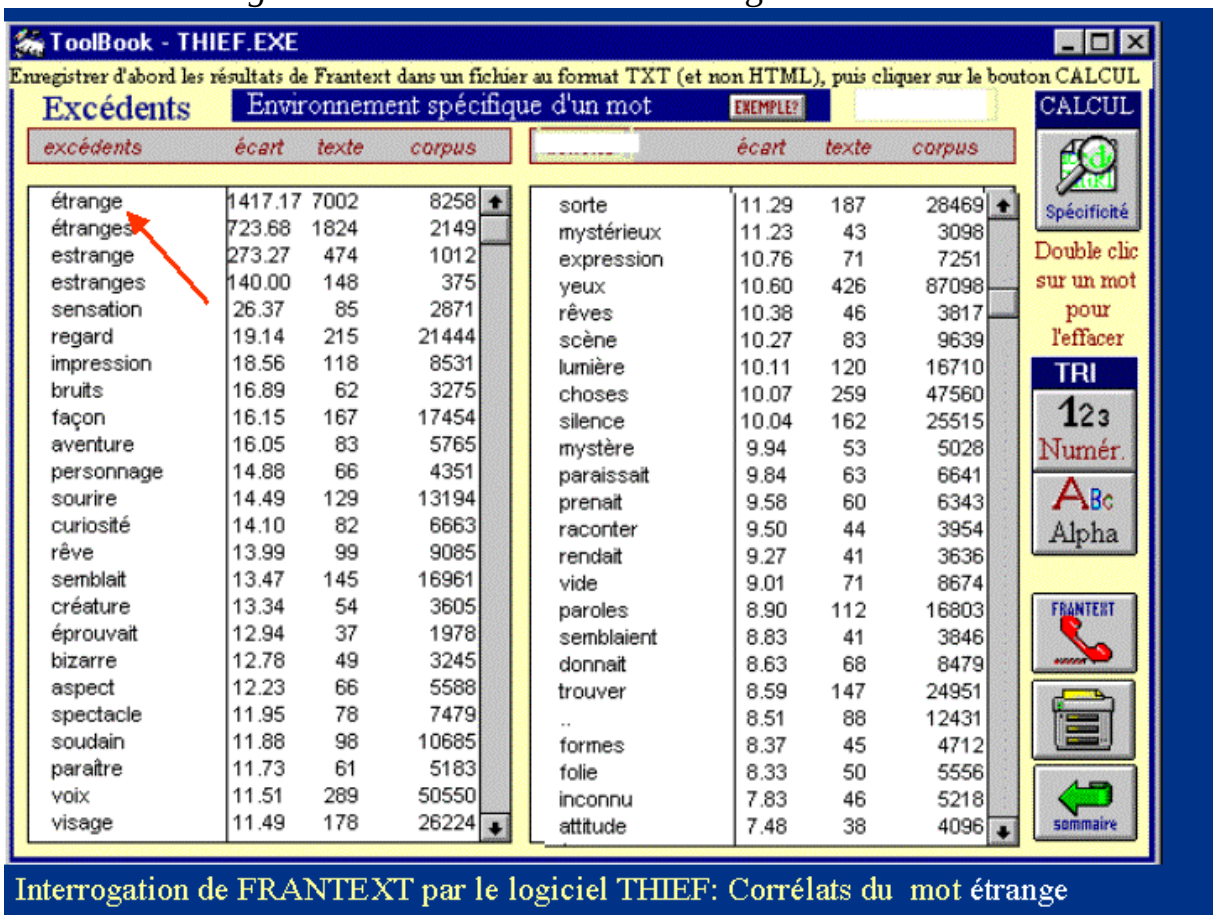
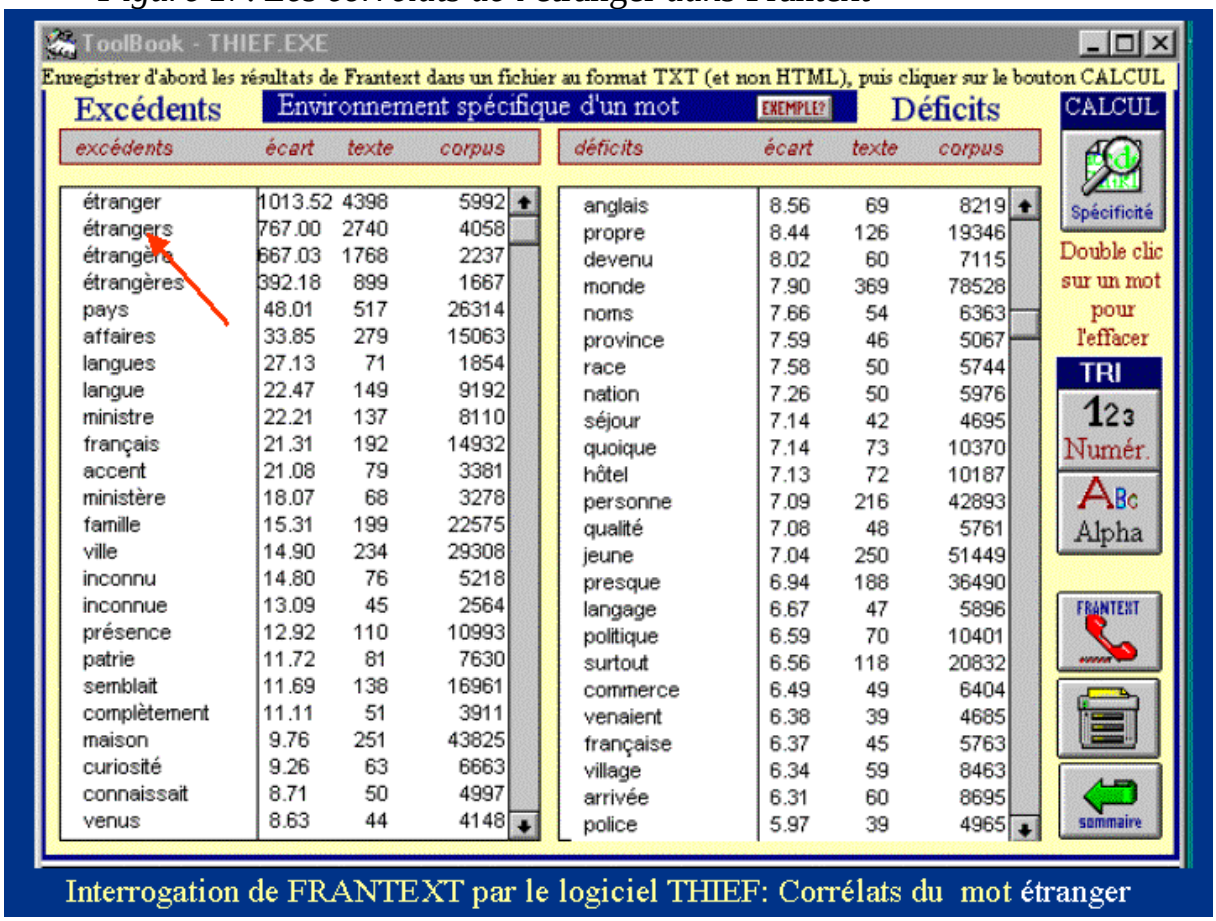


*Figure 17. Les corrélats de l'*étranger *dans* Frantext



Le champ sémantique constitué par les corrélats de l'*étranger* est bien différent (voir figure 17). Le terme ne cherche pas à suggérer, mais à identifier. Il entre de plain-pied dans la constellation lexicale qui circonscrit une communauté sociale, grande ou petite (*monde, race, pays, nation, patrie, province, ville, village, maison, famille*). Il s'applique au monde de la politique, des affaires et des voyages (*ministre, ministère, affaires, commerce, voyage, séjour, hôtel*), et plus étroitement encore aux barrières et aux échanges linguistiques (*langue, langage, noms, accent, français, anglais*).

Le parcours un peu cahotique (on hésite entre deux orthographes) que je viens d'accomplir ne risque guère la concurrence et la récurrence, et je crois qu'il a peu de chance d'être accepté comme un topos dans le cercle des sartoriens. Je crains que le chiffre introduit par effraction dans ce cercle lettré n'y paraisse bien étrange et ne soit rejeté comme un corps étranger. Mais s'il advenait que *Satorbase* non seulement multiplie les occurrences de topoi relevées dans les textes, mais aussi développe son système d'exploitation pour y faire entrer les textes mêmes qui ont été analysés, alors peut-être les techniques documentaires et quantitatives dont j'ai rendu compte trouveraient leur utilité et justifieraient l'exposé prématuré que je viens de faire.