Étienne BRUNET

HYPERBASE©

Logiciel hypertexte
pour le traitement documentaire et statistique des corpus textuels

MANUEL DE RÉFÉRENCE

Première partie

HYPERBASE pour WINDOWS

version standard 8.0 et 9.0

janvier 2011

La présente version d'Hyperbase (9.0, janvier 2011) offre peu de fonctions inédites, mis à part un nouveau module d'analyse factorielle, mais s'adapte aux contraintes des nouveaux systèmes Windows, principalement VISTA (64 bits) et SYSTEM 7 (64 bits), tout en restant compatible avec les systèmes précédents. Afin de ne pas confondre les fichiers, les bases disponibles affichent le suffixe .TBK au lieu de .EXE, tout en gardant le même nom.

L'installation peut être partielle ou complète, selon que l'on sollicite SETUPmin.EXE ou SETUPmax.EXE. Dans le premier cas le répertoire créé à la racine du disque dur, soit C:\HYPERBAS\, occupera 500 Mo, tout en contenant deux bases de démonstration (EXEMPLE.TBK, version standard, et GAULLE.TBK, version lemmatisée) et tous les outils nécessaires à la production de nouvelles bases. Dans le second cas, le contenu entier du DVD sera déversé sur le disque dur (soit plus de 3 Go), de telle sorte que le DVD ne sera plus nécessaire pour charger une base particulière.

Une seule contrainte demeure, dans les deux cas : préciser impérativement la localisation **C:\HYPERBAS** *et nulle autre, au moment où une boîte de dialogue demande la destination du logiciel.*
Quand l'installation est achevée, le répertoire C:\HYPERBAS\ s'ouvre sur un menu qui détaille les bases disponibles. En cliquant sur l'icône d'un écrivain, on ouvre la base correspondante, directement si l'installation a été complète, ou par copie automatique si l'installation a été partielle.

Pour créer une base nouvelle, utiliser TB100RUN.EXE en association avec l'un des modèles dont le nom commence par HYPER et se termine par .TBK

- hyperBAS pour une base standard, quelle que soit la langue

- hyperCOR pour une base en français avec lemmatisation Cordial

- hyperTAG et hyperVER pour une base en français avec lemmatisation TreeTagger (Attention ! TreeTagger n'acceptant ni Vista, ni les systèmes 64 bits, il y a lieu de lemmatiser les textes au préalable avec un système Windows XP et de tenir disponibles les fichiers-résultats (avec le suffixe.CNR) dans le répertoire C:\HYPERBAS\)

- hyperANG, hyperGER, hypESPAG, hyperPOR et hyperITA pour les textes respectivement anglais, allemands, espagnols, portugais et italiens (on peut avoir des exemples étrangers si l'on sollicite les bases BRITISH.tbk, GERMAN.tbk, ESPAGNE.tbk, PORTUG.tbk, ITALIE.tbk).

CHAPITRE 1

L'installation

AVERTISSEMENT

Le présent logiciel a été réalisé avec TOOLBOOK (version 10.5). Un double clic sur EXAMPLE.EXE suffit à un premier examen, ce qu'on peut faire aussi en lançant d'abord le programme (ou runtime) TB105RUN.EXE, puis en ouvrant la base EXAMPLE.EXE, ou bien encore en glissant l'icône EXAMPLE.tbk et en la déposant sur celle de TB105RUN.EXE. Pour une installation complète, voir ce qui suit.

Pour tout renseignement complémentaire, s'adresser à l'auteur:
Prof. Étienne Brunet,
Courriel: brunet@unice.fr

AUTRES VERSIONS

1 - Il existe une version pour Mac du même logiciel. Les fonctionnalités sont à peu près les mêmes. Mais la présentation diffère. Et le code est propre à ce standard (le dialogue y est régi par Hypercard, avec un recours constant à des commandes externes écrites en Pascal, C ou Fortran). Noter que le système MacOS X ignore l'environnement Hypercard, pourtant créé par Apple, et qu'en conséquence notre logiciel Hyperbase ne fonctionne sur ce système qu'en émulation CLASSIC. Mais maintenant les programmes Windows fonctionnent sur les machines Apple, que ce soit avec Bootcamp, Parallels ou VMWare, et il n'est plus nécessaire de maintenir la version proprement Mac d'HYPERBASE.

2 – Certaines implantations d'HYPERBASE ont fait un choix restrictif des fonctions disponibles, lorsqu'il y avait double emploi avec celles qu'offrait quelque autre moteur de recherche. C'est le cas d'une série de cédéroms inscrits au catalogue des éditions Champion (sur Proust, Rimbaud, Pascal…). Les fonctions documentaires y sont incomplètes et les ressources de création absentes.

3 – Inversement des versions spéciales du logiciel (*Hypercor.exe, HyperTag, HyperVer*) proposent des fonctions supplémentaires dont la version normale est

dépourvue et qui sont relatives à l'étiquetage et à la lemmatisation. Comme le lemmatiseur qu'elle met en œuvre ne peut être distribué sans condition, la version *Hypercor.exe* n'est utile qu'aux utilisateurs qui auraient acquis préalablement le lemmatiseur *ANALYSEUR* (version spéciale du correcteur *CORDIAL 7* ou ultérieure distribuée par Synapse Développement, 33 rue Maynard 31000 Toulouse, fax 05 61 63 69 09). En revanche le lemmatiseur TreeTagger étant librement téléchargeable, les versions HyperTag.exe et HyperVer.exe qui l'utilisent sont pleinement opérationnelles. Le lemmatiseur *WINBRILL*, dont HYPERBASE a fait usage quelques années, a été abandonné. S'il s'agit de textes anglais, allemands, espagnols portugais ou italiens, la lemmatisation est assurée par *TreeTagger*, proposé par Helmut Schmidt (université de Stuttgart). Une version particulière d'Hyperbase est fournie pour ces cinq langues : respectivement *HyperAng.exe, HyperGer.exe, HypEspag.exe, HyperPor.exe, HyperIta.ex*. On trouvera dans la dernière partie les particularités et les potentialités ouvertes par l'exploitation des données étiquetées.

4 – Hyperbase s'applique à toute langue qui utilise l'alphabet latin, ce qui exclut notamment l'arabe, le cyrillique, le grec et les idéogrammes chinois. Les caractères accentués (minuscules et majuscules) ont reçu un traitement adéquat. Le français a cependant un privilège: les dialogues et les messages visibles à l'écran sont dans cette langue. Si l'on a affaire à des textes français, la comparaison externe est faite avec les données du Trésor de la langue française ou avec celles du corpus français amassé par Google. Mais on a fourni des dictionnaires de référence différents s'il s'agit de textes anglais, italiens ou portugais.

INSTRUCTION SPÉCIALE EN CAS DE SYSTÈME NON-FRANÇAIS

Les paramètres régionaux qui accompagnent le système peuvent avoir une influence sur la gestion des caractères accentués, qu'il s'agisse du clavier, de l'affichage ou du tri. Il n'y a pas lieu de faire une réinstallation complète de Windows, si, par exemple, on utilise un système anglais et qu'on veuille traiter des données françaises. Windows a prévu qu'on puisse adapter aux besoins les réglages internationaux. La procédure est la suivante:

- cliquer sur POSTE DE TRAVAIL (ou l'équivalent anglais de ce terme)
- puis sur PANNEAU DE CONFIGURATION
- puis sur PARAMÈTRES RÉGIONAUX. On voit alors apparaître la carte du monde et au dessus de celle-ci un menu déroulant où l'on choisit FRANCAIS (STANDARD)
- puis cliquer sur OK et redémarrer .

INSTALLATION

Le logiciel HYPERBASE a perdu sa dernière lettre dans la version Windows, pour se conformer aux exigences du DOS. On le trouvera deux fois à

l'intérieur du présent DVD: sous le nom HYPERBAS.EXE (sans données) et sous le nom EXAMPLE.EXE (avec un jeu de données provisoire). L'installation consiste à créer un répertoire HYPERBAS à la racine du disque dur C, et à y transférer les fichiers nécessaires. Cela peut se faire par transfert à partir du DVD, en prenant soin de donner les droits d'écriture aux fichiers transférés. Comme ce transfert peut être estimé lourd, on peut se dispenser de copier les bases dont on n'a pas un besoin immédiat, sachant que le programme MENU.EXE peut le faire à tout moment.

Le logiciel INSTALLSHIELDS (version 2), qui réalisait jusqu'ici la contraction et la mise en place automatique des fichiers dans le répertoire C:/HYPERBAS/, n'est pas adapté aux nouvelles machines 64 bits. On utilisera donc la procédure usuelle de «décompression» sous Windows en sollicitant les boutons SETUPmin.EXE ou SETUPmax.EXE, comme indiqué plus haut.

Hyperbase utilise l'environnement TOOLBOOK, qui impose deux contraintes:

- la première est évidente et commune à la plupart des logiciels de traitement de données: c'est la nécessité d'avoir un espace où l'écriture soit autorisée, le DVD ne le permettant pas, non plus que le CD. En règle générale, un répertoire est dévolu au nouveau logiciel installé à l'adresse C:\PROGRAM FILES\. Lors de l'installation d'HYPERBASE, un tel répertoire n'est pas créé à cet endroit habituel, mais plutôt à la racine du disque dur, soit à l'adresse C:\HYPERBAS\ (noter que le nom a été raccourci à huit lettres).
- la seconde était naturelle dans les systèmes anciens de Windows, c'est l'accès, même en écriture, du répertoire temporaire: C:\WINNT\TMP\ (système NT) ou C:\WINDOWS\TEMP\ (système XP). Or cet accès, qui est maintenu quand on est l'administrateur de la machine, n'est plus autorisé pour un utilisateur dénué de droits (une zone temporaire lui est allouée qui n'est plus la zone commune).

En conséquence si l'utilisateur d'HYPERBASE est l'administrateur de la machine, HYPERBASE fonctionnera sans difficulté. Dans le cas contraire les droits d'écriture dans les deux répertoires intéressés doivent lui être alloués une fois pour toutes, en utilisant les commandes DOS ou Windows du système, ce que fait automatiquement le programme d'installation INSTALL.bat.

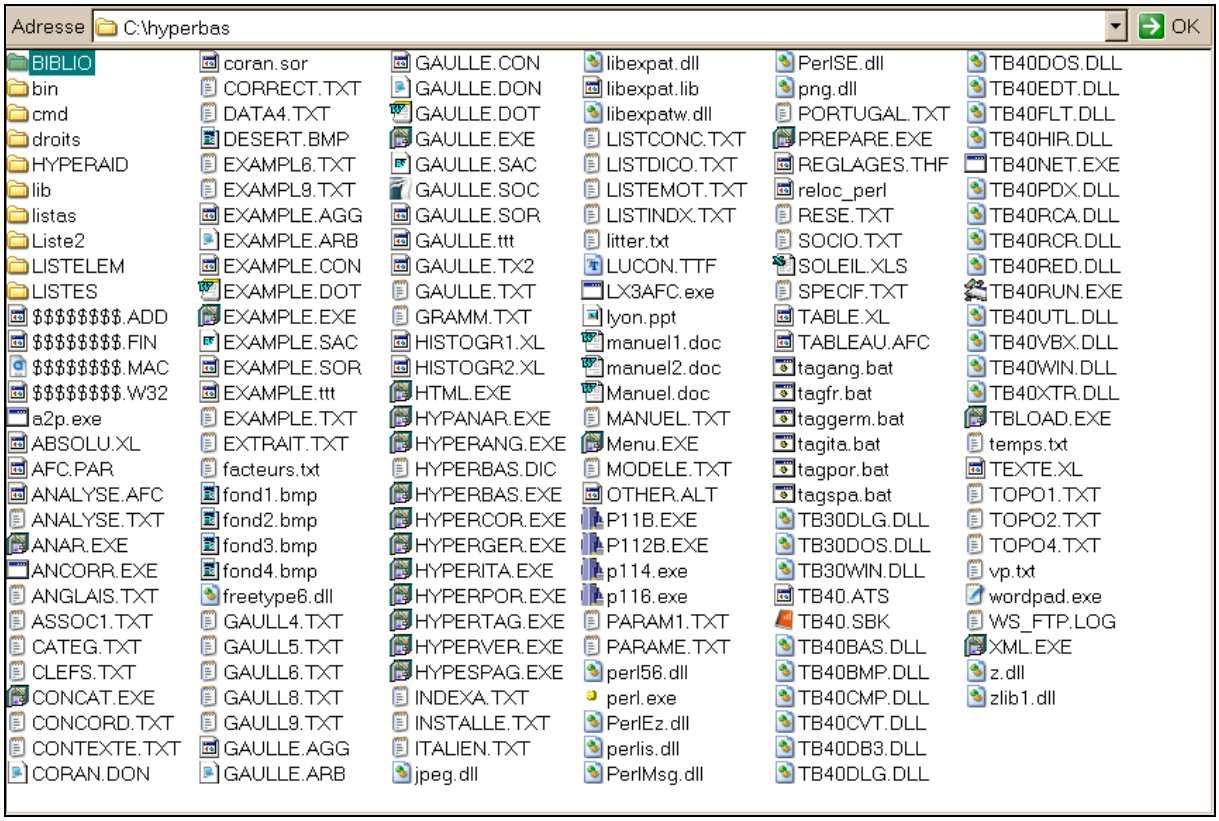
Une fois la copie réalisée, d'une ou d'autre façon, l'utilisateur est conduit sur un menu qui propose un grand choix de bases déjà disponibles. Quand la maîtrise du logiciel est suffisante, ces données laisseront la place à celles de l'utilisateur - ce qui se fait en sollicitant la base modèle HYPERBAS.TBK (ou quelque autre modèle dont le nom commence par HYPER). On est invité à donner un nom à la copie qui sera faite du modèle. La copie est alors lancée

automatiquement, montrant un bouton CREATION qui déclenchera le processus. Il y aura lieu, ultérieurement et une fois pour toutes, de compléter le paramétrage en sollicitant le bouton INSTALLE du menu principal, pour faire connaître l'emplacement des programmes externes: tableur, navigateur et éditeur.

Le menu

L'application s'accommode des différentes versions de Windows (3.xx, 95, 98 NT, Millenium, 2000, XP, VISTA, Système 7). Elle fonctionne sur toute la gamme, même sur les machines peu puissantes. Elle réclame peu de ressources. Mais la phase d'indexation et de préparation des données (qui heureusement n'a lieu qu'une fois pour un jeu de données) exige du temps, les programmes utilisant les fichiers plutôt que la mémoire centrale.

On trouvera ci-dessous les fichiers (bibliothèque, données ou programmes) qui doivent ou peuvent apparaître dans le répertoire HYPERBAS, quand le transfert est réalisé. Ces fichiers sont ceux de la version 8. Leur nom ne change pas dans la version 9 mais le suffixe .TBK est substitué au suffixe .EXE, et les routines de Toolbook apparaissent avec l'initiale TB100 et non plus TB40.



En réalité d'autres fichiers intermédiaires viendront s'ajouter, au gré de l'utilisateur, dès que la base sera exploitée ou renouvelée. En principe quelques noms sont réservés:

- EXTRAIT.txt pour recevoir les résultats que l'utilisateur a désignés au cours de la session comme devant être retenus. Ce fichier est au format ASCII (ou format TEXTE). Il est destiné à un éditeur ou traitement de texte.

- HISTOGR1.xl, HISTOGR2.xl, TEXTE.xl, TABLE.xl pour enregistrer les données destinées à un tableur. Le format est celui d'EXCEL. On a prévu l'emploi de la version française (avec la virgule tenant lieu de point décimal) et de la version internationale (qui maintient le point décimal dans les données numériques). Les quatre fichiers correspondent à quatre formats:

  - représentation graphique d'une série ligne (distribution d'un mot à travers les textes)
  - représentation graphique de deux séries-ligne (deux mots représentés conjointement dans les textes)
  - représentation graphique d'une colonne (profil d'un texte dessiné à travers une série de mots)
  - représentation graphique d'un tableau à deux dimensions (i mots en ligne et j textes en colonne)

Suivant les possibilités du tableur la représentation graphique peut prendre des formes diverses en deux ou trois dimensions (histogrammes, courbes, aires, secteurs, etc...)

- AFC.par, TABLEAU.afc, ANALYSE.afc, accompagnent le programme d'analyse factorielle ANCORR.exe. Un autre programme d'analyse factorielle LTX2AFC.exe est aussi fourni et s'entoure de deux fichiers d'entrée et de sortie : CORAN.don et CORAN.sor. Voir explications plus loin.

Veiller à ne modifier ni le nom, ni l'emplacement des fichiers ou dossiers essentiels à l'indexation ou à l'exploitation, en particulier MODELE.txt qui sert de référence pour le calcul des spécificités (ANGLAIS.txt, ITALIEN.txt et PORTUGAL.TXT servent pareillement de modèle pour les textes anglais, italiens ou portugais), et toute la bibliothèque propre à l'application TOOLBOOK, sur laquelle la présente base est établie. Le "Runtime" essentiel porte le nom de TB40RUN.exe et tous les programmes accessoires dont il a besoin ont un nom qui commencent par les mêmes initiales. Ne rien changer à cette bibliothèque. Par contre le nom de la pile ou des copies qu'on en fait peut et doit être changé (mais leur maintien à l'intérieur du même répertoire HYPERBAS est conseillé). Prendre garde que toute base que l'on crée est accompagnée d'un dictionnaire qui porte le même nom, accompagné du suffixe .TXT, et qui est constitué automatiquement au moment de l'indexation. Si l'on déplace la base, ou si on en modifie le nom, ne pas omettre de transmettre les mêmes mouvements ou les mêmes modifications au dictionnaire associé (et à quelques fichiers accessoires ayant le même nom). Dans la version lemmatisée, d'autres fichiers accompagnent la base, en lui empruntant son nom auquel divers suffixes sont ajoutés : .txt, .tx2, .cor, .cnr, .sor.

AIDE GÉNÉRALE

Dès que le curseur est dans le voisinage d'un bouton, une aide succincte apparaît au bas de l'écran pour indiquer la fonction de ce bouton. L'aide est beaucoup plus détaillée si l'on sollicite le bouton spécialisé AIDE. Elle correspond à l'essentiel du présent mode d'emploi. Un menu général est présenté d'abord, qui propose un choix de rubriques explicatives et qui permet de s'informer sur une action particulière. Si l'on choisit par un clic un des mots-clés du sommaire, on est renvoyé au paragraphe qui traite de la question. Cette aide en ligne peut être dissociée de la base: on la consulte de façon indépendante avec le lecteur *d'Acrobat*). Quant au présent manuel de référence, il se trouve aussi sur le DVD sous le nom de *Manuel.doc*.

AIDE PARTICULIÈRE À CHAQUE BOUTON

On peut obtenir une aide appropriée à certains boutons du menu principal en pressant la touche MAJUSCULE pendant qu'on clique sur le bouton qui fait problème. Suivant qu'on active ou non le bouton ANTIBULLES l'information est délivrée ou non, dès que la souris approche d'un bouton.

LE MENU PRINCIPAL

Le menu principal est le carrefour où l'on revient généralement après une opération. C'est là qu'on est introduit quand on ouvre la base. C'est là qu'on est conduit quand on sollicite le bouton SOMMAIRE, qui apparaît sur toutes les cartes (sous la forme d'une flèche verte coudée). C'est là que mène l'ordre spécial (appui simultané sur les touches ALT et flèche haut) que le langage a prévu pour faire apparaître la première page de la base.

Le menu principal

Avant d'expliquer le détail des actions qui peuvent être menées à l'intérieur de la base ainsi ouverte, on doit insister sur une nouvelle fonction qui donne accès non plus seulement à la base active, mais aussi à toutes celles qui ont été précédemment créées, dès lors que leur nom et leur adresse sont connus. Pour établir ces liens on doit activer le bouton du haut de l'écran *Choix des bases*. Puis pour passer d'une base à l'autre on déroulera le menu situé à droite du précédent bouton, en choisissant l'item correspondant à la base désirée. Le bouton *Choix des bases* permet les ajouts, les suppressions et la remise à zéro. Il n'est pas nécessaire (mais c'est recommandé) que toutes les bases disponibles soient situées dans le répertoire HYPERBAS, pourvu que le fichier qui tient à jour les adresses (il a pour nom OTHER.ALT) s'y trouve à demeure. Pour ne pas encombrer la mémoire, une seule base est activée à la fois. Mais le passage de l'une à l'autre est instantané.Le menu principal propose deux sortes de fonctions:

- les fonctions documentaires proposées horizontalement, au haut de l'écran
- les fonctions statistiques placées verticalement, à gauche de l'écran

À cela s'ajoutent quelques commandes de contrôle:

1 - les fonctions de **navigation**, relatives aux opérations de routine, qu'on trouve dans toute application: Sortie (bouton QUITTER), Retour à la page qu'on vient de quitter (bouton RETOUR). Une dérivation conduit du côté d'INTERNET. Une autre ouvre une base vide (ANAR.EXE ou HYPANAR.EXE) dans lesquels les calculs statistiques peuvent être réalisée à partir de tableaux de nombres. Ce sont les mêmes fonctions de calcul qu'on trouve dans HYPERBASE (histogrammes et analyses multidimensionnelles), mais elles s'appliquent à des données extérieures, issues ou non d'un corpus.

2 - les fonctions de **réglage**. Outre le bouton INSTALLER dont on vient d'expliquer l'emploi (on ne le sollicite généralement qu'une seule fois), on peut parfois solliciter le bouton INITIALISER, lorsqu'on désire modifier les titres qu'on a donnés aux textes du corpus. Attention! Cette action modifie les titres dans toute la base et déclenche la sauvegarde des modifications. D'autres réglages, provisoires, intéressent

- soit l'affichage, restreint ou large (bouton OPTION ÉCRAN),
- soit le mode de calcul, selon la loi normale ou le modèle hypergéométrique (bouton OPTION CALCUL),
- soit la façon dont on souhaite traiter les fins de ligne, lorsqu'on a affaire à des textes versifiés (bouton OPTION VERS), ou lorsque les fins de ligne ont été confondues avec les fins de paragraphe, le même retour de chariot ayant été utilisé pour ces deux balisages. Voir explications plus loin

3 - les fonctions d'**information:** Copyright (bouton en forme de timbre-poste) et AIDE.

4 - les fonctions de **sélection**, qui permettent de choisir la base à exploiter, comme indiqué précédemment, mais aussi de créer une nouvelle base avec d'autres données.

LE CORPUS DE TRAVAIL

Dans la base en cours d'exploitation, on peut aussi choisir les textes à explorer, en ignorant provisoirement les autres (bouton CHOIX DU CORPUS). La présente version offre en effet la possibilité de **choisir un corpus de travail** parmi les textes disponibles dans la base active. Cette fonction, peu utile lorsqu'un chercheur construit une base réduite (le choix des textes se fait avantageusement au moment de la saisie et du traitement initial), s'avère indispensable lorsqu'un corpus prend de l'ampleur et qu'il est proposé à des tiers. Un peu de liberté et de souplesse dans l'exploitation doit compenser la rigidité des choix imposés dans la phase de réalisation. L'étude globale reste pourtant possible et c'est même l'option par défaut, pour les fonctions quanti-

tatives comme pour les programmes documentaires. À l'opposé la sélection d'un texte unique parmi tous ceux qui constituent le corpus n'est nullement interdite ou déconseillée. Mais dans ce cas les comparaisons, qui sont à la base de la statistique, faisant défaut, seules les fonctions de recherche seront justifiées.

La sélection d'un corpus de travail s'opère par le bouton *Choix du corpus*. La liste des textes se présente alors au clic de la souris (les textes retenus passent dans le champ de droite), jusqu'à ce que l'utilisateur estime son panier suffisamment rempli (en sollicitant le bouton *Terminer*). Pour chaque sélection on est averti du nombre de de mots recueillis.

Choix du corpus de travail

Conformément à la sélection mise en oeuvre, le bouton *Lecture* donne accès au texte choisi parmi ceux du corpus courant (c'est-à-dire limité aux textes choisis). La lecture suivie, page après page, est possible grâce aux flèches de navigation, mais pour ce rôle traditionnel le papier offre un confort supérieur. L'écran prend l'avantage dans ses f**onctions hypertextuelles**: il suffit de cliquer sur un mot pour connaître sa répartition dans le corpus de travail et être conduit dans les passages où le mot se trouve employé. Ces excursions verticales peuvent se faire aussi à partir de l'index, c'est-à-dire de la liste alphabétique à laquelle un menu déroulant donne accès dans la page d'entrée (bouton *Index*). De la même façon, les autres programmes documentaires (*Concordance* et *Contexte*) ainsi que la plupart des programmes statistiques (répartis sur la gauche

de l'écran principal) s'appliqueront seulement au corpus de travail, en ignorant les textes non sélectionnés.

LE GENRE

La sélection du corpus de travail se fait sur les textes du corpus. Mais entre le corpus et les textes il peut y avoir un stade intermédiaire, qu'on a appelé genre, quoique le principe de regroupement puisse être commandé par d'autres critères que le genre: l'époque, le lieu, la signature, le registre…Cette division est facultative. Elle est activée par le bouton REINITIALISER du menu principal, qui propose une colonne GENRE pour tous les textes du corpus. On remplit cette colonne avec la rubrique qui convient à chaque ligne (par exemple roman, correspondance, thèâtre), en laissant vides les cases où aucune rubrique n'est souhaitée. Quand la grille est remplie, solliciter de nouveau le bouton REINITIALISER en indiquant qu'il n'y a pas d'autre modification à faire.

Le bouton REINITIALISER

La partition en "genres" peut se faire au moment de la réalisation de la base. Mais elle peut se faire en différé, à tout moment. On peut aussi la modifier et la compléter, comme aussi le titre des textes, en activant le même bouton REINITIALISER, qui pour Baudelaire propose l'écran qui précède.

L'exploitation de cette partition superposée aux textes se fait dans les programmes graphiques et multidimensionnels, soit que les genres y soient substitués aux textes, soit qu'ils imposent une couleur spécifique aux textes qu'ils regroupent.

CHAPITRE 2

La préparation

Ce chapitre s'adresse à ceux qui veulent créer une nouvelle base. Les lecteurs qui veulent seulement s'assurer la prise en main du logiciel peuvent l'ignorer provisoirement.

PRÉPARATION DE LA BASE

PRÉSENTATION DES DONNÉES

Dans les versions précédentes du logiciel standard HYPERBAS, un formatage minimum était requis, afin que le traitement puisse reconnaître le commencement d'une partie ou l'entrée dans une nouvelle page. Mais ces dispositions, quoique peu contraignantes, ne sont plus indispensables dans la version 9.

On peut proposer désormais à HYPERBAS un fichier de données textuelles dans son état originel, sans le moindre apprêt. Il suffit qu'il soit au format ASCII (ou « texte seulement ») et qu'il ait le suffixe .TXT. Si un seul fichier est ainsi proposé, il sera partitionné automatiquement en 9 parties. Si d'autres fichiers s'ajoutent à ce premier texte, le corpus les considérera comme autant de textes, constitutifs du corpus. Le nom du ces fichiers sera retenu pour désigner les textes.

Si le texte à traiter a le suffixe .DOC ou .PDF, une conversion est nécessaire, que le traitement de texte WORD accomplit aisément. Il est un cas cependant qui peut mettre HYPERBAS en difficulté : quand le suffixe est bien ce qu'il doit être et qu'il cache un texte écrit dans le format UNICODE. Les caractères accentuées du français sont alors transcrits sur deux octets et cela perturbe les tris et les opérateurs de comparaison alphabétique. Un test est effectué par HYPERBAS pour rejeter un tel codage et inviter l'utilisateur à convertir le format unicode (ou UTF-8) en Windows occidental (ce que Word accomplit sans problème).

Néanmoins les utilisateurs des anciennes versions peuvent garder leurs habitudes et préparer leurs données selon le format ancien, indiqué ci-dessous. La version 9 d'Hyperbase continue à reconnaître ce format, sans l'exiger (il n'était d'ailleurs plus en usage dans les versions lemmatisées).

ANCIEN FORMATAGE, MAINTENANT FACULTATIF

Les données textuelles doivent se trouver dans un fichier ASCII (ou "texte seulement"). On a pris en compte la plupart des alphabets européens. Aucun formatage particulier n'est obligatoire, le logiciel se chargeant de la pagination et de la partition, si elles sont absentes du fichier. En ce cas les cartes (ou pages) ont environ 200 mots et l'ensemble du texte est découpé en parties de longueur voisine, dont le nombre est à la discrétion de l'utilisateur (si le texte est de proportion restreinte, un dialogue propose 9 parties, s'il est plus long, 24 parties, mais toutes les segmentations de 2 à 75 sont acceptées). L'utilisateur est averti des dispositions requises et du résultat de l'expertise effectuée sur les données proposées. Si les conventions adoptées lui conviennent, il peut poursuivre le traitement ou l'interrompre, dans le cas contraire, pour corriger d'abord le fichier. Noter que l'expertise n'envisage que le début du fichier et ne garantit pas l'homogénéité et la fiabilité des données. L'erreur la plus commune est de laisser croire que les pages sont partout numérotées et partout précédées du symbole $, alors que ce code a été oublié à certains endroits du fichier. En une telle situation, le programme risque d'empiler les paragraphes les uns sur les autres, en attendant vainement la page suivante, et la saturation de la mémoire pourrait produire une erreur. Tout aussi grave est l'incohérence à l'endroit de la partition. S'il trouve une première partie, le logiciel s'attend à en trouver d'autres. En réalité le programme a prévu ces défauts graves et s'en accommode au mieux. Mais c'est au prix d'approximations qui ne seront pas toujours satisfaisantes. Les pages ou les parties dont le signalement aura été oublié dans les données seront bien intégrées à la base, mais elles seront annexées aux pages ou parties précédentes et porteront le même numéro.

Mais il vaut mieux suivre le découpage naturel des données, s'il existe. Deux conventions doivent alors être respectées:

1 - les parties doivent être précédées d'une ligne où l'on indiquera le titre ou les titres en utilisant, devant et derrière, le symbole composite &&& (sans blanc). Veiller à bien choisir le dernier mot du titre qui doit être unique et distinctif et qui sert d'abréviation lorsque la place manque, par exemple dans les graphiques.

Exemple: *&&&La vie en rose&&&*

On peut aussi proposer plusieurs titres pour un même texte, à condition de les séparer par une virgule (ce qui exclut la virgule dans le titre même). Le premier servira de référence explicite; le second, réduit à un mot, sera utilisé dans la légende des graphiques, là où la place manque pour un titre complet; le

troisième enfin, qui ne dispose que de deux lettres, indiquera le code qui symbolise chaque texte dans les concordances.

Exemple: *&&&La vie en rose, Rose, VR&&&*

Il arrive souvent que le choix des titres soit sujet à remords, une fois la base créée. Nul besoin de procéder à une nouvelle indexation, pour modifier les titres. Le bouton REINITIALISER du menu principal autorise les retouches, qu'il s'agisse des titres longs ou courts, ou des codes.

2 - les pages peuvent être expressément indiquées en ajoutant une ligne (au début de la page) et en y portant le numéro, immédiatement précédé d'un code spécial: le symbole $. Veiller à faire disparaître le symbole $, si ce code apparaît dans le texte même. En règle générale, les jalons de page alourdissent inutilement la phase de préparation. On s'en dispensera le plus souvent, en laissant au programme le soin de formater les pages à sa convenance, sans couper les phrases en fin de page.

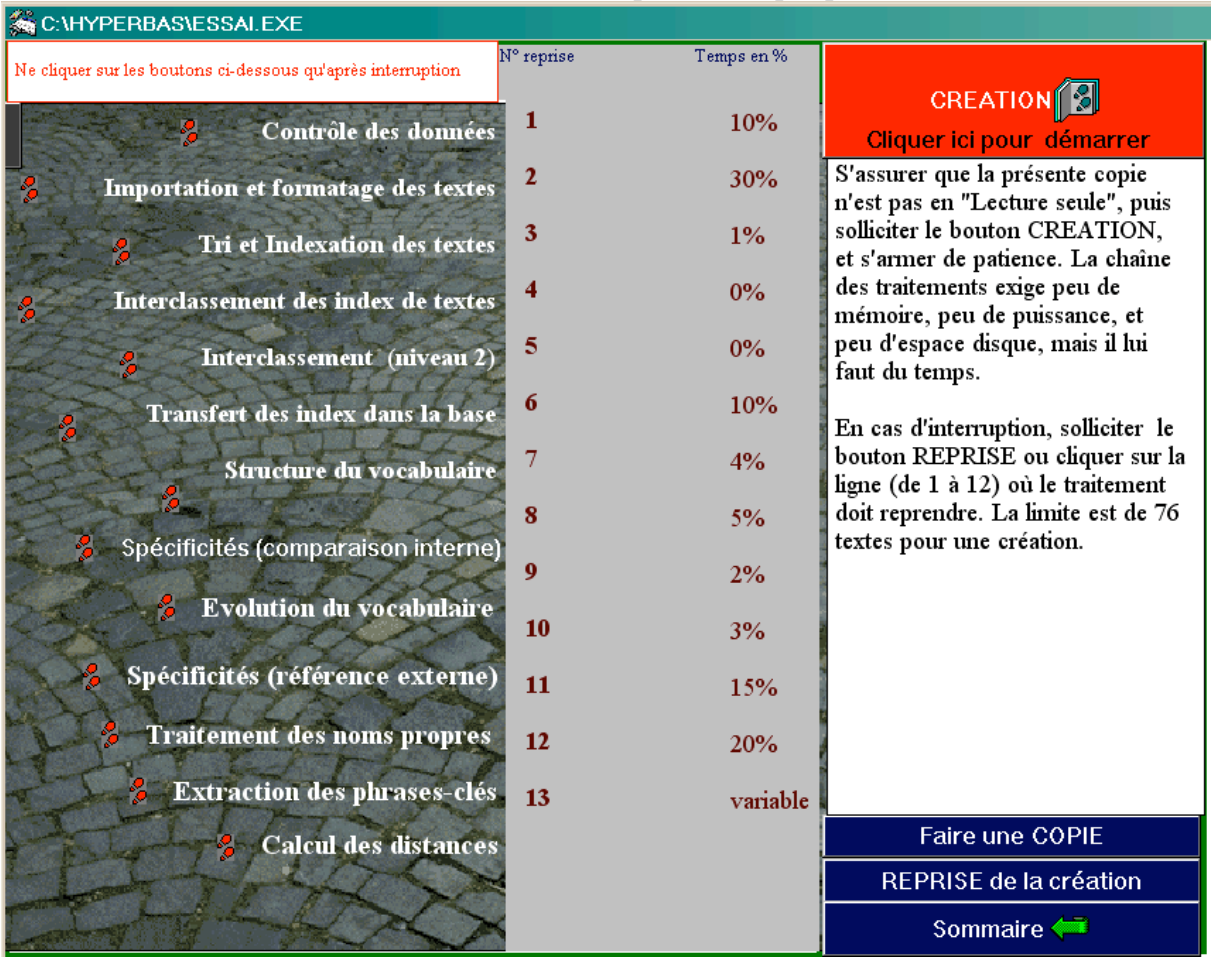
Exemple détaillé:	Exemple simplifié (recommandé) :
&&&La vie en rose, Rose, VR&&& $1 texte de la page 1 $2 texte de la page 2, etc. &&&Le travail au noir, Noir, TN&&& $62 texte de la page 62 $63 texte de la page 63, etc.	&&&La vie en rose, Rose, VR&&& texte en continu &&&Le travail au noir, Noir, TN&&& texte en continu

Veiller également à faire disparaître le code métalinguistique &&& s'il apparaît dans le texte même. Prendre garde également à consacrer une ligne entière à ces indications hors texte. Ne jamais les fondre à l'intérieur d'une ligne ou d'un paragraphe.

LE PROGRAMME ***CREATION***

Quand le fichier des données est dûment formaté et contrôlé, on crée une base nouvelle en lançant le programme standard HYPERBAS.EXE (ce sont d'autres noms dans les versions lemmatisées). Comme le fichier HYPERBAS.EXE est un modèle qui ne doit jamais être modifié, un dialogue propose qu'il en soit fait une copie utilisable sous le nom choisi par l'utilisateur. Cette copie est alors activée et présente l'écran ci-dessous.

L'écran au moment de la phase de préparation



Le lancement du programme CREATION (bouton rouge en haut de l'écran) fait apparaître un premier dialogue qui demande des précisions sur le nom du fichier à traiter (il est recommandé de le déposer dans le répertoire HYPERBAS). On s'attend à ce que ce fichier soit de type ASCII (suffixe .TXT), forme sous laquelle tous les traitements de texte peuvent présenter les données textuelles. Comme le type ASCII peut recouvrir diverses options selon les pays et selon les traitements de texte, il vaut mieux parler de la norme ANSI qui est non ambiguë et désigne partout le même code. Cette norme toutefois n'est rigoureuse et stable que pour les 128 premières places de l'alphabet, les lettres accentuées se situant au delà. Vérifier au préalable, par exemple avec l'éditeur Wordpad de Windows, que les choses sont bien en place. Si l'on travaille avec Word Microsoft se méfier de l'enregistrement dit "normal" qui non seulement conserve les codes typographiques mais aussi multiplie des codes de contrôle pour la gestion interne du document, spécialement au début et à la fin des fichiers. Toujours choisir le format "texte seulement".

Mais le format ASCII ne préjuge pas de la segmentation du corpus: en mots, en lignes, en paragraphes, en pages, en textes distincts. D'où quelques explications qui précisent les options retenues:

1 - Les pages

S'il existe un code de début de page le programme s'attend à trouver à cet endroit un numéro de page (si ce numéro n'existe pas ou si sa valeur n'est pas numérique, la numérotation de la page sera faite à partir de la page précédente, ou à partir de 1 si l'absence de numéro est systématique).

Si les pages (ou toute autre segmentation de même type) n'ont pas été distinguées, le programme procède de lui-même au découpage des pages à raison de 200 mots au moins par page (en s'abstenant de couper les phrases).

2 - Les paragraphes

Les paragraphes sont délimités par le retour de chariot (en réalité un code double: CR (ascii n° 13) suivi de LF (ascii n° 10). Si les paragraphes ainsi définis sont trop longs, le programme permet de les découper en unités plus petites (en s'abstenant de couper les phrases). Cette unité de segmentation doit nécessairement exister, sans quoi le programme d'importation, qui s'appuie sur ces délimiteurs, aura un fonctionnement perturbé.

3 - Les phrases

La segmentation en phrases ne trouve à s'appliquer que lorsqu'un paragraphe est jugé trop long, comme expliqué plus haut. Pour distinguer les phrases, le programme prend appui sur une ponctuation forte, principalement le point.

4 - Les lignes

La segmentation en lignes est rarement pertinente, sauf en poésie versifiée. Le programme n'en tient pas compte. Prendre garde à certaines options proposées par les scanners, qui maintiennent la mise en page originale des documents - ce qui est louable - mais confondent dans le même signe fins de ligne et fins de paragraphes - ce qui est moins heureux. Si les données sont de ce type, on aura intérêt à supprimer les retours de chariot intempestifs, en veillant en outre à recoller les mots coupés en fin de ligne. Cependant, si cette précaution n'a pas été prise, on peut rectifier le tir en fin de traitement, en sollicitant le bouton "OPTION VERS" du menu principal. Les retours de chariot qui finissent une ligne sans finir une phrase auront alors le code Lf seul (simple fin de ligne), tandis que les autres auront le code plein CrLf qui marque la rupture de paragraphe.

5 - Les mots

Le découpage des mots et leur classement obéissent aux exigences du français et plus généralement des langues pourvues de diacritiques. Les mots accentués prennent place au rang qu'ils ont dans le dictionnaire. La distinction

entre majuscule et minuscule est abolie dans les classements et dans les recherches ultérieures, mais non pas dans le texte, lequel restitue fidèlement la différence, ni même dans le dictionnaire, où la distinction est maintenue pour faire apparaître les noms propres. Comme cette décantation est entièrement automatique, elle est sujette à quelques bévues, notamment lorsqu'un mot du vocabulaire commun apparaît en tête de phrase, avec la majuscule, et qu'on ne le trouve nulle part ailleurs doté d'une minuscule. La définition des mots est dépendante de la liste établie pour les séparateurs, laquelle, outre le blanc, la tabulation et le retour de chariot, comprend les symboles suivants :

, . ; : ? ! " ' ( ) < > - — + / = { } [ ] …

Aucun de ces symboles n'est admis à l'intérieur d'un mot (à l'exception de l'apostrophe en position finale). Un blanc n'est pas significatif s'il accompagne un séparateur ou s'il est redoublé. Comme les guillemets appartiennent au métalangage, on leur a substitué un code non ambigu qui apparaît en position haute (“) et qui correspond au caractère ascii 148.

Quand le programme de préparation est lancé, il poursuit son cours jusqu'à la fin sans que l'opérateur ait beaucoup d'occasions d'intervenir. Mais l'utilisateur peut suivre sur l'écran le déroulement des opérations et attendre patiemment que la dernière étape arrive à son terme. Si les premières phases sont rapides (contrôle, importation et reformatage des données), l’étape suivante (indexation) comporte des opérations lourdes : un repérage des formes, puis un tri de ces formes suivi d'un tassement.

Cette phase était la plus longue dans la version précédente représentait la moitié du temps de traitement. Comme la mémoire des ordinateurs actuels est plus étendue que dans le passé, on a pris le risque d'indexer en une seule fois tout le corpus simultanément et non plus chaque texte isolément. Le gain en rapidité est considérable : le programme indexe maintenant un million de mots à la seconde sur une machine standard. Quant aux limitations d’étendue elles ont été repoussées aussi loin qu’on a pu (on a fait l'expérience d'un fichier de 30 millions de mots, sans rencontrer la limite).

L'algorithme de l'indexation est dû à Jean-Pierre Anfosso et extrait de sa thèse. Écrit en langage C, le programme P114.EXE est appelé au moment opportun. Pour assurer le synchronisme des opérations et poursuivre le traitement, l'utilisateur est invité à donner le signal de reprise, dès que l'indexation du corpus est achevée.

Limites à respecter

Il vaut mieux pour les traitements statistiques que les textes ne soient pas d'étendue trop inégale, même si des calculs de pondération redressent la perspective. On prendra garde aussi à ne pas accorder au même texte une place

trop exagérée. Il ne servirait à rien de fragmenter un corpus en textes indépendants si l'un de ces textes occupait la quasi totalité du corpus. Dans la pratique on évitera de dépasser la limite de 500 000 mots pour un même texte.

Dans sa version actuelle, le programme accepte 75 textes. La longueur de chacun des textes n'importe guère. Mais il est évidemment préférable que leur étendue, d'un texte à l'autre, ne soit pas trop déséquilibrée, quoique les calculs de pondération corrigent les différences de taille. Si les "textes" correspondent à des segments trop courts ou si leur nombre dépasse la limite 75, on procèdera à des regroupements, en satisfaisant non seulement aux contraintes du programme mais aussi aux impératifs méthodologiques de la statistique. Pour que les tests probabilistes puissent s'exercer, il faut en effet que la loi des grands nombres ait suffisamment d'espace pour se déployer - ce qui n'est pas le cas lorsqu'un texte n'a que quelques pages.

Il y a quelque avantage à ne pas trop morceler un corpus de taille étroite, afin que les sous-ensembles aient une étendue suffisante. Inversement, si l'on dispose d'un très vaste corpus, on aura des résultats plus fins en augmentant le nombre des parties.

Dernières phases

Le traitement s'emploie à comparer le corpus traité au corpus littéraire de FRANTEXT, qui comprend 117 millions de mots et s'étend sur cinq siècles. On a la liberté de choisir une époque limitée de ce corpus, afin de rapprocher les deux termes de la comparaison et de justifier, autant que faire se peut, ce recours à une référence externe. Mais on a aussi la possibilité de renoncer à ce traitement, s'il ne s'impose pas (et notamment lorsque le corpus étudié appartient à une autre langue que le français, l'anglais, l'italien ou le portugais). Dans le cas d'une langue étrangère, un dictionnaire de référence propre à la langue choisie peut être substitué au fichier MODELE.TXT, extrait du TLF. Pour l'établir on doit disposer d'une liste de fréquences rangée alphabétiquement, chaque ligne comprenant une forme précédée de sa fréquence. S'il s'agit de l'anglais, ce fichier de référence est déjà disponible, sous le nom ANGLAIS.TXT (il est extrait du British National Corpus qui compte 100 millions d'occurrences). Un fichier similaire, extrait du journal PUBLICO, est également proposé pour le portugais (PORTUGAL.TXT). Un autre semblablement constitué est prévu pour l'italien (ITALIEN.TXT). Pour ces trois langues, le choix de la référence externe se fait dans un dialogue, sans autre manipulation.

Les autres phases de la préparation sont généralement facultatives, notamment celle qui concerne le traitement des noms propres. Il s'agit de restituer la majuscule, dans le dictionnaire et les listes de résultats, aux vocables où cette propriété est attachée à la nature du mot et non à sa place en tête de phrase (dans le texte, les majuscules sont de toute façon respectées, où qu'elles se trouvent). Ce traitement est assez long, puisque les mots doivent être recherchés à l'intérieur de la phrase, pour vérifier si la majuscule est maintenue,

et il est incomplet car il reste le cas indécidable des mots qui ne se trouvent qu'en tête de phrase et pour lesquels l'ambiguïté demeure.

Quand le programme de préparation est terminé, on revient au menu principal et l'exploitation de la base peut commencer. Le programme prévoit cependant un rechargement préalable, afin de vérifier que tout est en place.

On peut verrouiller la base, ce qui prévient les accidents sans gêner l'exploitation. Le verrouillage sous Windows consiste à afficher les propriétés d'un fichier et à cocher l'option LECTURE SEULE. Une base transférée sur un support non réinscriptible, comme le cédérom, est par là même verrouillée et en lecture seule. Mais avant de verrouiller une base, d'une ou d'autre façon, il convient d'une part de la vider de résultats antérieurs et inutiles, notamment le contenu des champs des pages CONTEXTE, CONCORDANCE et LISTE, et d'autre part de la pourvoir des données complémentaires qui ne sont livrées que sur commande expresse (notamment les programmes LONGUEUR et GROUPE de la page LISTE). La façon la plus sûre de faire le ménage dans une base est d'utiliser le bouton VIDER du menu principal.

Au terme du parcours, divers traitements ont été exécutés sur lesquels on ne reviendra plus:

- la reconnaissance et le tri des formes
- l'indexation proprement dite
- le dictionnaire des fréquences, alphabétique et hiérarchique
- le calcul des spécificités, externes et internes
- les coefficients de corrélation
- le tableau de distribution des fréquences
- la mesure de la richesse lexicale, de l'accroissement du vocabulaire et de la proportion des hapax
- et divers tests statistiques qu'on détaillera par la suite.

Tous les résultats ont été communiqués à la base, dont l'exploration et l'exploitation peuvent désormais se faire de façon autonome sans recours aux fichiers externes.

CHAPITRE 3

L'exploration

EXPLORATION LIBRE

L'exploration a recours aux méthodes de l'hypertexte, de deux façons symétriques:

1 - NAVIGATION DANS LE DICTIONNAIRE

Dans une première démarche, en sollicitant le bouton MOT, on est renvoyé au dictionnaire des fréquences, à l'endroit où se trouve le mot proposé. On peut aussi utiliser le menu déroulant INDEX pour choisir la lettre de l'alphabet où le dictionnaire doit être ouvert. Une fois que le dictionnaire est sous les yeux, on peut parcourir la liste alphabétique en s'aidant de l'ascenseur et des flèches gauche et droite qui font défiler les pages. En cliquant sur une forme jugée intéressante, on voit apparaître tous les passages où cette forme est employée. Le mot repéré apparaît en vidéo inverse dans le texte et l'écran reste figé, en attendant un clic de la souris pour passer à l'occurrence suivante. On peut à tout moment interrompre la recherche en appuyant sur la touche ALT, qui arrête le défilement à la page courante ou, au choix, restitue à l'écran la page du dictionnaire où la forme avait d'abord été observée.

Chaque ligne du dictionnaire est consacrée à une forme (ou à un signe) et précise dans l'ordre:

a - la fréquence du mot dans le corpus

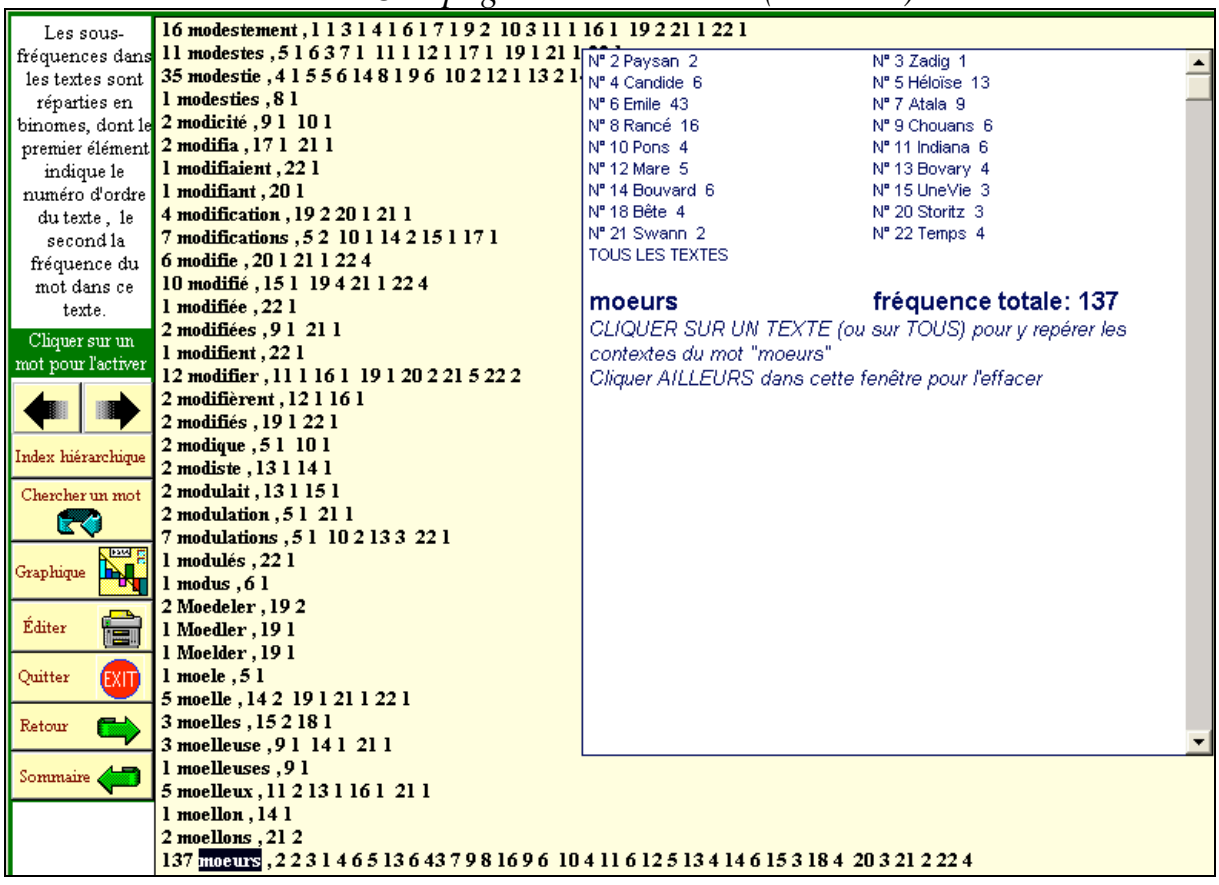
b - la forme elle-même

c - enfin, séparée du reste par une virgule, la liste des sous-fréquences de la forme dans les différentes parties du corpus. Noter que seules sont mentionnées les sous-fréquences non nulles, la lecture procédant par paires dont le premier membre indique le numéro d'ordre du texte concerné et le second la fréquence du mot dans ce texte. Ainsi la suite 3 6 5 2 8 3 signifie que le mot s'est rencontré 6 fois dans le troisième texte, 2 fois dans le cinquième et 3 fois dans le huitième. Ce mode d'enregistrement fait gagner de la place, même si le déchiffrement n'en est pas très aisé. Au reste ces indications ne s'adressent guère à

l'oeil humain, car un tableau plus explicite apparaît au premier plan, dès qu'on propose une forme particulière. L'exemple ci-dessous montre en clair (dans la zone bleue) la répartition des références du mot *mœurs* dans le sous-corpus considéré. Un clic sur l'un de ces textes permet de circonscrire la recherche des contextes à ce seul texte. En cliquant sur la ligne TOUS LES TEXTES, on déclenche la visualisation exhaustive (qu'on peut toutefois interrompre avec la touche ALT). Un clic dans toute autre partie de la fenêtre bleue la fait disparaître et annule la sélection, qui peut reprendre avec un autre mot de la liste ou être transférée à n'importe quelle page du dictionnaire (grâce au bouton *Chercher un mot*). Outre la navigation d'un mot à l'autre, d'une page à l'autre et du dictionnaire au texte, quelques actions sont disponibles:

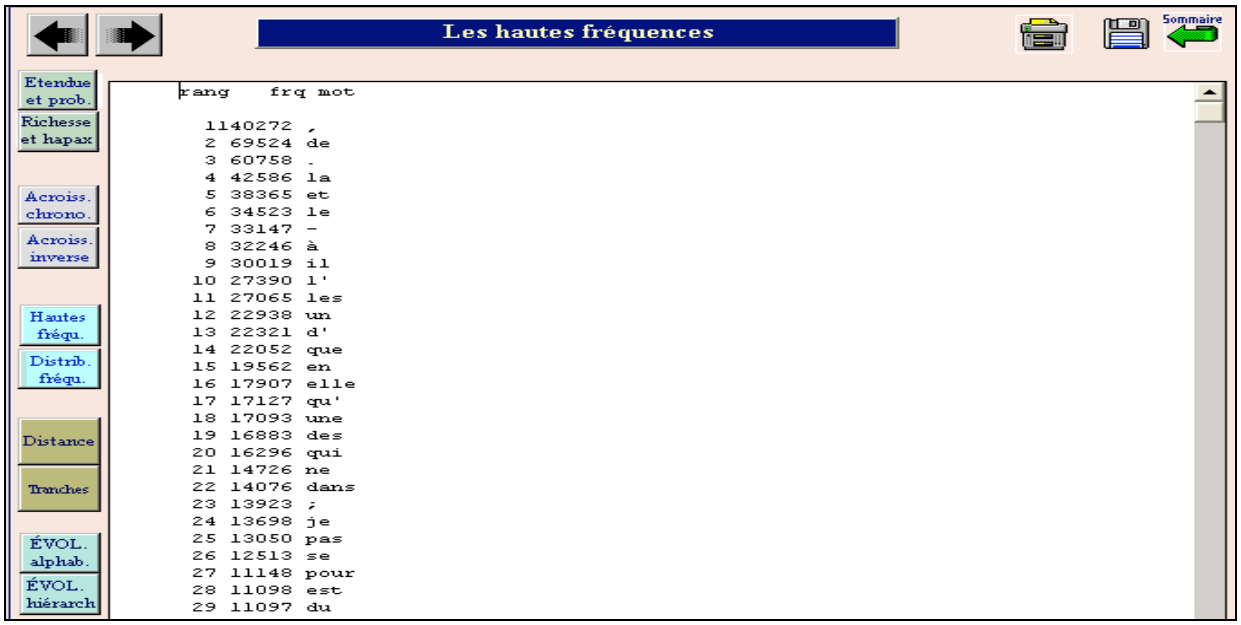
- l'impression de la page portée à l'écran
- la représentation graphique du mot sélectionné

Une page du dictionnaire (ou index)



Il existe une présentation hiérarchique du dictionnaire, par fréquences décroissantes, à laquelle on a accès par le bouton INDEX HIÉRARCHIQUE.

Le dictionnaire hiérarchique

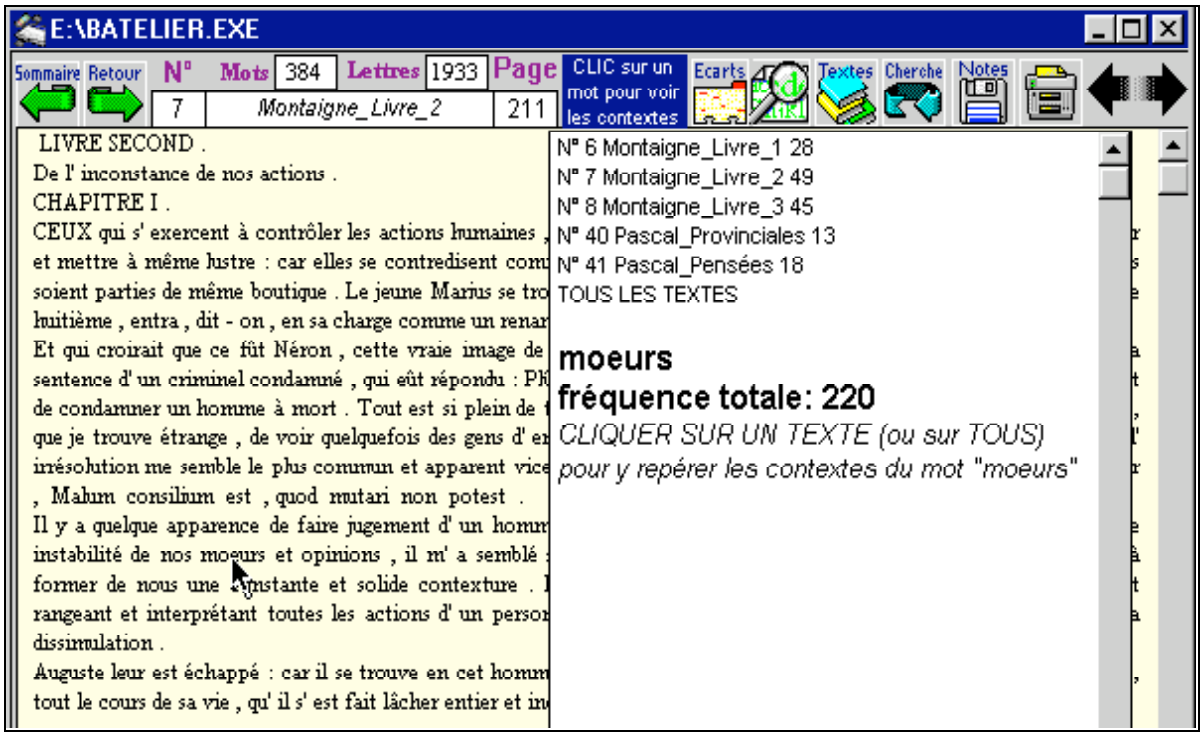


2 - NAVIGATION DANS LE TEXTE

La seconde démarche propre à l'hypertexte et symétrique de la précédente consiste à feuilleter les pages du texte (le bouton LECTURE permet d'ouvrir l'un des livres du corpus à la page qu'on veut). Dans cette exploration, on peut suivre la séquence des pages (en suivant les flèches DROITE ou GAUCHE) ou même rejoindre un autre texte du corpus (en sollicitant le bouton TEXTES). Deux boutons dont le symbolisme est clair sont prévus pour l'impression ou l'enregistrement de la page montrée à l'écran. L'enregistrement se fait dans un fichier au format *ascii* qui sert à empiler tout ce que l'utilisateur juge intéressant. Ce fichier, nommé EXTRAIT, est d'un type particulier, car il laisse voir le dernier enregistrement et propose le choix entre l'ajout ou l'effacement.

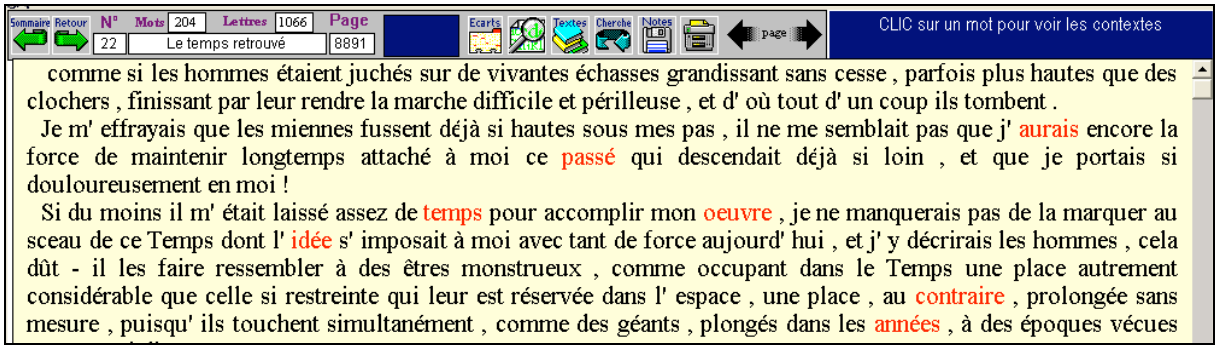
Pour une lecture plus confortable ou plus complète, on peut actionner le bouton LOUPE, qui grossit et rétrécit alternativement les caractères affichés à l'écran. Mais un hypertexte ne se réduit pas à la fonction lecture, dont le confort ne peut être comparé à celui du livre. Les fonctions proprement hypertextuelles sont celles de la recherche. Le bouton CHERCHE correspond à une recherche locale qui se limite à la page courante et montre en rouge les occurrences du mot proposé, s'il est présent dans la page. Une fonction plus puissante est attachée à chacun des mots de l'écran. Un clic sur l'un d'entre eux (dans l'exemple ci-dessous le curseur est pointé sur le mot *moeurs*) suffit à montrer sa fréquence et sa répartition dans le corpus de travail. Un second clic à l'intérieur du tableau de répartition déclenche la recherche et la visualisation des autres contextes où le même mot se rencontre dans le texte choisi ou dans l'ensemble du corpus, avant et après la page lue.

Une page du texte



Ces parcours peuvent être suspendus provisoirement (pression sur le bouton de la SOURIS) ou stoppés définitivement (pression simultanée sur la touche ALT et la SOURIS). A tout instant il est possible d'imprimer la page visible à l'écran.

Le voyage exploratoire dans les textes ne montre généralement que du texte. Les chiffres sont pourtant sous-jacents dans chaque page. Leur présence est révélée dès qu'on clique sur un mot, ce qui déclenche le tableau des fréquences pour le mot souligné. La statistique s'impose plus lourdement encore quand on sollicite le bouton *Écarts*, au haut de la page, qui provoque la **mise en relief** (en rouge) des mots qui, à l'intérieur de la page, reflètent les caractéristiques du texte considéré (par rapport au corpus d'ensemble). Ainsi sont soulignés dans la dernière page du *Temps retrouvé* les mots *longtemps, passé, temps, années, oeuvre*, qui sont récurrents dans le dernier roman de Proust (figure ci-dessous). On peut estimer gênante la mise en relief des mots grammaticaux, aussi sont-ils écartés s'ils ont moins de quatre lettres.



Les spécificités de la dernière page de Proust

3 - BIBLIOGRAPHIE

Le créateur d'une base textuelle doit pouvoir communiquer à l'utilisateur les éléments bibliographiques qui précisent et justifient le choix des textes et les éditions retenues. Ces informations qui concernent le corpus doivent rester extérieures au corpus et ne pas figurer dans le texte même. Elles seront recueillies dans un fichier qui doit porter le même nom que la base textuelle en question, le suffixe .DOC ou .INF étant substitué au suffixe .EXE. Il appartient au créateur de la base de créer un tel fichier avec le traitement de texte de son choix, en veillant à l'enregistrer dans le même répertoire que la base et à s'abstenir d'utiliser le suffixe .TXT (car un tel suffixe est déjà requis pour le fichier qui contient le dictionnaire de la base). Le menu principal de la base propose un bouton "BIBLIO" qui permet de consulter les informations enregistrées.

CHAPITRE 4

L'exploitation documentaire

LA FONCTION CONTEXTE

En dehors de la circulation libre à travers le texte et le dictionnaire, le logiciel propose dans le menu principal les outils propres à assurer une exploitation méthodique de la documentation. Les deux programmes essentiels CONCORDANCE et CONTEXTE obéissent aux mêmes principes et ne se distinguent que par la présentation des résultats:

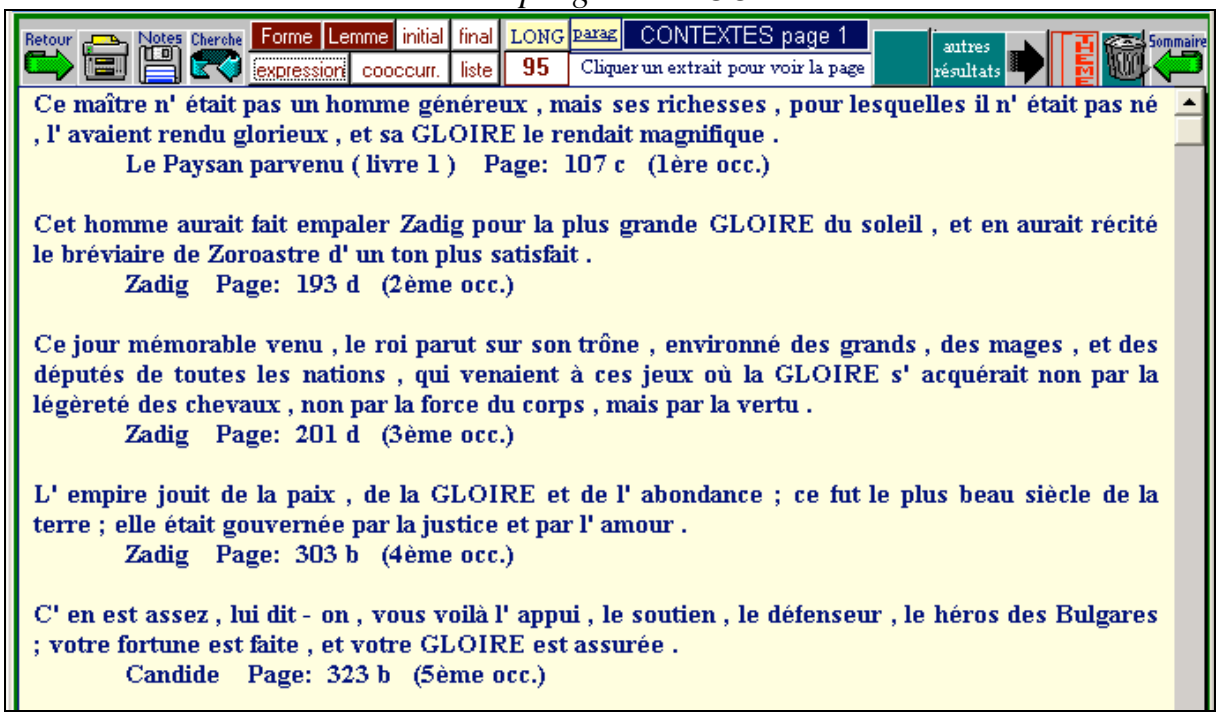
Si l'on met en oeuvre le bouton CONTEXTE (le résultat figure dans l'écran ci-dessous), chaque occurrence de ce qu'on cherche est située et montrée dans le contexte naturel du paragraphe. Pour permettre la reconnaissance aisée du mot dans le contexte, ce mot est converti en capitales dans le paragraphe où il est rencontré.

On peut refuser l'équivalence entre contexte et paragraphe. Si cette option par défaut ne convient pas, cliquer sur le bouton *longueur* et choisir la longueur désirée pour chacun des extraits, exprimée en nombre de caractères (de 50 à 1000). Noter que dans tous les cas on n'outrepassera pas la limite de cinq paragraphes consécutifs. Si l'extrait est raccourci, la contrainte de voisinage sera plus étroite et le bouton THEME (en rouge à droite dans la page CONTEXTE 1) subira ces contraintes où l'effet - souvent indésirable - de la syntaxe pourra être perçu. Dans tous les cas, le programme envisage la totalité des contextes (enregistrés dans le fichier CONTEXTE.TXT), et non pas seulement ceux qui sont affichés.

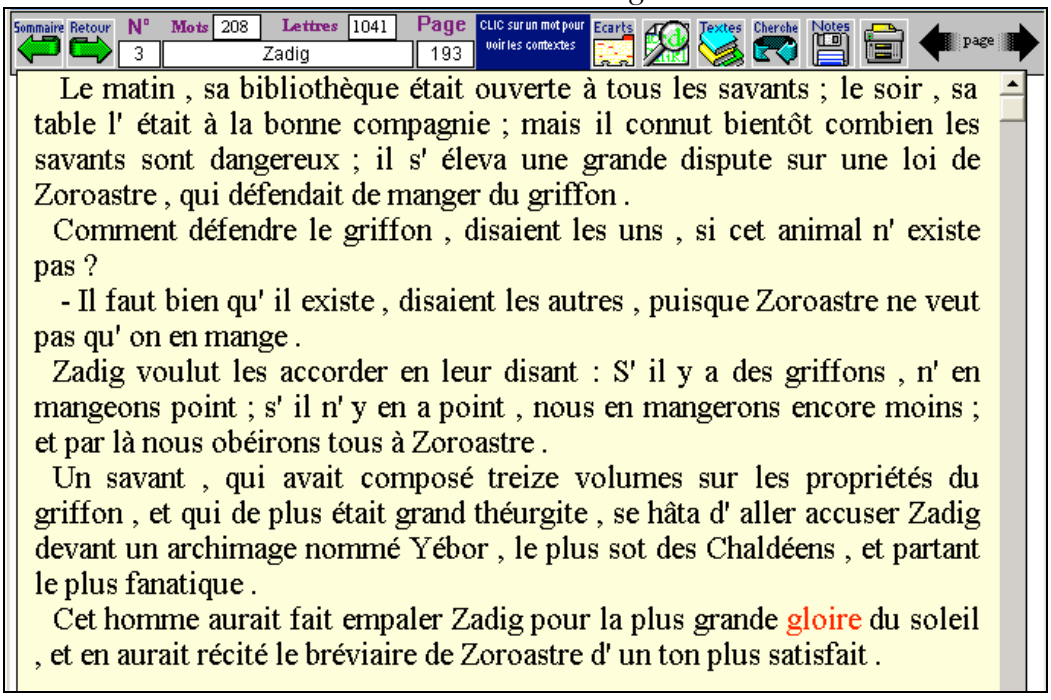
Quand le mot cherché a une fréquence élevée, les résultats deviennent vite encombrants. Ils sont alors répartis sur plusieurs pages auxquelles on accède par les flèches droite et gauche (en haut et à gauche de l'écran). Si ces flèches sont neutralisées, c'est que le terminus est atteint dans un sens ou dans l'autre. Dans tous les cas, les résultats affichés sont tronqués au bout de la sixième page, quand leur volume dépasse 200 000 caractères. Mais la recherche ne s'arrête pas pour autant et entrepose les résultats complets dans le fichier CONTEXTE.TXT. Là comme partout les deux boutons d'enregistrement et d'impression sont disponibles.

Le contexte restitué est généralement suffisamment explicite, d'autant que les références du passage sont livrées en clair, avec indication du texte, de la page, et de la zone dans la page (grâce à un code alphabétique qui commence à la lettre *a*, pour le début de page, et s'arrête à la lettre *f, g* ou *h* , pour la fin de page). Mais un lien est établi pour chaque extrait avec la page originale, où l'on est conduit instantanément lorsqu'on clique sur l'extrait en question. Ainsi le quatrième passage relevé dans l'exemple du mot *gloire* est développé dans une page du texte intégral, reproduite ci-dessous. Pour faciliter la localisation du passage, le mot cherché apparaît en rouge dans le texte.

Résultats du programme CONTEXTE



Renvoi au texte original

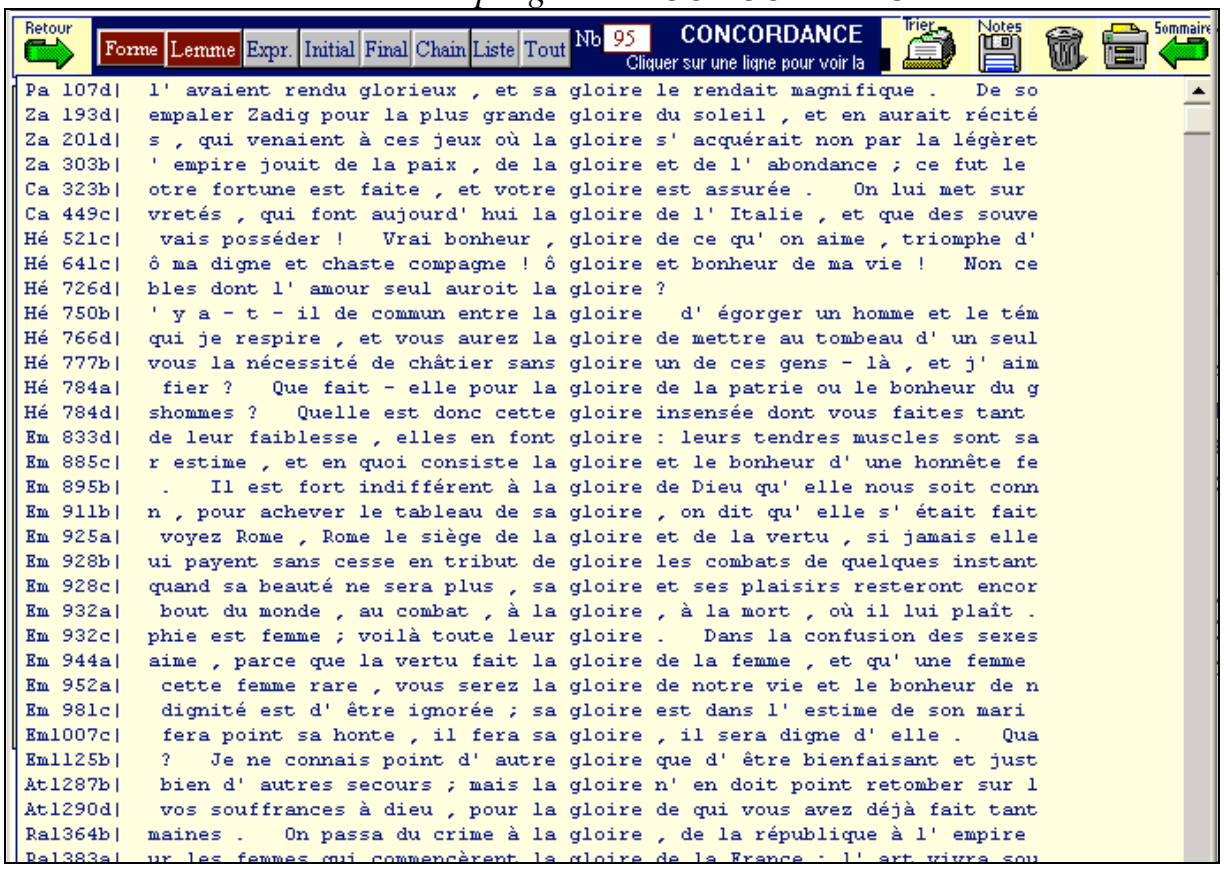


LA FONCTION CONCORDANCE

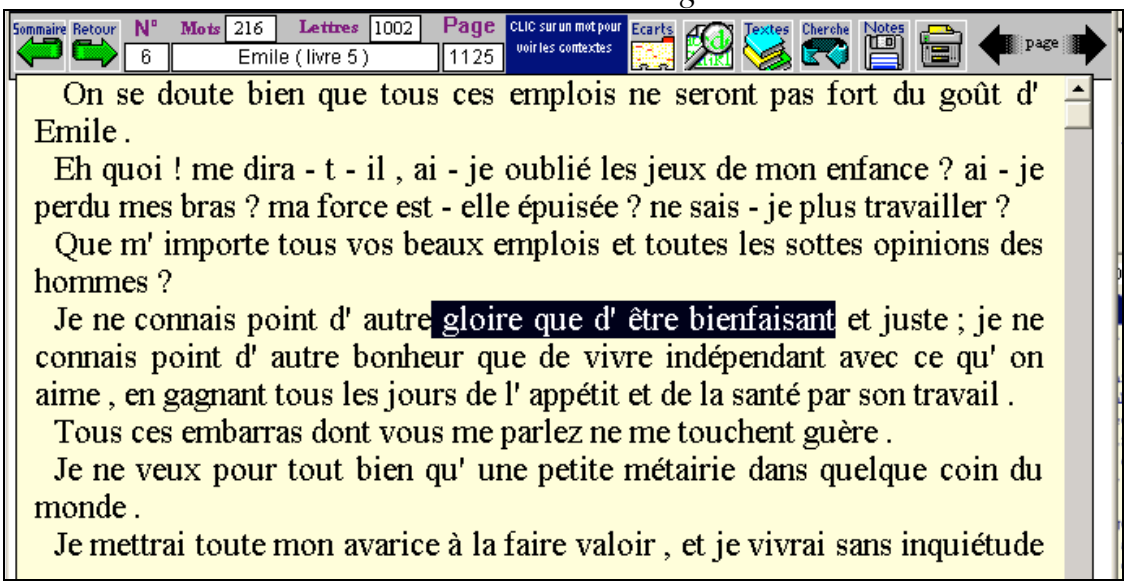
Si l'on fait appel à la fonction CONCORDANCE du menu principal (écran ci-dessous), on obtient un contexte étroit qui tient en une ligne et qui montre la forme (ou l'expression) cherchée, en position centrale, avec une demi-douzaine de mots à gauche et à droite. À côté des boutons habituels de navigation, d'enregistrement et d'impression, on remarquera une fonction de tri (bouton TRIER), qui est destinée à fournir une autre présentation des contextes. Au lieu de suivre l'ordre normal qui respecte la suite des textes, les contextes sont groupés selon leur environnement immédiat, à gauche ou à droite du mot-pôle. Cela souligne la résurgence de syntagmes répétitifs qui ressortissent souvent aux contraintes syntaxiques mais révèlent parfois aussi les tendances phraséologiques de l'auteur.

Si l'on estime trop étroite la fenêtre de concordance, un simple clic sur une ligne permet de retrouver la page concernée, qui reste exposée (avec mise en relief du mot) jusqu'au moment où un second clic la fait disparaître, comme on peut le constater ci-dessous pour le treizième exemple de la concordance du mot *gloire* dans le corpus *Exemple*. L'écran ne peut guère montrer que 200 lignes à la fois. Si la concordance obtenue dépasse cette limite, un bouton apparaît qui propose les 200 lignes suivantes, et, de proche en proche, tout le contenu du fichier CONCORDANCE.TXT où les résultats complets sont enregistrés.

Résultats du programme CONCORDANCE



Renvoi au texte intégral



Dans les deux procédures, CONTEXTE et CONCORDANCE, des options sont offertes à l'utilisateur pour qu'il puisse préciser la portée et l'objet de sa recherche.

Portée de la recherche

La recherche est limitée au corpus de travail (qui peut être coextensif au corpus total). Si le mot est fréquent, les résultats peuvent excéder la place disponible sur l'écran, même avec la ressource de l'ascenseur. En une telle situation, il convient soit de passer à l'écran suivant (bouton SUIVANT dans le programme CONCORDANCE, bonton noir en forme de flèche à droite dans le programme CONTEXTE), soit d'ouvrir les fichiers CONCORD.TXT ou CONTEXTE.TXT. qui contiennent les résultats exhaustifs, quelle qu'en soit la taille.

Objet de la recherche

La recherche peut porter non seulement sur une forme, mais aussi:

a - sur une EXPRESSION (une suite de mots ou de signes), par exemple *bout du tunnel*. Les signes de ponctuations sont acceptés à l'intérieur de ces expressions, à condition qu'ils soient précédés et suivis d'un blanc, cette règle s'appliquant aussi au trait d'union. Cas particulier: l'apostrophe, qui doit être suivie mais non précédée d'un blanc:

aujourd' hui ou *l' amour* qu'on opposera à *presse - bouton* ou *à la mode .*

Dans ce dernier cas, le point permet de choisir certaines clausules finales. Ces règles s'appliquent même si le texte d'origine obéit à des conventions différentes, car ce texte a été recomposé dans la phase de préparation, afin que le programme ne soit pas dépendant des particularités de la saisie.

b - sur un VOCABLE ou LEMME, par exemple toutes les formes conjuguées du verbe *aimer*. Dans la présente version l'option LEMME permet le

regroupement des formes derrière la forme canonique (ou lemme). Un dialogue demande qu'on précise s'il s'agit d'un verbe, d'un adjectif ou d'un substantif, ce qui ne donnera pas le même résultat quand on propose un homographe comme "bouche" ou "boucher" ou "bouchée". Les formes retenues figurent dans la page "LISTE" où des retouches sont possibles lorsque les formes homographes doivent être triées. Si la liste a subi des modifications, c'est l'option LISTE plutôt que l'option LEMME qu'il faut choisir. Dans certains corpus, le conjugueur est amené à tenir compte des graphies anciennes utilisées dans les siècles passés, mais sans remonter au-delà du XVIe siècle. Ainsi les formes en -*oi* peuvent être reconnues dans les désinences de l'imparfait ou du conditionnel, quand il s'agit d'un texte non modernisé, antérieur au XIXe siècle. Si on a affaire à un texte ancien et qu'on veuille retenir de telles graphies, appuyer sur la touche MAJUSCULE lorsque le programme demande qu'on choisisse entre les verbes, les substantifs et les adjectifs.

Prendre garde que la même forme peut appartenir à plusieurs lemmes. Si l'on veut éviter les mélanges indésirables, il suffit de rejoindre la carte LISTE et de cliquer sur les formes homographes qu'on veut écarter. Le bouton MODIF permet aussi des opérations d'ajout, d'élimination ou de regroupement. On notera qu'il ne s'agit pas ici de désambiguïsation, laquelle ne peut s'opérer que dans le texte, et non dans le dictionnaire. Des versions améliorées d'HYPERBASE (voir la deuxième partie de cet ouvrage) sont adaptées aux données désambiguïsées et lemmatisées. Elles sont associées aux logiciels d'étiquetage CORDIAL et TreeTagger.

c - sur les DÉBUTS DE MOT, ce qui résout en partie les problèmes de la lemmatisation ou du regroupement des formes. En fournissant la chaîne *aim*, on atteindra indirectement toutes les formes conjuguées du verbe *aimer*, mêlées il est vrai à des formes ou dérivés étrangers au paradigme proposé.

d - sur une CHAINE de caractères, où qu'elle se trouve dans un mot. Ainsi en recherchant la chaîne *thé*, on obtiendra non seulement la boisson anglaise, mais aussi *théâtre*, *athée* et *Léthé*.

e - sur les FINS DE MOT. Il s'agit là d'une forme particulière du cas précédent, un blanc étant ajouté - par le programme - à la chaîne recherchée. Ainsi en formulant une demande sous la forme *ment*, on filtrera tous les dérivés en *-ment* . Attention! il n'y a pas de catégorisation grammaticale dans ce programme simplifié, et le suffixe en *-ment* récoltera les adverbes (*sciemment)* aussi bien que les substantifs (*jugement*) et quelques verbes (*aiment*) ou adjectifs (*dément*). Si l'on veut filtrer finement les mots et les catégories (sans compter d'autres critères), c'est la version lemmatisée d'HYPERBASE qu'il faut employer.

f - sur les COOCCURRENCES (programme CONTEXTE). Seuls sont restitués les contextes qui contiennent la présence simultanée de deux formes

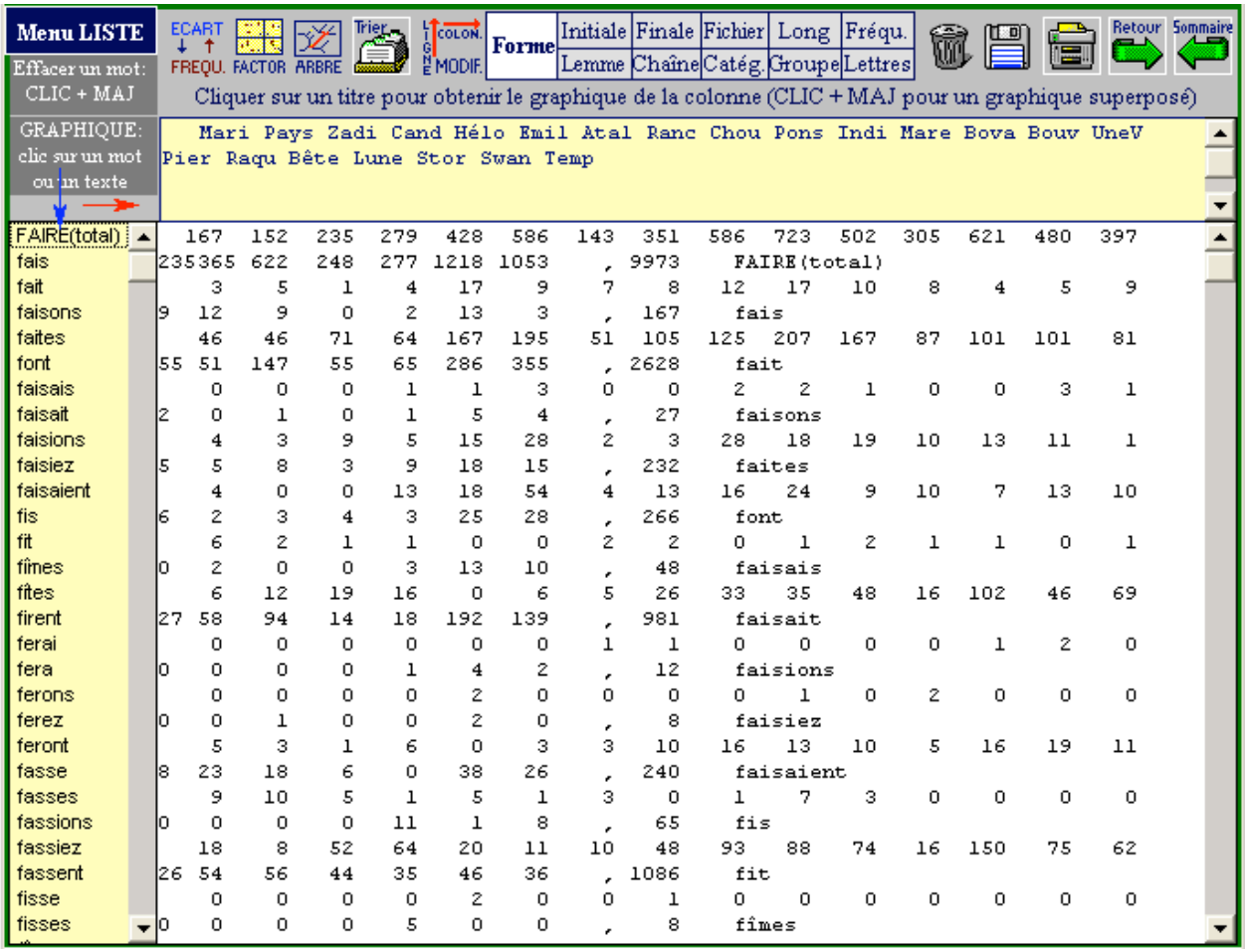
choisies. Par contexte il faut entendre la page. La cooccurrence n'est pas toujours vérifiable dans les extraits, parce que les extraits respectent la limite du paragraphe et non de la page, mais elle est contrôlable si l'on clique sur l'extrait et qu'on est renvoyé à la page originale, où les deux mots cooccurrents sont soulignés en rouge.

g - sur une LISTE DE MOTS, préalablement constituée.

LES LISTES DE MOTS

Solliciter d'abord le bouton LISTE de la carte principale, qui adresse l'utilisateur à une page spéciale où divers modules de sélection sont proposés: DEBUT DE MOT, FIN DE MOT, CHAINE dont l'effet est semblable aux options des commandes CONTEXTE ou CONCORDANCE. Le remplissage de la liste courante peut aussi être libre. L'utilisateur fournit alors autant de formes qu'il veut, en signifiant à la machine par un mot vide que la liste est close. Une fois constituée la liste peut être soumise d'un coup au programme de concordance, mais aussi à l'analyse factorielle (voir plus loin) ou à la représentation graphique, en sorte que son statut relève davantage des fonctions statistiques que nous allons examiner maintenant. Ci-dessous un exemple de liste empruntée au corpus *Example* et détaillant les formes du paradigme *faire*.

Le verbe faire *dans le corpus* Exemple



Noter que cette liste est modifiable et provisoire. On peut effacer un de ses éléments par un clic (associé à la touche MAJUSCULE). On efface la liste entière en en constituant une autre ou en sollicitant le bouton en forme de poubelle. Mais on peut aussi la compléter, la nouvelle liste s'ajoutant à la première. Un dialogue permet de choisir entre le remplacement ou le complément. Certaines limites liées toutefois à la contenance des pages doivent être respectées. Comme un même champ ne peut accepter plus de 32000 caractères, la liste aura autant de lignes que le champ peut en contenir, ce qui varie avec la longueur de la ligne, et en fin de compte avec le nombre de textes. Quand le détail risque de conduire au dépassement, il est possible de se contenter du cumul de tous les mots en un seul total, ce qui est avantageux en particulier avec les verbes. On a aussi la ressource d'utiliser le bouton MODIF, qui permet toutes sortes de regroupement et d'allègement.

CHAPITRE 5

L'exploitation statistique

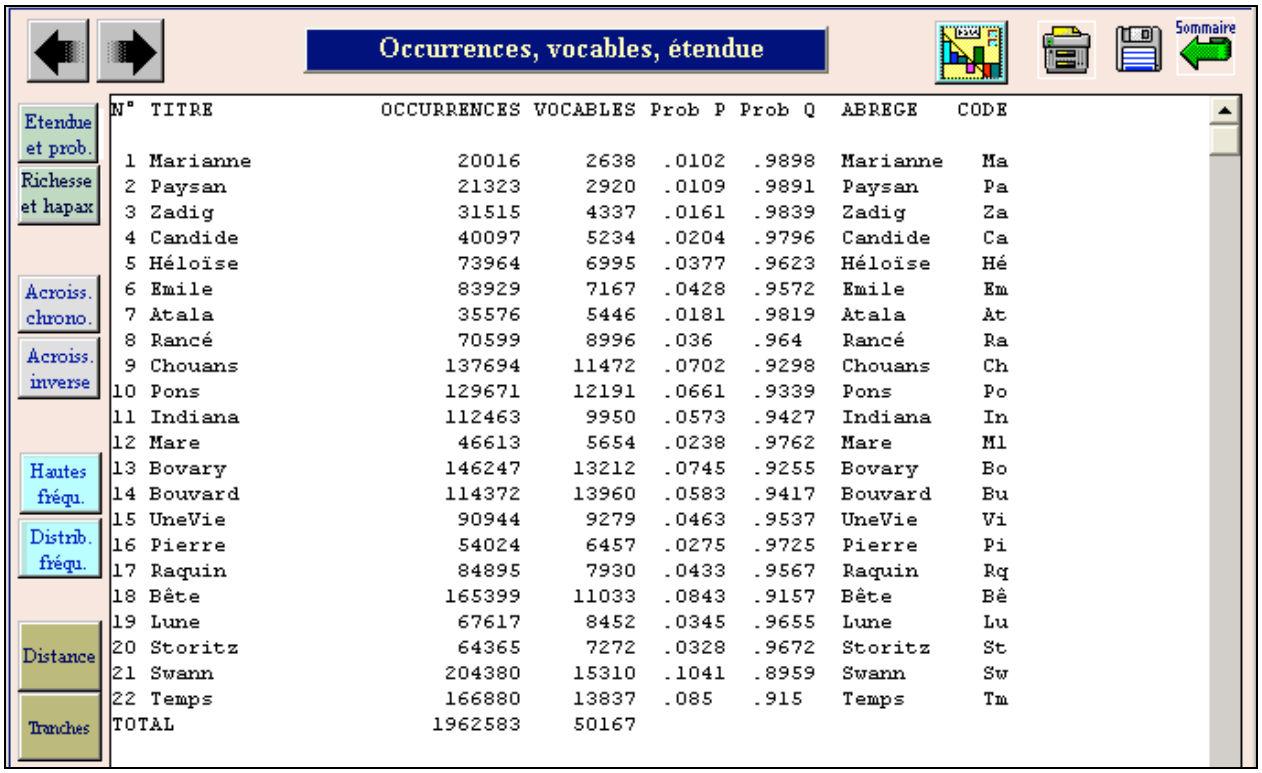
Les calculs

PARTITION et STATISTIQUE

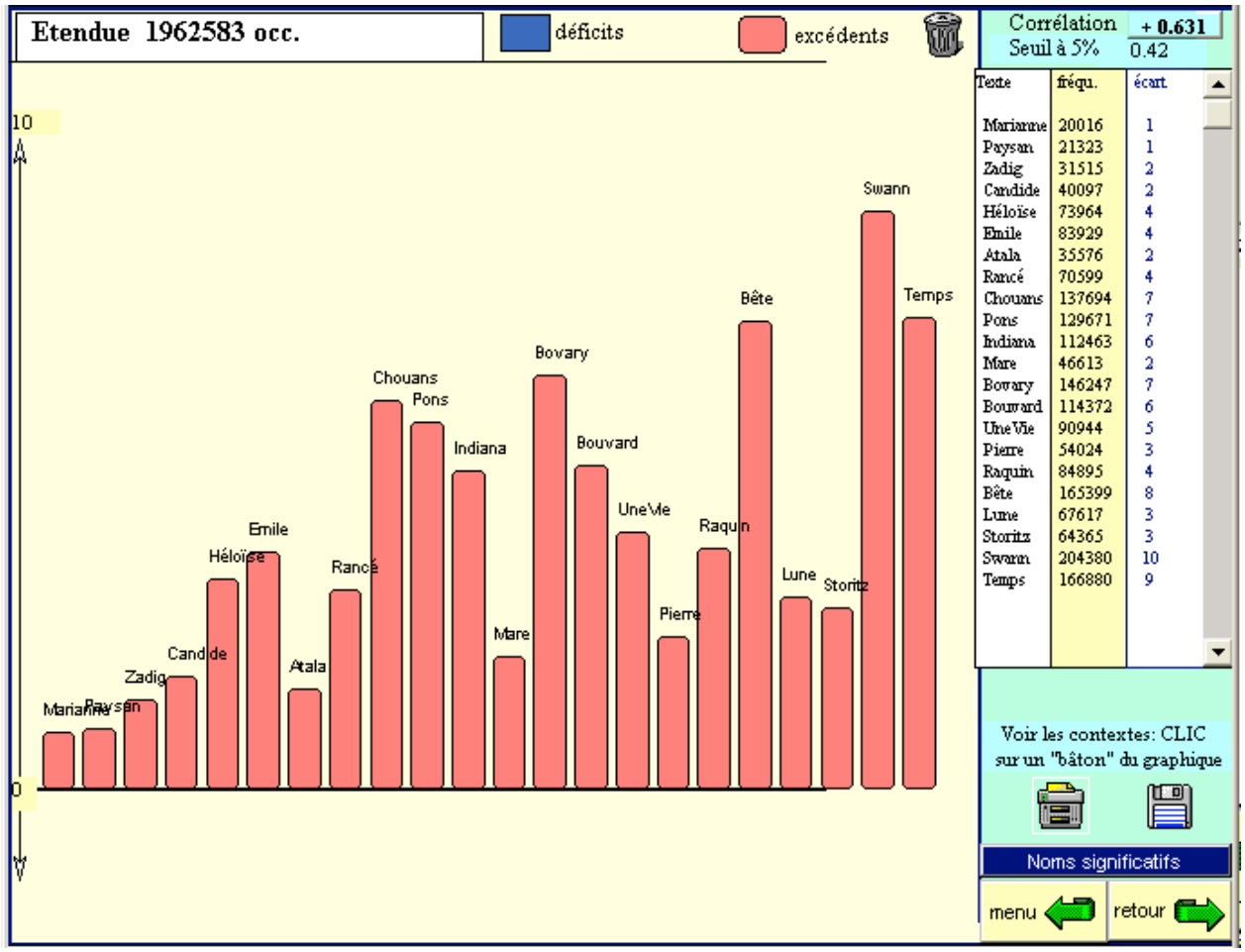
Si les données étaient d'un seul tenant, la seule comparaison - et donc la seule opération statistique - qu'on pourrait faire serait extérieure (et c'est précisément celle qui prend appui sur le Dictionnaire du TLF et qui produit la liste des spécificités externes).

En réalité, même si le corpus n'a pas de subdivisions naturelles, dûment jalonnées dans le fichier des données (comme expliqué précédemment), le traitement opère une segmentation artificielle en découpant neuf tronçons de longueur voisine dans le flux des données. Quand le texte représente un discours suivi ou que les données sont de type sériel ou chronologique, un découpage de cette sorte, même brutal et sommaire, peut délivrer des résultats suggestifs, qui inviteront à pratiquer une segmentation plus adéquate et à renouveler le traitement, sur des fondements mieux assurés. Que les divisions soient naturelles ou arbitraires, les calculs obéissent aux mêmes principes et s'appuient pareillement sur les lois classiques de la statistique linguistique, principalement la loi normale et la loi binomiale. Et les probabilités *p* et *q* qu'on lit dans la distribution ci-dessous servent à tous les calculs de pondération. Si l'on souhaite vérifier ces calculs de pondération et contrôler les tests statistiques, l'étendue et les caractéristiques de chacun sont montrées quand on sollicite le bouton DISTRIBUTION, puis le bouton ÉTENDUE et PROB.

Étendue relative des textes de la base EXAMPLE



N°	TITRE	OCCURRENCES	VOCABLES	Prob P	Prob Q	ABREGE	CODE
1	Marianne	20016	2638	.0102	.9898	Marianne	Ma
2	Paysan	21323	2920	.0109	.9891	Paysan	Pa
3	Zadig	31515	4337	.0161	.9839	Zadig	Za
4	Candide	40097	5234	.0204	.9796	Candide	Ca
5	Héloïse	73964	6995	.0377	.9623	Héloïse	Hé
6	Emile	83929	7167	.0428	.9572	Emile	Em
7	Atala	35576	5446	.0181	.9819	Atala	At
8	Rancé	70599	8996	.036	.964	Rancé	Ra
9	Chouans	137694	11472	.0702	.9298	Chouans	Ch
10	Pons	129671	12191	.0661	.9339	Pons	Po
11	Indiana	112463	9950	.0573	.9427	Indiana	In
12	Mare	46613	5654	.0238	.9762	Mare	Ml
13	Bovary	146247	13212	.0745	.9255	Bovary	Bo
14	Bouvard	114372	13960	.0583	.9417	Bouvard	Bu
15	UneVie	90944	9279	.0463	.9537	UneVie	Vi
16	Pierre	54024	6457	.0275	.9725	Pierre	Pi
17	Raquin	84895	7930	.0433	.9567	Raquin	Rq
18	Bête	165399	11033	.0843	.9157	Bête	Bê
19	Lune	67617	8452	.0345	.9655	Lune	Lu
20	Storitz	64365	7272	.0328	.9672	Storitz	St
21	Swann	204380	15310	.1041	.8959	Swann	Sw
22	Temps	166880	13837	.085	.915	Temps	Tm
TOTAL		1962583	50167				

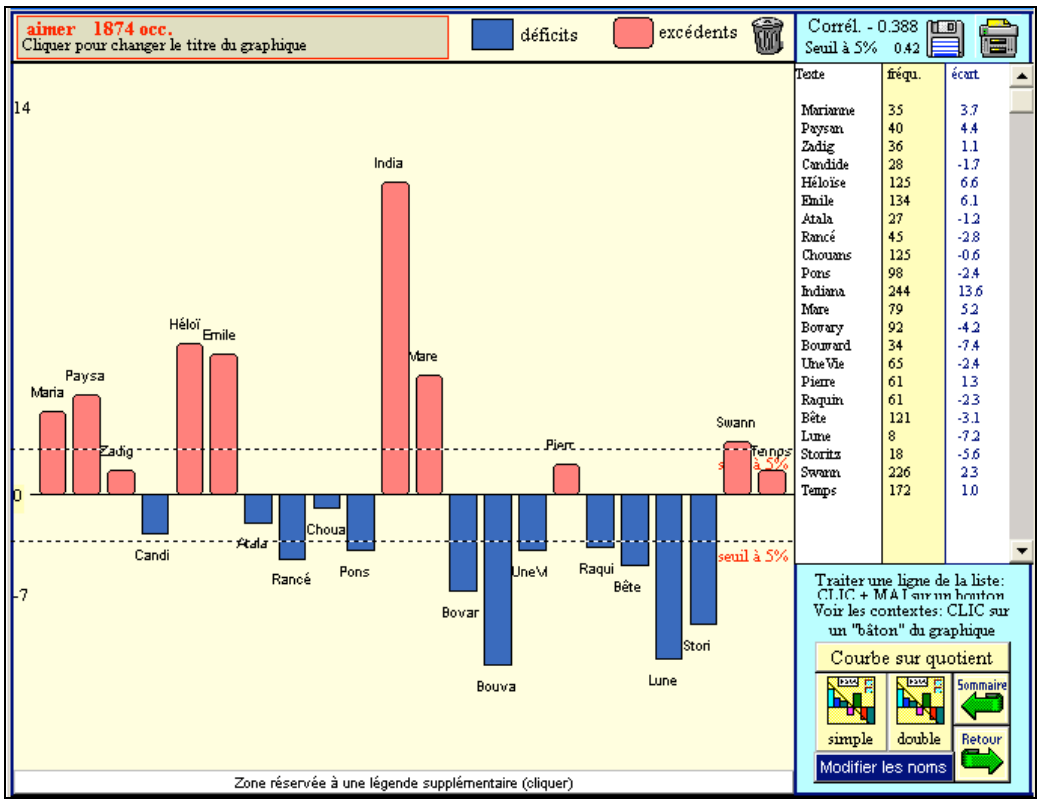


On distinguera deux sortes de résultats: ceux qui sont acquis à l'issue de la phase de préparation et ceux qui réclament un calcul spécial et supplémentaire. Les premiers sont délivrés par les boutons DISTRIBUTION, EVOLUTION, SPECIFICITES et PHRASES-CLES, les seconds par les programmes LISTES, THEMES, ASSOCIATIONS et TOPOLOGIE. Les premiers portent sur la totalité du corpus, les seconds sur une sélection. En réalité cette distinction n'est pas absolue. Certains programmes mis en œuvre dans la préparation peuvent être repris avec des options différentes (par exemple SPECIFICITES et PHRASES-CLES). Et certains autres, tout en portant sur la totalité du corpus, ne sont pas lancés dans la phase de préparation et attendent une demande expresse (par exemple CONNEXION). De plus les méthodes et les calculs ne sont pas spécifiques à l'une ou l'autre approche et il convient de les préciser tout d'abord.

ÉCART RÉDUIT et CALCUL HYPERGÉOMÉTRIQUE

La distribution d'un mot est rarement régulière à travers un corpus et des écarts s'y observent entre la fréquence d'un mot observée dans un texte et la fréquence théorique qu'on était en droit d'attendre, vu la proportion du texte dans l'ensemble, et qui s'établit avec une simple règle de trois (fréquence théorique d'un mot dans un texte = fréquence du mot dans le corpus pondérée par la probabilité *p* ou part du texte dans le corpus). Dans sa forme la plus simple, le calcul pondère cet écart absolu selon la formule de l'"écart réduit" (*q* étant la probabilité complémentaire 1-p):

*z = (réel-théorique) / racine carrée(théorique * q)*

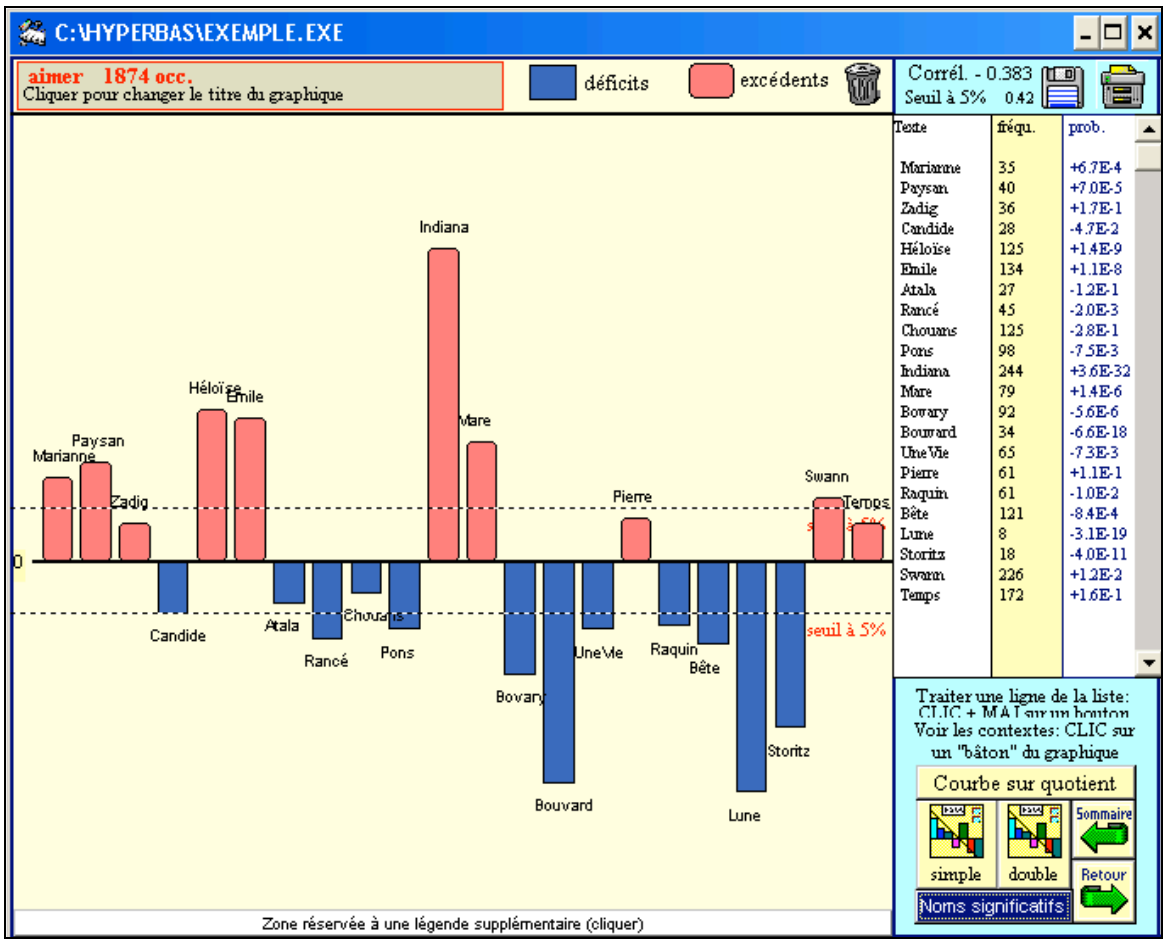


La distribution du verbe aimer, *selon la loi normale*

Les puristes qui redoutent les effets de trompe-l'œil auxquels donne lieu l'écart réduit dans les faibles effectifs sont en droit de préférer l'emploi de la loi hypergéométrique, qui s'applique impeccablement à tous les cas de figure, même dans les occasions où la loi normale offre un fondement mal assuré. L'inconvénient du calcul hypergéométrique n'est plus dans les temps de traitement, vu la puissance des machines actuelles, mais dans la lisibilité des résultats, lesquels parviennent sous forme de probabilités, souvent très faibles et toujours positives. On doit avoir recours au signe de l'écart et à la représentation logarithmique. HYPERBASE fait automatiquement le choix, plus exigeant, du modèle hypergéométrique lorsque la sécurité le recommande et que le corpus est de dimension restreinte. C'est le cas dans l'exemple du verbe aimer représenté ci-dessous. La colonne de droite indique les probabilités en virgule flottante.

En réalité, quand les effectifs sont suffisamment amples, les deux calculs convergent et les deux graphiques sont superposables.

Histogramme du verbe AIMER (calcul hypergéométrique)



Hyperbase à sa naissance faisait usage de la loi normale, notamment dans les calculs de spécificités et chaque fois que la distribution d'un mot, d'une classe ou d'une catégorie devait être pondérée pour tenir compte de l'inégale étendue des

textes comparés. Toute fréquence (ou effectif) observée était rapprochée de la fréquence attendue et convertie en écart réduit selon la formule:

$$z = \frac{k - fp}{2\sqrt{fpq}}$$

pour k = fréquence observée dans le texte,
f = fréquence observée dans le corpus
p = étendue du texte *(t)* / étendue du corpus *(T)*
$q = 1 - p$

Les spécificités, les histogrammes et les analyses factorielles étaient fondés sur ce calcul rapide et très classique. En réalité la loi normale qui le justifie est une approximation d'un calcul plus exact qui repose sur la distribution hypergéométrique.
P. Lafon, dans la revue *Mots* d'octobre 1980, a montré que ce modèle est celui qui convient le mieux à la lexicométrie parce qu'il s'applique à des données discrètes (alors que la loi normale traite de valeurs continues) et qu'il propose un tirage exhaustif (alors que la loi binomiale s'accommode d'un tirage non exhaustif). Et nous avions dès 1981 reconnu la supériorité du modèle hypergéométrique dans notre *Vocabulaire français de 1789 à nos jours* (p.31-50). Le gain en précision et en exactitude n'était cependant pas sensible dans les corpus de grande ampleur, qu'Hyperbase est appelé à traiter. Et cette précision entraînait un accroissement du temps de calcul. Aussi bien avions-nous renoncé au modèle idéal puisque le raccourci de la loi normale offrait des garanties suffisantes. La puissance multipliée des ordinateurs actuels nous amène à revoir notre position: même si le raccourci réduit les temps d'un facteur 100, les lourds calculs exigés par la loi hypergéométrique n'ont rien de prohibitif puisque le résultat apparaît de façon instantanée. En outre l'avantage de l'exactitude n'est pas négligeable dans les corpus de faible dimension, et notamment lorsque le calcul s'applique aux mots peu fréquents, et plus encore lorsqu'il s'agit d'un déficit.

Rappelons que le modèle a besoin, comme la loi normale, de quatre paramètres:

T = taille du corpus,
t = taille du texte,
f = fréquence du mot dans le corpus,
k = fréquence du mot dans le texte,

et que la formule, en mettant en œuvre des factorielles, calcule la probabilité pour un mot d'avoir la fréquence k dans un texte:

$$prob\ (x=k) = \frac{f!\ (T-f)!\ t!\ (T-t)!}{k!\ (f-k)!\ (t-k)!\ (T-f-t+k)!\ T!}$$

Bien entendu le calcul des factorielles est facilité par l'utilisation des logarithmes et l'approximation de Stirling:

$$log\ n! = n\ log\ (n) - n + \frac{log\ (2\ \pi\ n)}{2} + \frac{1}{12\ n}$$

et une itération est à faire pour tenir compte de la probabilité que la fréquence du mot se situe entre 0 et k (si l'on a affaire à un déficit) ou entre k et f (s'il s'agit d'un excédent). Dans le premier cas la probabilité s'établit selon la formule:

$$prob\ (x = k\text{-}1\ ) = prob\ (x{=}k)\ \frac{k\ (T\text{-}f\text{-}t{+}k)}{(f\text{-}k{+}1)(t\text{-}k{+}1)}$$

et dans le second le calcul est pareillement récurrent :

$$prob(\ x = k{+}1\ ) = prob\ (x{=}k)\ \frac{(f\text{-}k)\ (t\text{-}k)}{(k{+}1)(T\text{-}t\text{-}f{+}k{+}1)}$$

Les histogrammes désormais restituent les probabilités ainsi calculées dans le champ jadis réservé aux écarts réduits (colonne de droite). Ainsi dans l'exemple ci-dessus du verbe *aimer*, la fréquence 35 relevée dans *La Vie de Marianne* est estimée supérieure à ce qui était attendu et la probabilité de cet excédent est évaluée à 6.7 E-04, soit prob = 0.00067. La lisibilité du graphique a imposé une mise à l'échelle, que l'on a obtenue par un procédé logarithmique, selon la formule (où *prob* est la probabilité fournie par le calcul hypergéométrique et et où le logarithme est à base *beta*):

$alpha = 1 / prob$
$beta = 2.37 + (prob * 3)$
$gamma = 0.57 + (prob / 5)$
$z = log_{beta}\ (alpha)^{gamma}$

C'est à dessein que nous utilisons le symbole z de l'écart réduit car la conversion est calculée de façon à retrouver l'échelle de l'écart réduit et la représentation adoptée dans les versions précédentes d'Hyperbase. Mais qu'on ne s'y trompe pas, l'abscisse de chaque point ne doit rien à la loi normale.

Pour rendre la lecture plus sûre, nous avons représenté en pointillés la plage où le seuil de 5% est atteint dans la zone des excédents, comme dans celle des déficits. Il se trouve en effet que le graphique occupe tout l'espace disponible, même lorsque les écarts et les probabilités sont faibles. Et cette distorsion de la perspective aurait pu induire en erreur les utilisateurs peu attentifs aux valeurs numériques. En particulier un message explicite est délivré quand le seuil significatif n'est pas atteint.

Il arrive que les probabilités soient extrêmement faibles. La tradition est d'arrêter les calculs quand l'exposant atteint la valeur -99. Nous avons repoussé

cette limite jusqu'à -217. Et dans ce cas la conversion en écart atteint la valeur -37, qu'on observe parfois pour les noms propres dans les gros corpus.

Le même traitement (calcul hypergéométrique, puis conversion en écart réduit) a lieu également dans la page LISTE chaque fois qu'un tableau est constitué. Un tableau de probabilités est calculé parallèlement, aussitôt transformé en écarts. C'est sur ces écarts que se fondent ensuite les histogrammes de ligne ou de colonne et les analyses factorielles ou arborées.

Enfin le modèle hypergéométrique s'applique aux spécificités, même si là encore l'apparence est celle d'un écart réduit. Il reste toutefois un vestige de la loi normale: quand le corpus traité est comparé à la base *Frantext*, on n'a pas jugé utile d'avoir recours au modèle hypergéométrique. En effet ce corpus de référence dépasse 100 millions d'occurrences et permet de bénéficier de la loi des grands nombres. Comme le calcul n'est entrepris que dans la zone sûre où la fréquence théorique est suffisamment élevée, la loi normale offre ici les mêmes garanties et la même précision que le modèle hypergéométrique.

Toutefois, l'utilisateur peut en dernier recours imposer la méthode de son choix, modèle hypergéométrique ou loi normale, du moins lorsque les calculs sont faits à la demande, pour un graphique, un tableau ou une analyse factorielle. Le menu principal offre une option qui demeure valable le temps d'une session pour tous les calculs ultérieurs. L'option par défaut est celle du calcul hypergéométrique, sauf lorsqu'il s'agit de gros corpus.

L'option de calcul dans le menu principal

LE COEFFICIENT DE CORRÉLATION. LE PROGRAMME *ÉVOLUTION*.

Lorsque les textes qui constituent le corpus s'échelonnent dans le temps, dans l'espace ou dans quelque succession logique, en suivant l'ordre imposé par une structure sérielle, le coefficient de corrélation peut être calculé, en comparant, pour chaque mot, les valeurs de l'écart réduit au rang de chaque élément. Le programme de préparation présuppose que les données sont de type sériel ou chronologique et établit le coefficient de Bravais-Pearson quand la fréquence d'une forme est suffisante pour permettre le calcul probabiliste.

Rappelons la formule de ce coefficient en renvoyant le lecteur aux manuels de >Charles Muller pour les explications complémentaires:

$r = \Sigma ((x_i - \tilde{x}) (y_i - \tilde{y})) / n\sigma_x\sigma_y$

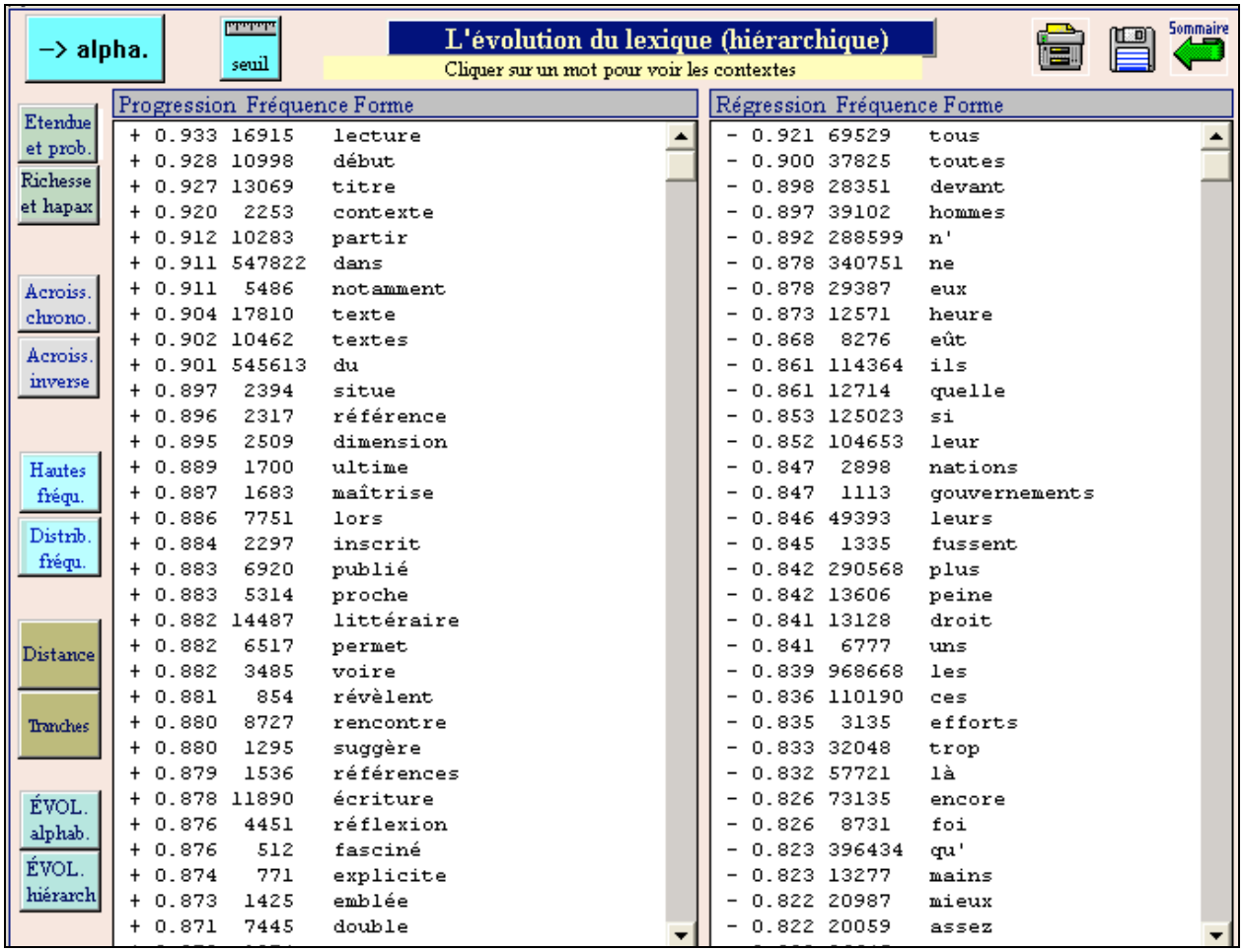
$\tilde{x}$ et σ_x étant respectivement la moyenne et l'écart-type de la série en *x*,

$\tilde{y}$ et σ_y la moyenne et l'écart-type de la série en *y*.

Dans l'exemple de la revue *Europe* représenté ci-dessous, l'évolution est sensible qui parcourt une série de 900 numéros de 1923 à 2000. Tous les mots qui atteignent un seuil approprié au nombre de textes considérés (la limite apparaît quand on sollicite le bouton SEUIL) sont catalogués dans deux pages spéciales qu'on atteint par le bouton EVOLUTION du menu principal et où les résultats peuvent être lus dans l'ordre alphabétique ou dans l'ordre hiérarchique. Deux colonnes sont alors visibles qui reproduisent, à gauche, la liste des mots de plus en plus employés et, à droite, ceux qui sont abandonnés progressivement.

Ce calcul initialement appliqué à tous les mots du corpus est renouvelé chaque fois qu'on établit un graphique sur un mot ou quelque autre objet du corpus. Il s'applique aussi, indépendamment de la chronologie, à deux distributions quelconques, quand on veut comparer le profil de deux mots dans le corpus.

La corrélation chronologique (corpus de la revue EUROPE)



ANALYSE FACTORIELLE

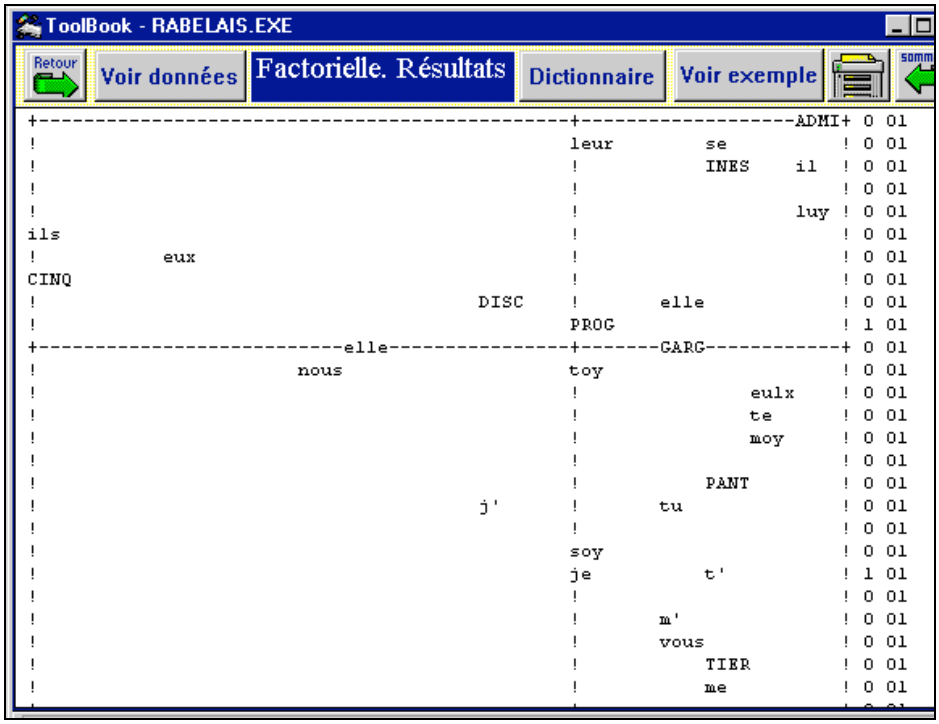
Il est possible d'utiliser des procédures statistiques plus synthétiques que de simples histogrammes. Le programme FACTORIELLE permet de soumettre au calcul une série de formes, qui seront traitées ensemble selon les méthodes multidimensionnelles. Le programme utilisé a été fourni par l'association ADDAD, qui distribue un logiciel complet pour l'analyse des données. Le module ici mis en oeuvre est celui de l'analyse de correspondance, qui suit l'algorithme proposé par Jean-Paul Benzécri et dont l'adaptation à Windows a été réalisée par André Salem.

L'analyse factorielle est réalisée par un programme extérieur: ANCORR.exe. Elle prend appui sur le fichier des paramètres (AFC.par) et celui des données (TABLEAU.afc) et produit celui des résultats (ANALYSE.afc). Ces trois fichiers sont créés - et donc d'abord effacés - à chaque lancement du programme FACTORIELLE. Il suffit de leur donner un autre nom pour les conserver. Mais cela n'est pas nécessaire, car les résultats reviennent dans un champ, que l'on peut éditer et imprimer. Au reste le lancement du programme peut être indépendant d'Hyperbase (il suffit de lancer le programme externe ANCORR.EXE après avoir correctement rempli le fichier des paramètres (AFC.PAR) et celui des données (TABLEAU.AFC), le résultat étant écrit dans le fichier ANALYSE.AFC).

Noter que le graphique signale l'emplacement des points doubles où se produit un recouvrement et que la désignation de chaque point (qui peut être une forme ou un texte) occupe quatre lettres au maximum dans le fichier des résultats (ANALYSE.afc). On peut rendre le graphique plus lisible en complétant les noms (sans modifier l'emplacement) ou en explicitant les symboles. Pour faciliter la localisation et l'interprétation, le programme ANCORR fournit les coordonnées de chaque individu (de chaque ligne) et de chaque texte (de chaque colonne), ainsi que la contribution de la variable considérée à chacun des facteurs isolés.

On trouvera ci-dessous un premier exemple d'analyse factorielle, réalisé à partir du corpus de Rabelais. Elle est fondée sur la distribution des pronoms personnels. Ce n'est pas le lieu de le commenter. Mais les lignes de force y sont fort visibles, qui gouvernent les situations pragmatiques, et qui opposent les genres les uns aux autres et Rabelais (*Pantagruel, Gargantua, Tiers Livre, Quart Livre et Cinquième livre*) aux autres auteurs du corpus.

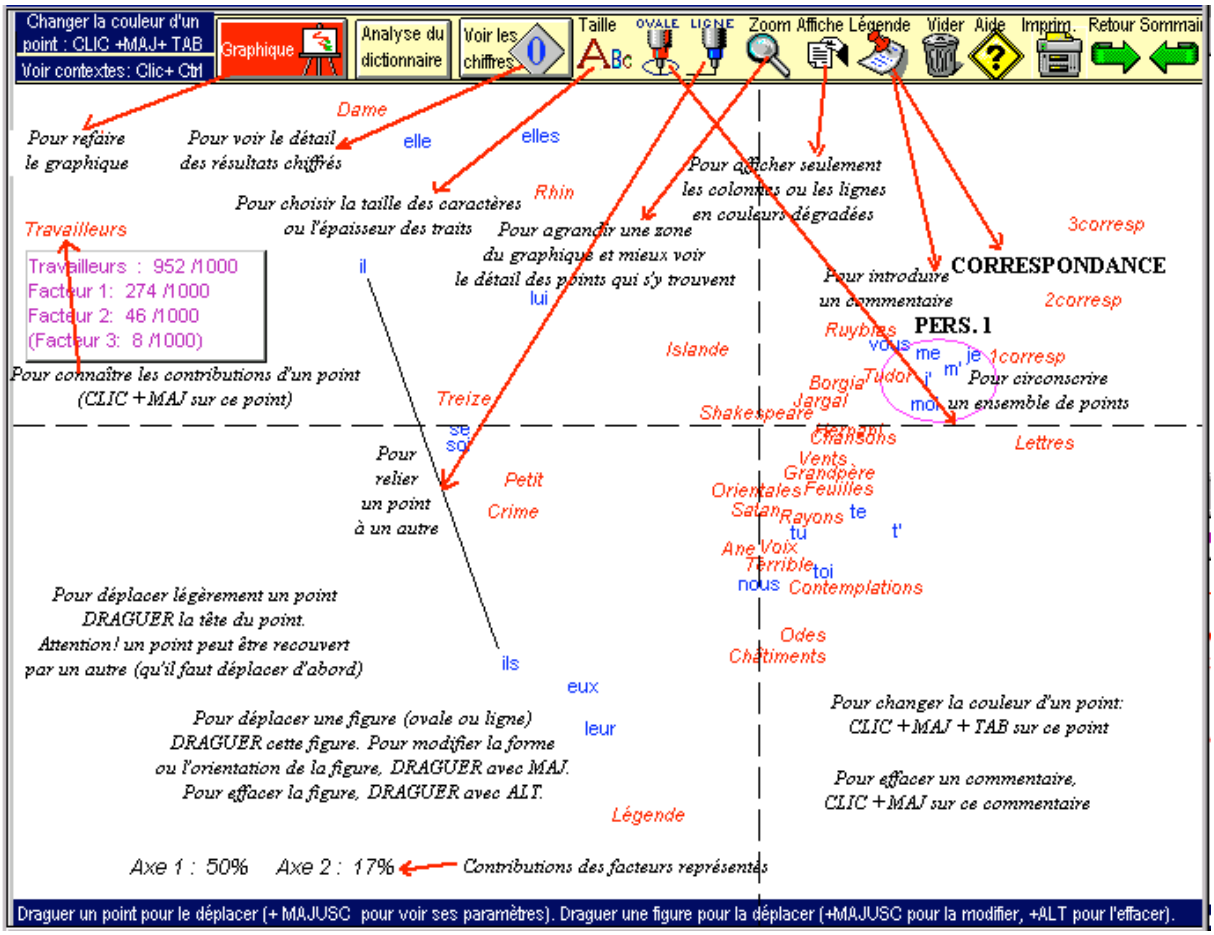
Facteurs 1 et 2 de l'analyse factorielle. Les pronoms personnels chez Rabelais.



On constate cependant qu'un tel graphique n'est pas de lecture aisée, car les noms sont réduits à quatre lettres et les points doubles se cachent l'un l'autre. Si Hyperbase continue à utiliser ce vénérable programme écrit en Fortran il y a plus de trente ans (sous le nom de *Tabet*) c'est parce qu'il est très rapide et qu'il est familier à beaucoup d'utilisateurs. Quoique les résultats qu'il propose aient un graphisme rudimentaire, qui date de l'époque où les périphériques de sortie ne délivraient que des caractères, ils restent accessibles si l'on actionne le bouton *Voir les chiffres*, qui restitue le détail des coordonnées, corrélations et contributions pour chaque variable, chaque individu et chaque facteur.

Mais la lisibilité a été grandement améliorée par les techniques modernes qui donnent au graphisme la transparence, la couleur et une meilleure définition. Les points ne sont plus limités à quatre caractères; ils peuvent partager la même localisation (en cas de recouvrement la souris peut opérer un léger déplacement), et se distinguer par la couleur, la taille ou le style des caractères, selon qu'on a affaire à une variable (un texte) ou à un individu (un mot). Des aides à l'enrichissement et à l'interprétation sont accessibles à tout endroit du graphique, par l'ajout de titres, légendes, commentaires (bouton *Légende*) ou figures graphiques (boutons *Ovale* et *Ligne*) aptes à rendre plus clairs les regroupements et les oppositions. Une explication spécifique - reproduite ci-dessous - détaille le mode d'emploi des outils proposés.

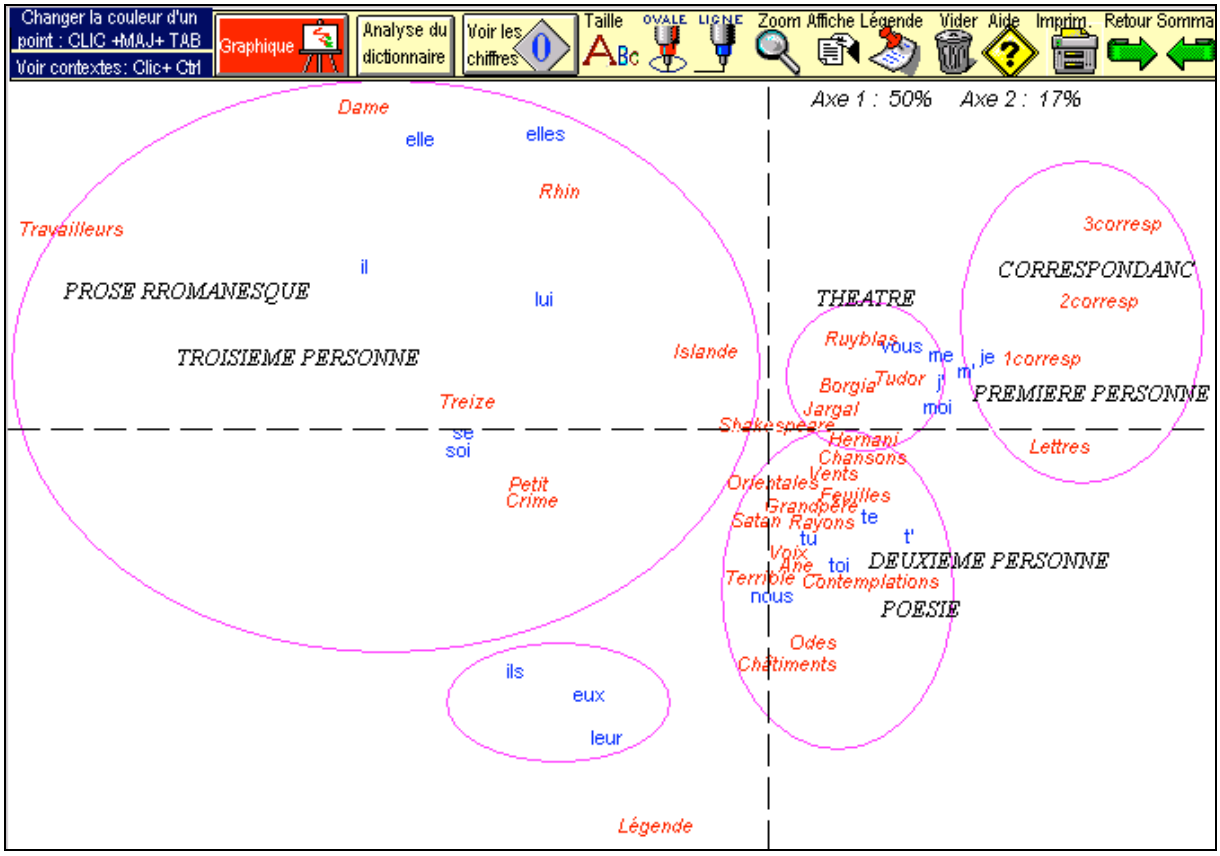
L'aide à l'analyse factorielle



Quand les points à représenter sont trop nombreux pour être lisibles, on peut adopter une taille de caractères plus petite (bouton TAILLE), ou bien agrandir la zone trop encombrée (bouton ZOOM), ou bien consulter le détail des résultats chiffrés (bouton VOIR LES CHIFFRES). Choisir alors une police à espacement fixe (LUCIDA CONSOLE ou COURIER NEW, en petite taille) et le format “paysage”.

On trouvera ci-dessous un exemple emprunté au corpus Hugo. Les pronoms personnels y montrent des choix exclusifs qui sont en relation avec les genres littéraires. La troisième personne a évidemment partie liée avec le récit, tandis que la première personne est attendue dans la correspondance, mais aussi au thèâtre et dans la poésie lyrique, où elle rencontre la seconde personne. Tous les corpus que nous avons étudiés montrent un semblable partage, qui donne la priorité au genre. Mais l'analyse fournit aussi des informations moins triviales, comme l'opposition du singulier et du pluriel, la tendance de la poésie épique (*Légende des siècles*) à rejoindre la troisième personne, ou l'indépendance des formes *nous* et *vous*.

Analyse factorielle des pronoms personnels chez Hugo



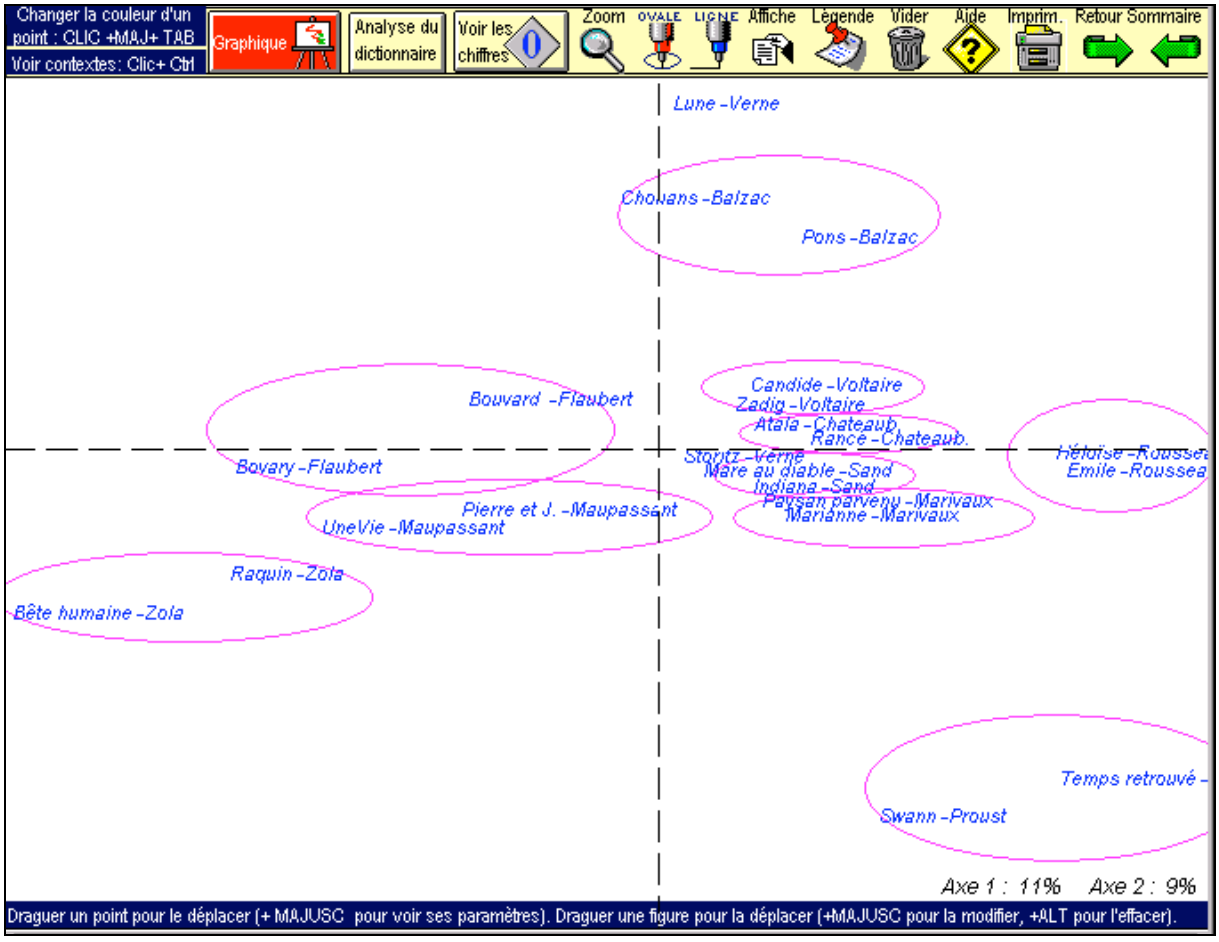
ANALYSE FACTORIELLE DU DICTIONNAIRE

Le champ ouvert par l'analyse factorielle est illimité. Les lignes du tableau (les mots de la liste) peuvent être aussi nombreuses qu'on le souhaite, si la taille du champ qui reçoit les résultats ne dépasse pas 30000 caractères. Cela est suffisant pour autoriser plusieurs centaines de lignes, et davantage même, si le nombre de colonnes est réduit. Il est un cas cependant où ces bornes sont franchies: lorsqu'on souhaite prendre en compte tous les mots du corpus, ou tout au moins tous ceux qui ont une fréquence suffisante pour autoriser les calculs. Le tableau à analyser comprend alors plusieurs milliers de lignes. Cette éventualité a été prévue. Pour y faire face, un dialogue exige qu'on précise si l'objet de l'analyse se trouve dans le dictionnaire ou dans une liste. En choisissant la première proposition, on détourne le programme vers le dictionnaire où il va puiser ses données.

Le traitement aboutit à un résultat global dont la lecture risquerait d'être difficile si la représentation graphique des mots était maintenue. Dans un tel cas on se contente de représenter les variables ou colonnes, c'est à dire les textes du

corpus. Et l'on acquiert ainsi une vue synthétique des alliances ou oppositions qui se manifestent parmi les textes à travers l'ensemble de leur vocabulaire. Voici ce qu'on obtient dans le corpus *Example*, quand près de 3000 mots interviennent. L'analyse ici élimine les mots rares et considère les fréquences plus élevées, parmi lesquelles les mots grammaticaux occupent une place de choix. Le point de vue est donc plutôt stylistique. On voit que les deux textes d'un même écrivain sont voisins, à l'exception de ceux de Jules Verne, dont l'écriture est en effet fort différente. On voit aussi les écrivains se rapprocher ou s'opposer selon l'époque et l'école littéraire. Ceux qui prennent leur modèle chez Flaubert occupent le flanc gauche (Flaubert, Maupassant, Zola), les autres se dispersent sur le flanc droit.

Analyse factorielle du dictionnaire (corpus Example*)*

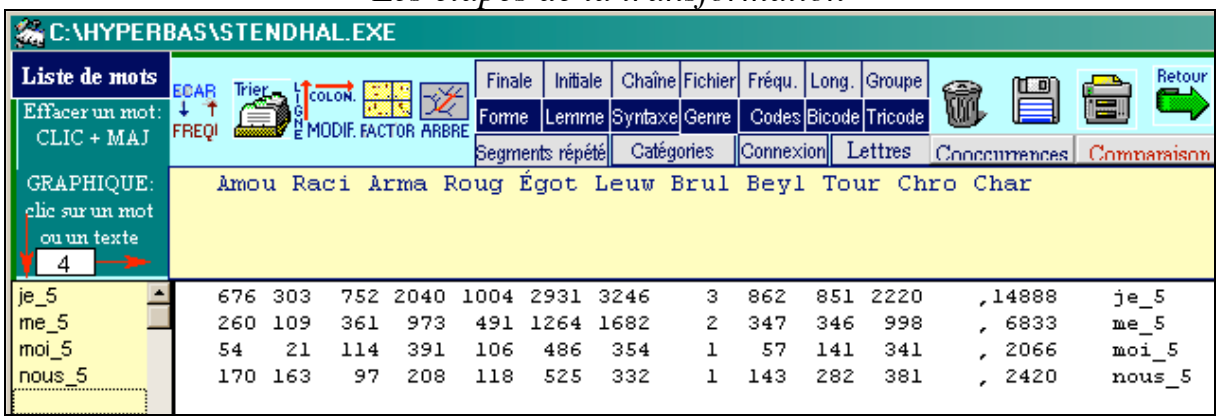


DONNÉES BRUTES OU PONDÉRÉES

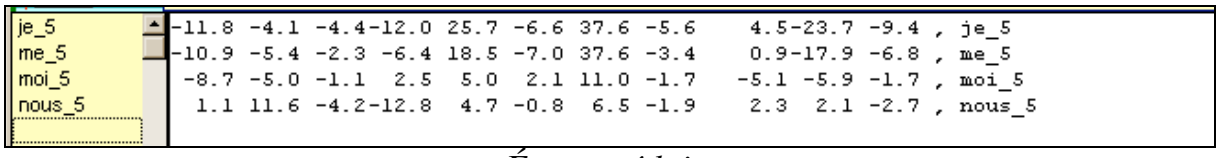
On a prévu la distorsion que peut amener dans les données linguistiques l'effet de taille, c'est-à-dire une trop grande disproportion entre les lignes (les mots peuvent avoir des fréquences très inégales dont le rapport peut être de 1 à 1000), ou entre les colonnes (les textes - ou parties de texte - peuvent avoir des étendues fort déséquilibrées). Afin d'atténuer ces inégalités, le programme calcule les écarts réduits (ou ce qui en tient lieu, quand le modèle hypergéométrique est appliqué), puis les translate dans la zone positive, le plus

grand nombre négatif s'alignant sur zéro et les autres éléments gardant leurs distances respectives (car l'analyse n'accepte pas les données négatives). On trouvera ci-dessous les étapes de la transformation. Noter que la translation finale s'accompagne d'un facteur 10 afin de supprimer le point décimal et d'alléger le tableau. Cette multiplication ne pose aucun problème théorique, mais devant la translation elle-même, fondée sur une addition, les puristes froncent les sourcils. Ils admettent certes, quoique avec réticence, les autres formes de pondération qui mettent en œuvre des logarithmes, des fréquences relatives, des racines carrées. Ils accepteraient même la simplification extrême qui ne garderait que deux chiffres: 0 pour les déficits et 1 pour les excédents. Mais pour eux l'ajout d'une tare, même légère, même égale, fausse la balance.

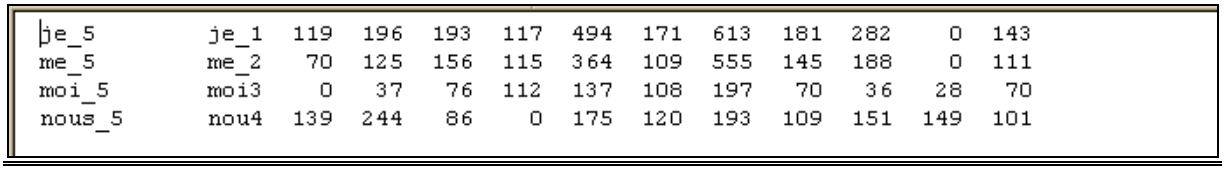
Les étapes de la transformation



Données brutes



Écarts réduits

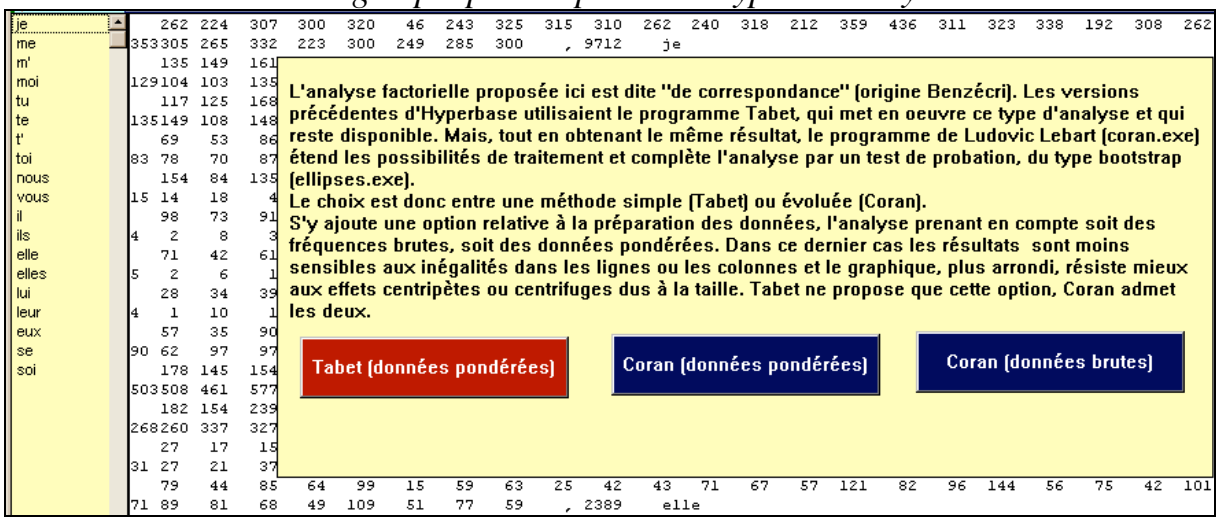


Translation des écarts dans la zone positive

Exemple: Je dans le premier texte
Fréquence: 676 écart réduit: -11.8
Ecart négatif maximum dans la série: -23.7
*Donnée pondérée pour Je dans le premier texte: (-11.8 + 23.7) * 10 = 119*

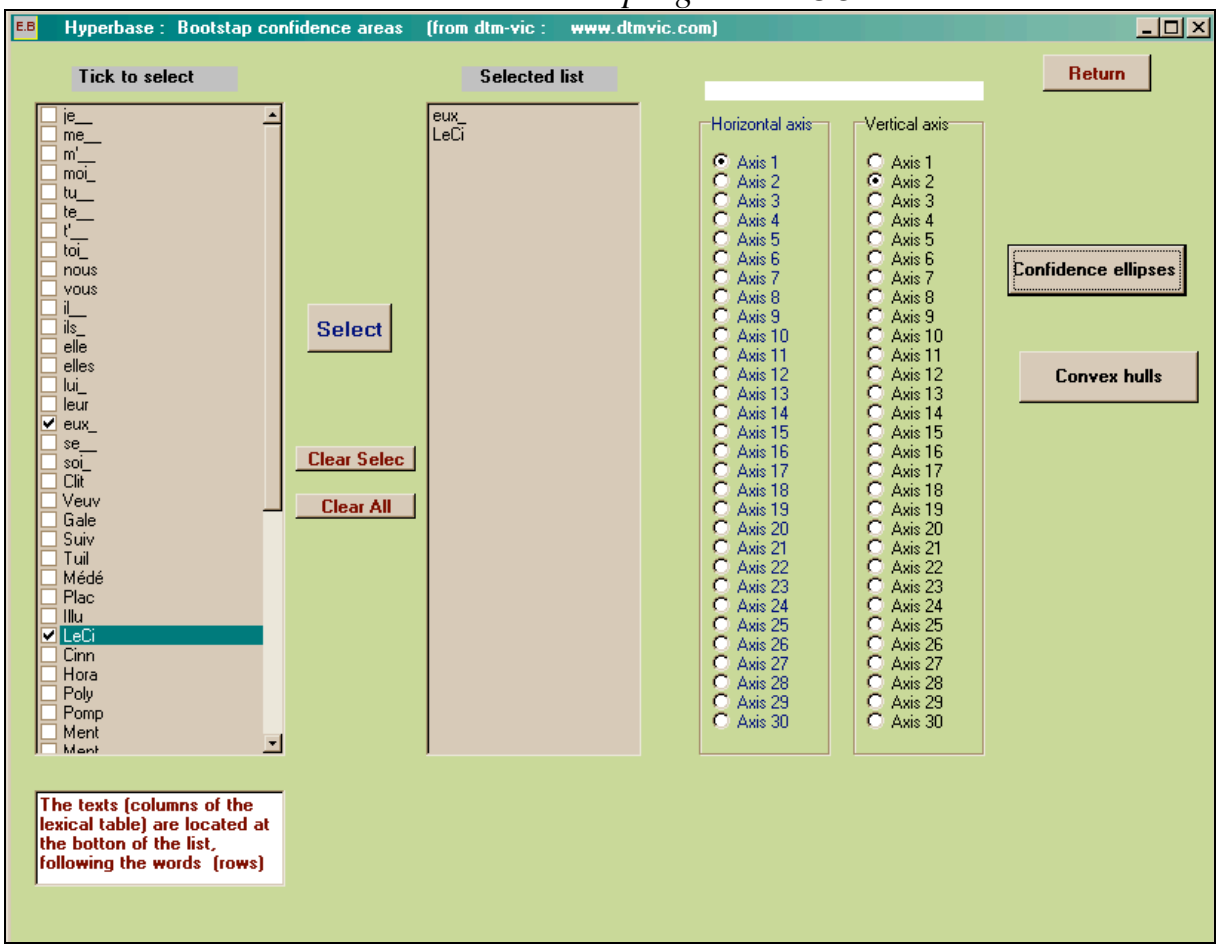
Aussi pour contenter leurs scrupules, on a fait appel au maître de la discipline, Ludovic Lebart, dont les ouvrages font autorité en matière de statistique et particulièrement d'analyse de données. Notre ami a bien voulu extraire de son logiciel DTM-VIC (adresse: http://www.dtmvic.com) les fonctions dont nous avions besoin en les adaptant à notre environnement. Pour y accéder, au lieu de choisir l'analyse traditionnelle *Tabet* (en rouge à gauche dans le dialogue ci-dessous), on sollicitera le bouton de droite (en bleu) *Coran (données brutes)*.

Dialogue proposant plusieurs types d'analyse



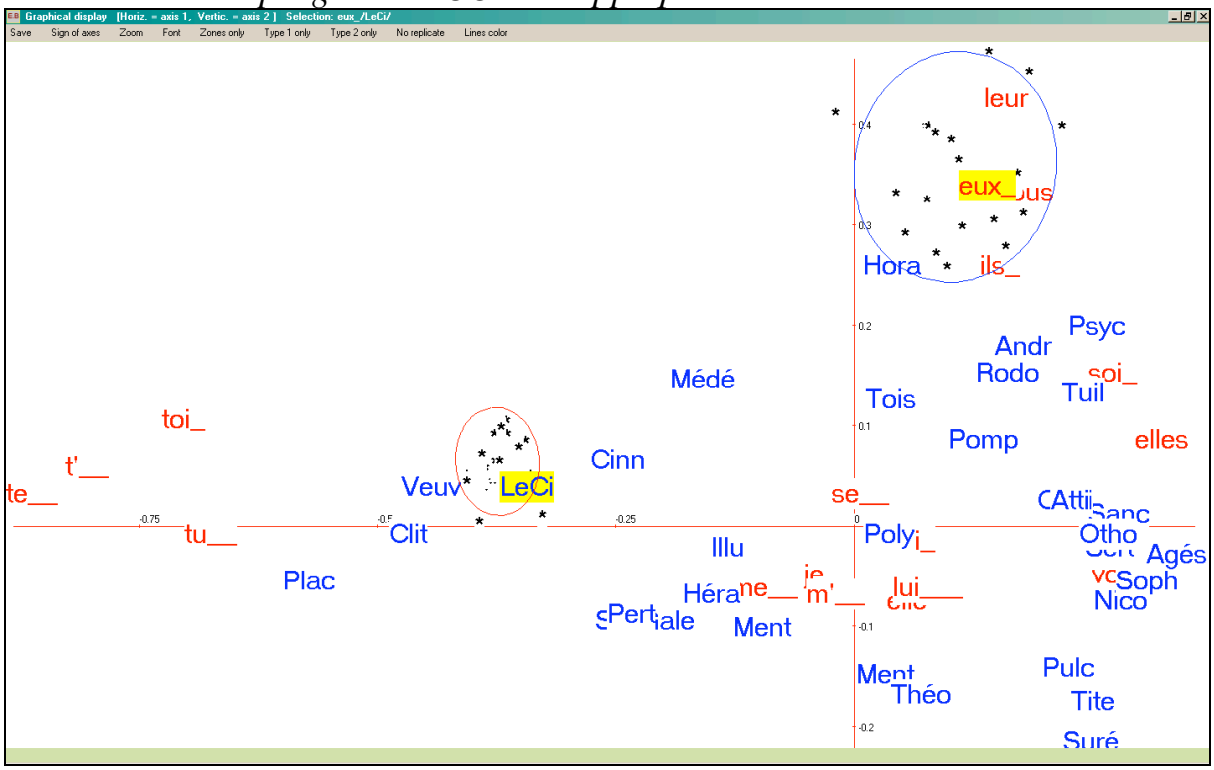
Cela conduit à un écran où l'utilisateur est invité à choisi les facteurs qu'il veut croiser (colonnes de droite) et les lignes (mots) ou colonnes (textes) sur lequel il veut attirer l'attention (colonne de gauche). Cette sélection est facultative: les facteurs 1 et 2 sont pris par défaut et la liste des éléments sélectionnés peut rester vide. Le graphique est obtenu en cliquant sur le bouton *Confidence ellispses*.

Écran de sélection du programme CORAN



Ce n'est pas le lieu de commenter un tel tableau consacré aux pronoms personnels dans les 34 pièces de Corneille et où les trois personnes jouent aux quatre coins. Le tutoiement s'arroge l'extrême gauche du graphique et s'oppose au *vous*, isolé à droite. La première personne se réserve le quadrant inférieur gauche, tandis que la troisième personne se divise, sur la moitié droite, en deux lots, selon qu'il s'agit du singulier ou du pluriel. Cette opposition du nombre s'exerce aussi sur le *nous* qui est presque toujours un pluriel et qui se situe en haut, et le *vous*, qui est souvent un faux pluriel et qui campe en bas. Quant aux textes ils se répartissent suivant les accointances qui les rapprochent ou les éloignent des camps ainsi identifiés. Les comédies ayant tendance à employer le *tu* s'installent de préférence à gauche. Comme elles se situent en début de carrière, elles amorcent un mouvement chronologique qui va de droite à gauche et qui de poursuit encore, cette fois du haut vers le bas, lorsque les tragédies sont seules en cause.

Le programme CORAN appliqué aux données brutes

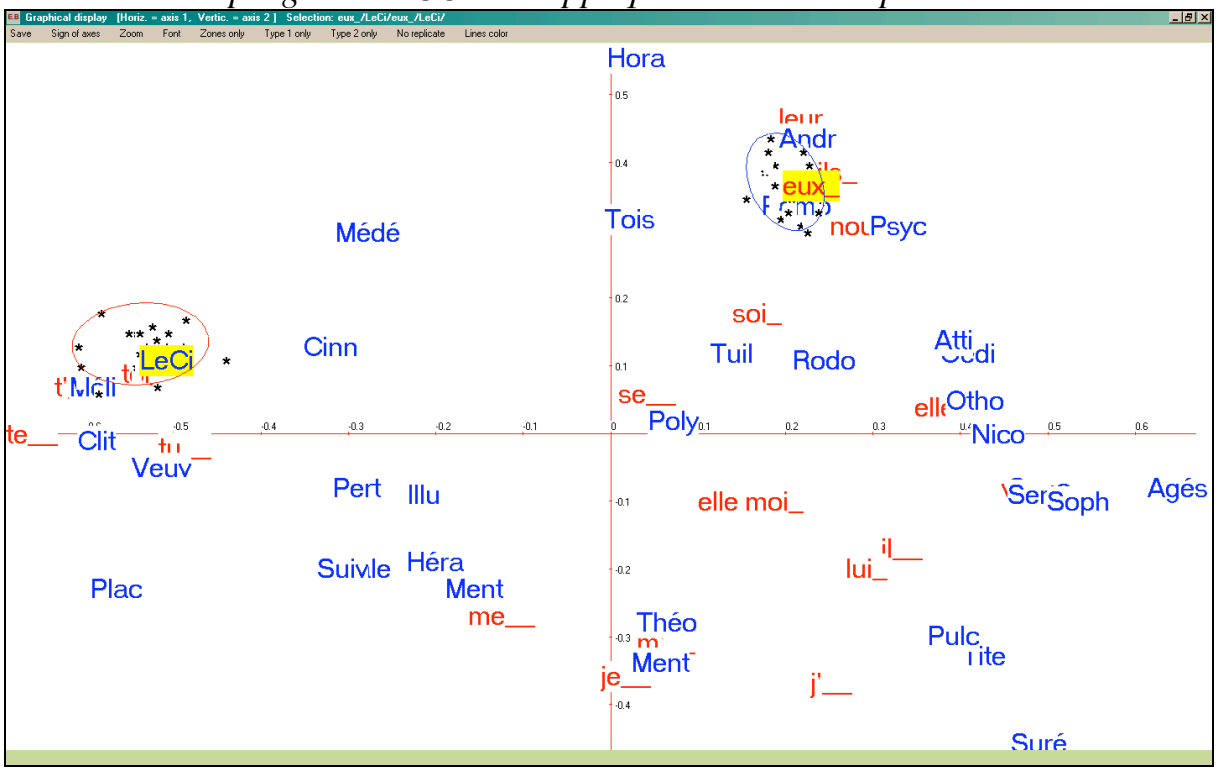


Reste à apprécier la fiabilité de telles analyses. On a parfois reproché à l'analyse factorielle de toujours révéler des divisions et des oppositions même là où les données sont homogènes et presque aléatoires. Or le programme CORAN permet précisément de simuler le hasard et d'apprécier la sûreté ou la faiblesse des résultats. Pour cela jetons dans l'urne tous les emplois des pronoms personnels, chacun ayant la couleur du texte où on le trouve. Si l'on procède à autant de tirages qu'il y a de boules dans l'urne - mais en replaçant la boule à chaque fois - on obtient un tableau qui ressemble ou non à la distribution réelle et qui donne lieu à une analyse, semblable ou différente. En réitérant ce

processus vingt fois on obtient pour chaque variable et chaque individu une position flottante, une constellation de vingt points dans l'espace factoriel. Si le tir est groupé, c'est que l'analyse est solide. Si au contraire l'ellipse qui contient les impacts s'étend largement, le point en question a une définition fragile et incertaine, au moins pour les deux facteurs considérés. Dans l'exemple proposé plus haut, les deux ellipses représentées autour d'un texte (*Le Cid*) et d'un mot (*eux)* permettent de conclure à la robustesse de l'analyse.

Ce test de vérification porte le nom de *bootstrap*[1]. Il peut aussi s'appliquer à des données pondérées, comme le montre le graphique ci-dessous obtenu avec les mêmes données que le précédent, préalablement réduites.

Le programme CORAN appliqué aux données pondérées



Les deux méthodes et les deux graphiques disent la même chose, même si dans le détail les mots les plus fréquents tendent, dans les données pondérées, à s'écarter un peu du centre (par exemple *vous, je* et *il,* qui ont près de 10000 occurrences) et inversement les mots les moins fréquents à s'en rapprocher (notamment *soi, elles, eux, toi, t'* et *ils*, qui en ont moins de 1000). La lisibilité est meilleure dans le traitement pondéré, l'équilibre étant mieux respecté dans les textes comme dans les mots, mais aussi entre les textes et les mots. Le centre du graphique y est dégagé et la périphérie plus arrondie. C'est encore plus vrai lorsque les effectifs sont de faible étendue, ou que l'inégalité s'accroît entre les riches et les pauvres. Il suffit parfois d'un seul mot très excentrique, par exemple

[1] Cette image humoristique qui évoque les lacets de chaussures, qu'il suffirait de tirer vers le haut pour s'élever dans les airs, donne le sens de la démarche: en s'appuyant sur la distribution réelle et les lignes de force souterraines qui la gouvernent, l'application répétitive du hasard en tire une lisibilité fiable.

d'un hapax, pour élargir tellement le champ des données brutes qu'on ne voit plus rien au centre, dans la zone utile, par suite de l'encombrement. Certains diront qu'il suffit de mettre ce mot en élément supplémentaire. Mais un autre mot rare ou mal distribué risque de se présenter pour semer à nouveau la panique. Quand faudra-t-il arrêter l'épuration? Lorsqu'on met le doigt dans le choix des élus et des exclus, on entre dans l'arbitraire et dans la chaîne sans fin des jugements en appel. La déontologie la plus sage est de considérer le jugement des chiffres comme sans appel. Si avant l'analyse le chercheur a tous les droits sur ses données, et notamment celui de les pondérer, il n'en a guère après l'analyse, quand le jugement est rendu, sinon celui d'interpréter l'arrêt.

À la différence du paramétrage de *Tabet* qui s'inscrit dans un fichier indépendant des données (comme expliqué page 136), le choix des paramètres pour *CORAN.exe* et *ELLIPSES.exe* se fait dans le fichier des données, à savoir *CORDON.TXT*. *Hyperbase* assure le remplissage automatique mais l'utilisateur a le loisir de modifier les options en intervenant directement dans ce fichier avec un éditeur. Voici, précisée par Ludovic Lebart, la structure de ce fichier:

Première ligne: **Titre**
Deuxième ligne: (format libre)
iexa nqexa nvidi lfmt modig
iexa = nombre de lignes, nqexa: nombre de colonnes, nvidi: longueur (en a4) de l'ident des lignes
lfmt = longueur du format en cartes (!) = multiple de 20a4 (en général, lfmt = 1)
modig = 0 si tous les individus sont actifs (pas de ligne de sélection du type 00001110111011 pour les individus)
Troisième ligne: **identificateurs des colonnes en a4** (4 caractères par identificateur) (max: 900)
Quatrième ligne: **selection des variables** (colonnes), ex: 000001111111111122222 (0 abandonnée, 1 active, 2 supplémentaire)
Cinquième ligne (seulement si modig = 1) **sélection des individus** (lignes)
Ligne suivante (cinquième ou sixième selon que modig = 0 ou 1): **format fortran de lecture des données**.
Les "iexa" lignes suivantes contiennent les données
Dernière ligne: **nfac list3 (= 1) ngraf npage (=1) nlign(=60) jbase (= 0) niter (= 0) nsimu nebru**
nfac= nombre de facteurs désirés
list3= 1 si on veut les coordonnées des lignes (en général), = 0 sinon (cas de 20 000 lignes...)
ngraf= nombre de graphiques factoriels (ngraf = 3 signifie: plans (1,2), (2,3),(3,4))
npage (paramètre fossile des graphiques sur imprimante = laisser à une page)
nlign = nbre de lignes des graphiques (souvent 60)
Laisser jbase et niter à 0.
nsimu= nombre de simulations bootstrap (< 31)
nebru = 1 signifie: graphiques avec les racines des contributions absolues, pour bien voir les variables qui participent (sinon nebru = 0)

Les résultats s'inscrivent dans deux fichiers: corsor.txt et nboot.txt

A part les sorties classiques, sur "corsor.txt",
le fichier "nboot.txt" contient les coordonnées factorielles
des nsimu réplications bootstrap + l'analyse originale
(donc (nsimu + 1) fois (les lignes actives + les colonnes actives))
La première ligne de ce fichier contient:
itot, iact, kv, nidmax, nsimtot, nboot (=1)
itot = nombre total d'éléments (lignes actives + colonnes actives)
iact = nombre de lignes actives
kv = nombre d'axs factoriels
nidmax= longueur maximale de l"identificateur.
nsimtot = nombre de réplications + original (en premier) (nsimu +1)
nboot = 1 (pour le bootstrap partiel, le plus courant, par défaut ici)

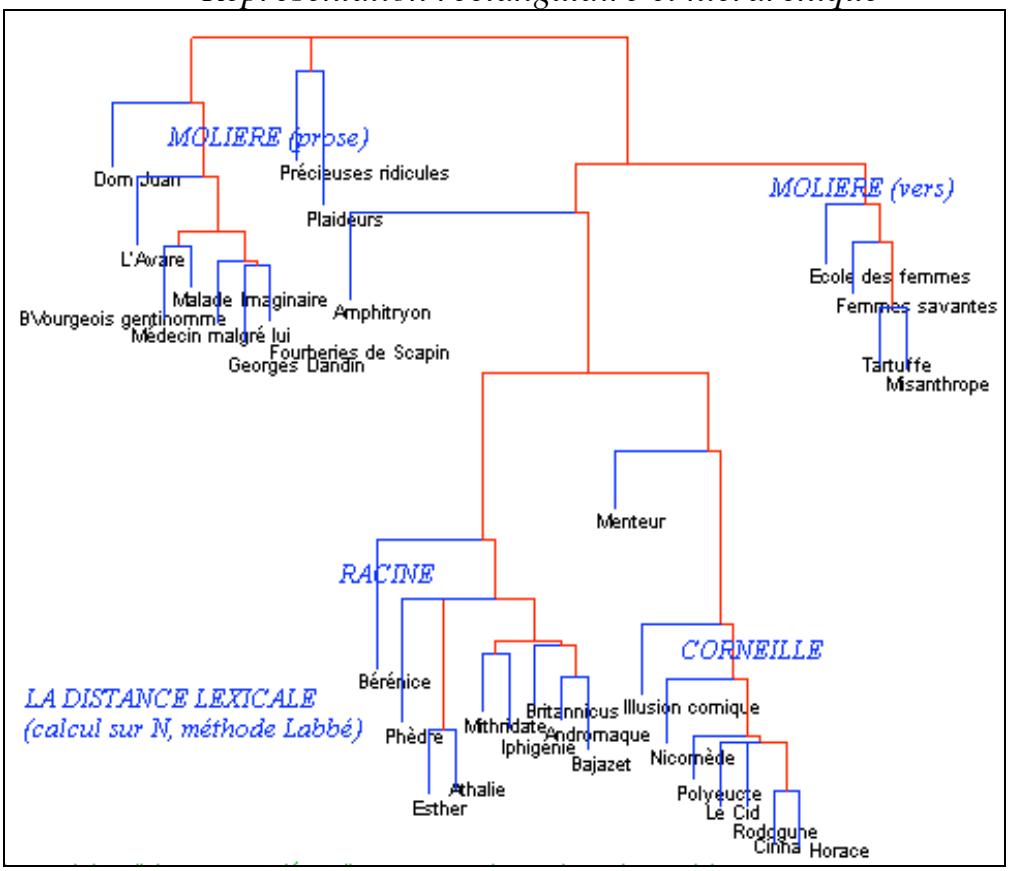
L'ANALYSE ARBORÉE (MÉTHODE LUONG)

Nous empruntons ici à Xuan Luong une technique de classification qu'il a appelée analyse arborée. Nous renvoyons à l'auteur et à sa thèse (*"Méthodes d'analyse arborée. Algorithmes. Applications"*, 1988, Université de Paris 5) pour le détail des calculs de topologie. Bornons-nous à expliquer que l'algorithme produit des graphes qui rendent compte de la proximité des objets étudiés (ici des textes) à partir d'une distance (ici celle de Labbé). L'avantage de cette technique, par rapport à l'analyse factorielle, est qu'on n'a plus à distinguer et à croiser des facteurs, dont chacun n'explique qu'une partie de la variance. Toute l'explication se résume ici en une seule représentation graphique, qui peut prendre deux formes: rectangulaire ou radiale.

La première est la plus habituelle car c'est sous cette forme que se traduit le programme bien connu de classification hiérarchique. On y voit Racine se distinguer de Corneille, tous les deux conjugués s'opposant à Molière et les comédies du *Menteur* et des *Plaideurs* servant de transition. Les vers se distinguent de la prose dans l'espace occupé par Molière. Le détail de l'évolution se lit dans le graphe réservé à Racine, les dernières pièces *Athalie* et *Esther*, et à un moindre degré *Phèdre*, se détachant des autres.

Analyse arborée de la distance lexicale (calculée sur N)

Représentation rectangulaire et hiérarchique

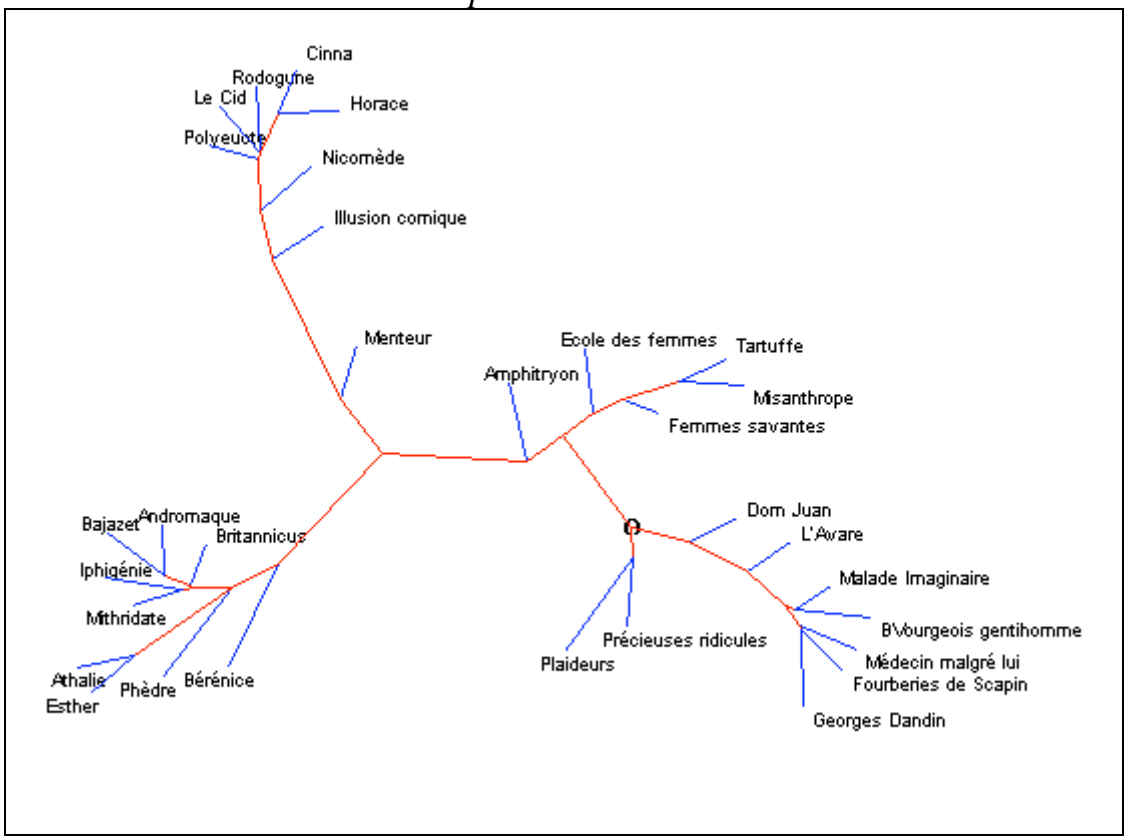


On prendra garde toutefois à ne tenir aucun compte de l'écartement latéral qui sépare les groupes. Il s'agit d'un artifice de présentation qui tend à répartir les textes sur la surface du plan. Mais seules les distances verticales sont à interpréter, en sorte que le groupe des œuvres en vers de Molière, déporté à l'extrême droite, est en réalité proche des textes en prose du même auteur, que le graphique inscrit à l'extrême gauche. Les segments verticaux qu'il faut emprunter pour joindre ces deux groupes sont de faible ampleur. Ils sont beaucoup plus importants lorsque le chemin part de Racine ou de Corneille pour rejoindre Molière et surtout Molière prosateur.

Les distances sont plus faciles à interpréter dans la présentation radiale des résultats de l'analyse arborée. Car elles sont directement proportionnelles à la longueur des parcours dessinés en rouge sur le graphique ci-dessous. À chaque bifurcation le chemin emprunte une direction dont le sens importe peu, c'est la distance qui seule compte et qui se mesure par l'addition des segments de jonction. Le danger ici serait de mesurer les distances à vol d'oiseau et de prendre visuellement des raccourcis, comme on fait en montagne. Dans le cas du théâtre classique, la carte obtenue est limpide, avec trois branches partant du centre et conduisant respectivement à Corneille, Molière et Racine, les *Plaideurs*, pour des raisons qui tiennent au genre, constituant la seule exception à cette division tripartite.

Analyse arborée de la distance lexicale (calculée sur N)

Représentation radiale

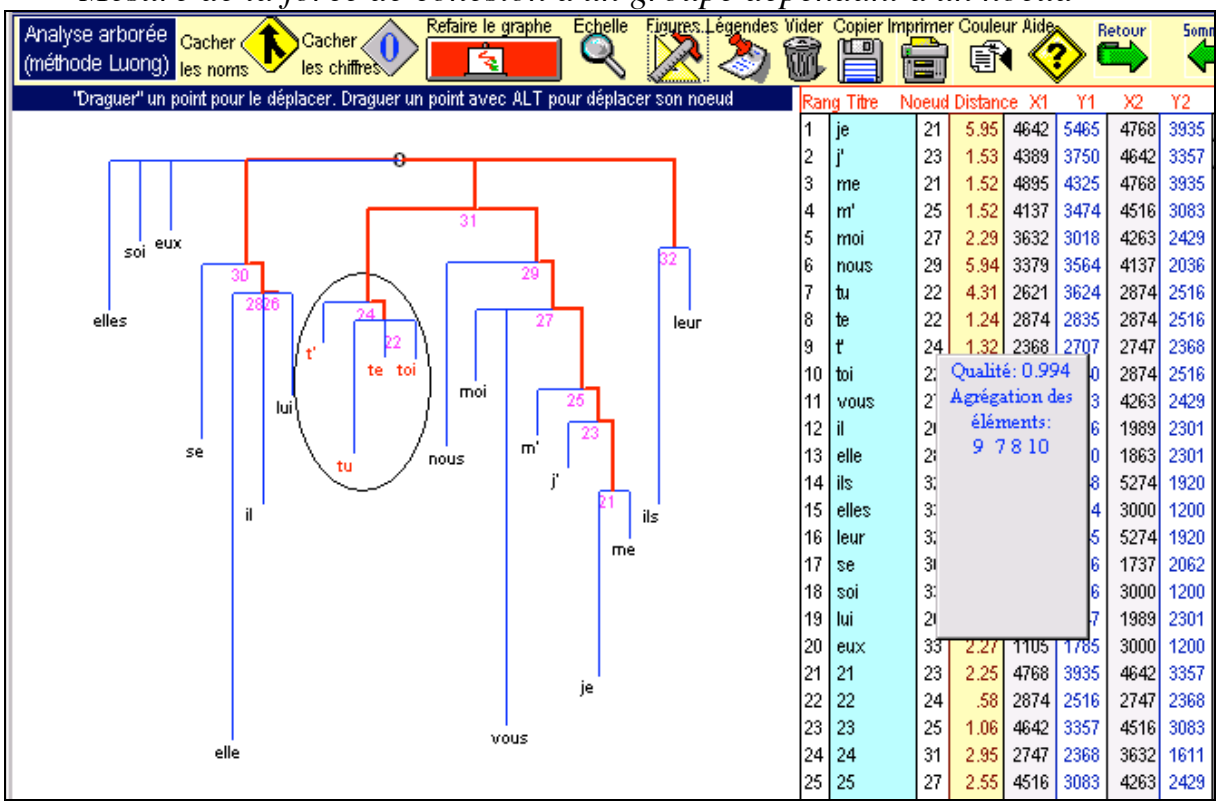


Quelques ajustements ont été ménagés en vue d'améliorer la lisibilité du graphique:

1 - Quand il y a recouvrement, le programme donne à la souris le droit de déplacer les textes (ou feuilles du graphe), à condition de respecter le point d'ancrage (ou nœud) auquel ils se rattachent et la longueur du segment terminal (en bleu) qui leur est propre. Il suffit de draguer le nom du texte à déplacer et de le faire pivoter jusqu'à l'endroit désiré, en s'arrêtant quand la longueur requise est atteinte et que le programme l'autorise (l'icône du curseur change de forme quand le trajet est trop court ou trop long).

2 - Les nœuds apparaissent ou non selon que le graphique est encombré ou non. Mais il est toujours possible de les montrer ou de les cacher. Lorsqu'on sollicite (en draguant) un de ces nœuds (par exemple le noeud 24 dans la figure ci-dessous), on fait apparaître en rouge toutes les feuilles (ou textes) qui dépendent de ce nœud (soit *tu, te, toi* et *t'*), en même temps qu'une fenêtre indique le degré de cohésion (mesuré de 0 à 1) du groupe agrégé autour de ce nœud (ici 0,994).

Mesure de la force de cohésion d'un groupe dépendant d'un noeud



Rang	Titre	Noeud	Distance	X1	Y1	X2	Y2
1	je	21	5.95	4642	5465	4768	3935
2	j'	23	1.53	4389	3750	4642	3357
3	me	21	1.52	4895	4325	4768	3935
4	m'	25	1.52	4137	3474	4516	3083
5	moi	27	2.29	3632	3018	4263	2429
6	nous	29	5.94	3379	3564	4137	2036
7	tu	22	4.31	2621	3624	2874	2516
8	te	22	1.24	2874	2835	2874	2516
9	t'	24	1.32	2368	2707	2747	2368
10	toi	2				2874	2516
11	vous	2				4263	2429
12	il	2				1989	2301
13	elle	2				1863	2301
14	ils	3				5274	1920
15	elles	3				3000	1200
16	leur	3				5274	1920
17	se	3				1737	2062
18	soi	3				3000	1200
19	lui	2				1989	2301
20	eux	33	2.27	1105	1785	3000	1200
21	21	23	2.25	4768	3935	4642	3357
22	22	24	.58	2874	2516	2747	2368
23	23	25	1.06	4642	3357	4516	3083
24	24	31	2.95	2747	2368	3632	1611
25	25	27	2.55	4516	3083	4263	2429

3 - En désignant successivement par un clic deux textes dont on veut mesurer la distance (enfoncer en même temps la touche MAJUSCULE), on dessine en vert tous les jalons qui permettent de joindre les deux textes, tandis qu'une fenêtre indique la distance totale à parcourir (on retrouve d'ailleurs par ce moyen les données initiales).

4 - Le choix est proposé entre une représentation étroite ou large. Le zoom s'étend de 50% à 200%.

5 - Un bouton permet alternativement de montrer ou de cacher les résultats quantifiés qui servent à la représentation graphique. Divers champs apparaissent alors qui pour chaque segment précisent le numéro d'ordre, le nom du texte, le nœud voisin, la distance du segment et les coordonnées du départ et de l'arrivée. La figure ci-dessus illustre ce type de présentation.

6 - Les mêmes outils qui servent à illustrer l'analyse factorielle se retrouvent ici, permettant d'ajouter des commentaires, de circonscrire des points qu'on veut regrouper, ou de choisir des couleurs en dégradé si on a affaire à des données sérielles dont on veut montrer l'évolution.

CHAPITRE 6

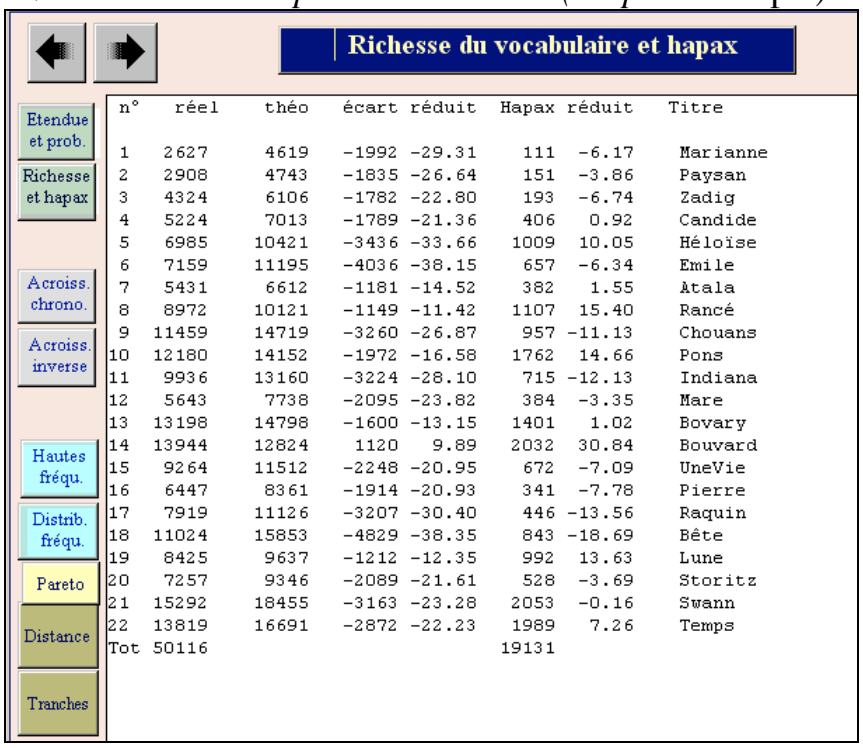
Le menu Distribution

RICHESSE LEXICALE, HAPAX, ACCROISSEMENT

Le programme de préparation, entre autres tâches, constitue le tableau de distribution des classes de fréquences, le relevé des hapax (ou mots employés une seule fois) et bien d'autres résultats qui intéressent la structure du vocabulaire. Pour voir et imprimer ces tableaux, solliciter le bouton DISTRIBUTION qui conduit à une page spécifique où sont consignés les résultats statistiques acquis dans cette perspective. Le bouton RICHESSE fait le dénombrement des formes différentes relevées dans chaque texte. Et en s'appuyant sur le tableau de distribution des fréquences (voir ci-dessous) et sur l'étendue relative des textes, un calcul est exécuté par le programme, qui suit la loi binomiale (méthode de Charles Muller) et mesure la richesse lexicale des sous-ensembles. Voir graphique page suivante.

Un calcul plus classique est appliqué aux hapax, c'est-à-dire aux formes qui ont été rencontrées une seule fois dans le corpus, et conséquemment dans un seul texte. La méthode est ici plus simple et se rattache à la loi normale. On aboutit pareillement à des écarts réduits qui servent d'ordonnées au programme de courbe.

Vocabulaire et hapax. Les données (corpus Example*).*



n°	réel	théo	écart réduit		Hapax réduit		Titre
1	2627	4619	-1992	-29.31	111	-6.17	Marianne
2	2908	4743	-1835	-26.64	151	-3.86	Paysan
3	4324	6106	-1782	-22.80	193	-6.74	Zadig
4	5224	7013	-1789	-21.36	406	0.92	Candide
5	6985	10421	-3436	-33.66	1009	10.05	Héloïse
6	7159	11195	-4036	-38.15	657	-6.34	Emile
7	5431	6612	-1181	-14.52	382	1.55	Atala
8	8972	10121	-1149	-11.42	1107	15.40	Rancé
9	11459	14719	-3260	-26.87	957	-11.13	Chouans
10	12180	14152	-1972	-16.58	1762	14.66	Pons
11	9936	13160	-3224	-28.10	715	-12.13	Indiana
12	5643	7738	-2095	-23.82	384	-3.35	Mare
13	13198	14798	-1600	-13.15	1401	1.02	Bovary
14	13944	12824	1120	9.89	2032	30.84	Bouvard
15	9264	11512	-2248	-20.95	672	-7.09	UneVie
16	6447	8361	-1914	-20.93	341	-7.78	Pierre
17	7919	11126	-3207	-30.40	446	-13.56	Raquin
18	11024	15853	-4829	-38.35	843	-18.69	Bête
19	8425	9637	-1212	-12.35	992	13.63	Lune
20	7257	9346	-2089	-21.61	528	-3.69	Storitz
21	15292	18455	-3163	-23.28	2053	-0.16	Swann
22	13819	16691	-2872	-22.23	1989	7.26	Temps
Tot	50116				19131		

Histogramme de la richesse lexicale (corpus Example*)*

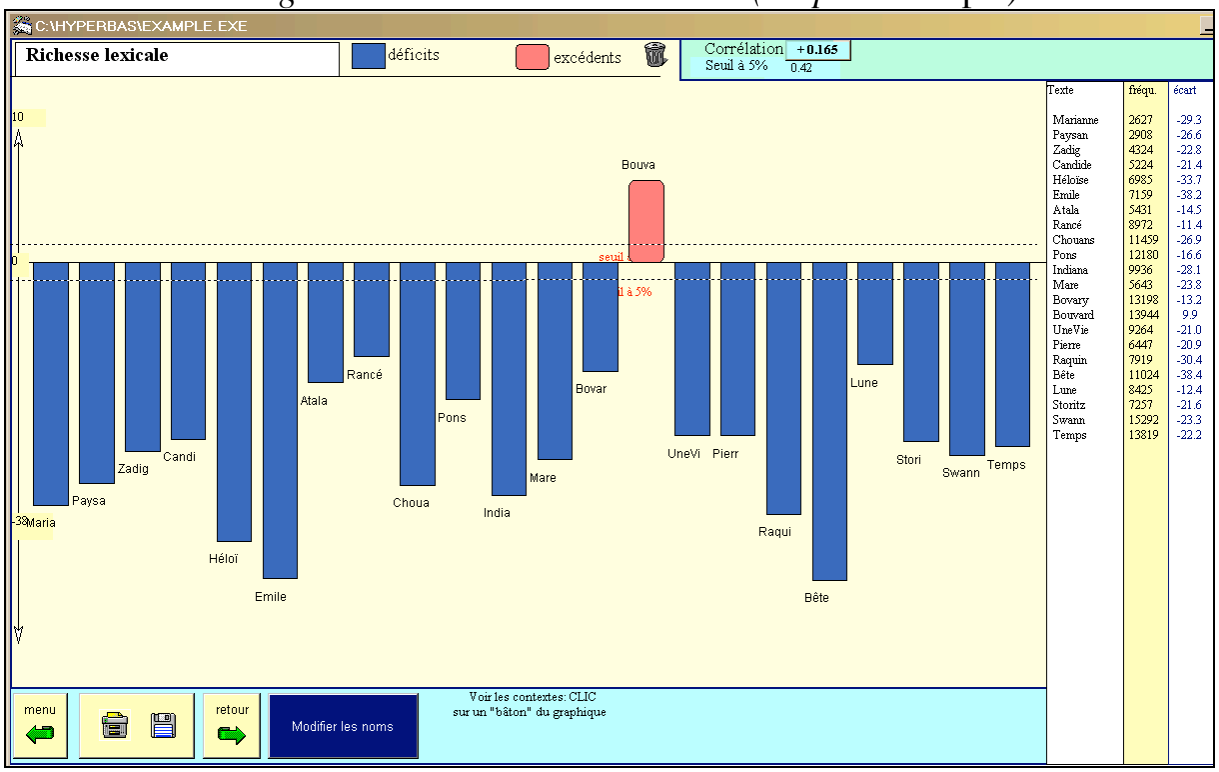
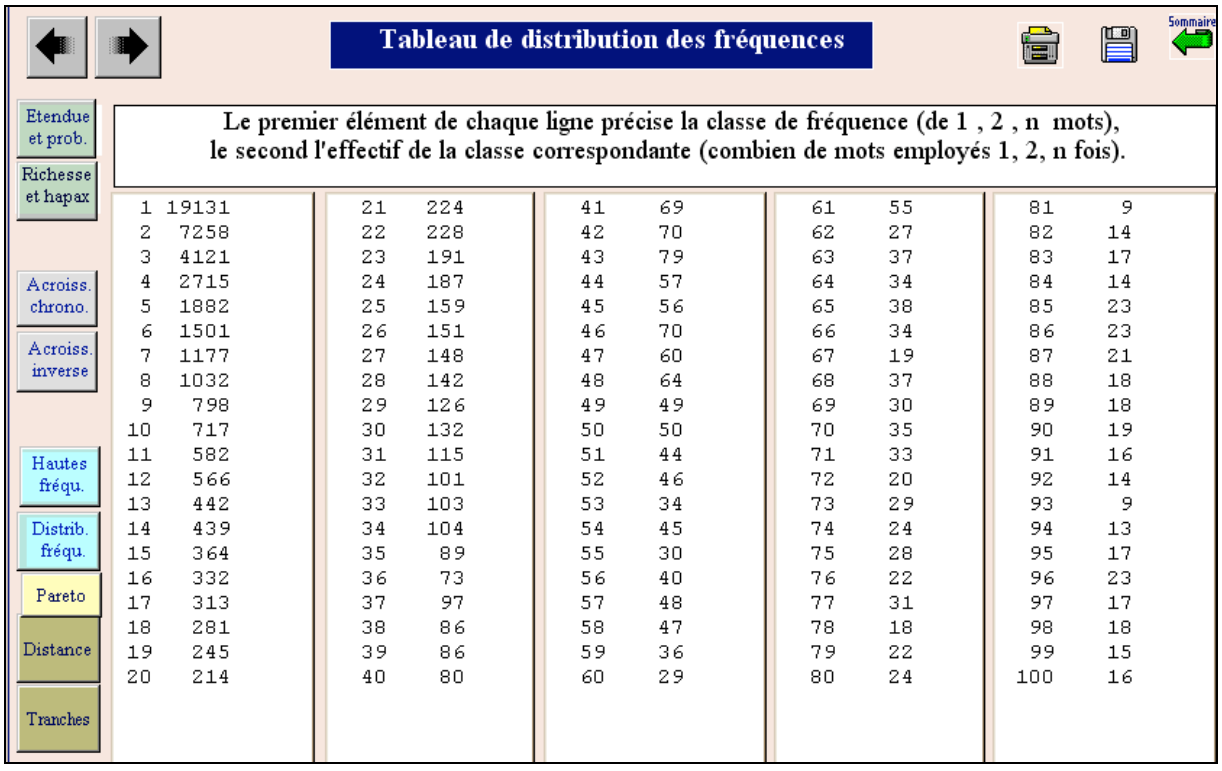


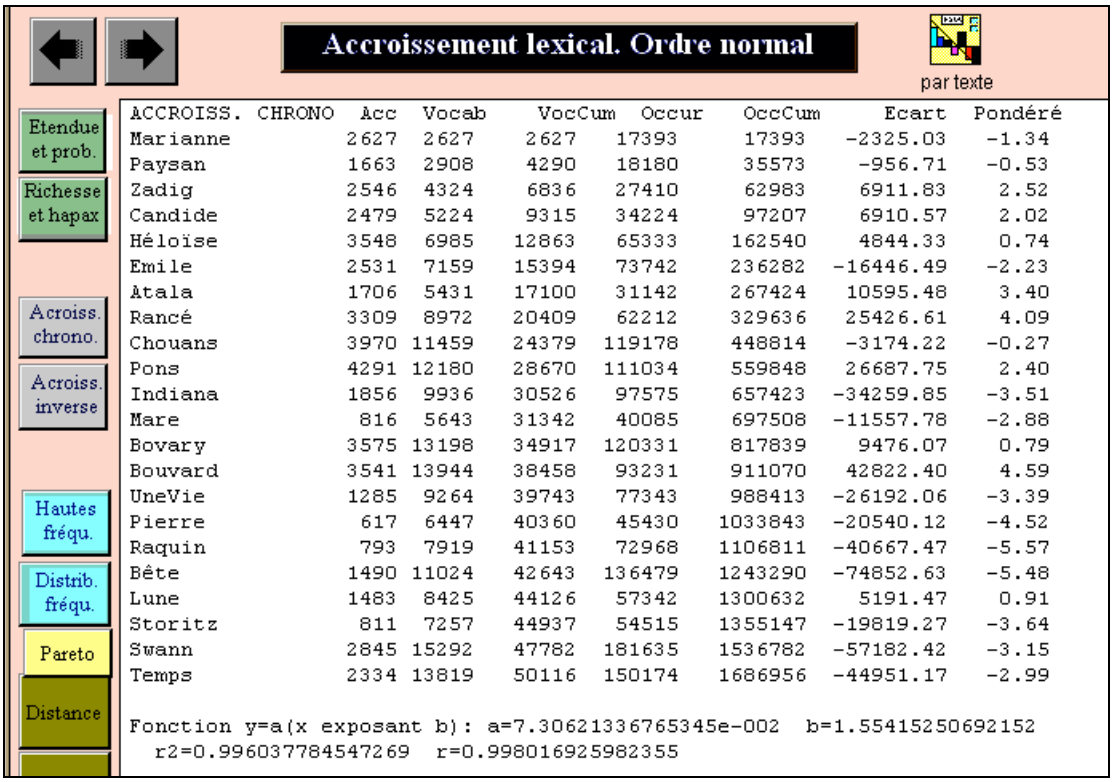
Tableau de distribution des fréquences



C'est par contre une approximation qui rend compte au mieux de l'accroissement du vocabulaire (par un ajustement de fonction puissance, selon la formule : $y = ax^b$). Cette fois la visée est dynamique, puisqu'on évalue le cumul progressif des formes et le renouvellement de plus en plus ralenti du

vocabulaire. La direction naturelle est celle qui suit la chronologie mais le chemin inverse qui prend le temps à rebours peut révéler des ruptures également intéressantes. Les deux trajets sont empruntés successivement et donnent lieu à deux boutons (CHRONO et INVERSE), à deux tableaux et à deux graphiques.

L'accroissement lexical. Données et calcul.



ACCROISS. CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
Marianne	2627	2627	2627	17393	17393	-2325.03	-1.34
Paysan	1663	2908	4290	18180	35573	-956.71	-0.53
Zadig	2546	4324	6836	27410	62983	6911.83	2.52
Candide	2479	5224	9315	34224	97207	6910.57	2.02
Héloise	3548	6985	12863	65333	162540	4844.33	0.74
Emile	2531	7159	15394	73742	236282	-16446.49	-2.23
Atala	1706	5431	17100	31142	267424	10595.48	3.40
Rancé	3309	8972	20409	62212	329636	25426.61	4.09
Chouans	3970	11459	24379	119178	448814	-3174.22	-0.27
Pons	4291	12180	28670	111034	559848	26687.75	2.40
Indiana	1856	9936	30526	97575	657423	-34259.85	-3.51
Mare	816	5643	31342	40085	697508	-11557.78	-2.88
Bovary	3575	13198	34917	120331	817839	9476.07	0.79
Bouvard	3541	13944	38458	93231	911070	42822.40	4.59
UneVie	1285	9264	39743	77343	988413	-26192.06	-3.39
Pierre	617	6447	40360	45430	1033843	-20540.12	-4.52
Raquin	793	7919	41153	72968	1106811	-40667.47	-5.57
Bête	1490	11024	42643	136479	1243290	-74852.63	-5.48
Lune	1483	8425	44126	57342	1300632	5191.47	0.91
Storitz	811	7257	44937	54515	1355147	-19819.27	-3.64
Swann	2845	15292	47782	181635	1536782	-57182.42	-3.15
Temps	2334	13819	50116	150174	1686956	-44951.17	-2.99

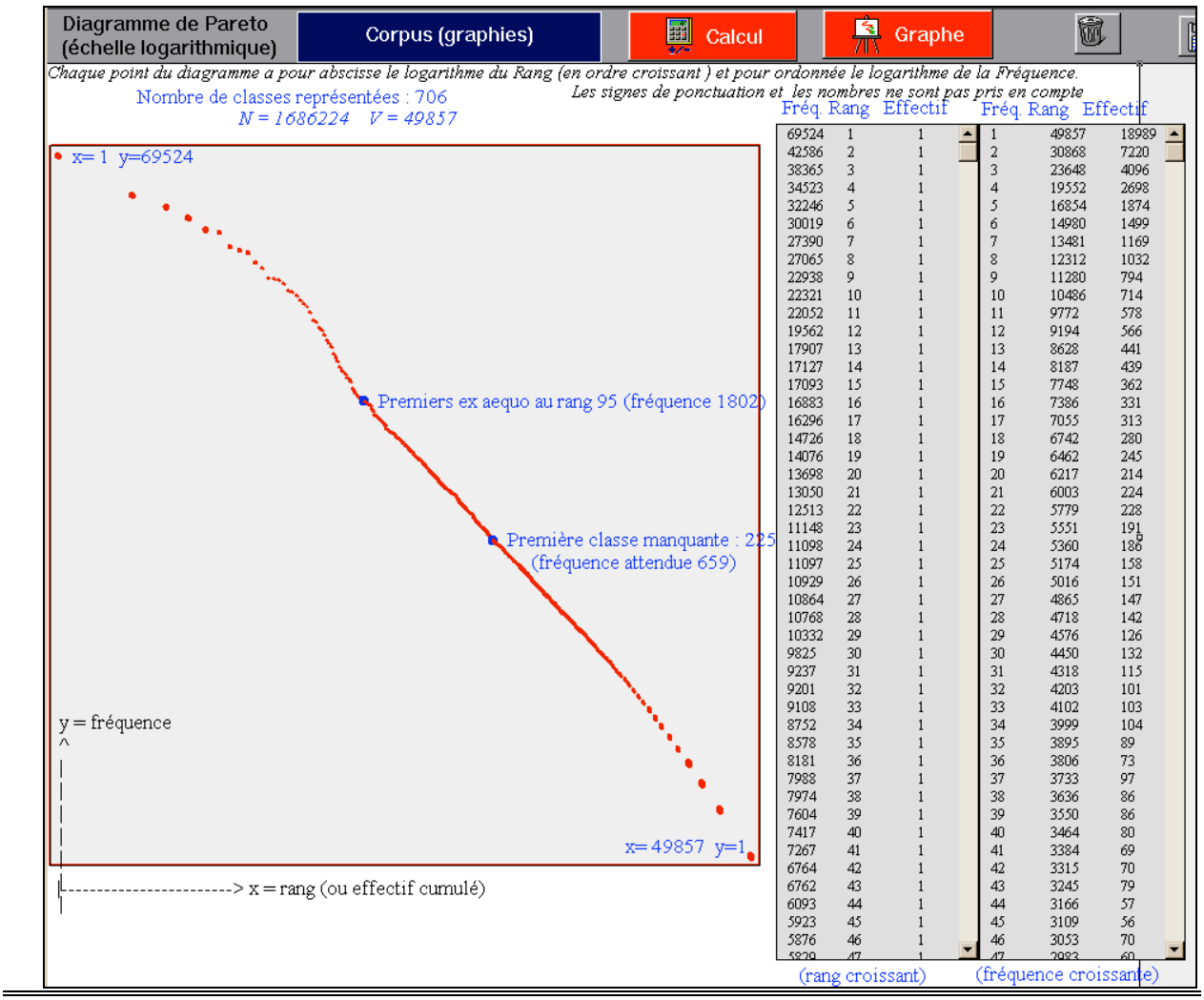
Fonction y=a(x exposant b): a=7.30621336765345e-002 b=1.55415250692152
r2=0.996037784547269 r=0.998016925982355

LE DIAGRAMME DE PARETO

La fameuse loi de Zipf a souvent été appliquée au domaine lexical, quoiqu'elle rende compte aussi de beaucoup d'autres phénomènes et plus généralement de toutes les distributions "à longue queue", comme le classement de la population selon le revenu ou le relevé des ventes en librairie. La fréquence des mots relève de cette répartition inégalitaire où peu ont beaucoup et beaucoup ont peu.

Le programme range la population des mots par classes de fréquence, et les dispose par rang croissant (les hautes fréquences en tête) et le produit du rang par la fréquence tend vers une constante, ce qui produit une droite quand rang et fréquence sont portés sur une échelle logarithmique. En réalité une légère distorsion apparaît aux deux extrémités de la courbe et le dessin de la droite n'est pur qu'entre deux points d'inflexion remarquables: celui où, venant du haut, on observe les premiers ex æquo et celui où, venant du bas, on bute sur la première classe manquante. Le graphique est proposé pour le corpus et chacun des textes.

Le diagramme de Pareto établi sur le corpus Example



LA CONNEXION LEXICALE (ou distance intertextuelle)

1 - Dans une première approche Hyperbase suit la **méthode Jaccard** qui ne se préoccupe pas de fréquence et pour un mot donné ne considère que sa présence - ou son absence - dans le texte considéré. Ou plus exactement, pour deux textes dont on cherche à apprécier la connexion, un mot contribue à rapprocher ces deux textes s'il est commun aux deux et à augmenter la distance s'il est privatif et ne se rencontre que dans un seul. La collection des données est assez lourde parce qu'il faut considérer tous les mots sans exception et que pour chacun on doit prendre en compte tous les appariements de textes deux à deux (le nombre des confrontations pour *n* textes étant égal à *n* * (*n*-1) / 2). Elle est réalisée dans la phase d'indexation et le résultat auquel on aboutit est délivré par le bouton DISTANCE de la page DISTRIBUTION.

Pour chaque paire considérée, la distance obtenue tient compte de l'étendue de l'un et l'autre vocabulaires, selon la formule: *d = ((a-ab)/a) + ((b-ab)/b)*, où *ab* désigne la partie commune aux vocabulaires *a* et *b* (*a-ab* et *b-ab* recouvrant les parties privatives). C'est cette distance que montre le tableau dans sa partie supérieure, les éléments du calcul (parties communes et privatives) étant détaillés dans la suite (du moins lorsque la place est suffisante).

Tableau de la distance lexicale des textes pris deux à deux (corpus GRACQ)

	ARGO	TENE	GRAN	PECH	BRET	SYRT	PENT	FORE	PREF	LETT	ILE	LETT	ETRO	ECRI	VILL	COLL	CHEM
ARGO	1000	1113	1308	1326	1267	1096	1401	1244	1209	1286	1213	1269	1335	1240	1303	1394	1243
TENE	1113	1000	1174	1163	1195	1005	1246	1153	1109	1168	1134	1168	1245	1166	1236	1308	1168
GRAN	1308	1174	1000	1366	1410	1194	1409	1257	1286	1282	1219	1214	1359	1295	1295	1390	1238
PECH	1326	1163	1366	1000	1289	1149	1213	1301	1193	1277	1330	1290	1407	1266	1365	1427	1276
BRET	1267	1195	1410	1289	1000	1197	1403	1385	1054	1254	1385	1286	1346	1111	1292	1381	1213
SYRT	1096	1005	1194	1149	1197	1000	1238	1073	1136	1181	1063	1184	1221	1197	1220	1309	1170
PENT	1401	1246	1409	1213	1403	1238	1000	1344	1278	1344	1362	1349	1448	1334	1431	1480	1347
FORE	1244	1153	1257	1301	1385	1073	1344	1000	1289	1239	1057	1209	1286	1301	1292	1385	1237
PREF	1209	1109	1286	1193	1054	1136	1278	1289	1000	1147	1290	1181	1247	1053	1209	1284	1146
LETT	1286	1168	1282	1277	1254	1181	1344	1239	1147	1000	1230	1153	1260	1146	1222	1292	1144
ILE	1213	1134	1219	1330	1385	1063	1362	1057	1290	1230	1000	1165	1261	1278	1224	1341	1180
LETT	1269	1168	1214	1290	1286	1184	1349	1209	1181	1153	1165	1000	1179	1159	1154	1214	1096
ETRO	1335	1245	1359	1407	1346	1221	1448	1286	1247	1260	1261	1179	1000	1224	1248	1369	1181
ECRI	1240	1166	1295	1266	1111	1197	1334	1301	1053	1146	1278	1159	1224	1000	1167	1230	1100
VILL	1303	1236	1295	1365	1292	1220	1431	1292	1209	1222	1224	1154	1248	1167	1000	1234	1123
COLL	1394	1308	1390	1427	1381	1309	1480	1385	1284	1292	1341	1214	1369	1230	1234	1000	1194
CHEM	1243	1168	1238	1276	1213	1170	1347	1237	1146	1144	1180	1096	1181	1100	1123	1194	1000

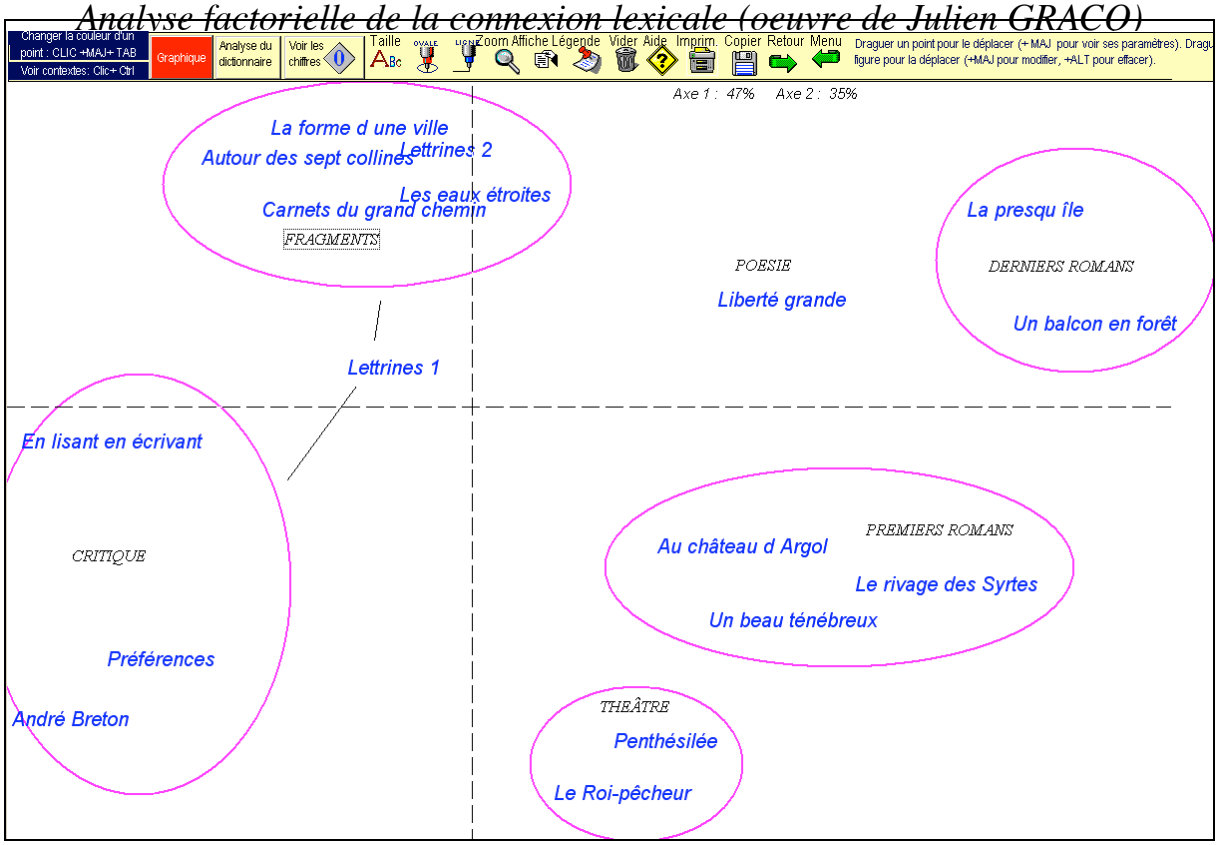
Chacun des deux quotients (dont la somme constitue la mesure de la distance) est le rapport, pour un texte donné, du vocabulaire exclusif au vocabulaire total. Il évolue nécessairement entre 0 et 1. La somme a donc pour limites 0 et 2 et la moyenne 0 et 1[2]. En réalité la somme se situe autour de 1 et reste insensible aux différences d'étendue des deux textes mis en parallèle. Observons en effet que les deux quotients évoluent en sens inverse et d'un même pas, quand s'accroît l'inégalité d'étendue des textes. En une telle situation le plus petit texte aura du mal à affirmer son indépendance face au plus gros, et son quotient d'exclusivité se rapprochera de zéro. Mais pour la même raison, le texte le plus long aura un gros contingent de termes exclusifs qui échapperont par la force des choses au plus petit, et son quotient d'exclusivité tendra vers 1. Au total on observera une neutralisation mutuelle de ces deux mouvements opposés[3]. Quand les distances lexicales sont visibles à l'écran, le programme GRAPHIQUE permet la représentation de la distance variable qu'un texte établit avec tous les autres (il y a donc autant de profils que de textes), tandis que le programme ANALYSE envisage l'ensemble de ces distances et propose une

[2] C'est cette moyenne qui est proposée dans la version la plus récente du logiciel

[3] Le calcul est en réalité un peu plus complexe dans la dernière version d'Hyperbase. Il intègre non seulement les mots commun aux textes A et B et les mots privatifs qui se trouvent dans A sans être dans B et réciproquement, mais aussi les mots du corpus qui ne se trouvent ni dans A, ni dans B. Ces mots pareillement rejetés par les deux textes contribuent dans une certaine mesure à rapprocher, même négativement, les deux textes, puisqu'ils partagent les mêmes répulsions ou les mêmes désintérêts.

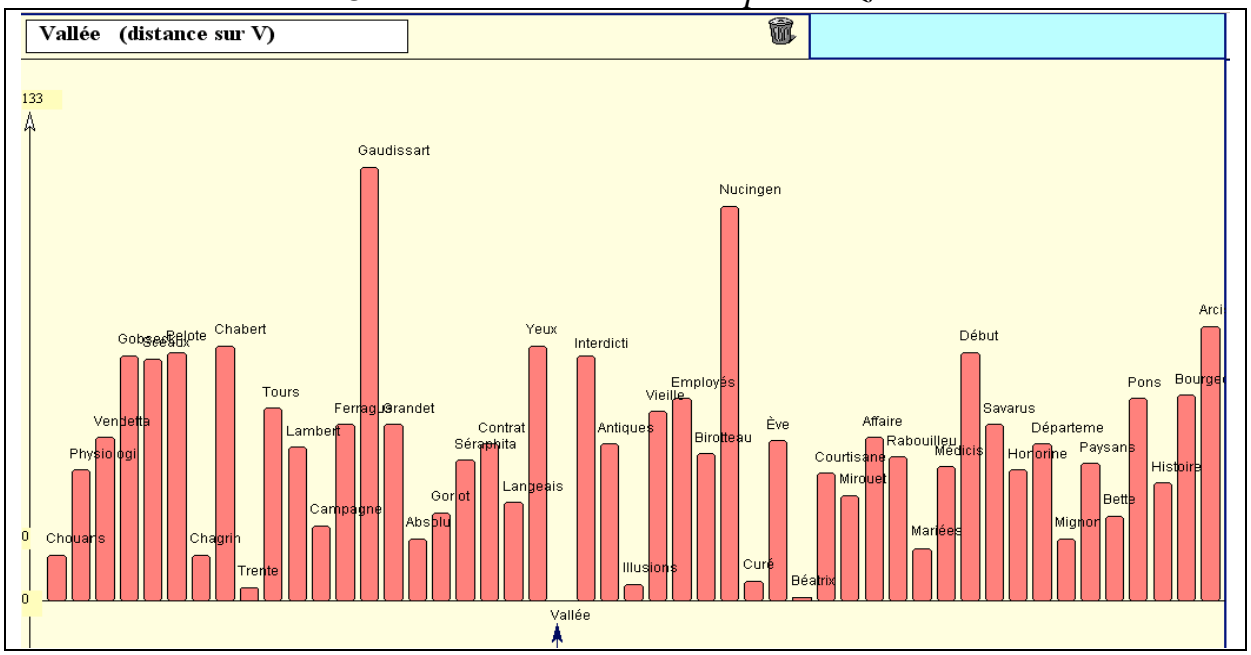
typologie des textes selon ce critère (voir ci-dessous les forces du genre et du temps clairement établies dans l'œuvre de Gracq).

Analyse factorielle de la connexion lexicale (oeuvre de Julien GRACO)



On trouvera ci-dessous une illustration de la distance intertextuelle, empruntée au corpus Balzac. Elle concerne le *Lys dans la Vallée*, dont la thématique est proche de la *Femme de trente ans*, de *Béatrix*, du *Curé de campagne* et des *Illusions perdues*, tous ces textes se situant au bas du graphique, là où la distance est la plus courte.

Connexion lexicale. Le corpus Balzac



D'aucuns ont observé que la méthode Jaccard faisait la part belle aux raretés du vocabulaire et particulièrement aux hapax, au détriment des fréquences plus courantes. Les classes de fréquence élevée perdent ainsi tout poids dans le calcul, puisqu'elles se trouvent nécessairement dans la partie commune et inévitable du vocabulaire (*ab*). Et on avait estimé que la distance ainsi mesurée était surtout sensible aux variations thématiques, les paramètres stylistiques s'attachant plutôt aux mots de fréquence supérieure. Mais ce calcul peut être jugé trop sensible aux artefacts que peuvent produire l'inconstance de l'orthographe, les fautes de frappe, l'abondance des noms propres, bref tous les phénomènes, parfois mineurs et négligeables, qui engendrent la multiplication des formes. Certains considèrent que c'est donner trop d'importance à l'excentricité et qu'une véritable appréciation de la distance entre deux textes doit considérer, pour un même mot, le dosage des fréquences dans les deux textes comparés. Il est évident que si le partage des fréquences est inégal (par exemple 1 occurrence dans le texte A et 19 dans le texte B), il contribue moins à rapprocher les deux textes que si la répartition était équilibrée, soit 10 occurrences dans chacun (en considérant que les deux textes sont de même étendue). Dans les deux cas le calcul précédent rangeait le mot à l'intersection des deux textes, ne tenant compte que de la présence/absence, en ignorant les disparités des fréquences.

2 - Dominique Labbé, bien connu pour ses travaux sur les hommes politiques, principalement Mitterand et de Gaulle, a proposé récemment (au Colloque *JADT 2000*, Lausanne, p. 85-94) un algorithme efficace qui pour chaque mot apprécie la distribution réelle des fréquences dans les deux textes en la comparant non plus à la répartition théorique mais à l'écart maximal possible dans cette distribution:

$D_{(A,B)} = \sum d_i \ / \sum dmax_i$

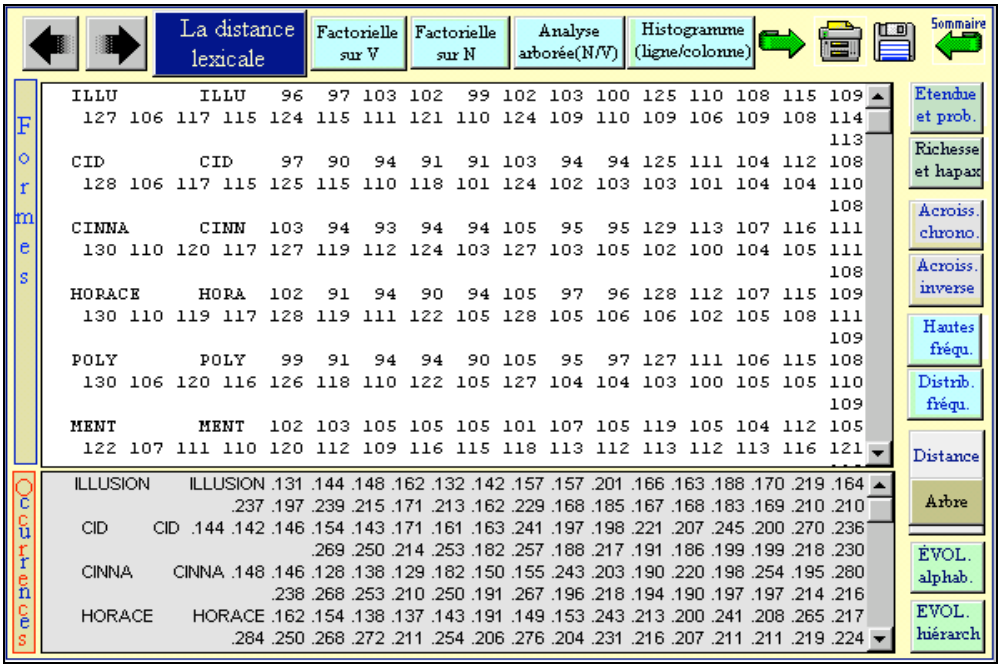
pour i variant du premier au dernier mot du vocabulaire des textes A et B.

Le calcul est simple à programmer mais son exécution est redoutable quand le corpus est de grande taille. Il faut disposer de ce qu'on appelle, avec Salem, le tableau lexical entier, dont le nombre de lignes correspond au nombre de mots et les colonnes aux textes distingués dans le corpus. Dans un grand corpus ces deux dimensions peuvent s'étendre au delà de la mémoire disponible (même au delà des possibilités de la mémoire virtuelle). Pour éviter ce dépassement, on a morcelé les données en paquets cumulatifs, mais pour chaque paquet le calcul est répété autant de fois que l'on compte de combinaisons de textes deux à deux (soit n (n-1)/2 pour un nombre n de textes). Et le temps de traitement s'accroît en conséquence. C'est pourquoi nous avons rendu facultative cette fonction du programme dans Hyperbase. Son déclenchement s'opère dès qu'on sollicite une opération qui met en cause de

telles données dans la page "DISTANCE": soit le bouton *Factorielle sur N*, soit le bouton *Analyse arborée (N/V)*, soit le bouton *Histogramme (ligne/colonne)*. Naturellement ce long calcul ne se fait qu'une fois et ses résultats sont enregistrés dès qu'ils sont acquis (si la base est sur un support inscriptible). Quant à la distance précédemment établie sur V, elle est disponible dès que la base est constituée et aucun calcul complémentaire n'est nécessaire, lorsqu'on lance sur de telles données un programme de représentation graphique.

On remarquera que ces deux tableaux de distance (incomplètement visibles dans l'écran qui suit) sont en réalité des matrices carrées symétriques, les valeurs se reflétant en miroir de chaque côté de la diagonale principale (car la distance de A vers B est la même que de B vers A). Quant à la diagonale principale, elle devrait ne comporter que la valeur 0; la distance d'un texte à lui-même est nécessairement nulle, même en cas d'autodérision. Si pourtant une valeur non nulle s'y trouve, c'est par un artifice nécessité par le traitement des profils. Dans un tel traitement, les valeurs nulles de la diagonale ne devraient avoir aucune importance. Elles en ont pourtant dans les calculs multidimensionnels (analyse de correspondance, ou analyse arborée), et leur influence est d'autant plus nuisible que leur valeur s'écarte plus fortement des autres éléments de la même ligne (ou de la même colonne). C'est pourquoi on a porté sur la diagonale une valeur qui correspond au minimum de la série en question, en admettant que la distance de soi à soi est inférieure ou au plus égale à toute autre distance observée dans la série.

Calcul de la distance lexicale, sur V et sur N. Les relevés.

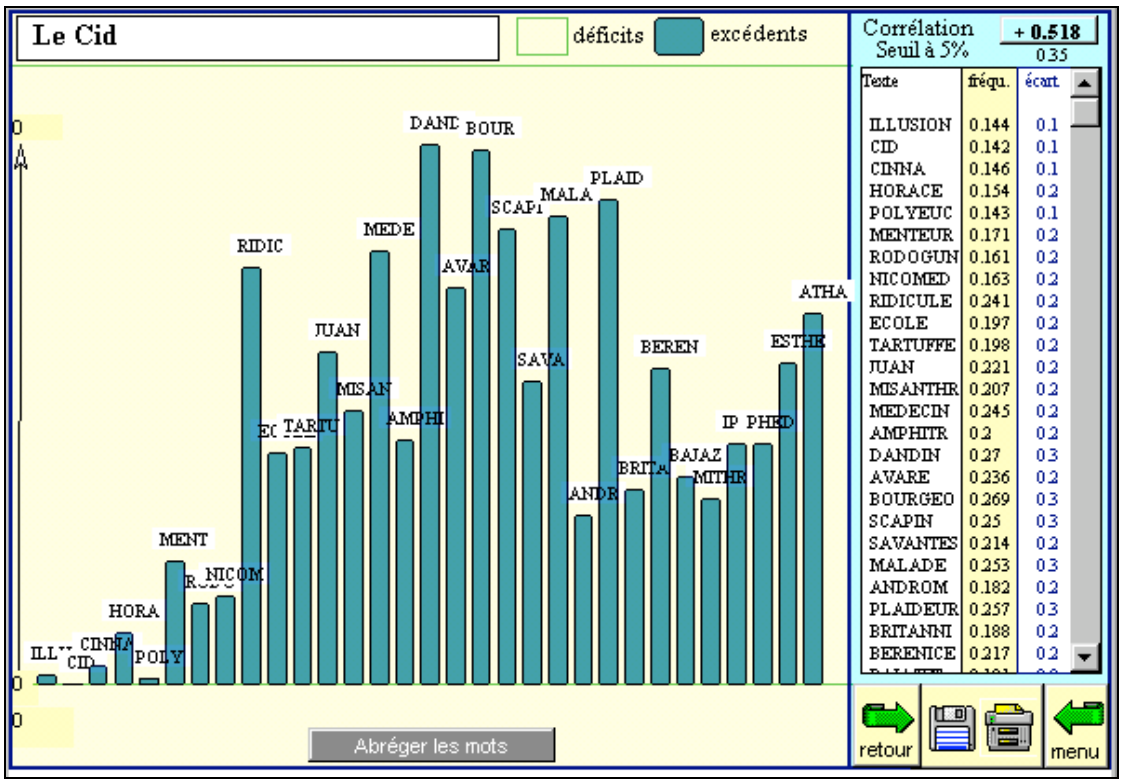


Les résultats peuvent être représentés graphiquement, texte par texte, sous forme d'histogramme. Chaque bâton de la figure a une longueur proportionnelle à la distance qui sépare le texte correspondant du texte pris pour pôle d'observation. Voir l'exemple du *Cid* ci-dessous.

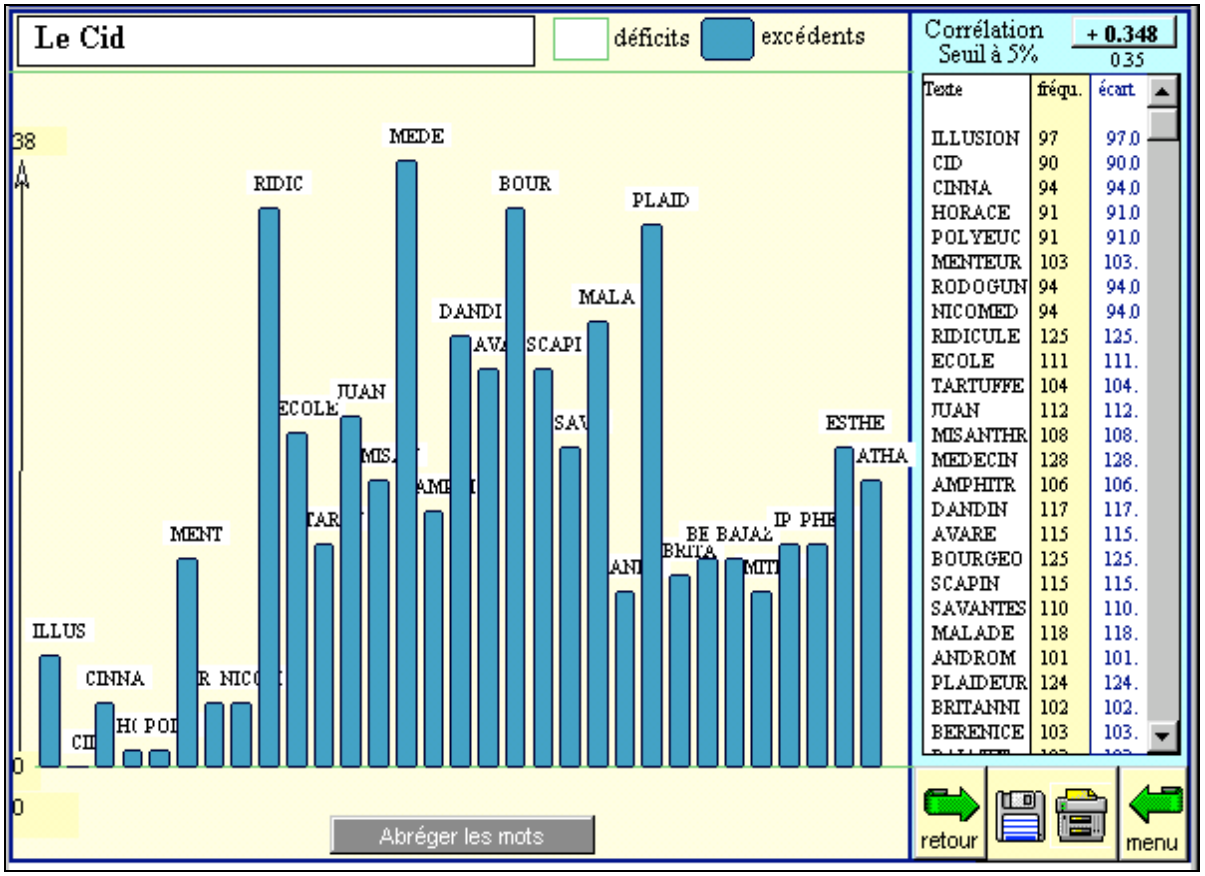
La distance reste faible tant qu'on a affaire à une pièce de Corneille, même s'il s'agit d'un genre différent (dont relève par exemple le *Menteur*). Elle est plus longue quand la pièce est de Racine et cela d'autant plus que l'on s'éloigne dans le temps, d'*Andromaque* à *Athalie*. Mais la distance s'accroît encore quand les deux paramètres genre et auteur changent en même temps, ce qui est le cas des *Plaideurs* de Racine et des textes de Molière.

Est-ce que le portrait diffère beaucoup lorsqu'on établit la distance sur les seules formes, en ignorant leurs fréquences? On prendra la mesure de l'écart en considérant la deuxième figure qui reprend l'exemple du *Cid*. Il faut reconnaître qu'en l'occurrence l'écart est faible et que les mêmes lignes de force s'y dessinent. C'est généralement le cas, quand les corpus atteignent une taille suffisante. Car l'expérience montre que, quel que soit l'objet choisi: formes brutes, formes étiquetées, lemmes véritables, avec ou sans considération de fréquences, l'étude d'ensemble du vocabulaire montre les mêmes influences s'exerçant dans les textes dans le même sens et avec la même intensité.

Le Cid *de Corneille. Distances d'autres textes du corpus* Théâtre classique
(le calcul est établi sur les **occurrences**, *selon la méthode de Labbé)*



Le Cid *de Corneille*. *Distances des autres textes du corpus* Théâtre classique
(le calcul est établi sur les **formes** *sans considération de fréquence)*



Toute une galerie de portraits peut ainsi être dressée en enfilade, avec les 31 textes du corpus. Les conclusions seront très proches d'un portrait à l'autre, chacun portant témoignage d'un air de famille commun à toutes les pièces d'un même auteur. Le genre n'apparaît qu'en second lieu comme principe discriminant dans ce corpus et son influence arrive à contrebalancer le facteur premier dans le cas du *Menteur* ou des *Plaideurs*. Mais nous nous trouvons dans une situation où les deux critères ne s'opposent guère, les trois auteurs en question n'ayant guère cultivé qu'un seul genre littéraire. Dans d'autres corpus le genre apparaît souvent comme l'élément le plus important le la classification.

3 - La connexion lexicale de Charles Muller

La méthode, réputée complexe, est pourtant clairement établie, dès 1968, dans *l'Initition à la statistique linguistique* (Larousse). Nous y renvoyons le lecteur:

La connexion lexicale selon la loi binomiale (Initiation, p. 211)

Connaissant l'étendue respective des deux textes A et B, qui est N_a et N_b, on calculera aisément la probabilité pour qu'une occurrence prise au hasard soit dans A ou dans B ; nous appellerons conventionnellement la première p et la seconde q, pour retrouver des notations déjà employées :

$$p = \frac{N_a}{N_a + N_b} \qquad q = \frac{N_b}{N_a + N_b} \qquad p + q = 1 .$$

Il suffira ensuite d'appliquer les développements du binôme $(p + q)^f$ pour construire un modèle ; on appellera f la fréquence du vocable dans l'ensemble, et V_f l'effectif qui lui est associé ; f'_a et f'_b les sous-fréquences dans les deux textes, et V'_{fa}, V'_{fb} leurs effectifs.

La probabilité, pour un vocable de fréquence f, d'avoir les sous-fréquences 0, 1, 2... f dans l'un ou l'autre des textes est alors :

	Probabilité d'une sous-fréquence dans A					id. dans B				
f	0	1	2	3	4...	0	1	2	3	4...
1	q	p	0	0	0	p	q	0	0	0
2	q^2	$2pq$	p^2	0	0	p^2	$2pq$	q^2	0	0
3	q^3	$3pq^2$	$3p^2q$	p^3	0	p^3	$3p^2q$	$3pq^2$	q^3	0
4	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4
etc.										

Selon la démarche habituelle chez Muller, la méthode conduit d'abord à un modèle, puis à un relevé des faits dans le texte et enfin à un écart entre le modèle et l'observation. Prenons pour exemple les deux dernières pièces de Corneille, *Pulchérie* et *Suréna*. Leur taille est à peu près la même, soit 19235 et 19148 mots. Les probabilités p et q sont donc très voisines p = 0.5011 et q = 0.4989. On trouvera dans le tableau ci-dessous les effectifs théoriques auxquels conduit la loi binomiale, sachant que les totaux pour chaque classe de fréquence sont les suivants :

fréquence 1 : 827, fréquence 2: 329, fréquence 3 : 206, fréquence 4 : 132, fréquence 5 : 92, fréquence 6 : 82, fréquence 7 : 50, fréquence 8 : 47, fréquence 9 : 45.

On vérifiera que le total de la ligne 9 des tableaux réel et théorique est bien ce qu'il doit être : 45. Le tableau des écarts confirme la tendance des textes à s'arroger une part privative plus grande que ne le ferait un hasard impartial, ce que Muller appelle la spécialisation lexicale. Dans la zone centrale du tableau, celle où l'on partage, les écarts sont négatifs, tandis que sur les marges, là où règne l'exclusivité, les écarts sont plus souvent positifs. Bien entendu les écarts absolus doivent être convertis en CHI2, et comptabilisés dès que l'effectif théorique atteint ou dépasse la valeur 5. On voit que la dernière ligne qui s'arrête

à la fréquence 9 n'épuise pas le calcul puisque pour celle classe l'effectif théorique de certaines cases dépasse encore 10. Il a donc fallu pousser plus loin la chaîne des calculs sans s'effrayer si les probabilités p et q ont des exposants terrifiants (en réalité nul besoin de recourir à l'exponentiation : on passe d'une classe à l'autre en ne recourant qu'à la multiplication). On a donc poursuivi le calcul jusqu'à la classe 50. Lorsque l'effectif théorique n'atteint pas le seuil, on procède au regroupement des effectifs trop minces. Reste à totaliser les CHI2 partiels et à les confronter aux degrés de liberté. Pour l'exemple ci-dessous, pour un ddl = 36, on obtient un CHI2 total de 93 et , lorsque le calcul s'étend jusqu'à la classe 50, le CHI2 s'établit à 360 pour un ddl de 48. Dans les deux cas, que le calcul soit limité ou étendu, la valeur du CHI2 laisse une chance infinitésimale au hasard: les deux pièces puisent dans des zones distinctes du lexique. Et il en est ainsi de toutes les pièces de Corneille, confrontées deux à deux. Mais ce qui importe n'est pas de prouver la spécialisation lexicale, mais de la mesurer et de se servir de cette mesure pour établir une distance entre les textes.

Calcul de la connexion lexicale pour les basses fréquences

```
Probabilité p  et q 0.501133314227653    0.498866685772347

Tableau theorique
   0.00
 414.44 412.56
  82.62 164.50  81.88
  25.93  77.42  77.07  25.58
   8.33  33.15  49.50  32.85   8.18
   2.91  14.47  28.81  28.68  14.28   2.84
   1.30   7.76  19.31  25.62  19.13   7.62   1.26
   0.40   2.77   8.26  13.70  13.64   8.15   2.70   0.38
   0.19   1.49   5.19  10.33  12.85  10.23   5.09   1.45   0.18
   0.09   0.80   3.20   7.43  11.10  11.05   7.33   3.13   0.78   0.09

Tableau réel
   0.00
 435.00 392.00
  86.00 144.00  99.00
  28.00  68.00  62.00  48.00
   8.00  39.00  37.00  31.00  17.00
   7.00  10.00  22.00  23.00  19.00  11.00
   1.00   9.00  19.00  19.00  18.00  10.00   6.00
   3.00   4.00   5.00  12.00   9.00  11.00   5.00   1.00
   2.00   2.00   4.00  12.00  10.00   9.00   4.00   0.00   4.00
   2.00   1.00   5.00   6.00   4.00  14.00   7.00   2.00   3.00   1.00

Tableau des écarts
   0.00
  20.56 -20.56
   3.38 -20.50  17.12
   2.07  -9.42 -15.07  22.42
  -0.33   5.85 -12.50  -1.85   8.82
   4.09  -4.47  -6.81  -5.68   4.72   8.16
  -0.30   1.24  -0.31  -6.62  -1.13   2.38   4.74
   2.60   1.23  -3.26  -1.70  -4.64   2.85   2.30   0.62
   1.81   0.51  -1.19   1.67  -2.85  -1.23  -1.09  -1.45   3.82
   1.91   0.20   1.80  -1.43  -7.10   2.95  -0.33  -1.13   2.22   0.91
```

Tableau des CHI2

```
    .00
   1.02                                    1.02
    .14   2.55  3.58
    .17   1.15  2.95  19.66
    .01   1.03  3.16    .10  9.53
    .00   1.38  1.61   1.13  1.56  .00
    .00    .20   .00   1.71   .07  .74   .00
    .00    .00  1.29    .21  1.58  1.00  .00  .00
    .00    .00   .27    .27   .63   .15  .23  .00  .00
    .00    .00   .00    .28  4.54   .79  .02  .00  .00  .00
```

Reste à prolonger le calcul dans les classes de fréquence supérieures. À partir de ce point, on pourra traiter les vocables non plus par classes de fréquence, mais isolément, et le modèle deviendra très simple : fa' = pf fb'= qf, chaque mot donnant lieu au calcul d'un CHI2 partiel qui s'ajoute à tous les autres, tandis que les degrés de liberté s'accroissent d'une unité. Pour le couple *Pulchérie-Suréna*, le résultat est un CHI2 de 250 pour un ddl de 97. On l'ajoute au résultat antérieur pour obtenir en fin de compte :

CHI2 = 610 ddl = 145

Mais comme les résultats doivent se lire sur une table à des endroits différents, il est préférable de faire la conversion en un écart réduit dont l'interprétation est directe. Là encore Muller propose la formule de conversion:

$$z = \sqrt{(2 * chi2)} - \sqrt{((2 * ddl) - 1)}$$

Les résultats - CHI2 ou écarts réduits- appartiennent à l'univers probabiliste, dans lequel on mesure les chances d'obtenir les faits constatés. La valeur de l'écart observé est prise en compte mais aussi la taille des observations. Pour un même écart en pourcentage, par exemple 10% en plus ou en moins, le CHI2 aura des valeurs très différentes si l'effectif considéré est de l'ordre des centaines, des milliers ou des millions. Il y a donc lieu de procéder à une légère pondération qui fait entrer la taille des deux textes considérés dans le calcul de leur distance. Après de multiples essais, nous n'avons pas trouvé mieux que la racine carrée de leur vocabulaire.

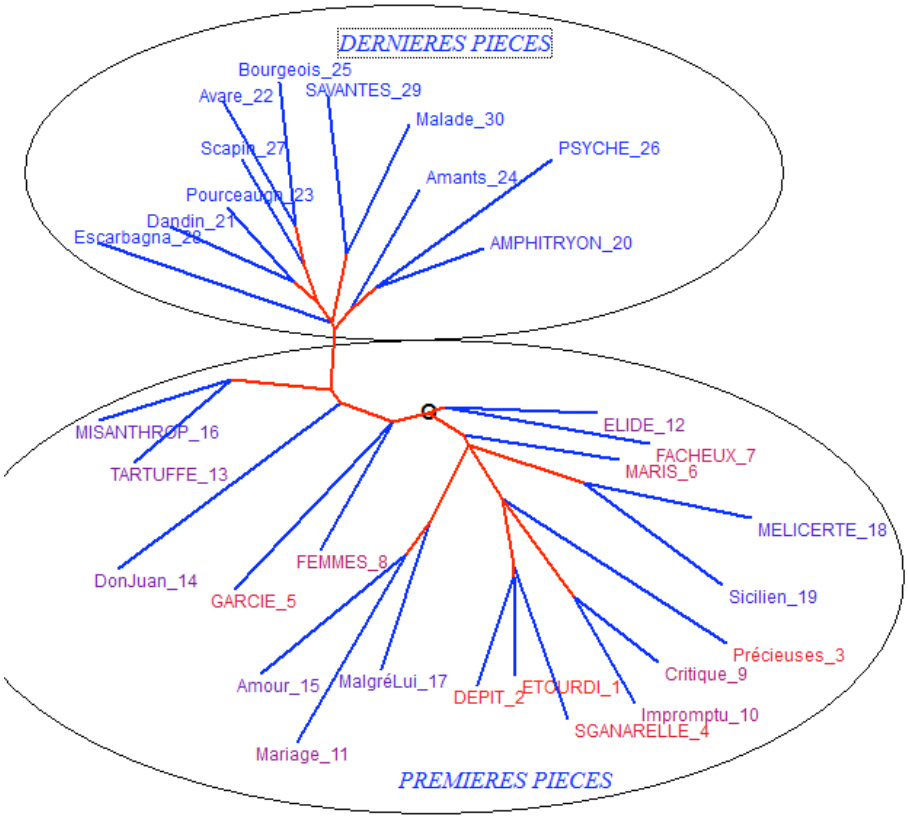
Hautes et basses fréquences

Dans la connexion lexicale on ne sait trop quels facteurs agissent. Le choix des mots est gouverné par diverses influences qui se combattent ou s'appuient : le genre, l'époque, le sujet. Un indice global ne peut que les mêler sans permettre la décantation. Nous avons donc introduit dans le calcul un filtre qui aide à isoler les basses et les hautes fréquences, et peut-être, à travers cette distinction, les facteurs visés. Nous prendrons pour exemple cette fois le corpus de Molière. La connexion lexicale ainsi livrée au spectroscope est décomposée en deux coupes. La première ne tient compte que des fréquences basses (de 1 à 50)[4]. La seconde ne s'intéresse qu'aux autres fréquences Dans les deux cas l'interprétation est aisée, mais la convergence n'est pas au rendez-vous.

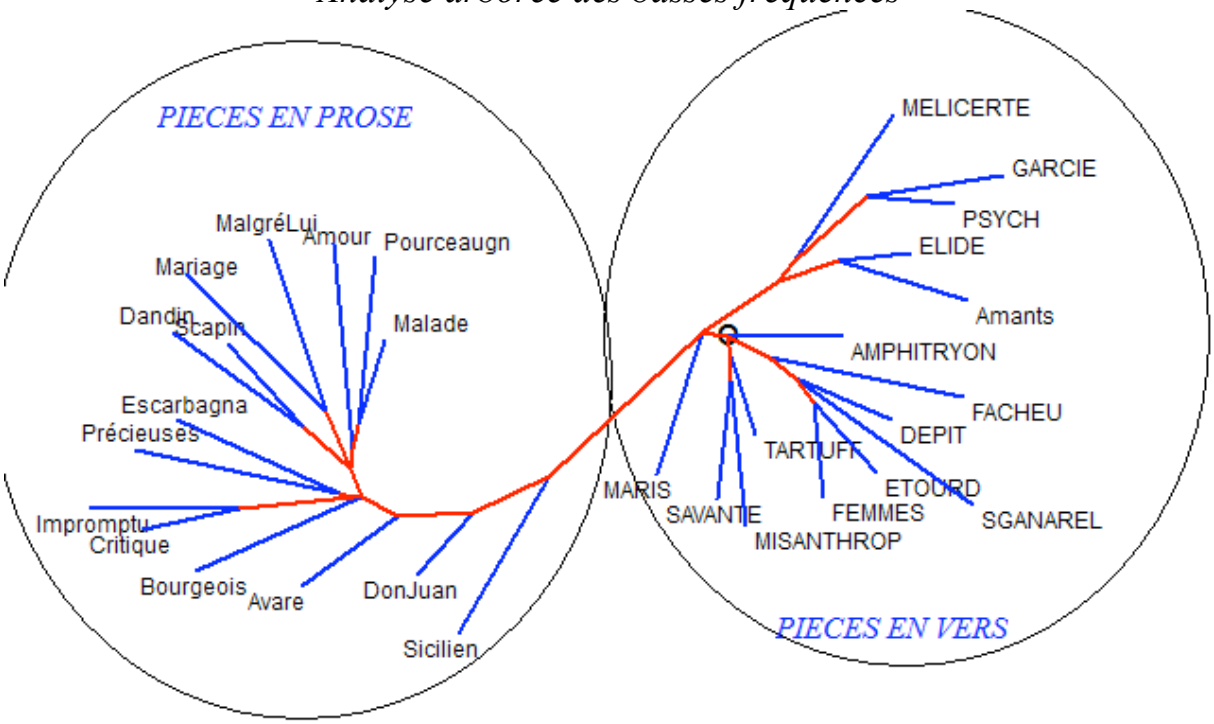
Dans la première figure deux sous-genres se font face. Les pièces en vers (leur nom est en majuscules) ne se compromettent pas avec les pièces en prose (en minuscules). Il n'y a qu'une exception, facilement explicable : si les *Amants magnifiques* se retrouvent dans le camp noble du vers, c'est parce que ce "divertissement royal" est rempli d'intermèdes versifiés. Dans la seconde figure au contraire les majuscules se mêlent aux minuscules et les vers à la prose. Un autre regroupement s'y manifeste qui suit la chronologie : les onze dernières pièces occupent le haut de la figure, les premières se groupent en bas.

[4] Cette limite n'intéresse pas la fréquence d'un mot dans le corpus, mais celle d'un mot dans l'assemblage de deux textes, comme expliqué précédemment.

Analyse arborée des hautes fréquences



Analyse arborée des basses fréquences



La conclusion, confirmée par d'autres monographies (par exemple Balzac, Verne, Zola, Proust, Anatole France), est que le choix du genre et du sujet impose davantage sa loi dans les basses fréquences : il y a des mots qui n'ont pas leur place dans une pièce en vers, d'autres qui sont permis dans une comédie mais non dans une tragédie. La poésie a son univers lexical qui n'est pas celui de

la correspondance ou du récit, etc. Ces interdits et ces privilèges concernent moins directement les mots fréquents et encore moins les mots-outils parce que leur emploi est inévitable dans tout discours et que l'ostracisme est plus difficile à leur endroit[5]. Mais les fréquences hautes n'en sont pas moins animées de mouvements qui paraissent plus lents mais plus profonds et qui décrivent sourdement l'évolution de l'écriture. Ces mouvements de fond sont sans doute moins conscients ou moins volontaires que les choix clairs que l'écrivain fait parmi les genres et les sujets. Plus stylistiques que thématiques, ils sont davantage le reflet de la structure que du contenu.

Reste à comparer la connexion lexicale aux deux autres méthodes, celle de Jaccard et celle de Labbé. Cette dernière, que nous avons intégrée depuis longtemps à notre logiciel Hyperbase, tient compte, comme la méthode binomiale, des faits de fréquence. Et, comme la distance de Jaccard, elle se réduit à enregistrer des quotients ou des rapports, en dehors des lois probabilistes. Comme le vote de chaque mot est individuel, le poids de chacun tend à être le même, ce qui affaiblit l'influence des mots fréquents et puissants, perdus dans la foule des petits. C'est pourquoi, même si son auteur écarte certains mots rares, et notamment les hapax du texte le plus long, et si d'autres précautions sont prises quand l'écart est trop faible, le test de Labbé privilégie les basses fréquences et dans les faits son témoignage s'accorde très souvent avec la méthode Jaccard, sans permettre d'isoler ce qui tient aux basses fréquences et ce qui tient aux hautes. Nous avons maintenu cet outil dans notre logiciel, car en multipliant les approches, on peut espérer de leur convergence plus de fiabilité dans les résultats. Mais nous ne cachons pas notre préférence pour la connexion de Muller, en regrettant de n'avoir pas introduit plus tôt cette mesure dans notre logiciel. On craignait que la mémoire ne vienne à manquer pour recueillir les observations ou que le temps du traitement soit prohibitif. Il n'en est rien. Certes la distance de Jaccard est plus rapide à calculer mais celle de Labbé est plus lente.

[5] Il y a cependant des mots-outils extrêmement sensibles à la situation du discours et aux contraintes ou préférences du genre. Les pronoms personnels ou possessifs sont sujets à de fortes variations de cet ordre, comme les démonstratifs, les subordonnants, les relatifs et certaines prépositions.

CHAPITRE 7

Les menus GRAPHIQUE et LISTE

GRAPHIQUES

Le fonction GRAPHIQUE utilise les probabilités générées par la partition du corpus pour établir des écarts et les représenter graphiquement sur un plan. La distribution d'un mot est rarement régulière à travers un corpus et des écarts s'y observent entre la fréquence d'un mot observée dans un texte et la fréquence théorique qu'on était en droit d'attendre, vu la proportion du texte dans l'ensemble, et qui s'établit avec une simple règle de trois (fréquence théorique d'un mot dans un texte = fréquence du mot dans le corpus pondérée par la probabilité *p* ou part du texte dans le corpus). Dans sa forme la plus simple, le calcul pondère cet écart selon la formule de l'"écart réduit" (*q* étant la probabilité complémentaire 1-p):

*z = (réel-théorique) / racine carrée(théorique * q)*

Une fois calculés les écarts réduits, le programme présente une illustration graphique de la distribution, sous forme d'histogramme. Un dialogue s'établit avec l'utilisateur qui doit fournir le mot à étudier (comme ci-dessous le verbe *aimer*).

Les "bâtons" de l'histogramme se répartissent de part et d'autre de la ligne médiane qui représente la valeur 0 de l'écart réduit. Chacun de ces "bâtons" est explicité par le titre du texte correspondant. Si la série représentée se limite à une seule forme, les effectifs absolus sont détaillés sur la marge droite, la colonne voisine détaillant les écarts réduits qui servent d'ordonnées à la représentation graphique. Veiller à contrôler la valeur des écarts réduits, afin de s'assurer qu'on peut ajouter foi aux différences constatées. Si ces valeurs restent à l'intérieur de la plage –2 +2, l'hypothèse nulle ne peut être écartée et les écarts peuvent être considérés comme pouvant s'expliquer par le hasard. Cette précaution s'impose particulièrement lorsque les effectifs sont faibles et qu'on représente la distribution d'un mot peu fréquent. Noter que l'échelle des ordonnées est variable d'un graphique à l'autre (sauf si deux graphiques sont superposés) et que tout l'espace du graphique est occupé, même si l'écart est faible en valeur absolue.

Retouches et variantes

Quand un grand nombre de textes sont représentés dans le même graphique, l'encombrement peut gêner la lisibilité, les zones se recouvrant qui désignent chaque texte. Plusieurs moyens permettent d'y voir clair.

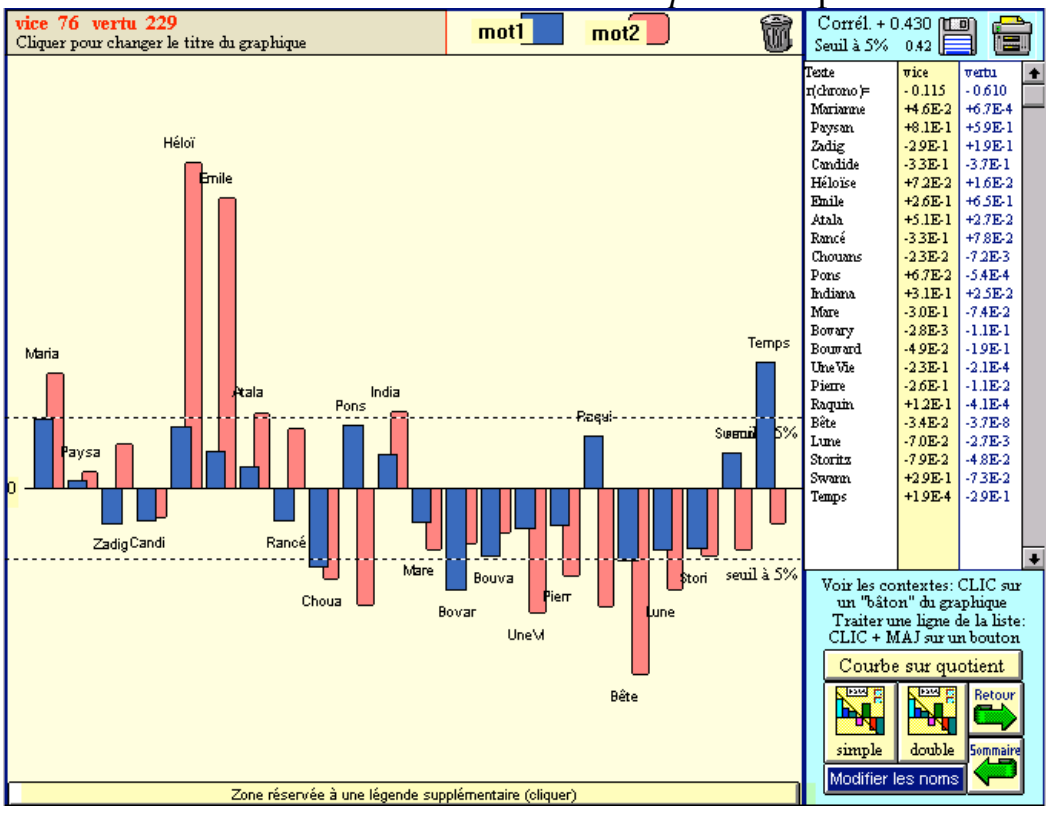
Le premier consiste à effacer ces zones pour ne laisser place qu'aux bâtons de l'histogramme. On activera alors le bouton "*Modifier les noms*", situé en bas de l'écran à droite. Et pour expliciter la lecture, on remplira à sa guise la légende prévue en bas de l'écran et protégée contre l'atteinte des bâtons. Un clic à cet endroit suffit à permettre l'écriture.

Le second artifice est d'accepter les autres propositions du même bouton, soit "*Rétablir noms longs*", soit "*Rétablir noms courts*", soit surtout "*Noms significatifs*". Dans ce dernier cas, les textes situés près de la ligne médiane ne sont pas étiquetés, puisque leur indifférence les situe hors du débat. Seuls sont expressément nommés les textes qui manifestent un goût marqué (ou une répulsion mal dissimulée) à l'endroit du mot (ou de la série) étudié.

Reste enfin la possibilité de faire apparaître ou disparaître alternativement et individuellement le nom de chaque texte représenté, en cliquant sur la zone qui lui est dévolue, au dessus du bâton correspondant, s'il s'agit d'un excédent, au dessous si l'on a affaire à un déficit.

Le bouton DOUBLE est destiné à superposer une seconde distribution à la première et à représenter deux séries sur le même graphique. Si tel est le cas, les deux séries d'écarts réduits seront visibles sur la marge droite de l'écran.

Le vice *et la* vertu *dans le corpus* Example



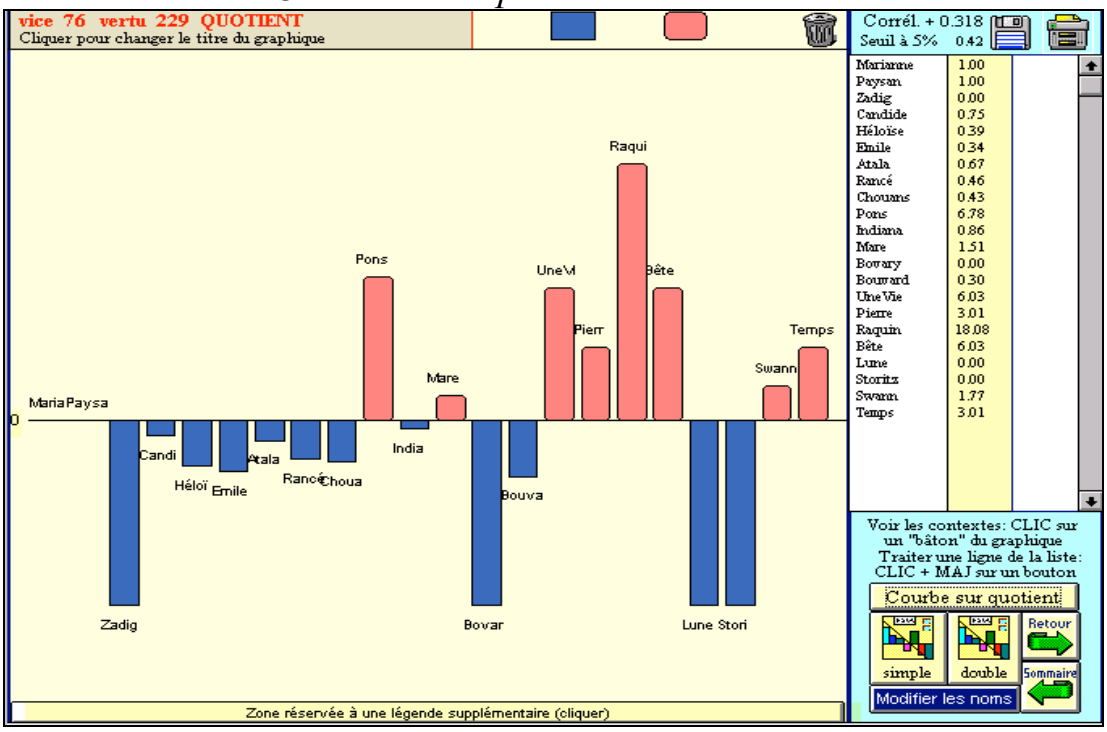
Le symbolisme des couleurs prend alors une autre signification: le bleu est réservé à la première série, le rouge à la seconde. Les étiquettes, si on les conserve, sont attachées à la première. Pour mesurer la force d'attraction mutuelle des deux mots, un calcul de corrélation (c'est le coefficient de Bravais-Pearson) est établi et apparaît en haut et à droite de l'écran. Ce même calcul de corrélation est appliqué aussi à toute série dont on souhaite suivre la distribution à travers le corpus, même si l'on ne compare pas, comme dans le cas présent, deux séries, mais une seule. La deuxième série est alors constituée par le rang des textes échelonnés dans le corpus. On aboutit ainsi à un coefficient de corrélation chronologique (ou sériel) qui mesure la progression ou la régression d'un mot dans la suite des textes. On trouvera ci-dessus un exemple où la *vertu* est mise en parallèle avec le *vice*.

Si deux courbes sont projetées simultanément, deux coefficients apparaissent au haut des deux colonnes réservées aux deux séries, à quoi s'ajoute le troisième coefficient qu'on a évoqué d'abord et qui, mettant en rapport les deux séries, établit leur corrélation mutuelle, en dehors de toute chronologie. Pour une lecture plus facile des résultats, on a indiqué quelle valeur les tables fournissent pour le coefficient de Bravais-Pearson au seuil de 5%. Bien entendu ce seuil est calculé en tenant compte du nombre de paires étudiées.

Remarquons dans l'exemple du *vice* et de la *vertu* que la corrélation entre ces deux éléments est établie, avec un coefficient de 0,43 , juste au dessus du seuil requis (0,42). Cela signifie qu'on retrouve ces deux antonymes dans les mêmes passages, sinon dans les mêmes âmes. On voit en effet que les deux courbes suivent les mêmes inflexions. C'est le sort des antonymes d'être enchaînés l'un à l'autre, bon gré mal gré, dans les mêmes contextes. Dans ces couples désunis qui ne peuvent se détacher l'un de l'autre, comme les deux pôles d'un même aimant, on peut cependant établir une distinction. Ainsi dans l'exemple considéré, le *vice* résiste mieux que la *vertu* à la désaffection qui frappe les notions morales dans notre littérature: le déclin de la *vertu* est mis en évidence par un coefficient très significatif (-0.61), alors que la courbe du vice est hésitante (-0.11).

Il peut arriver qu'on souhaite non pas souligner le parallélisme de deux termes mais l'écart qui tend ou non à se creuser entre eux. En établissant un rapport direct entre les deux séries comparées, le bouton COURBE SUR QUOTIENT permet de savoir si l'une prend le pas sur l'autre - ce qui se vérifie pour le *vice* au détriment de la *vertu,* dans la courbe ci-dessus. On montrerait de la même façon que les deux termes de la négation *ne pas*, si soudés qu'ils soient, tendent à se désolidariser et que le second prend de plus en plus l'avantage sur le premier.

Courbe sur le quotient de deux séries



La fonction graphique est largement disponible. Elle apparaît dans le menu principal, dans toutes les pages de l'index, dans le menu DISTRIBUTION et aussi, sous une forme particulière, dans le menu LISTES. Dans ce menu le bouton graphique n'apparaît pas, mais les marges de lignes ou de colonnes sont sensibles au clic de la souris et cela génère l'histogramme souhaité.

LE TRAITEMENT DES LISTES

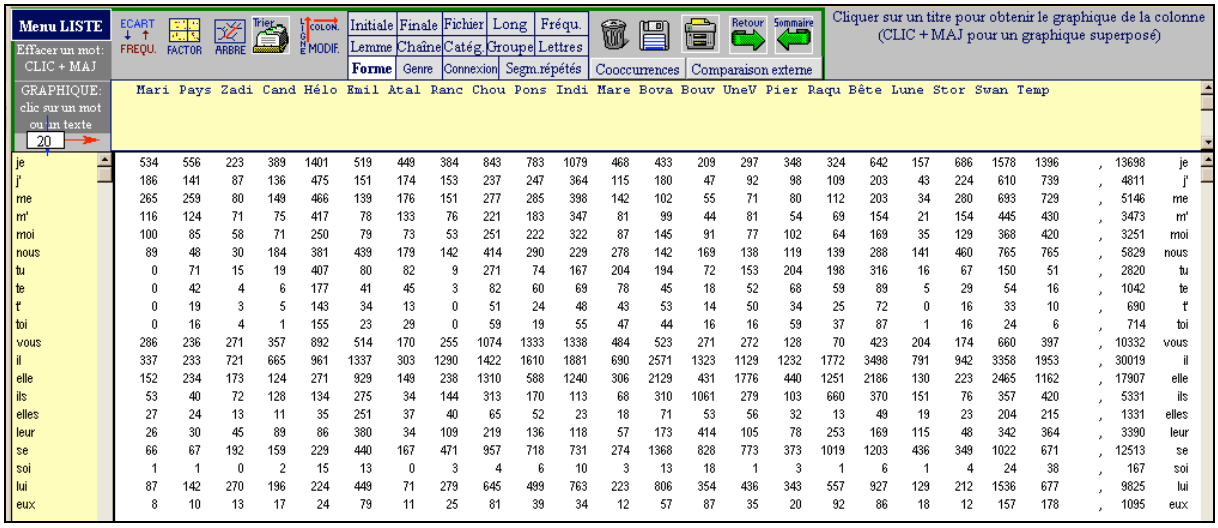
On a vu précédemment comment l'on pouvait constituer des listes de mots, afin de faire une concordance limitée aux formes contenues dans cette liste. Le tableau à deux dimensions qui en résulte se prête aussi et plus encore à l'exploitation statistique. De tels tableaux représentent le plus souvent des mots individuels, nommément désignés (grâce au bouton FORME) ou répondant à des critères de sélection:

- l'appartenance au même lemme
- le partage de la même initiale
- le partage de la même finale
- la présence d'une chaîne quelconque
- ou bien encore un seuil de fréquence
- ou l'appartenance à certaines catégories grammaticales à effectif restreint.

Mais certains tableaux regroupent des ensembles constitués autour d'un caractère particulier, en particulier: la longueur du mot, la répartition des lettres, la classe de fréquence des mots, les segments répétés, le tableau de la connexion lexicale, le genre ou regroupement des textes.

À chacun de ces modes de sélection est réservé un bouton au haut de l'écran.

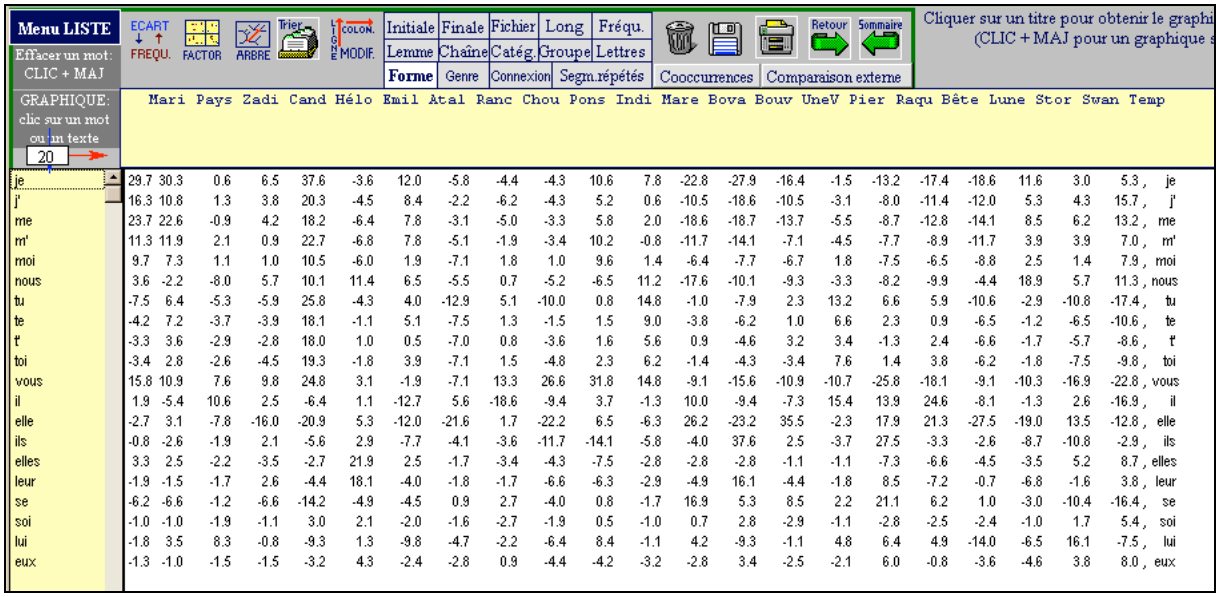
Exemple de liste: les pronoms personnels dans le corpus Example



	Mari	Pays	Zadi	Cand	Hélo	Emil	Atal	Ranc	Chou	Pons	Indi	Mare	Bova	Bouv	UneV	Pier	Raqu	Bête	Lune	Stor	Swan	Temp	
je	534	556	223	389	1401	519	449	384	843	783	1079	468	433	209	297	348	324	642	157	686	1578	1396	, 13698 je
j'	186	141	87	136	475	151	174	153	237	247	364	115	180	47	92	98	109	203	43	224	610	739	, 4811 j'
me	265	259	80	149	466	139	176	151	277	285	398	142	102	55	71	80	112	203	34	280	693	729	, 5146 me
m'	116	124	71	75	417	78	133	76	221	183	347	81	99	44	81	54	69	154	21	154	445	430	, 3473 m'
moi	100	85	58	71	250	79	73	53	251	222	322	87	145	91	77	102	64	169	35	129	368	420	, 3251 moi
nous	89	48	30	184	381	439	179	142	414	290	229	278	142	169	138	119	139	288	141	460	765	765	, 5829 nous
tu	0	71	15	19	407	80	82	9	271	74	167	204	194	72	153	204	198	316	16	67	150	51	, 2820 tu
te	0	42	4	6	177	41	45	3	82	60	69	78	45	18	52	68	59	89	5	29	54	16	, 1042 te
t'	0	19	3	5	143	34	13	0	51	24	48	43	53	14	50	34	25	72	0	16	33	10	, 690 t'
toi	0	16	4	1	155	23	29	0	59	19	55	47	44	16	16	59	37	87	1	16	24	6	, 714 toi
vous	286	236	271	357	892	514	170	255	1074	1333	1338	484	523	271	272	128	70	423	204	174	660	397	, 10332 vous
il	337	233	721	665	961	1337	303	1290	1422	1610	1881	690	2571	1323	1129	1232	1772	3498	791	942	3358	1953	, 30019 il
elle	152	234	173	124	271	929	149	238	1310	588	1240	306	2129	431	1776	440	1251	2186	130	223	2465	1162	, 17907 elle
ils	53	40	72	128	134	275	34	144	313	170	113	68	310	1061	279	103	660	370	151	76	357	420	, 5331 ils
elles	27	24	13	11	35	251	37	40	65	52	23	18	71	53	56	32	13	49	19	23	204	215	, 1331 elles
leur	26	30	45	89	86	380	34	109	219	136	118	57	173	414	105	78	253	169	115	48	342	364	, 3390 leur
se	66	67	192	159	229	440	167	471	957	718	731	274	1368	828	773	373	1019	1203	436	349	1022	671	, 12513 se
soi	1	1	0	2	15	13	0	3	4	6	10	3	13	18	1	3	1	6	1	4	24	38	, 167 soi
lui	87	142	270	196	224	449	71	279	645	499	763	223	806	354	436	343	557	927	129	212	1536	677	, 9825 lui
eux	8	10	13	17	24	79	11	25	81	39	34	12	57	87	35	20	92	86	18	12	157	178	, 1095 eux

Tout d'abord il est loisible de représenter graphiquement une ligne (c'est à dire un mot) de ce tableau. Il suffit de cliquer sur ce mot dans la colonne de gauche. Parallèlement aux opérations effectuées sur les effectifs absolus, un tableau d'écarts est tenu à jour et peut être porté à l'écran, si on sollicite le bouton ECART/FREQU, comme ci-dessous.

Transformation en écarts réduits



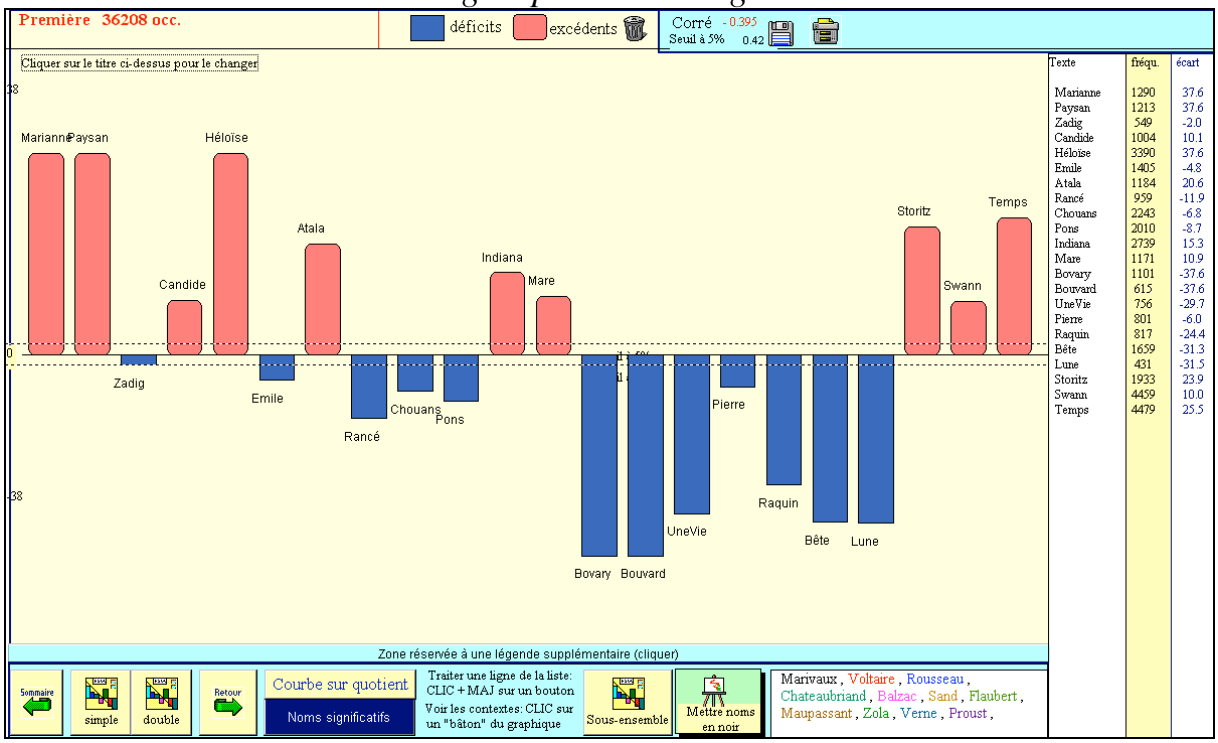
	Mari	Pays	Zadi	Cand	Hélo	Emil	Atal	Ranc	Chou	Pons	Indi	Mare	Bova	Bouv	UneV	Pier	Raqu	Bête	Lune	Stor	Swan	Temp	
je	29.7	30.3	0.6	6.5	37.6	-3.6	12.0	-5.8	-4.4	-4.3	10.6	7.8	-22.8	-27.9	-16.4	-1.5	-13.2	-17.4	-18.6	11.6	3.0	5.3 ,	je
j'	16.3	10.8	1.3	3.8	20.3	-4.5	8.4	-2.2	-6.2	-4.3	5.2	0.6	-10.5	-18.6	-10.5	-3.1	-8.0	-11.4	-12.0	5.3	4.3	15.7 ,	j'
me	23.7	22.6	-0.9	4.2	18.2	-6.4	7.8	-3.1	-5.0	-3.3	5.8	2.0	-18.6	-18.7	-13.7	-5.5	-8.7	-12.8	-14.1	8.5	6.2	13.2 ,	me
m'	11.3	11.9	2.1	0.9	22.7	-6.8	7.8	-5.1	-1.9	-3.4	10.2	-0.8	-11.7	-14.1	-7.1	-4.5	-7.7	-8.9	-11.7	3.9	3.9	7.0 ,	m'
moi	9.7	7.3	1.1	1.0	10.5	-6.0	1.9	-7.1	1.8	1.0	9.6	1.4	-6.4	-7.7	-6.7	1.8	-7.5	-6.5	-8.8	2.5	1.4	7.9 ,	moi
nous	3.6	-2.2	-8.0	5.7	10.1	11.4	6.5	-5.5	0.7	-5.2	-6.5	11.2	-17.6	-10.1	-9.3	-3.3	-8.2	-9.9	-4.4	18.9	5.7	11.3 ,	nous
tu	-7.5	6.4	-5.3	-5.9	25.8	-4.3	4.0	-12.9	5.1	-10.0	0.8	14.8	-1.0	-7.9	2.3	13.2	6.6	5.9	-10.6	-2.9	-10.8	-17.4 ,	tu
te	-4.2	7.2	-3.7	-3.9	18.1	-1.1	5.1	-7.5	1.3	-1.5	1.5	9.0	-3.8	-6.2	1.0	6.6	2.3	0.9	-6.5	-1.2	-6.5	-10.6 ,	te
t'	-3.3	3.6	-2.9	-2.8	18.0	1.0	0.5	-7.0	0.8	-3.6	1.6	5.6	0.9	-4.6	3.2	3.4	-1.3	2.4	-6.6	-1.7	-5.7	-8.6 ,	t'
toi	-3.4	2.8	-2.6	-4.5	19.3	-1.8	3.9	-7.1	1.5	-4.8	2.3	6.2	-1.4	-4.3	-3.4	7.6	1.4	3.8	-6.2	-1.8	-7.5	-9.8 ,	toi
vous	15.8	10.9	7.6	9.8	24.8	3.1	-1.9	-7.1	13.3	26.6	31.8	14.8	-9.1	-15.6	-10.9	-10.7	-25.8	-18.1	-9.1	-10.3	-16.9	-22.8 ,	vous
il	1.9	-5.4	10.6	2.5	-6.4	1.1	-12.7	5.6	-18.6	-9.4	3.7	-1.3	10.0	-9.4	-7.3	15.4	13.9	24.6	-8.1	-1.3	2.6	-16.9 ,	il
elle	-2.7	3.1	-7.8	-16.0	-20.9	5.3	-12.0	-21.6	1.7	-22.2	6.5	-6.3	26.2	-23.2	35.5	-2.3	17.9	21.3	-27.5	-19.0	13.5	-12.8 ,	elle
ils	-0.8	-2.6	-1.9	2.1	-5.6	2.9	-7.7	-4.1	-3.6	-11.7	-14.1	-5.8	-4.0	37.6	2.5	-3.7	27.5	-3.3	-2.6	-8.7	-10.8	-2.9 ,	ils
elles	3.3	2.5	-2.2	-3.5	-2.7	21.9	2.5	-1.7	-3.4	-4.3	-7.5	-2.8	-2.8	-2.8	-1.1	-1.1	-7.3	-6.6	-4.5	-3.5	5.2	8.7 ,	elles
leur	-1.9	-1.5	-1.7	2.6	-4.4	18.1	-4.0	-1.8	-1.7	-6.6	-6.3	-2.9	-4.9	16.1	-4.4	-1.8	8.5	-7.2	-0.7	-6.8	-1.6	3.8 ,	leur
se	-6.2	-6.6	-1.2	-6.6	-14.2	-4.9	-4.5	0.9	2.7	-4.0	0.8	-1.7	16.9	5.3	8.5	2.2	21.1	6.2	1.0	-3.0	-10.4	-16.4 ,	se
soi	-1.0	-1.0	-1.9	-1.1	3.0	2.1	-2.0	-1.6	-2.7	-1.9	0.5	-1.0	0.7	2.8	-2.9	-1.1	-2.8	-2.5	-2.4	-1.0	1.7	5.4 ,	soi
lui	-1.8	3.5	8.3	-0.8	-9.3	1.3	-9.8	-4.7	-2.2	-6.4	8.4	-1.1	4.2	-9.3	-1.1	4.8	6.4	4.9	-14.0	-6.5	16.1	-7.5 ,	lui
eux	-1.3	-1.0	-1.5	-1.5	-3.2	4.3	-2.4	-2.8	0.9	-4.4	-4.2	-3.2	-2.8	3.4	-2.5	-2.1	6.0	-0.8	-3.6	-4.6	3.8	8.0 ,	eux

En outre la totalisation complète ou partielle des lignes du tableau peut être réalisée en agissant sur le bouton MODIF, lequel peut en outre opérer des manipulations soustractives. L'exemple ci-dessous illustre un tel regroupement : celle des mots (ou lignes) qui relèvent de la première personne.

La totalisation symétrique des colonnes n'a pas à être calculée puisqu'elle apparaît dans le tableau, à la lin de chaque ligne. La représentation graphique de

ces totaux marginaux n'a de sens que si on les compare à une référence extérieure. Cette comparaison est assurée par le bouton "comparaison extérieure", au moins pour le français, l'anglais, le portugais et l'italien.

Regroupement des lignes



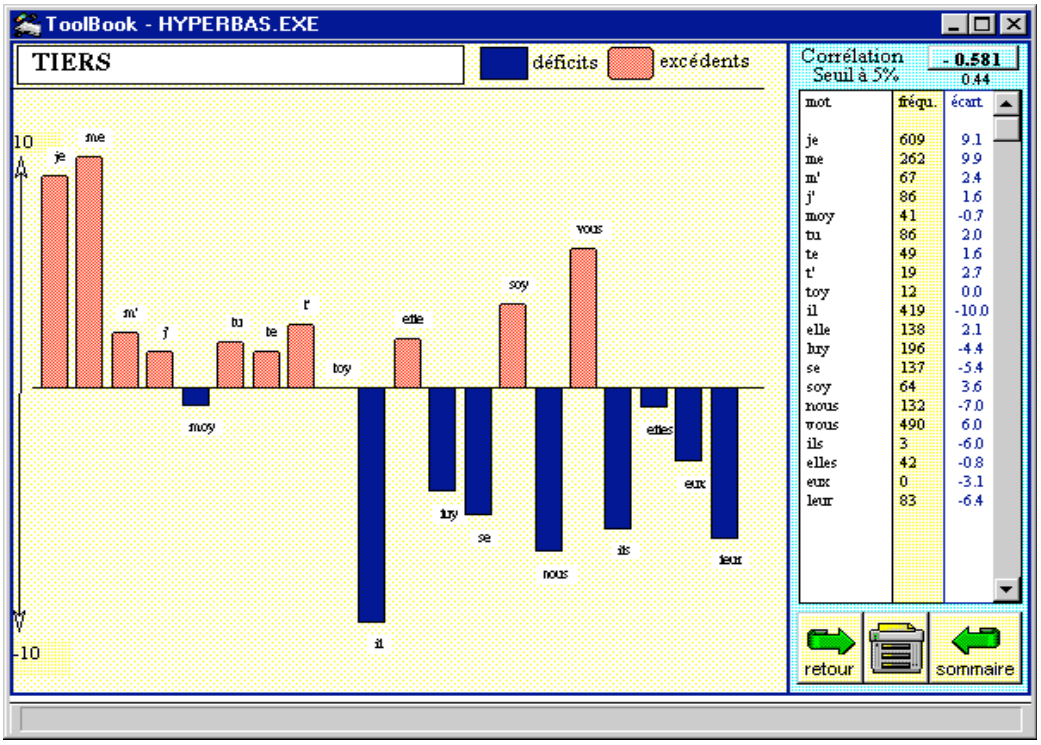
LA REPRÉSENTATION GRAPHIQUE DES COLONNES

Le programme d'illustration graphique offre une variante qui analyse les éléments d'une colonne, dans un tableau de fréquences où généralement les lignes désignent les mots et les colonnes les textes. Un tel graphique permet non plus de suivre la distribution d'un mot à travers les textes, mais de dresser le profil d'un texte (ou d'un groupe de textes) à travers les mots qui s'y trouvent employés ou les faits linguistiques qu'on a soumis au dénombrement.

Ici comme précédemment, c'est l'écart qui sert d'ordonnée à l'histogramme. Le clic portera cette fois sur le texte qui correspond à la colonne souhaitée et qui figure avec les autres textes sur la ligne supérieure de la page LISTE. On veillera à ignorer le coefficient de corrélation qui dans cette perspective n'a guère de signification. Dans le cas le plus fréquent, l'ordre des mots d'une liste est alphabétique, c'est-à-dire arbitraire, et ne peut donner lieu à aucune découverte. Mais si l'on dispose les mots selon un ordre logique, comme dans l'exemple suivant, où l'on passe progressivement de la première à la troisième personne, le coefficient retrouve son efficience.

Histogramme d'une colonne , c'est-à-dire d'un texte

(ici le Tiers Livre *de Rabelais)*



FONCTIONS ÉTENDUES DU MENU LISTE

Les listes sont des tableaux de contingence où l'on assemble les données de son choix. Il s'agit de grouper, en lignes successives, des mots, des catégories ou quelque objet que l'on veut, les colonnes correspondant aux textes du corpus (ou à certains textes choisis dans ce corpus). Cette fonction essentielle à l'analyse statistique existe depuis toujours dans Hyperbase. On a seulement ajouté quelques critères pour constituer de tels tableaux, ou pour les présenter. Par exemple un bouton TRIER est maintenant disponible pour redisposer les lignes dans l'ordre alphabétique, ou dans l'ordre hiérarchique (par fréquences décroissantes), ou dans l'ordre initial de saisie.

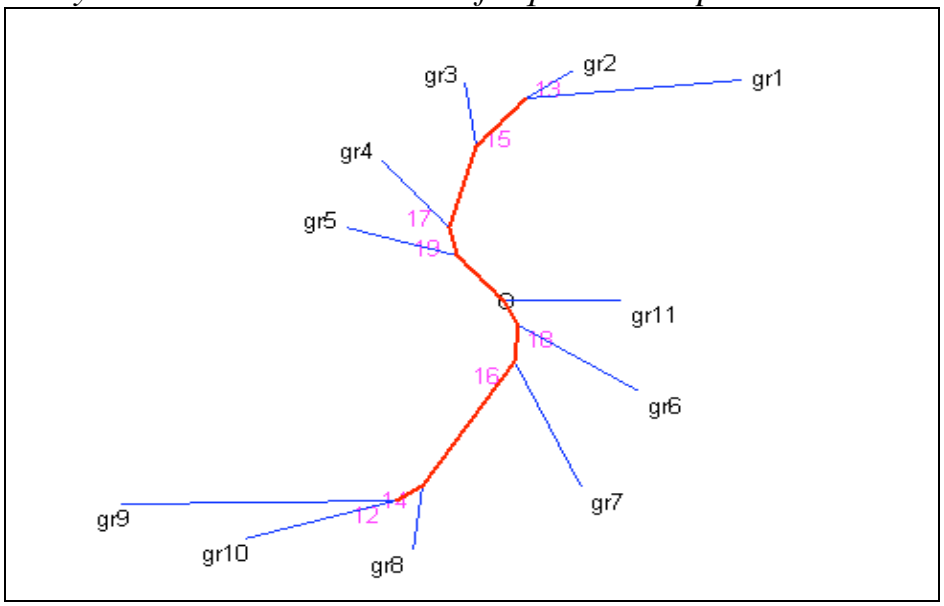
Ce qui change aussi, c'est la manière d'exploiter les données. Certes on garde la possibilité de dessiner un histogramme sur une ligne (on clique sur un mot) ou sur une colonne (on clique sur un texte). Et on a tout loisir, comme auparavant, de faire les manipulations que l'on veut par élimination ou par regroupement. Mais un nouveau bouton apparaît qui propose l'application de l'analyse arborée. Et le bouton propre à l'analyse factorielle a été remodelé. Précisons que les données apparentes restituent les effectifs absolus et que les traitements s'appuient de préférence sur un tableau parallèle où les effectifs réels sont convertis en écarts, afin que la pondération neutralise les effets de taille qu'on observe souvent dans les résultats quand les lignes ou les colonnes ont des

poids trop inégaux. Mais cette conversion n'est pas obligatoire, au moins dans le cas de l'analyse arborée.

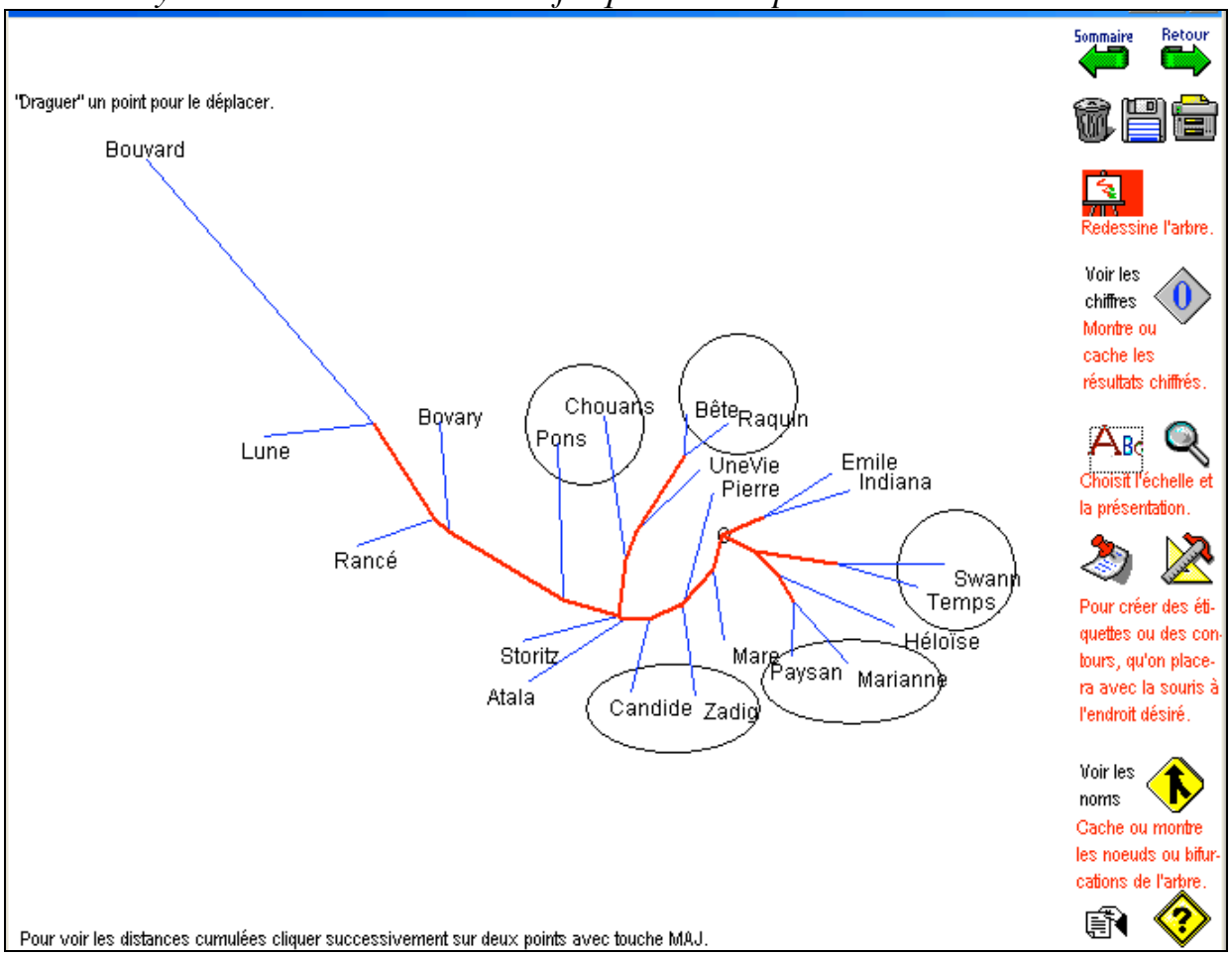
Pour illustrer ce dernier traitement on choisira un nouvel exemple, qui ne concerne plus des mots individuels mais des classes de mots. Hyperbase en effet au moment où la base est réalisée constitue des groupes de fréquence et y range les mots selon qu'ils sont rares ou fréquents. Onze classes sont ainsi distinguées dont la première est consacrée aux mots rares (qui ont moins de 500 occurrences dans le corpus FRANTEXT), la suivante aux mots dont la fréquence ne dépasse pas 1024, puis de 1024 à 2048, de 2048 à 4096 on gravit l'échelle jusqu'à la dernière marche où se retrouvent les mots très fréquents qui ont plus de 262000 occurrences.

Ces onze classes sont représentées sur l'arbre de la figure ci-dessous, qui propose une dichotomie très claire: au sommet du graphe se réunissent les fréquences basses (classes 1 à 5) tandis que la base du graphe est réservée aux fréquences élevées (classes 6 à 10), à l'exception de la classe 11 où se retrouvent la plupart des mots-outils et qui occupe une position médiane. Le chemin est court qui joint deux classes contiguës (un seul nœud suffit généralement à opérer la jonction), mais il est maximal lorsque les classes extrêmes sont mises en relation. On a apparemment affaire à des données sérielles, qui se répartissent selon un ordre progressif et hiérarchique.Reste à projeter cette distribution sur les textes, ce qu'on réalise en sollicitant de nouveau l'analyse arborée, sans changer les données. Mais cette fois on demande que le graphe représente les colonnes et non plus les lignes du tableau. Le graphe obtenu maintient la distinction entre les auteurs et le rapprochement des deux textes qui appartiennent à la même plume.

Analyse arborée des classes de fréquences. Représentation des classes.



Analyse arborée des classes de fréquences. Représentation des textes.

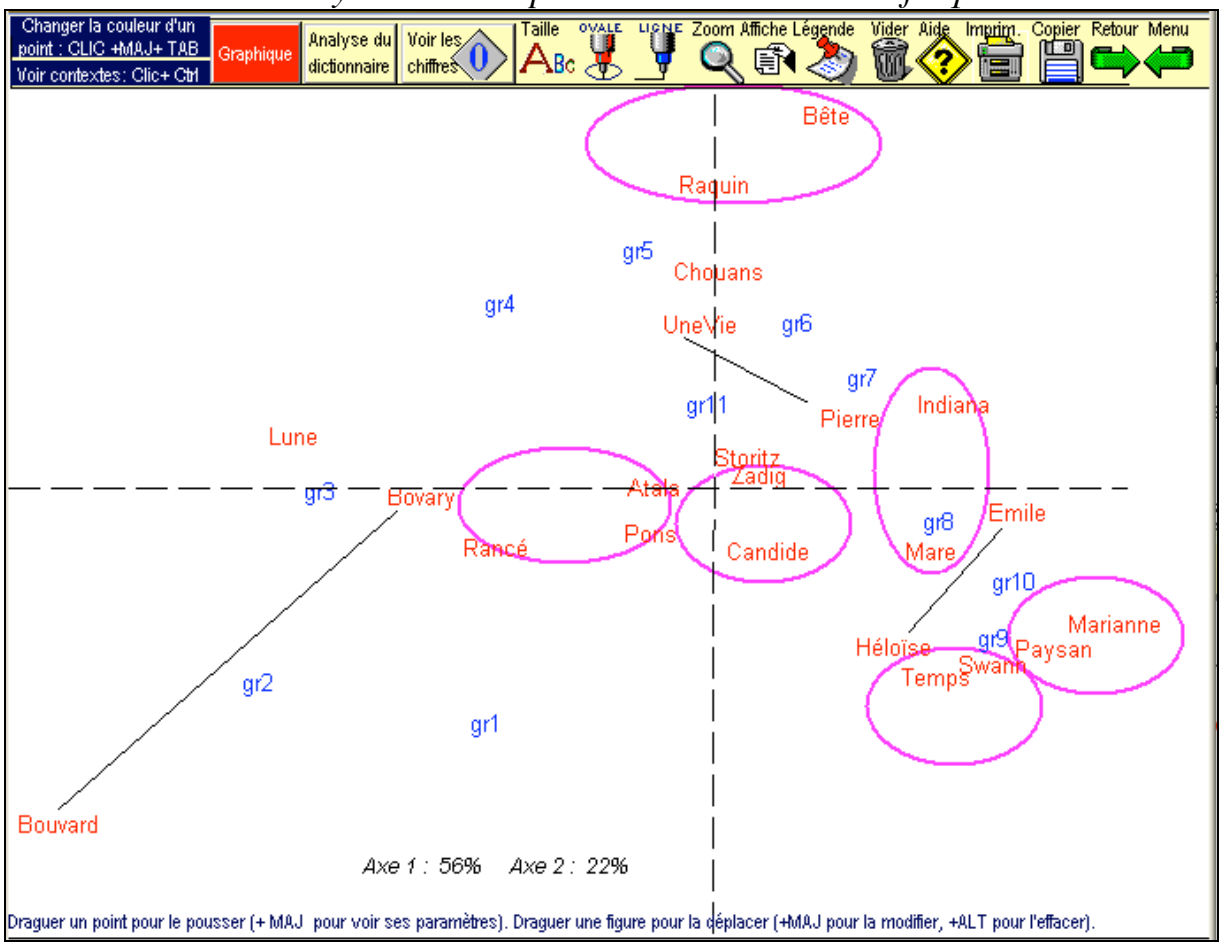


Cependant demeure un certain embarras de l'interprétation qui tient au fait que les mêmes données engendrent deux graphes qui ne sont pas superposables.

L'analyse de correspondance résout élégamment cette difficulté dans la figure ci-dessous. Les classes de fréquences, réparties en arc de cercle, y parcourent successivement les quatre quadrants du plan, ce qui est caractéristique des données sérielles. Et les deux textes du même écrivain occupent comme prrécédemment une position voisine. Ce qui est nouveau et précieux pour interpréter la distribution, c'est que la superposition des classes (en bleu) et des textes (en rouge) donne une explication aux phénomènes observés. Si les textes du XVIIIe se cantonnent dans les fréquences hautes, à droite sur le graphique, les mots rares et les basses fréquences sont concentrées à gauche dans le voisinage des textes de plus grande technicité, même s'il s'agit de fantaisie comme dans le *Voyage dans la lune* de Jules Verne ou de dérision comme dans *Bouvard et Pécuchet*. Le goût de Chateaubriand pour les curiosités de la nature, de la société et de la langue le porte aussi de ce côté.

Ainsi les outils de l'analyse se complètent et s'épaulent. Si l'analyse de correspondance doit céder le pas à l'analyse arborée dans les tableaux de distance où la symétrie confond les lignes et les colonnes, elle reprend l'avantage quand les tableaux sont rectangulaires et que les lignes et les colonnes désignent des objets différents.

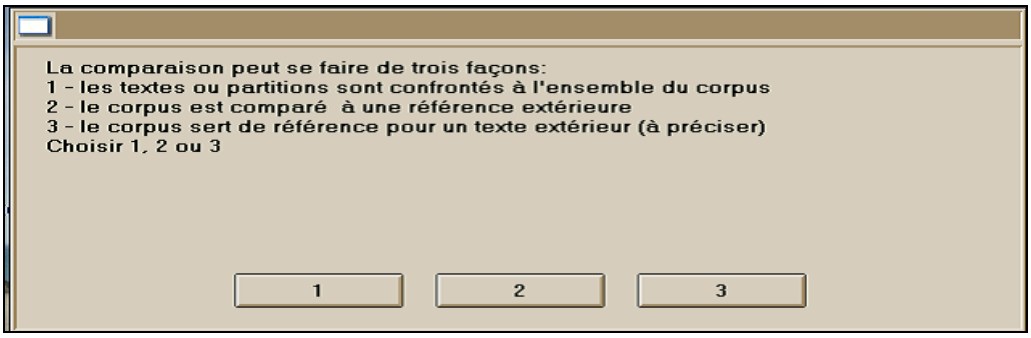
Analyse de correspondance des classes de fréquence



CHAPITRE 8

Le menu SPÉCIFICITÉS

Le bouton *SPÉCIFICITÉS* du menu principal invite à choisir entre plusieurs options, selon qu'on envisage une référence interne (le corpus pour ses parties), externe (le TLF pour le corpus) ou mixte (le corpus pour un nouveau texte).



1 - LA COMPAISON INTERNE

Quand le corpus comporte une segmentation en textes ou parties (cette segmentation de toute façon est introduite si elle ne figure pas expressément dans les données), le programme d'indexation calcule le vocabulaire spécifique de chaque texte du corpus en se fondant sur la loi normale et en prenant pour norme l'ensemble du corpus (sans procéder au calcul pour les mots de basse fréquence et en ne retenant que les écarts réduits supérieurs à 2). On peut faire apparaître le profil caractéristique de chaque texte, si l'on déroule au bon endroit le menu déroulant du centre de l'écran (pourvu d'un triangle). Pour gagner de la place, deux textes sont présentés à la fois, les listes étant triées par ordre décroissant de l'écart réduit. Ci-dessous un extrait des spécificités de *Zadig* (à gauche) et de *Candide*(à droite). La liste est beaucoup plus longue et plus instructive que ce court extrait où les noms propres, trop étroitement liés à l'intrigue, occupent ostensiblement et peu légitimement les premières places.

On peut enfin consulter les listes de spécificités à propos d'une forme particulière. On dessine alors le profil du mot parmi les sous-ensembles, si du moins l'emploi de ce mot est suffisamment caractéristique pour franchir le seuil

significatif dans au moins un des textes du corpus. Pour cette recherche (ici le mot *que* dans le corpus *Francil)* on pressera le bouton CHERCHE dont l'effet peut s'observer dans le t ableau ci-dessous (le même effet est obtenu si on clique sur un mot de l'écran).

Le vocabulaire spécifique de Zadig *et de* Candide *dans le corpus* Example

Refaire résumé | Zadig | Voir les passages spécifiques et cacher les mots spécifiques | Trier | Cherche | Retour | Sommaire | Candide

CLIC sur un mot: Recherche du mot dans les spécificités | Zadig Candide | CLIC+MAJ: Recherche du mot dans les textes

N°	écart	corpus	texte	mot
3	37.3	287	287	Zadig
3	25.1	68	64	Babylone
3	21.7	383	98	roi
3	21.4	157	71	reine
3	20.4	44	43	Astarté
3	17.1	31	31	Sétoc
3	15.1	25	25	Cador
3	12.5	46	25	ermite
3	11.3	25	18	envieux
3	11.3	15	15	Nabussan
3	11.3	15	15	basilic
3	10.9	14	14	egyptien
3	10.5	30019	721	il
3	10.5	27	17	pêcheur
3	10.4	19	15	onces
3	10.1	14	13	armure
3	9.9	15	13	arabie
3	9.0	102	22	seigneur
3	8.4	27	13	chienne
3	8.3	9825	270	lui
3	8.2	1952	86	fut
3	7.5	10332	271	vous
3	7.5	102	18	destinée
3	7.5	19	10	Egypte
3	7.4	4412	141	dit
3	7.2	515	36	belle
3	7.2	16	9	arabe
3	6.7	1086	52	fit
3	6.6	34523	719	le
3	6.6	15	8	brigand
3	6.5	54	12	majesté
3	6.4	1901	72	point
3	6.4	17	8	fromages
3	6.3	372	27	dame
3	6.3	26	9	esclaves
3	6.2	213	20	cheval

N°	écart	corpus	texte	mot
4	37.3	391	379	Candide
4	36.7	125	125	Cunégonde
4	32.3	129	112	Martin
4	32.0	98	98	Pangloss
4	29.5	85	85	Cacambo
4	21.7	4412	324	dit
4	15.7	53	36	vaisseau
4	13.7	77	35	chapitre
4	13.7	33	26	jésuite
4	13.4	200	48	mademoiselle
4	12.8	23	21	bulgares
4	12.2	352	55	vieille
4	12.0	70	29	Venise
4	11.8	17	17	paquette
4	11.4	43	23	moutons
4	11.4	21	18	inquisiteur
4	10.5	14	14	Pococuranté
4	10.5	14	14	eldorado
4	10.3	297	43	baron
4	10.3	40	20	diamants
4	10.0	16	14	piastres
4	9.8	14	13	buenos
4	9.7	10332	357	vous
4	9.2	26	15	pendu
4	9.0	3305	149	deux
4	8.9	19	13	Constantinople
4	8.7	21	13	révérend
4	8.6	64	19	souper
4	8.5	22	13	espagnol
4	8.4	492	45	pays
4	8.3	15	11	hollandais
4	8.2	940	63	monsieur
4	8.2	178	27	fus
4	8.1	66	18	patron
4	7.8	421	39	maître
4	7.8	37	14	monseigneur

Spécificités du mot *que* dans le corpus EXEMPLE

N°	écart	corpus	texte	mot	
1	15.7	22052	473	que	Marianne
2	11.2	22052	420	que	Paysan
5	22.5	22052	1454	que	Héloïse
6	9.0	22052	1214	que	Emile
12	4.1	22052	616	que	Mare
21	19.5	22052	3130	que	Swann
22	29.6	22052	3014	que	Temps
	-10.05	121826522052		que	corpus

CLIQUER DANS CE CHAMP POUR LE FAIRE DISPARAITRE

2 - LA COMPAISON EXTERNE

Si les conditions d'une comparaison justifiée sont remplies au moment de la création de la base (corpus en français, en anglais, en italien ou en portugais), la spécificité du corpus est détaillée par rapport au dictionnaire de référence choisi. S'il s'agit du français, le choix de la norme peut être précisé: telle ou telle période du TLF, du 16ème siècle à l'époque contemporaine. On a ici le choix entre une présentation alphabétique ou hiérarchique (par valeurs décroissantes de l'écart réduit). L'écran montre alors deux champs dont l'un à gauche livre les formes en excédent et l'autre à droite les formes déficitaires.

Là aussi on peut demander au programme (bouton CHERCHER) de vérifier si un mot figure parmi le vocabulaire spécifique, positif ou négatif. Si l'on choisit la présentation hiérarchique, les deux listes sont triées d'après la valeur absolue de l'écart réduit, de façon à mettre en relief ce qui est le plus significatif, dans un sens ou dans l'autre. Comme les listes dépassent généralement les possibilités de l'écran, le bouton EDITER permet de les restituer sur l'imprimante dans leur intégralité. Rappelons que le seuil significatif généralement admis est aux alentours de la valeur 2 (en laissant 5 chances sur 100 au hasard). Nous nous sommes arrêté à la valeur 3 en adoptant un seuil plus sévère.

Au reste la comparaison avec l'usage observé dans le Trésor de la langue française doit être interprétée prudemment. D'une part le TLF reflète l'usage littéraire de la langue, dans un registre relevé, et si le corpus qu'on traite se trouve éloigné de ce niveau de langue, la valeur de la comparaison en est amoindrie. D'autre part toutes les formes n'ont pas été soumises à la comparaison, parce que le calcul de l'écart réduit perd de sa légitimité quand la fréquence théorique est trop faible, ce qui dépend certes de la taille du corpus traité, mais aussi de la fréquence du mot en question. Le fichier MODELE.txt qui sert de référence est largement dimensionné puisqu'il rend compte de 100 000 formes distinctes, chacune étant dotée de 12 sous-fréquences, du XVI^{e} siècle au XX^{e}. Cela permet, en fixant la date de départ et la date d'arrivée, de rapprocher les deux corpus comparés et de justifier leur confrontation. Mais il peut se faire que des mots soient très significatifs d'un corpus donné, même s'ils ne figurent pas dans la liste des 100000, soit parce qu'ils sont trop récents, soit parce qu'ils sont trop techniques.

Depuis décembre 2010 Google peut se substituer à Frantext comme référence externe. Même réduit au domaine français, le corpus de Google est gigantesque et s'étend sur 44 milliards de mots (y compris ceux qui ont été estropiés par la lecture optique). Il s'agit là d'un modèle plus neutre, qui n'a plus la tonalité littéraire de Frantext. Mais la taille, si grande soit-elle, garantit-elle la représentativité d'un corpus et l'autorité d'une norme ? On peut en douter pour

Google comme pour Frantext. Le choix entre ces deux modèles est désormais proposé.

Spécificité de Hugo par rapport au TLF (19e siècle) (extrait très partiel)

N°	écart	corpus	texte	mot
103.22		3431	1638	Victor
92.25		1895	1065	Adèle
80.78		500	458	Hernani
78.49		402	397	Hauteville
78.45		655	515	Meurice
75.02		10579	2461	ombre
71.71		3302	1174	h
66.20		2085	841	v
66.15		335	307	Guernesey
64.99		2521	924	Hugo
63.55		606	407	dona
58.57		825	448	Bruxelles
57.73		1068	511	gouffre
57.38		11607711569		c'
55.80		625	368	Ruy
55.58		533	336	Gringoire
55.34		2583	821	auguste
55.32		5460	1295	sombre
52.58		760	389	House
52.52		28847223921		vous
52.26		2172	707	Bonaparte
44.65		12220	1900	roi
43.20		72267251882		l'
42.56		39892230206		est
42.42		3140	751	rois
42.26		2486	648	cieux
42.24		830	338	hideux
41.74		1512	475	aube
41.73		8186	1384	jusqu'
41.20		24674	3008	toi
39.91		11988	1752	ô
39.87		1933	531	César

N°	écart	corpus	texte	mot
-69.02		275841	7889	elle
-44.29173863090036				de
-38.62		168387	6304	mais
-36.56		32846414659		ne
-35.52		64479131763		il
-35.05		182705	7365	lui
-31.66		40246719285		une
-31.60		24712	299	ça
-29.90		30376014243		pas
-28.39		39044	1003	chez
-26.81		43756	1285	)
-25.41		127426	5463	ses
-25.40		52091626778		d'
-25.17		42884	1327	(
-25.08		24920211922		se
-24.73		160135	7222	sa
-24.39		37815	1135	très
-24.17		21280210071		plus
-24.01		32644	922	autres
-22.50		24677012095		n'
-21.80		195391	9390	son
-21.32		166470	7884	me
-21.00		160516	7596	avec
-20.16		45243	1687	vie
-20.07		138275	6493	avait
-19.99		29962	970	trop
-19.30		18220	471	gens
-19.22		25605	801	ah
-18.75		27433	903	elles
-18.71		79559243579		à
-18.38		33978317762		qu'
-17.69		78012	3490	ou

Tout le monde reconnaîtra Hugo dans le portrait-robot que dessine le calcul dans le tableau ci-dessus. *Victor* est le premier des noms de personne, *Hauteville* des noms de lieux, *ombre* des substantifs et *sombre* des adjectifs. Ajoutons que la rime *ombre*/*sombre* est aussi la plus fréquente.

3 - LA SITUATION MIXTE

Habituellement les spécificités d'un texte sont calculées à partir d'un corpus où ce texte figure parmi d'autres. Mais il peut arriver qu'on veuille soumettre à la comparaison un texte extérieur au corpus déjà constitué. Faut-il alors refaire la base en incorporant ce nouveau texte et en l'ajoutant aux autres? Ce n'est pas toujours souhaitable si le texte en question risque de nuire à l'équilibre du corpus et à l'homogénéité requise. Et surtout cette procédure risque d'être trop lourde, si le but à atteindre est seulement le calcul des spécificités.

La méthode la plus économique (quelques secondes suffisent) est d'indexer le texte extérieur en le réduisant à une liste de fréquences. En comparant cette liste à celle du corpus, on obtient le résultat souhaité. C'est ce que réalise le bouton "Traiter des textes extérieurs", qu'on peut solliciter autant de fois que l'on

veut, en proposant des fichiers différents. Ces fichiers-textes n'ont nul besoin de préparation.

Si plusieurs fichiers doivent être réunis, il suffit de préciser leur nombre et leur adresse. La concaténation sera réalisée et les spécificités calculées pour l'ensemble. Cette dernière méthode est particulièrement utile lorsqu'on veut modifier provisoirement le découpage d'un corpus et considérer un sous-ensemble englobant plusieurs textes de ce corpus. Solliciter alors le bouton "Redécoupage du corpus".

Dans le cas particulier où le sous-ensemble envisagé comporte à la fois des textes du corpus et des textes extérieurs au corpus, on doit d'abord exporter les textes à partir de la base et solliciter le bouton "Exporter" du menu principal, en choisissant l'option "un par un" qui reconstitue autant de fichiers qu'il y a de textes, sous les noms tex1.txt, tex2.txt, tex3.txt, etc.. On peut alors proposer au calcul des spécificités un texte composite réunissant certains de ces fichiers et d'autres qui n'appartiennent pas au corpus. Solliciter le bouton "Traiter des textes extérieurs".

Le calcul des spécificités fait appel à la loi hypergéométrique, même si les mesures probabilistes prennent l'apparence d'écarts réduits (ce sont en réalité des probabilités hypergéométriques converties en écarts réduits). Quand l'écart atteint la valeur 37, c'est le signe qu'on dépasse les possibilités de calcul et que la probabilité est infiniment faible. Dans chaque champ les écarts occupent la première colonne, les deux colonnes suivantes étant réservées aux fréquences du mot dans le corpus de référence et dans le sous-ensemble.

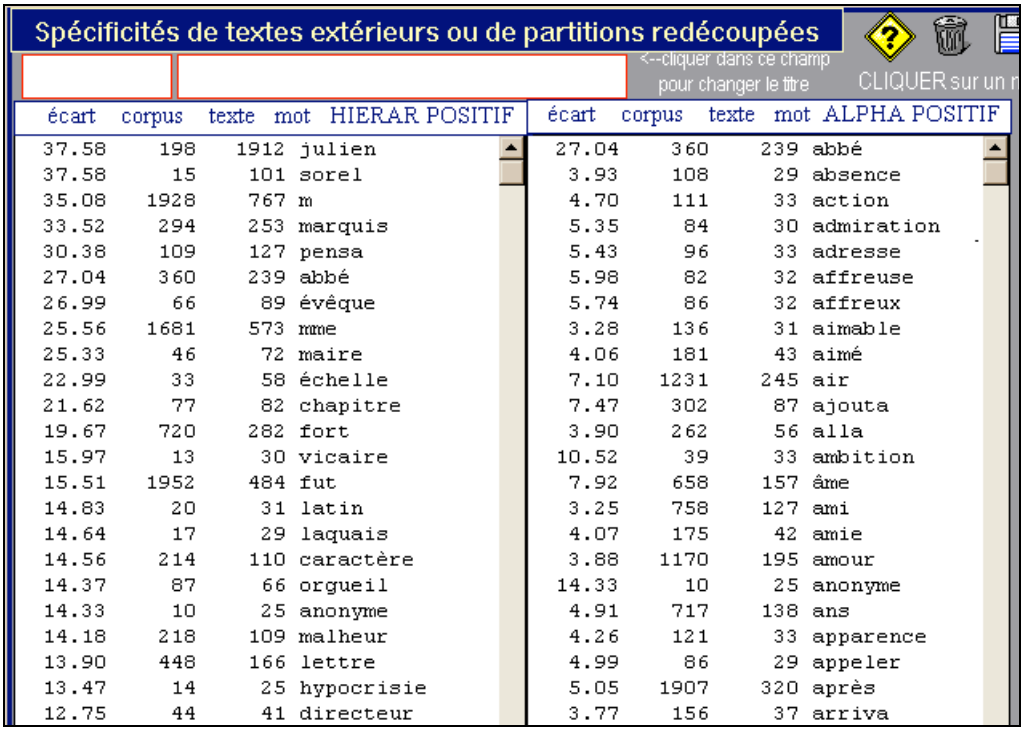
Une fois les calculs réalisés, les résultats apparaissent dans quatre champs: ceux de gauche restituent les spécificités positives (les excédents), en ordre hiérarchique, puis alphabétique, ceux de droite les spécificités négatives (les déficits). On peut enregistrer, imprimer ou effacer ces listes.

On peut aussi les modifier

- soit en fixant les limites minimale et maximale pour la fréquence (bouton "Modifier le seuil")
- soit en excluant les mots-outils, à savoir les mots de moins de 4 lettres et une liste de mots-outils plus longs (bouton "Exclure les mots-outils")
- soit en intervenant directement sur les listes (un clic sur un mot l'efface partout)

À tout moment les résultats d'origine peuvent être rappelés (bouton "Rétablir les résultats complets").Par défaut le nom des fichiers traités apparaît dans le champ du titre (en rouge, en haut à gauche). En cliquant dans ce champ, on peut changer le titre. On trouvera ci-dessous un court extrait des spécificités, ainsi calculées pour le *Rouge et le noir* , le corpus *Example* servant de référence.

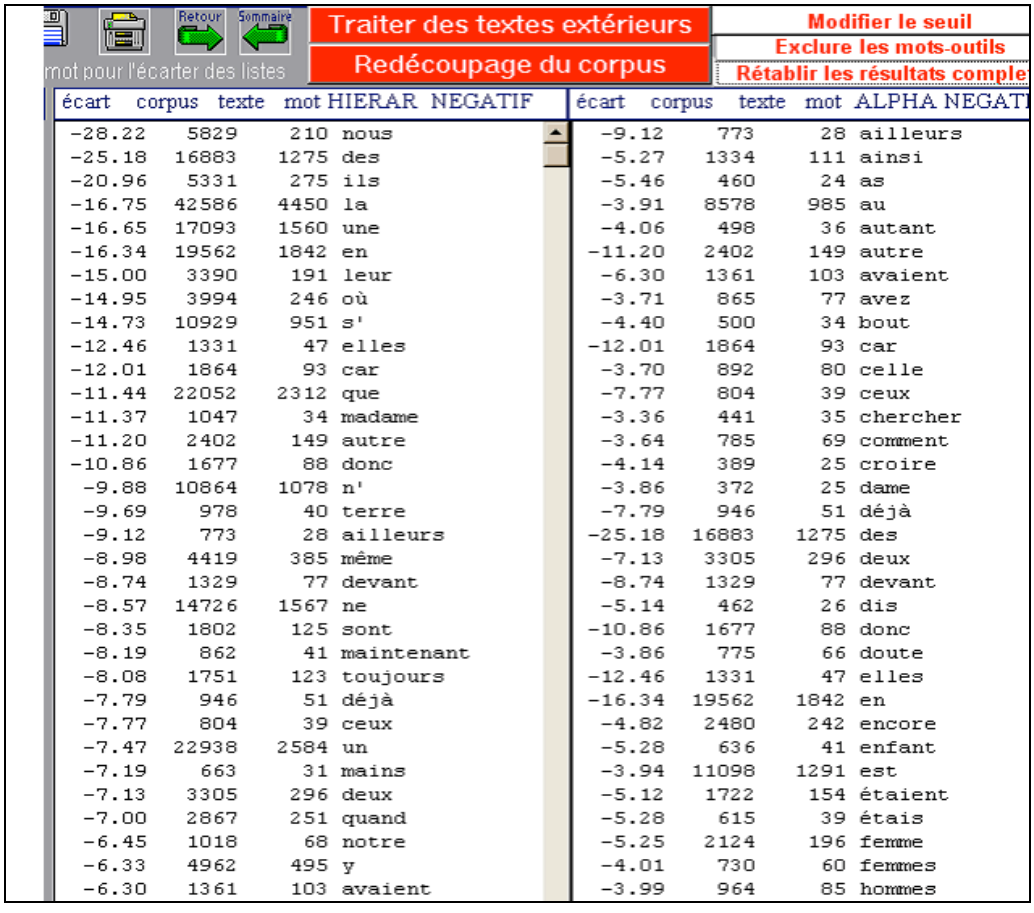
Les spécificités positives (*Le Rouge et le noir* comparé au corpus *Example*)



écart	corpus	texte	mot	HIERAR POSITIF
37.58	198	1912	julien	
37.58	15	101	sorel	
35.08	1928	767	m	
33.52	294	253	marquis	
30.38	109	127	pensa	
27.04	360	239	abbé	
26.99	66	89	évêque	
25.56	1681	573	mme	
25.33	46	72	maire	
22.99	33	58	échelle	
21.62	77	82	chapitre	
19.67	720	282	fort	
15.97	13	30	vicaire	
15.51	1952	484	fut	
14.83	20	31	latin	
14.64	17	29	laquais	
14.56	214	110	caractère	
14.37	87	66	orgueil	
14.33	10	25	anonyme	
14.18	218	109	malheur	
13.90	448	166	lettre	
13.47	14	25	hypocrisie	
12.75	44	41	directeur	

écart	corpus	texte	mot	ALPHA POSITIF
27.04	360	239	abbé	
3.93	108	29	absence	
4.70	111	33	action	
5.35	84	30	admiration	
5.43	96	33	adresse	
5.98	82	32	affreuse	
5.74	86	32	affreux	
3.28	136	31	aimable	
4.06	181	43	aimé	
7.10	1231	245	air	
7.47	302	87	ajouta	
3.90	262	56	alla	
10.52	39	33	ambition	
7.92	658	157	âme	
3.25	758	127	ami	
4.07	175	42	amie	
3.88	1170	195	amour	
14.33	10	25	anonyme	
4.91	717	138	ans	
4.26	121	33	apparence	
4.99	86	29	appeler	
5.05	1907	320	après	
3.77	156	37	arriva	

Les spécificités négatives (*Le Rouge et le noir* comparé au corpus *Example*)



écart	corpus	texte	mot	HIERAR NEGATIF
-28.22	5829	210	nous	
-25.18	16883	1275	des	
-20.96	5331	275	ils	
-16.75	42586	4450	la	
-16.65	17093	1560	une	
-16.34	19562	1842	en	
-15.00	3390	191	leur	
-14.95	3994	246	où	
-14.73	10929	951	s'	
-12.46	1331	47	elles	
-12.01	1864	93	car	
-11.44	22052	2312	que	
-11.37	1047	34	madame	
-11.20	2402	149	autre	
-10.86	1677	88	donc	
-9.88	10864	1078	n'	
-9.69	978	40	terre	
-9.12	773	28	ailleurs	
-8.98	4419	385	même	
-8.74	1329	77	devant	
-8.57	14726	1567	ne	
-8.35	1802	125	sont	
-8.19	862	41	maintenant	
-8.08	1751	123	toujours	
-7.79	946	51	déjà	
-7.77	804	39	ceux	
-7.47	22938	2584	un	
-7.19	663	31	mains	
-7.13	3305	296	deux	
-7.00	2867	251	quand	
-6.45	1018	68	notre	
-6.33	4962	495	y	
-6.30	1361	103	avaient	

écart	corpus	texte	mot	ALPHA NEGATI
-9.12	773	28	ailleurs	
-5.27	1334	111	ainsi	
-5.46	460	24	as	
-3.91	8578	985	au	
-4.06	498	36	autant	
-11.20	2402	149	autre	
-6.30	1361	103	avaient	
-3.71	865	77	avez	
-4.40	500	34	bout	
-12.01	1864	93	car	
-3.70	892	80	celle	
-7.77	804	39	ceux	
-3.36	441	35	chercher	
-3.64	785	69	comment	
-4.14	389	25	croire	
-3.86	372	25	dame	
-7.79	946	51	déjà	
-25.18	16883	1275	des	
-7.13	3305	296	deux	
-8.74	1329	77	devant	
-5.14	462	26	dis	
-10.86	1677	88	donc	
-3.86	775	66	doute	
-12.46	1331	47	elles	
-16.34	19562	1842	en	
-4.82	2480	242	encore	
-5.28	636	41	enfant	
-3.94	11098	1291	est	
-5.12	1722	154	étaient	
-5.28	615	39	étais	
-5.25	2124	196	femme	
-4.01	730	60	femmes	
-3.99	964	85	hommes	

4 - LA MESURE DE L'ÂGE

La conscience du temps et de l'âge remplissent les dernières pages de la *Recherche du temps perdu* : "Non seulement tout le monde sent que nous occupons une place dans le Temps, mais cette place, le plus simple la mesure approximativement comme il mesurerait celle que nous occupons dans l'espace. Sans doute, on se trompe souvent dans cette évaluation, mais qu'on ait cru pouvoir la faire, signifie qu'on concevait l'âge comme quelque chose de mesurable." Proust évoque ici les traits physiques, la voix, la démarche qui changent avec l'âge. Lui qui est si sensible aux traits stylistiques des écrivains qu'il en fait même des pastiches ne semble pas pourtant avoir cherché dans l'écriture même, au moins dans celle des autres, les marques du temps. Comme l'explique Gracq, le lecteur cherche "à ramener les parties successives de l'oeuvre sous un éclairage uniforme et intemporel; sa préférence va au constat réitéré de l'identité, acquiesce avec délectation à la tyrannie unificatrice de la signature (c'est bien de lui !). L'écrivain, devant ses livres, est sensible surtout à son évolution, le lecteur à ses constantes."

L'ordinateur n'est sensible à rien de tout cela mais il peut être utilisé pour analyser ce qui change et ce qui reste quand l'œuvre se déroule. Peut-on lui demander plus encore? Non pas seulement le fait et les modalités de l'évolution, mais, si l'on peut dire, la constance de cette évolution, d'un écrivain à l'autre. Encore ne s'agit-il pas de lier cette constance à l'influence commune et parallèle que l'époque, elle même changeante, exerce sur les écrivains contemporains. On voudrait en déceler l'origine dans la loi physiologique de l'âge. Puisque tout vieillit dans un homme, son corps, sa voix, son allure, mais aussi ses sentiments et ses pensées, n'y a-t-il pas aussi des rides et des raideurs dans l'écriture quand l'âge fait trembler la main et la plume?

Il est peu de profit, dans cette perspective, à attendre d'une étude historique qui s'attacherait à mesurer l'évolution générale constatée dans la littérature et la langue française, de siècle en siècle. On y observe le mouvement des mots, et des réalités qu'ils désignent, dans un temps universel, où les écrivains interviennent à l'époque que le sort leur a désignée, sans que leur âge soit pris en compte. C'est le temps individuel qu'on veut saisir et non pas la place que chaque écrivain occupe dans la suite des temps[6].

Or une fonction est fournie par notre logiciel qui permet dans chaque monographie de mesurer les variations chronologiques et qu'on a expliquée page 41. Le bouton *EVOLUTION* repère et trie dans un corpus les mots qui progressent et ceux qui régressent. Deux listes, positive et négative, sont proposées et ordonnées selon la valeur du coefficient de corrélation. Elles n'ont de sens que si les textes du corpus sont disposés initialement selon la

[6] On trouvera cependant, aux pages 199 et 200 de ce manuel, deux bases où le point de vue historique est adopté. L'une, *CHRONO*, découpe la littérature en tranches, du XVIe au XXe siècle. L'autre, *AUTEURS*, réunit dans une étude comparative les 70 écrivains les mieux représentés dans Frantext.

chronologie et si elles appartiennent à la même plume. Encore faut-il que l'empan soit suffisamment large pour qu'une évolution significative puisse s'y déployer. Enfin on peut craindre des ruptures dans la chaîne, liées au genre ou au sujet, même si l'on peut admettre que les choix thématiques ou génériques s'inscrivent dans un parcours et peuvent eux aussi révéler une tendance. Giraudoux est un exemple de pareil retournement qui lui fait délaisser le roman pour le théâtre. Mais la constance générique peut accompagner l'inconstance de la composition parfois dissimulée, comme chez Proust, derrière le paravent de la publication. Malgré ces aléas et malgré leurs différences initiales, ces deux écrivains semblent suivre des chemins parallèles. D'un univers plus poétique tous deux glissent aux préoccupations morales. Chez l'un comme chez l'autre le temps éteint le sourire et durcit le visage, tandis que l'abstraction gagne du terrain. Parlera-t-on de rapprochement? Ce serait imprudent. On a tout lieu de penser qu'il s'agit là d'une loi de la nature, d'un effet de l'âge et de la maturité. Mais cette hypothèse même est imprudente tant qu'elle n'a pas été obervée et confirmée sur un nombre suffisant d'écrivains[7].

On a donc patiemment constitué une trentaine de monographies littéraires, de Corneille à Proust, et relevé dans chacune les listes significatives des mots en progrès et en régression. Puis on a regroupé ces listes en un seul fichier, où les valeurs individuelles ont été cumulées. En tête se retrouvent les mots que les écrivains apprécient quand l'âge se fait sentir, en queue ceux qu'ils ont tendance à abandonner. On a alors tronqué le milieu de la liste, où se concentraient les mots indécis, livrés aux choix contradictoires des écrivains. En fin de compte il reste une série, alphabétique ou hiérarchique, où figurent les choix unanimes ou majoritaires, qu'ils s'exercent dans un sens ou dans l'autre selon le signe du test. L'extrait très partiel, livré ci-dessous, donne une idée des zones du vocabulaire qui s'effacent ou s'estompent progressivement au fil des ans. La désaffection qui frappe le premier de la liste négative, le point-virgule, est une tendance générale qui s'incrit dans le cours des siècles mais se reconnaît aussi à l'échelle d'une vie et d'une œuvre. En revanche le déclin des mots qui suivent ne s'observe pas sur le long terme, au niveau de la langue ou de la production collective, qu'il s'agisse des ponctuations affectives (*! ...*), du tutoiement (*toi, t'*), du corps (*bouche, bras, poitrine, lèvres, sang, pieds, cheveux, genoux*), de la perception (*regards, voix, silence, bruit*) , ou du sentiment (*pitié, désirs, pleurs, sourire, vanité, sentir, aimer*), ou du monde extérieur *(feu, flots, soleil, ciel, ombre, nature, vent*). C'est entre la naissance et la mort de l'écrivain que se limite la désaffection. Chaque génération découvre ces promesses de la vie et les abandonne avec l'âge. Ce mouvement ondulatoire, lié à la physiologie et

[7] Ces derniers propos ont été tenus en 1983, à l'occasion du Colloque du Centenaire de Giraudoux au Collège de France (article paru dans la *Revue d'Histoire Littéraire de la France*, p.832-841). Il aura fallu près de vingt ans pour amasser les matériaux et permettrre la synthèse.

répété de génération en génération, mériterait une étude attentive qui n'a pas de place ici.

Extrait partiel des mots généralement en régression chez les écrivains

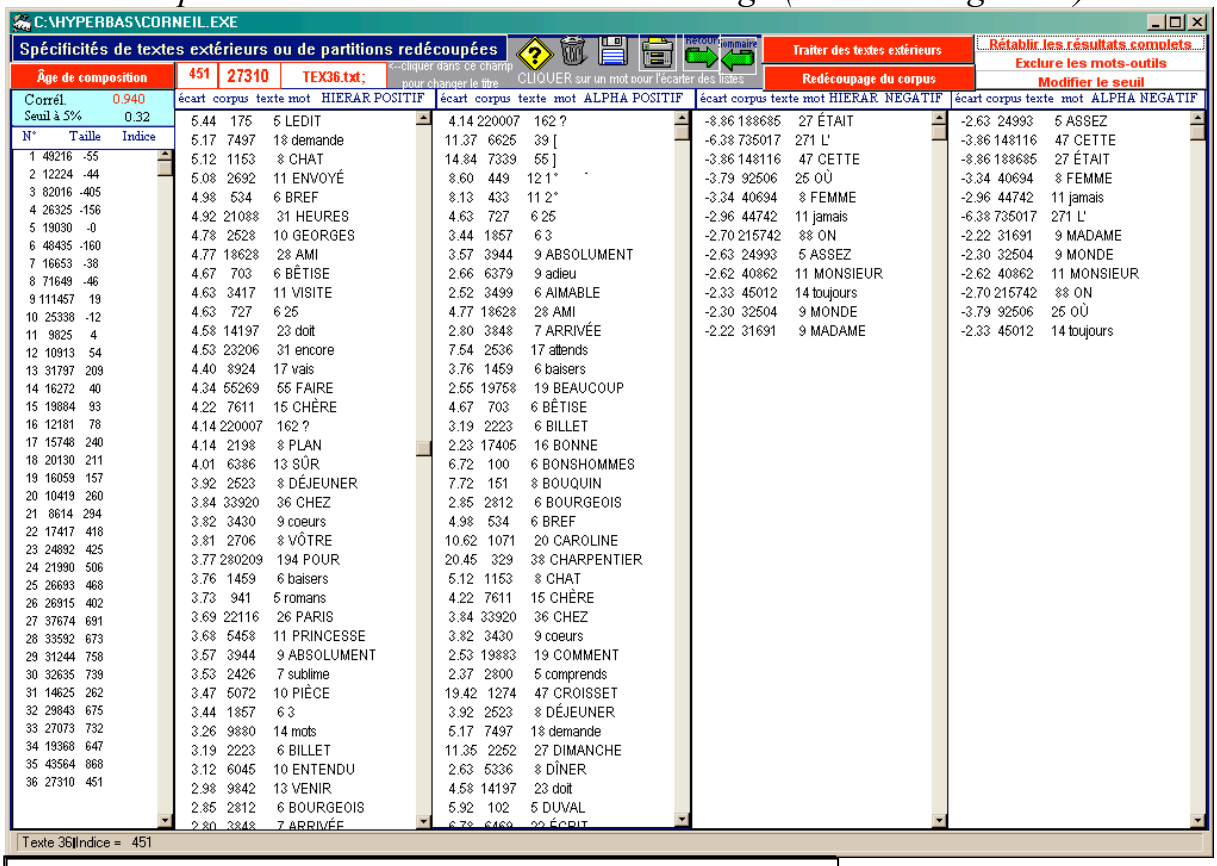
-3.50	;	-2.70	pieds	-2.43	bras
-3.32	bientôt	-2.68	lieux	-2.40	flots
-3.21	toi	-2.67	voix	-2.38	aimer
-3.16	t'	-2.65	pitié	-2.38	toujours
-3.01	cependant	-2.62	donc	-2.35	légèrement
-2.99	chaque	-2.61	:	-2.35	lèvres
-2.97	bouche	-2.58	bruit	-2.35	tantôt
-2.96	feu	-2.57	lentement	-2.34	désormais
-2.95	regards	-2.55	bout	-2.33	chacun
-2.94	instant	-2.54	...	-2.32	cheveux
-2.89	adieu	-2.53	poitrine	-2.31	sourire
-2.87	misérable	-2.52	longs	-2.30	sentir
-2.87	tandis	-2.50	autour	-2.29	douce
-2.86	traits	-2.50	calme	-2.29	puisse
-2.82	pensées	-2.48	désirs	-2.29	sang
-2.82	silence	-2.48	lève	-2.28	mouvements
-2.80	loin	-2.48	tomber	-2.27	genoux
-2.71	!	-2.46	pleurs		
-2.71	combat	-2.44	vanité		

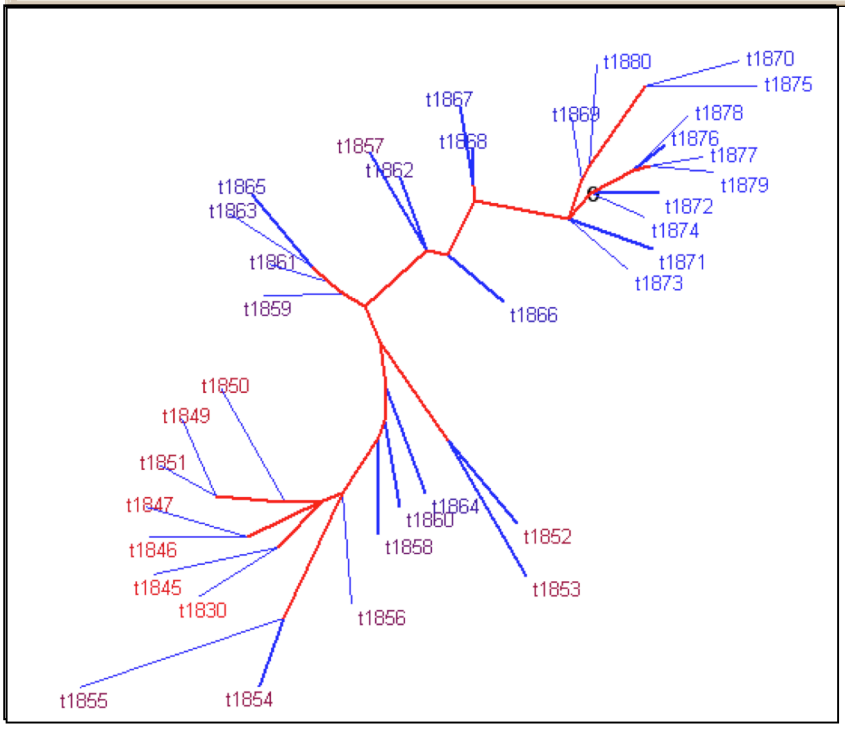
Mais le fichier qui recueille les observations du sondage est disponible sous le nom EVOL11.TXT. Il peut servir en outre de référence pour dater les textes. On ne parle pas ici de datation absolue, comme fait le carbone 14. Mais de datation relative, comme fait l'archéologue devant des ossements d'adulte ou d'enfant. À quel signe reconnaît-on qu'un texte a des chances d'être une œuvre de jeunesse ou de maturité? Il suffit de comparer son vocabulaire à cette liste de référence, où chaque mot est pourvu non seulement d'un test de tendance, négatif ou positif, mais aussi d'une fréquence cumulant l'ensemble des 31 monographies explorées (soit plus de 50 millions de mots). Pour chaque mot du texte présent dans la liste de référence, on calcule ainsi un écart par rapport à cette norme. Négatif ou positif, cet écart (z) est cumulé en tenant compte du signe et le la valeur du test de tendance (r) selon la formule:$\sum(z*r)$. Si z et r ont le même signe, la contribution est positive et cela signifie que le texte examiné répugne à l'emploi d'un mot habituellement en régression ou qu'au contraire il cultive un mot qu'on apprécie volontiers avec l'âge. Au total la somme est positive si ce choix est maintenu pour la plupart des mots de référence et l'on en conclut que le texte a des chances d'appartenir à la maturité de l'écrivain. Un résultat négatif est au contraire l'indice d'une œuvre de jeunesse.

Là encore le genre et le sujet influent sur l'indice et peuvent contredire occasionnellement l'effet de l'âge. Le genre poétique a ainsi tendance à produire des valeurs négatives, comme certains thèmes, par exemple le paradis de l'enfance ou la saison des amours. Un résultat isolé peut être sujet à caution et mêler des facteurs dont le dosage est difficile à mesurer. Mais si l'on soumet au test successivement tous les textes du même écrivain, la décantation est significative et le test passe du négatif au positif à mesure que l'œuvre se déroule dans le temps. Il en est ainsi des 36 années de la correspondance de Flaubert, représentées dans la colonne de gauche où le test de l'âge suit l'évolution attendue, que souligne un coefficient de corrélation extrêmement

fort (r = 0.94 pour 35 degrés de liberté). Ce test externe, pratiqué à l'aide d'une grille de lecture fixe, est heureusement confirmé par l'analyse interne et notamment par la mesure de la distance intertextuelle (voir figure suivante rendant compte de la connexion de Muller).

La correspondance de Flaubert. L'indice de l'âge (colonne de gauche)





La Correspondance de Flaubert. Distance intertextuelle (indice de Muller)

La correspondance de Flaubert est un exemple privilégié, où le genre n'intervient pas, non plus que le thème. Mais si l'enquête envisage toute l'œuvre du même auteur, le seuil à 5% est encore atteint (r = 0,50 pour n = 15). Et c'est le cas pour presque toutes les monographies dont les données nous sont connues, pourvu que l'étendue chronologique de l'œuvre et le nombre de textes pris en compte soient suffisamment vastes pour admettre une évolution significative, ce qui exclut quelques écrivains comme Racine, Pascal et Rimbaud. En pratique on s'abstiendra d'engager la procédure si le corpus a moins de dix titres et si moins de dix ans séparent le premier texte du dernier. Avec ces réserves le tableau ci-dessous montre que peu d'auteurs échappent à l'emprise de l'âge. Mis à part Diderot, il s'agit d'écrivains qui se sont illustrés en vers et en prose (Baudelaire, La Fontaine Aragon) et le heurt des genres a créé un remous dans la dérive du temps.

La dérive du temps chez les écrivains

écrivain	*nb-textes*	*corrél.*	*seuil 5%*	*écrivain*	*nb-textes*	*corrél.*	*seuil 5%*
Corneille	34	0.75	0.66	Nerval	11	0.62	0.60
Molière	30	0.42	0.36	Balzac	49	0.74	0.28
La Fontaine	38	-0.07	0.36	Dumas	42	0.50	0.30
Marivaux	54	0.69	0.25	Sand	62	0.69	0.25
Montesquieu	13	0.56	0.55	Flaubert	15	0.50	0.51
Rousseau	35	0.71	0.22	Flaubert (corresp)	36	0.94	0.31
Voltaire	52	0.53	0.27	Maupassant	38	0.58	0.32
Diderot	24	-0.02	0.40	Baudelaire	15	0.01	0.51
Chateaubriand	16	0.62	0.49	Verne	42	0.63	0.30
Stendhal	11	0.56	0.60	Zola	20	0.63	0.44
Hugo (vers)	14	0.79	0.53	Proust	18	0.42	0.46
Hugo (prose)	30	0.68	0.36	A.France	23	0.41	0.41
Lamartine	15	0.71	0.51	Gracq	17	0.47	0.47
Musset	38	0.57	0.32	Aragon	23	0.23	0.41
Vigny	15	0.49	0.51				

5 - LES PHRASES-CLÉS

On appelle phrases-clés les passages caractéristiques par quoi on tente de donner une idée du contenu d'un texte. À l'heure d'Internet, cette démarche se justifie par l'abondance de l'information et la nécessité de réduire, par des raccourcis, le temps de la consultation. Nous avions répugné jusqu'ici à proposer cette fonction, parce que, en l'absence d'une intelligence humaine, seule capable de comprendre et de synthétiser un texte, une série discontinue d'extraits ne saurait constituer un véritable résumé. À la réflexion cette approche est du même type que celle qui justifie la recherche des spécificités lexicales ou des segments répétés et les extraits retenus peuvent être considérés comme des spécificités phrastiques. Certes il n'y a pas de statistique possible sur les

paragraphes, parce que chacun est quasiment unique dans un texte. La technique de calcul des segments répétés ne peut en effet être étendue à la dimension de la phrase et encore moins au delà, parce que la combinatoire des mots devenant exponentielle, la répétition ne s'observe plus à ce niveau. Il faut donc utiliser des procédures indirectes. La plus évidente consiste à prendre appui sur les spécificités lexicales et à retenir les paragraphes qui contiennent les spécificités les plus nombreuses et les plus significatives. Encore faut-il doser, affiner et combiner ces deux critères.

Le nombre absolu des spécificités enregistrées dans un paragraphe dépend du seuil adopté pour en établir la liste. Généralement ce seuil correspond à la probabilité de 0,05 laissée au hasard. Nous l'avons adopté également, avec un correctif qui limite à 200 le nombre des éléments considérés pour un même texte. Ces éléments sont évidemment les premiers de la liste triée des spécificités, une fois que cette liste a été épurée. On a écarté en effet les noms propres parce que leur apport, trop prévisible, a une valeur amoindrie pour l'analyse de contenu. On a cru devoir rejeter aussi les mots grammaticaux, parce que leur influence mêle à la thématique des effets stylistiques, qu'il vaudrait mieux étudier à part. Pour simplifier la sélection, les mots-outils ont été assimilés aux mots courts (ayant de 1 à 3 lettres), avec une liste complémentaire des mots grammaticaux de longueur supérieure. Il y a quelques laissés pour compte dans cette mesure: les substantifs très courts comme *an* ou *âme* ont été négligés et inversement quelques adverbes ont été épargnés.

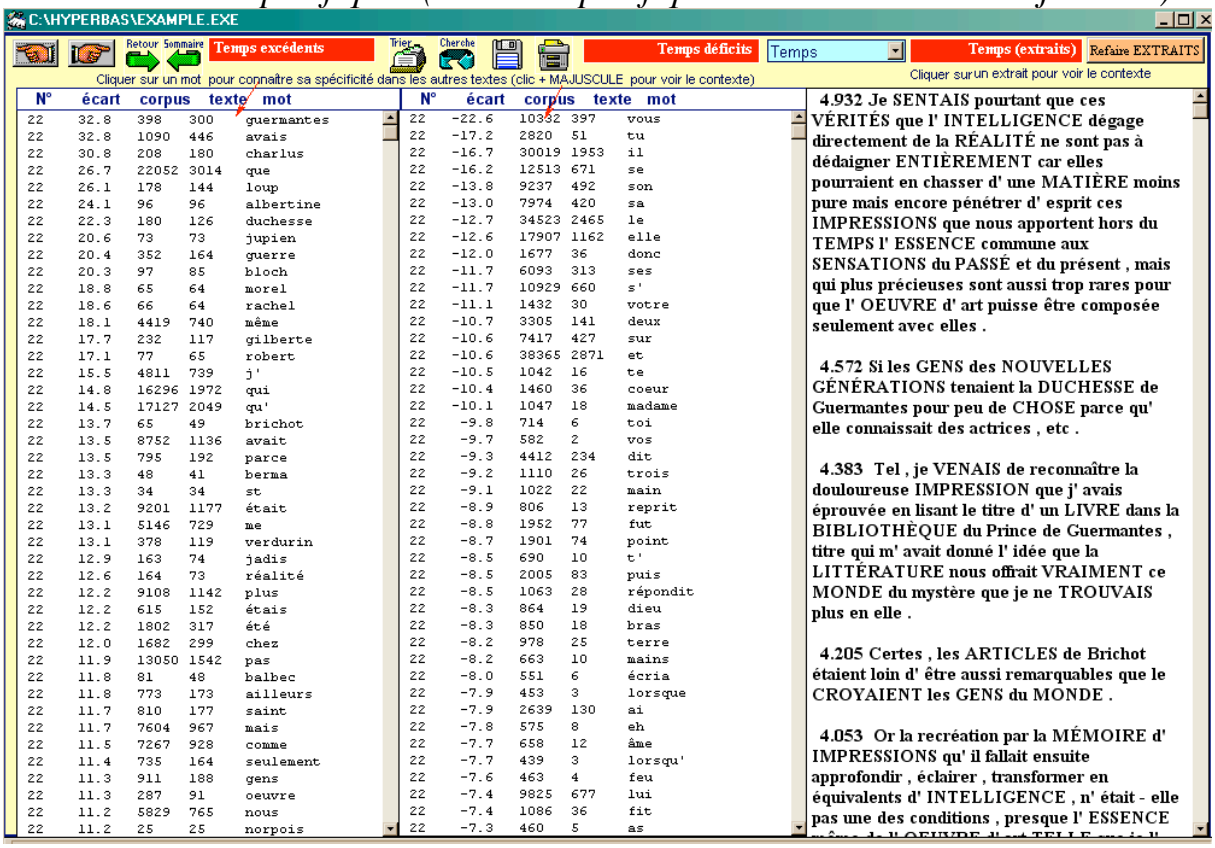
Mais le filtrage des spécificités n'est pas suffisant. Leur nombre dans une phrase doit être pondéré par la longueur de la phrase, sans quoi les phrases les plus longues auraient un avantage injustifié. Encore faut-il compter dans la phrase les candidats à la spécificité, lesquels sont soumis aux mêmes critères de sélection et doivent n'être ni des noms propres, ni des mots grammaticaux. Le score obtenu pour un paragraphe prend donc la forme d'un quotient reçus/candidats.

Les deux termes du quotient doivent encore subir des retouches préalables. D'une part le nombre de spécificités (établi au numérateur) doit tenir compte de la spécificité plus ou moins forte des termes relevés. On fera donc la somme des écarts relevés pour chacun (rappelons que ce sont des probabilités calculées selon le modèle hypergéométrique et converties ensuite en écarts réduits). Un même total peut ainsi être obtenu par un nombre restreint de spécificités fortes ou par un nombre plus grand de spécificités moindres. On a tout de même diminué l'amplitude des écarts, en les réduisant à leur racine carrée. D'autre part le nombre de candidats (affiché au dénominateur) est aussi soumis à pondération (là encore la racine carrée). Enfin quelques précautions ont été prises pour éliminer les résultats excentriques: paragraphes trop courts (moins de 20 mots) ou trop longs (plus de 120 mots), nombre de spécificités

insuffisant, paragraphes trop nombreux à franchir le seuil (la limite est de 150). Ces derniers critères dépendant de l'étendue des textes doivent être adaptés en conséquence.

Au terme de cette sélection précautionneuse, les paragraphes sont triés selon le score obtenu, et enregistrés dans les pages consacrées aux spécificités. Pour chaque texte on obtient en parallèle la liste des spécificités lexicales (positives et négatives) et celle des spécificités phrastiques. Ces dernières sont sensibles au clic de la souris et renvoient aux pages originales du texte. On en trouvera une illustration ci-dessous, empruntée à un texte de Proust: *Le temps retrouvé*.

Les extraits spécifiques (les mots spécifiques sont restitués en majuscules)



Lorsqu'on n'est pas satisfait du résultat, en particulier lorsqu'aucune phrase n'a été relevée, on peut déplacer les filtres, admettre des critères moins sévères et recommencer l'opération. On sollicitera alors le bouton REFAIRE RESUMÉ. Cette opération s'impose parfois quand des retours de chariot intempestifs ont été placés malencontreusement à la fin de chaque ligne. Dans un tel cas la phrase, étant assimilée à la ligne, devient trop courte pour que le calcul soit entrepris. La solution consiste d'abord à solliciter le bouton CORPUS EN VERS du menu principal, puis à refaire l'extraction des résumés.

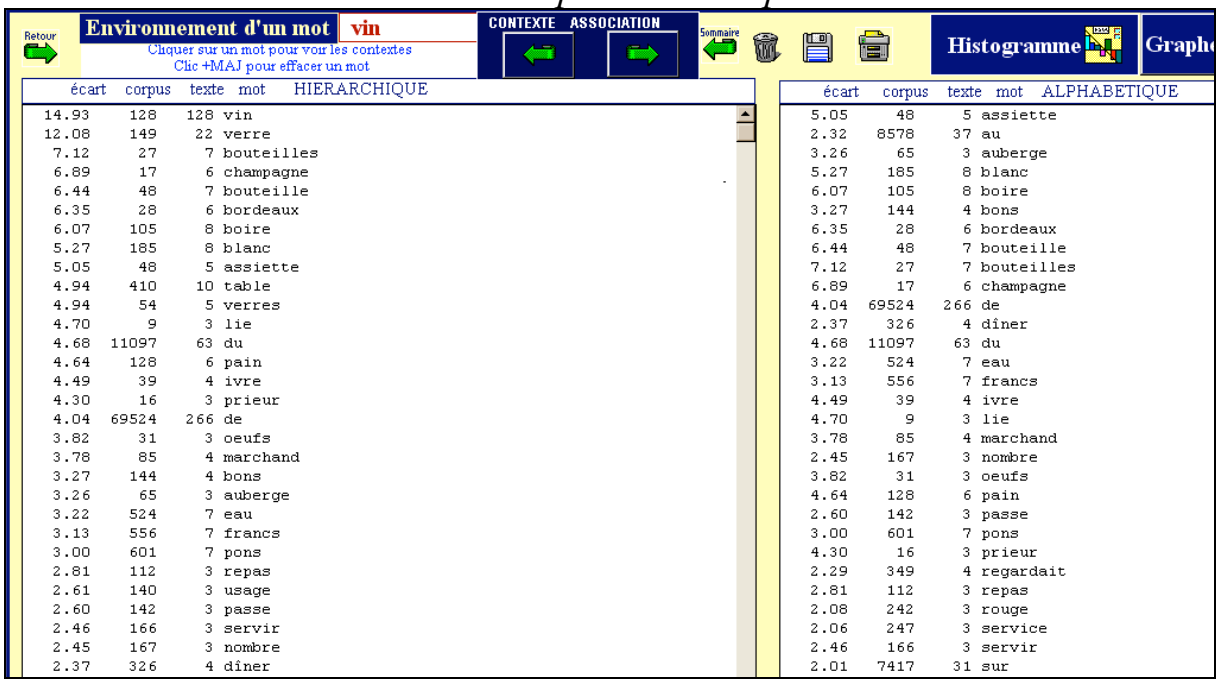
CHAPITRE 9

Les menus THÈME et ASSOCIATIONS

La recherche thématique est un calcul de spécificité particulier, puisqu'on ne recherche plus une accointance entre un mot et un texte, mais une relation privilégiée entre les mots eux-mêmes - ce que mesure aussi le calcul de corrélation, quand deux séries sont juxtaposées dans le même graphique. Mais la procédure ne se réduit pas ici à deux mots confrontés, mais à l'ensemble indéfini de tous les mots qui peuvent se trouver dans l'entourage d'un mot (ou d'un groupe de mots) qu'on définit comme étant le pôle. On peut changer de pôle aussi souvent qu'on veut, si bien que la procédure se trouve multipliable et réversible.

Cette procédure se trouve dans la page CONTEXTE. On la déclenche par un clic sur le bouton THEME (bouton rouge à droite de l'écran). Elle prend en considération la liste des contextes recensés pour une recherche donnée et restitués dans la page CONTEXTE et les suivantes.

La recherche thématique: les mots que le vin *attire*



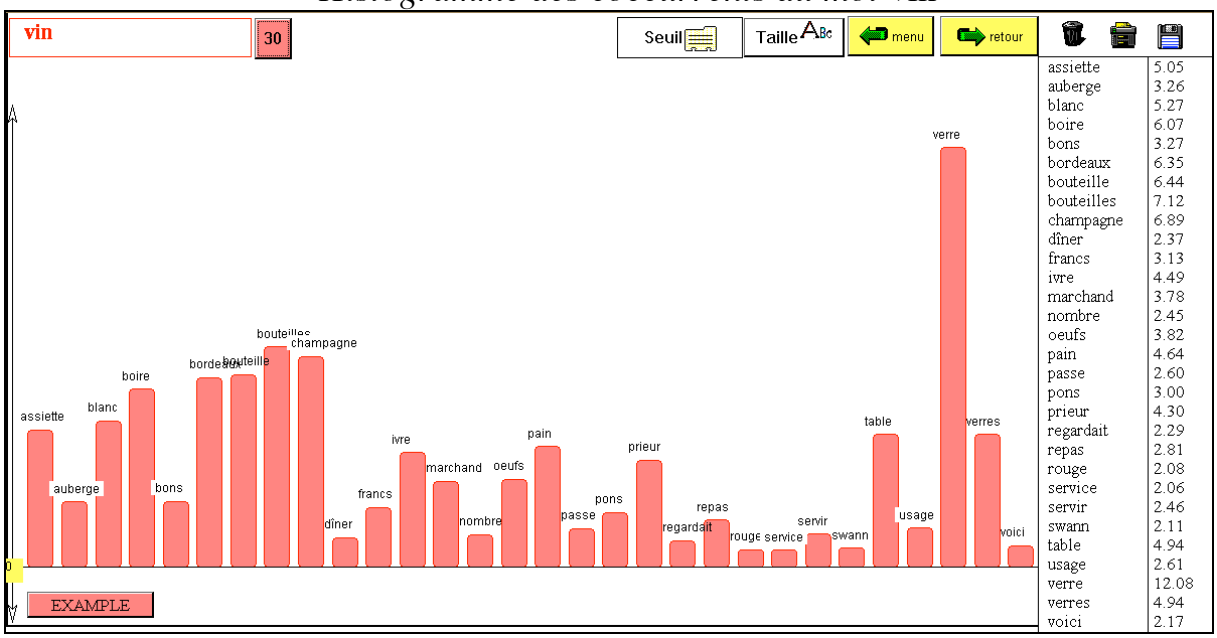
Tous les mots trouvés là (sauf ceux qui appartiennent aux références) sont soumis au tri et aux calculs de fréquence, et la liste obtenue est confrontée au dictionnaire du corpus. Il s'agit en somme de la comparaison classique de deux corpus, l'un englobant l'autre. La liste des corrélats est établie de façon

alphabétique (à droite) et hiérarchique (à gauche). Le nombre des éléments peut être réduit si on adopte un critère plus sévère (bouton SEUIL).

Première approche simplifiée

Les résultats apparaissent dans une liste représentée ci-dessus et illustrée par un histogramme joint ci-dessous. L'exemple choisi est relatif au *vin*. La relation avec le *verre* et la *bouteille* coule ici de source. Mais lorsqu'il s'agit d'une notion plus trouble, les résultats peuvent être moins triviaux et moins prévisibles et leur intérêt herméneutique plus affirmé.

Histogramme des coccurrents du mot vin



C'est le cas du mot *nature* dont le sémantisme est ambigu: tantôt il accompagne les spéculations abstraites sur l'homme et le monde, tantôt il se fait plus concret, surtout à l'époque de Balzac, et évoque les paysages agrestes. On s'attend que les mots associés à ce double sens dessinent une constellation bipolaire. C'est ce qu'on observe si l'on active le bouton GRAPHE, qui conduit à la figure ci-dessous. De la liste des spécificités le programme commence par éliminer les mots grammaticaux (au moins ceux qui ont moins de quatre lettres), puis constitue le tableau général des cooccurrences, et procède au calcul et au tri de tous les indices qui évaluent la distance entre les mots de la liste pris deux à deux.

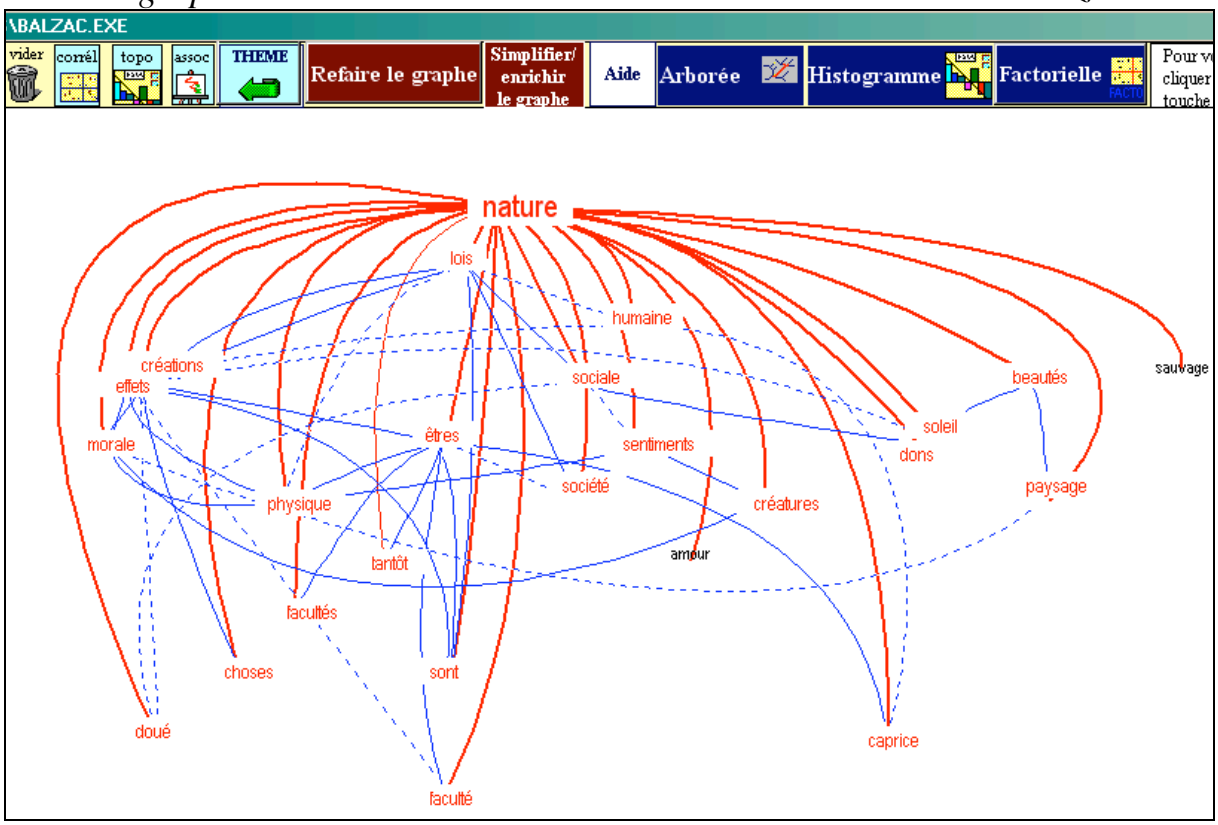
Les liens représentés dans le graphe sont en rouge s'ils concernent le mot-pôle, en bleu s'ils relient deux mots liés au pôle, en noir dans les autres cas. Quant aux mots, ceux qui sont des noeuds fréquentés sont en rouge, ceux qui ont peu de liaisons (moins de cinq) sont en noir. Si les relations collatérales (en noir) encombrent sans profit le graphique, on peut les faire disparaître (ou les rétablir) en faisant appel au bouton SIMPLIFIER/ENRICHIR LE GRAPHE.

Le retour au texte est assuré, si l'on clique sur un mot avec appui de la touche MAJUSCULE. Apparaissent alors en vidéo inverse les paragraphes successifs où s'observe la cooccurrence du mot en question avec le mot-pôle.

Le calcul du graphe arborescent et de la position des noeuds et des arcs est assuré par le logiciel libre GRAPHVIZ (licence GNU) aimablement communiqué par Serge Heiden. Les données sont fournies à ce programme selon les spécifications du langage DOT et les résultats bruts, enregistrés dans un fichier au suffixe .DOT, sont repris par Hyperbase dans une présentation graphique qui tient compte non seulement des positions mais aussi des pondérations.

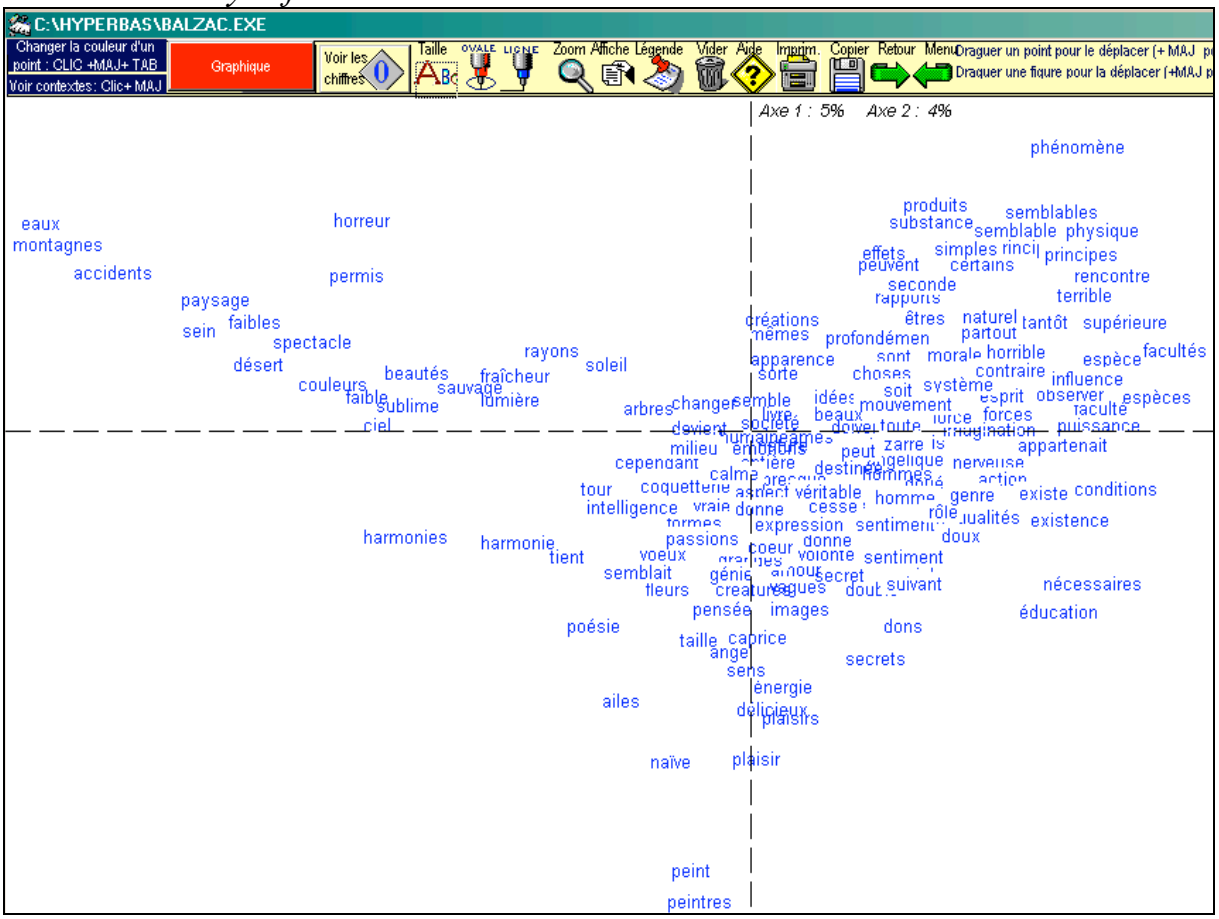
La mesure de la cooccurrence est assurée par le calcul hypergéométrique. La valeur de l'écart prend place dans un tableau carré et se substitue à la mesure brute des cooccurrences. Dès lors un tel tableau peut être soumis aux méthodes habituelles. Trois boutons HISTOGRAMME, FACTORIELLE et ARBOREE sont disponibles à cet effet et s'appliquent au mot qui a été choisi pour pôle.

Le graphe des cooccurrents du mot nature *dans 50 romans de Balzac*

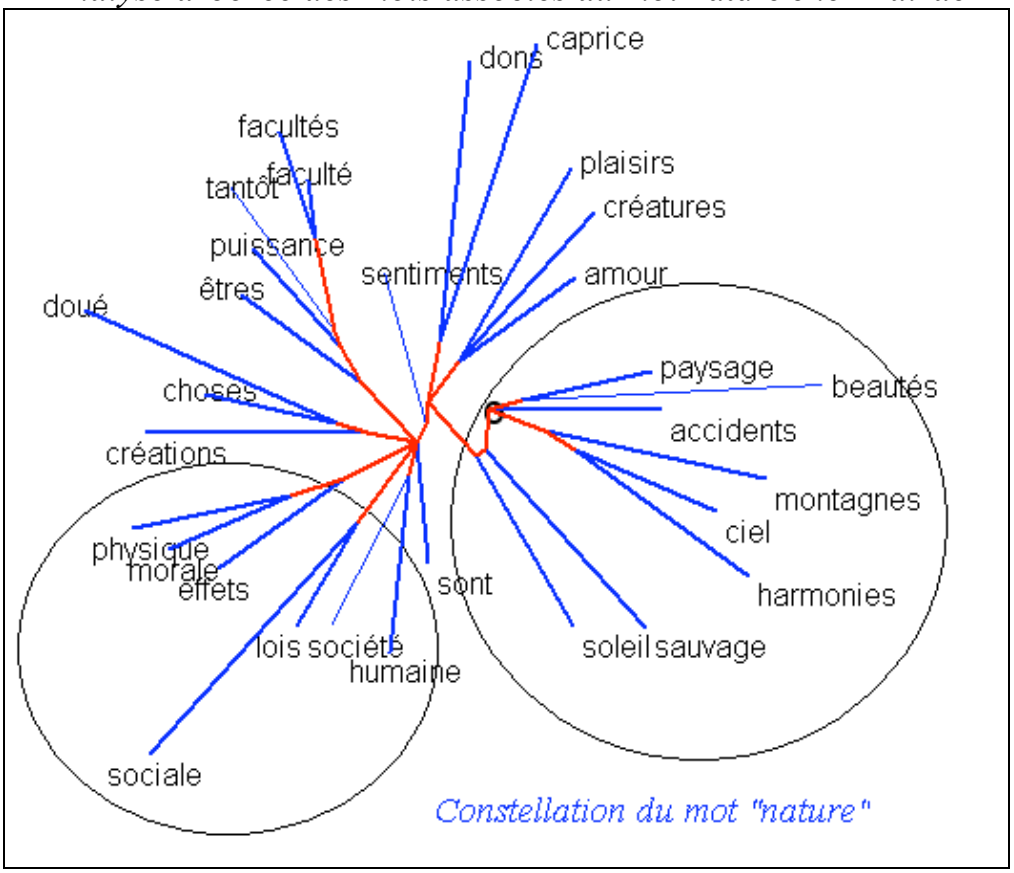


Le graphe obtenu distingue un réseau à droite qui contemple le *paysage sauvage*, le *soleil*, les *beautés*, les *dons*, les *caprices* de la nature, tandis qu'à droite on disserte sur la *création*, les *êtres*, les *facultés humaines*, les *lois physiques*, *sociales* et *morales*. Confirmation en est donnée par d'autres méthodes : analyse factorielle et analyse arborée. À partir du même tableau de cooccurrences, les graphiques ci-dessous opposent les diverses acceptions de la nature.

Analyse factorielle des mots associés à la nature *chez Balzac*



Analyse arborée des mots associés au mot nature *chez Balzac*



Seconde approche

La démarche qu'on vient de proposer a des limitations. La principale vient du rétrécissement du champ d'observation. Une fois qu'ont été isolés les passages où apparaît le mot-pôle, tous les calculs ultérieurs ne considèrent que ce seul sous-corpus, cet assemblage d'extraits aimantés par le pôle. En gardant la liste des spécificités, on pourrait étendre au corpus entier l'examen des corrélations du thème isolé, ce qui aurait l'avantage de fournir des effectifs plus larges, plus fiables et moins étroitement orientés. D'autre part, vu la faiblesse des effectifs, on a eu des scrupules à utiliser des méthodes lourdes de pondération: ainsi le programme de représentation graphique ne retient du tableau de cooccurrences qu'une information binaire: présence ou absence. L'analyse de correspondance n'applique qu'une transformation bénigne (une racine carrée) au tableau des cooccurrences. Quant à l'analyse arborée, elle prend en compte des fréquences relatives: tout élément du tableau est pondéré par le total de la ligne ou de la colonne où il se trouve, ou mieux par les deux totaux marginaux à la fois.

On a donc offert un chemin plus long et plus exigeant à ceux qui préfèrent des méthodes plus sûres et plus rigoureuses. Ce chemin,dans la page THEME, est indiqué par un panneau vert, portant la mention PREPARE.

Il prend pour point de départ le point d'arrivée de l'étape précédente, c'est à dire la liste des spécificités thématiques. Tous les mots de cette liste sont alors recherchés séquentiellement dans tous les paragraphes du corpus, en notant à chaque fois les cooccurrences rencontrées. Il faut quelques secondes de patience pour obtenir le tableau complet de ces rencontres, enregistré dans un fichier qui porte le nom de la base avec le suffixe .SUC. Deux autres fichiers sont aussi constitués quand le calcul hypergéométrique est appliqué à ce tableau pour en extraire des enseignements statistiques (suffixe 7.TXT pour le tableau des écarts et 3.TXT pour la liste triée des écarts).

La mise en œuvre du calcul hypergéométrique exige ici quelques éclaircissements. Habituellement les paramètres appliquées sont les suivants :
T = nombre total de mots du corpus,
t = nombre de mots de la partie considérée,
f = fréquence du mot dans le corpus
et k = fréquence du mot dans la partie. Ces paramètres sont conservés, et s'appliquent successivement aux deux mots intéressés par la cooccurrence. Mais k vaut 0 (on veut évaluer la probabilité d'une absence) et t n'est plus la taille

d'une partie ou d'un texte, mais celle d'une page ou d'un paragraphe. La cooccurrence peut en effet être évaluée à ces deux niveaux. Les deux méthodes sont proposées simultanément. En proposant pour *t* la taille moyenne d'une page ou d'un paragraphe, on ne ne procède au calcul élémentaire qu'une seule fois, quitte à étendre le résultat au nombre de pages ou de résultats (*n*), comme s'il s'agissait d'un tirage accompli *n* fois.

Le calcul de la cooccurrence théorique s'appuie sur deux probabilités: celle qui est attachée à l'absence du premier mot (*p1*) et celle qui est propre à l'absence du second (*p2*) dans l'espace considéré. Chacune de ces probabilités relève du calcul hypergéométrique, posé comme suit:

$$prob\ (x=k) = \frac{f!\ (T-f)!\ t!\ (T-t)!}{k!\ (f-k)!\ (t-k)!\ (T-f-t+k)!\ T!}$$

soit en simplifiant, puisque *k* doit être nul

$$prob\ (x=0) = \frac{f!(T-f)!t!(T-t)!}{f!t!(T-f-t)!T!}$$

Le complément à l'unité de chacune de ces deux probabilités sert à mesurer les chances de rencontrer le mot dans le segment considéré, quelle que soit sa fréquence (1 ou plus): *q1 = 1 - p1 , q2 = 1 - p2*. Le produit des deux résultats

$$q = q1 * q2$$

mesure alors les chances de rencontrer la cooccurrence des deux mots à la fois dans la même page (ou le même paragraphe). Par cooccurrence on entend la coprésence, ni l'ordre des deux éléments ni leur répétition éventuelle n'étant pris en compte. En multipliant cette probabilité élémentaire par le nombre de pages (ou de paragraphes), on obtient l'effectif théorique des cooccurrences. Reste à comparer l'effectif réel[8] à l'effectif théorique, ce dont rend compte le calcul classique de l'écart réduit.

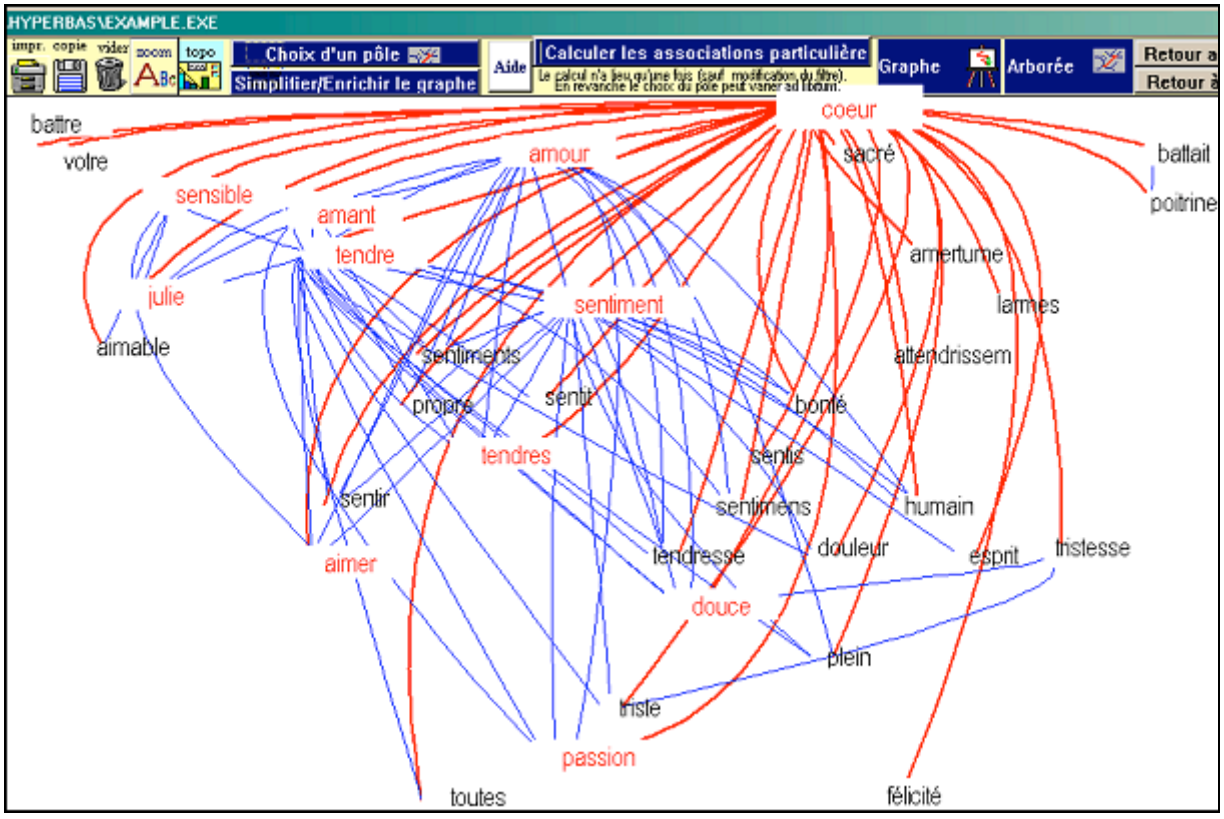
Le premier résultat du traitement est une liste classée des cooccurrences significatives, à l'image de celle qui suit, établie autour du mot *cœur* dans la base EXAMPLE. Comme il s'agit ici de graphies, on voit souvent le singulier accompagner le pluriel pour le même vocable (par exemple *tendre* et *tendres*, *sentiment* et *sentiments*), cet artéfact un peu trivial disparaissant dans les bases lemmatisées. La liste, qui est beaucoup plus riche que l'extrait représenté ci-dessous, est moins accueillante aux verbes qu'aux catégories nominales et cela tient aussi au statut des graphies et à leur dispersion dans le paradigme verbal. Là aussi l'équilibre est mieux respecté quand le lemme est pris en considération.

[8] Bien entendu le dénombrement de l'effectif réel relève les coprésences, en négligeant pareillement l'ordre et les répétitions. Après avoir dans un premier temps utilisé la méthode de Pierre Lafon, exposée p; 163 de sa thèse, on a renoncé à cet algorithme parce qu'il s'applique à des couples orientés et non à des paires où l'ordre est indifférent. Il y a bien aussi dans cet ouvrage un second algorithme qui pourrait s'appliquer à des paires (c'est à dire à la coprésence). Mais le calcul est complexe et nous avons préféré plus de simplicité.

Liste des cooccurrences significatives

LISTE ALPHABÉTIQUE			LISTE HIÉRARCHIQUE		
test	mot1	mot2	test	mot1	mot2
17.56	coeur	battait	642.18	tendre	tendres
18.59	coeur	amour	549.38	tendres	tendresse
15.72	coeur	sensible	278.86	tendre	tendresse
15.32	coeur	battre	227.10	tristesse	triste
9.10	coeur	votre	221.68	sentiment	sentiments
11.66	coeur	amant	40.38	tendres	douce
25.81	coeur	tendre	37.68	amour	tendres
11.05	coeur	tristesse	35.51	tendre	douce
22.52	coeur	sentiment	34.71	amour	tendre
15.63	coeur	julie	32.48	sentiment	passion
10.49	coeur	sentiments	31.25	amour	propre
7.35	coeur	sophie	30.53	bonté	tendres
8.84	coeur	larmes	28.41	amour	passion
8.38	coeur	poitrine	28.38	sentiment	tendres
10.07	coeur	propre	27.98	humain	passion
8.64	coeur	attendrissem	26.83	coeur	tendres
4.31	coeur	combien	25.81	coeur	tendre
8.53	coeur	sentis	25.31	tendres	passion
8.11	coeur	sentit	24.79	sentiments	passion
8.09	coeur	amertume	22.52	coeur	sentiment
8.41	coeur	bonté	21.34	sacré	droits
11.84	coeur	sacré	20.31	humain	douleur
7.50	coeur	sentait	19.91	amour	sentiment
26.83	coeur	tendres	19.42	tendre	sentiment
12.43	coeur	humain	19.17	coeur	passion
9.91	coeur	esprit	18.73	tendres	aimer
8.19	coeur	sentimens	18.72	bonté	douce
5.52	coeur	malgré	18.59	coeur	amour
8.04	coeur	aimable	18.58	tendre	bonté
10.40	coeur	douleur	18.01	coeur	triste
9.36	coeur	sentir	17.66	tendre	passion

À la fin du processus, on est conduit à une page où sont regroupés les outils propres au traitement des associations ou cooccurrences. Voir ci-dessous le graphe des cooccurrences du mot *cœur* dans le corpus *Example.*

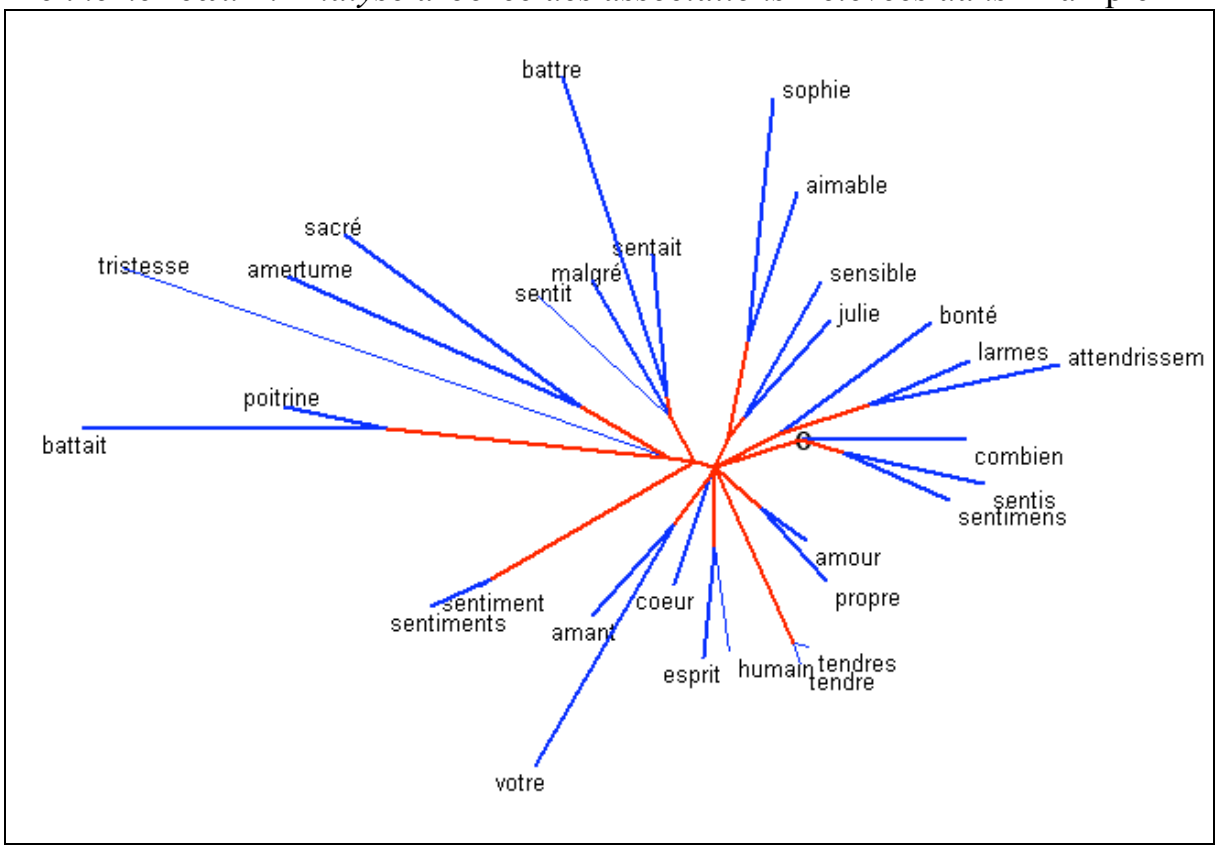


1 - Le bouton CALCULER LES ASSOCIATIONS traite la liste des mots retenus dans la page THEME (ou dans la page LISTE), et procède au calcul et au tri de tous les indices qui évaluent la distance entre les mots de la liste pris deux à deux. Le calcul est déjà fait quand la page est montrée. Il n'a pas à être renouvelé sauf en cas d'écrasement des fichiers qu'il génère (avec suffixe .SUC pour le tableau des cooccurrences brutes, 7.TXT pour le tableau des écarts et 3.TXT pour la liste triée des écarts). Ces trois fichiers sont écrasés et mis à jour dès qu'une nouvelle liste de mots est proposée, soit par la fonction THEME, soit par la fonction LISTE.

2 - Solliciter le bouton CHOIX D'UN POLE pour établir le graphe d'un mot de la liste des cooccurrents (le bouton GRAPHE permet de reproduire le graphique en gardant le même pôle). La liste des mots disponibles apparaît et attend le clic de la souris. Si les relations collatérales (en noir) encombrent sans profit le graphique, on peut les faire disparaître (ou les rétablir) en faisant appel au bouton SIMPLIFIER/ENRICHIR LE GRAPHE. Le retour au texte est assuré, si l'on clique sur un mot avec appui de la touche MAJUSCULE. Apparaissent alors en vidéo inverse les paragraphes successifs où s'observe la cooccurrence du mot en question avec le mot-pôle.

3 - Le programme ARBOREE prend appui sur le tableau des écarts, soit en explicitant le réseau de tous les mots qui s'y trouvent, soit en ne retenant que les mots qui sont le plus souvent associés au mot choisi pour pôle.

Le thème "cœur". Analyse arborée des associations relevées dans Example



Là aussi l'exemple du *cœur* (ci-dessus) serait plus clair si l'on avait affaire à des lemmes. Les verbes *battre* et *sentir* distribuent leurs formes à différents endroits pour des raisons qui tiennent à la conduite du récit ou des dialogues et non au sémantisme du mot.

CHAPITRE 10

Le menu TOPOLOGIE

Les fonctions statistiques supposent habituellement une segmentation du corpus en textes séparés et sont fondées sur des fréquences ou des effectifs observés dans ces textes. Cette segmentation se justifie souvent parce que les textes rassemblés se distinguent par la date, le genre, l'auteur, le thème ou tout simplement le titre. Quand il s'agit d'œuvres distinctes, la partition du corpus semble aller de soi. Pourtant cela ne va pas parfois sans quelque arbitraire. Les nécessités du traitement posent souvent des problèmes de sélection, de jalons et de frontières et imposent des équilibres, des regroupements ou des sectionnements. Or paradoxalement la décision doit être prise avant que le traitement puisse l'éclairer. Au stade du traitement, la statistique, étant essentiellement comparative, est amenée à durcir les oppositions entre les parties de l'ensemble. Et la segmentation initiale, bonne ou mauvaise, se trouve alors artificiellement justifiée.

Il serait de meilleure méthode de n'imposer aucune segmentation au départ de l'entreprise et de laisser à l'analyse du contenu le soin d'en suggérer une. Ce n'est pas le cas malheureusement de notre logiciel Hyperbase, dont les jalons, une fois définis, ne peuvent être déplacés, sauf à refaire le traitement. Il est toutefois possible d'observer les mots dans le détail de leur répartition tout au long du corpus, en négligeant les frontières des textes, et sans se soucier de constituer des effectifs, des fréquences et des sous-fréquences, chaque occurrence étant considérée individuellement, indépendamment du texte où elle se trouve. On rejoint là une perspective ouverte par Pierre Lafon dans la dernière partie de sa thèse, où il oppose les séquences aux fréquences.

Représentation graphique

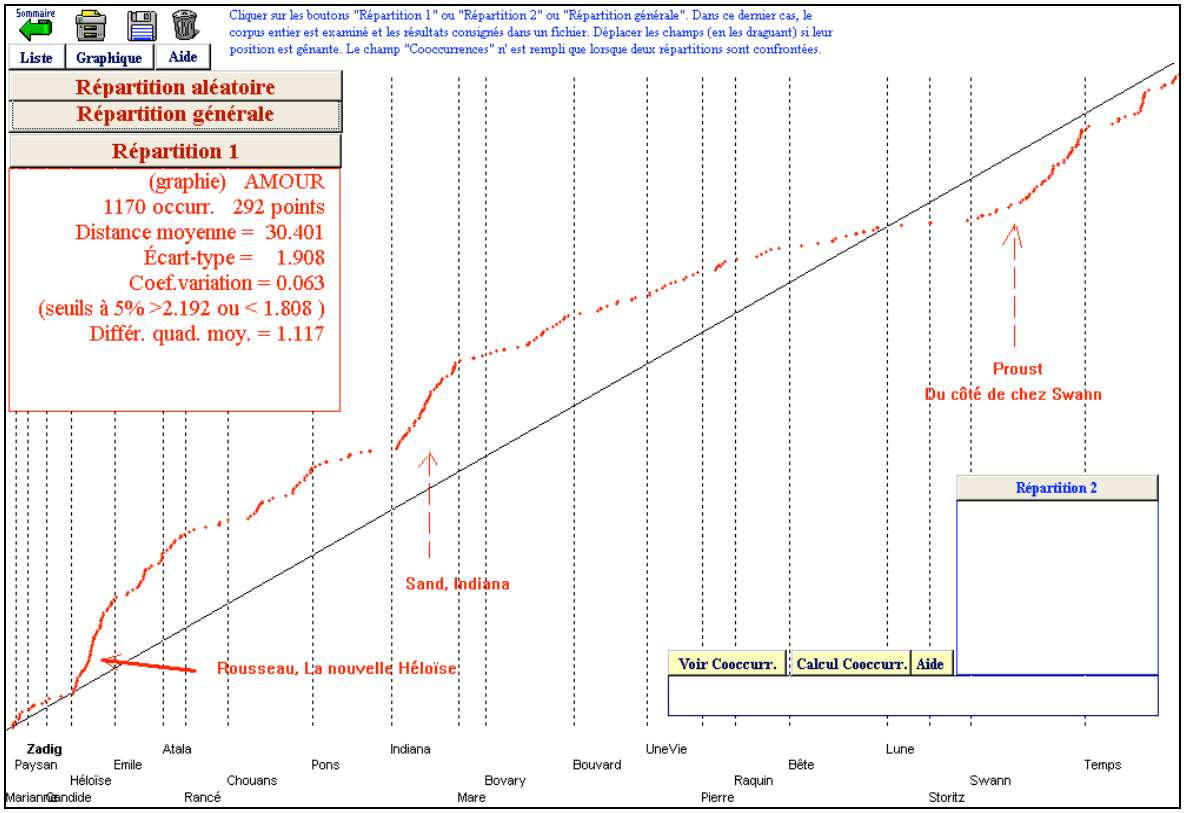
Graphiquement la chose est aisée à représenter, soit qu'on établisse sur un plan la suite linéaire des paragraphes du corpus, en les répartissant symboliquement à la queue leu leu, de place en place et de ligne en ligne, comme l'écriture le fait pour les mots, soit qu'on adopte tout bonnement la ligne droite pour représenter dans l'espace la séquence temporelle du discours. Mais, dans ce dernier cas, la largeur réduite de l'écran impose une unité plus large que le paragraphe. Et nous avons choisi la page, d'autant que, contrairement aux paragraphes, les pages sont de longueur constante. Et pour être moins serré, le graphique emprunte la diagonale. Le mot recherché est représenté séquentiellement de la première occurrence (en bas à gauche) à la dernière (en

haut à droite), chaque point étant déterminé par la position du mot dans le corpus - c'est l'abscisse - et le numéro de l'occurrence - c'est l'ordonnée. La diagonale obtenue est plus ou moins régulière selon que l'objet représenté est plus ou moins régulièrement distribué. Quand les points se rapprochent et s'orientent vers la verticale, il s'agit d'une "rafale", c'est-à-dire d'une concentration des occurrences due à quelque cause locale, thématique ou stylistique. Quand les points s'espacent et s'inclinent à l'horizontale, cela correspond à une raréfaction momentanée de l'objet recherché. Le nombre de pages parcourues d'une occurrence à l'autre donne la mesure de la distance. Les pages étant de longueur voisine, cette mesure est plus fiable que celle des paragraphes et presque aussi précise que celle des mots. La distance est alors convertie en pixels. Mais, le nombre de pixels étant limité sur l'écran, on a réduit à un échantillon les mots très fréquents (une occurrence sur 2 ou 3 ou plus), de telle sorte qu'on ait au maximum 360 points à représenter et ainsi une lisibilité acceptable.

Le graphique ci-dessous, précisément, représente un mot dont la fréquence est naturellement élevée dans un corpus romanesque. L'*amour* n'est pourtant pas équitablement partagé. Il jaillit verticalement dans la *Nouvelle Héloïse*, dans *Indiana* de Georges Sand et dans *Un Amour de Swann* et il s'étiole en un mince filet languissant de Flaubert à Verne et Zola. Cela l'histogramme l'aurait dit aussi, mais de façon abrupte et carrée, sans la finesse et la fluidité des détails, sans les changements de rythme qu'on observe à l'intérieur même d'un texte, par exemple dans *Les Chouans*, *Madame Bovary*, *Du côté de chez Swann*, ou *Le Temps retrouvé*. Précisons que si les lignes en pointillés symbolisent le passage d'un texte à un autre, cette grille a été surajoutée au graphique, comme les méridiens sur le globe terrestre, la segmentation du corpus n'entrant nullement en ligne de compte dans le traitement.

Cependant si l'œil est satisfait de la représentation, des variations de la pente et de la gradation des pleins et déliés, la raison a des raisons de s'inquiéter. Car elle ne sait trop si les sinuosités observées sont ou non le fruit du hasard. Alors que la hauteur des « bâtons » de l'histogramme donne une réponse brutale, mais claire et chiffrée, on a besoin ici d'outils supplémentaires pour apprécier en probabilité l'orientation de la courbe et confirmer l'impression visuelle. On aura recours aux calculs traditionnels qui s'offrent à toute distribution : la moyenne des intervalles, l'écart-type de leur répartition et le coefficient de variation qui combine moyenne et écart-type. Mais leur témoignage est décevant car s'il reflète les irrégularités du profil, il dépend aussi de la fréquence du mot, rendant délicate la comparaison de deux distributions inégales. On a donc eu recours à d'autres mesures propres aux données sérielles.

Représentation graphique du mot amour*, de Marivaux à Proust*



Le test de la « différence quadratique moyenne successive »

La première est connue sous le nom de "différence quadratique moyenne successive". Elle correspond à la formule:

$$\delta 2 = \frac{1}{f} \sum_{i=1}^{i=f} (x(i+1) - xi)^2$$

pour *i* variant de 1 à *f* (*f* = nombre d'occurrences de l'objet recherché).

Pour apprécier sa valeur, on doit la rapprocher de la variance σ2, où la sommation est celle des carrés des écarts à la moyenne[9]. Or δ2 est moins sensible que σ2 aux variations lentes, quand des intervalles de même type se suivent, soit courts, soit longs. Ce genre de distribution, qui correspond aux rafales, est marqué par un rapport δ2/σ2 plus faible que le seuil convenu. C'est le cas du mot *amour* où le test propose la valeur 1.117, qui est très inférieure à la limite basse 1.808 où l'on franchit le seuil de 5%.

La distribution inverse, où les distances sont soumises à une alternance rapide et régulière, peut rarement être observée dans les textes (sauf quand s'exerce quelque contrainte syntaxique ou prosodique). Le rapport δ2/σ2 est alors plus élevé que la limite dévolue au hasard. Les deux seuils qui fixent

[9] Est-il utile d'en rappeler la formule $\sigma 2 = \frac{1}{f} \sum_{i=1}^{i=f} (xi - \bar{x})^2$?

l'espace de l'hypothèse nulle sont calculés pour chaque fréquence[10] et la valeur du quotient δ2/σ2 est considérée comme significative lorsqu'elle échappe à la fourchette indiquée.

Le test est malheureusement dévoyé dans certains cas où toutes les occurrences d'un mot sont concentrées dans un seul passage. Le cas de telles distributions déséquilibrées n'est pas exceptionnel dans la réalité. Il arrive qu'en accord avec le thème traité un même texte accapare toutes les occurrences d'un mot. Force est donc de corriger cette anomalie, ce qu'on peut faire avec une simple mesure d'écrêtement : toute valeur extrême qui dépasse un seuil convenu est ajustée à la valeur limite précisée par ce seuil. Certes ce correctif introduit dans notre programme amortit les secousses les plus brutales, mais il vaut sans doute mieux recourir à un test plus fiable et plus constant.

Le test de Lafon

La mesure la plus appropriée semble être celle que propose Pierre Lafon, dans "*Dépouillements et statistiques en lexicométrie*. Elle repose elle aussi sur un calcul de variance un peu particulier, ainsi formulé: $\frac{1}{f}\sum_{i=1}^{i=f} xi(xi-1)/2$

On calcule la moyenne (ou espérance mathématique) et l'écart-type de cet indicateur, ce qui permet d'apprécier la probabilité du résultat obtenu, au moyen de l'écart réduit, soit :

z = (valeur observée - valeur théorique) / écart-type

La valeur théorique est calculée par la formule: $m = \frac{(T-f)(T-f-1)}{f(f+1)}$

et l'écart-type par : $\sigma = \frac{mT(T+1)(f-1)}{f(f+1)(f+2)(f+3)}$

où *f* désigne la fréquence du mot et *T* l'étendue du texte, évaluée avec l'unité de mesure choisie pour les distances xi, c'est-à-dire ici le nombre de pages du corpus.

Cette mesure est plus sûre que les précédentes dans les cas extrêmes où toutes les occurrences d'un mot sont concentrées dans un même texte du corpus. L'écart réduit atteint alors des valeurs très élevées, qui signalent les monopoles exclusifs. Ce sont souvent les noms propres.

En dehors de ce cas particulier, les différentes mesures accordent généralement leurs témoignages, surtout lorsque les fréquences sont élevées[11].

---

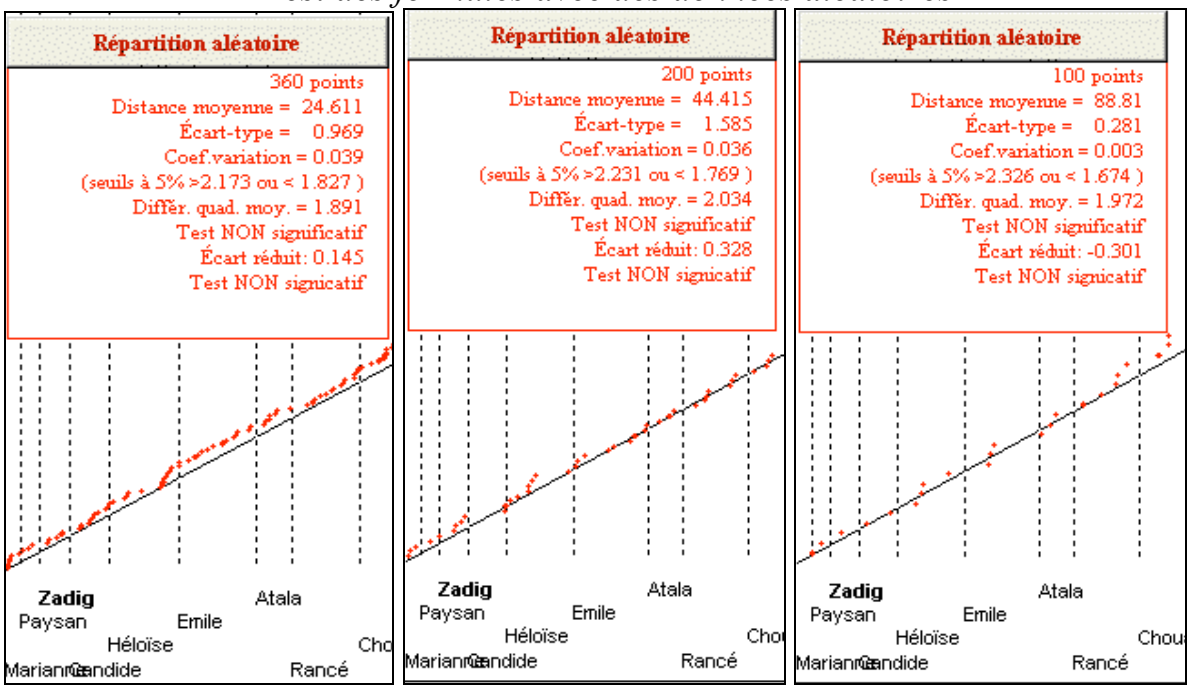
[10] Selon la formule $\eta g, \eta d = 2 \pm 2 \mid ca \mid \sqrt{\frac{N-2}{(N-1)(N+1)}}$, la valeur *ca* étant lue dans les tables de la distribution normale, au seuil de 5%, soit 1,6449.

Dans le cas inverse, celui des fréquences faibles, toutes les méthodes sont pareillement fragilisées et l'on évitera, dans cette situation, de leur faire une confiance absolue.

Application à des données aléatoires

Avant d'aller plus loin il convient de vérifier la validité des méthodes, en proposant au calcul un mot virtuel dont la répartition serait aléatoire. Un générateur fournit des nombres au hasard, qu'on va considérer comme les positions du mot dans le corpus. Le résultat montre alors les points s'aligner sur la diagonale et les tests s'accorder avec le hasard: l'écart réduit a tendance à se rapprocher de zéro et le calcul de *delta* se maintient généralement dans la fourchette où l'hypothèse nulle ne peut être écartée. La figure ci-dessus en montre quelques exemples pour les fréquences 360, 200 et 100.

Test des formules avec des données aléatoires



Liste des distributions irrégulières

Fort de cette expérience concluante, on peut soumettre l'ensemble des mots du corpus au test de Lafon, ce que propose une fonction du programme. Le calcul exclut toutefois les mots de trop basse fréquence pour lesquels il perd son sens et ceux de haute fréquence (f > 360) pour lesquels il perd sa légitimité pour les raisons contingentes qu'on a dites. Les résultats, enregistrés dans un fichier ASCII accessible à l'éditeur, apparaissent successivement en ordre alphabétique,

[11] Précisons que les calculs portent sur les séries entières, même si elles s'étendent à des milliers d'emplois. L'échantillonnage n'est nécessaire que pour la représentation graphique qui ne peut guère aligner plus de 360 points distincts.sur l'écran.

puis selon un tri hiérarchique, fondé sur la valeur décroissante de l'écart réduit. La liste ne retient que les mots distribués en rafales, où le seuil significatif est atteint. Les premières places de la liste hiérarchique sont évidemment occupées par les personnages principaux de chaque roman, le plus souvent par les noms propres et parfois même ceux qui apparaissent dans le titre, comme c'est le cas pour *Zadig, Raquin, Storitz, Émile, Bovary, Chouans, Indiana, Atala*. La vertu heuristique est plus intéressante dans la suite de la liste où l'on rencontre des thèmes supportés par des noms communs, comme le *projectile* dans le roman de Verne *De la Terre à la lune* ou le *train* de la *Bête humaine*. On ne peut s'étonner de l'absence presque systématique des mots-outils puisque le critère de la fréquence a éliminé la plupart. Mais la rareté des verbes, des adverbes et des adjectifs s'explique autrement. Ces catégories ont des privilèges ou des exclusives moins affirmés que les substantifs. Ils s'accommodent plus aisément de situations, de thèmes, ou de genres différents, au lieu que le substantif a tendance à s'attacher, voire à s'identifier au texte où il apparaît, en ignorant les autres.

Cependant un corpus trop hétérogène comme celui que nous avons choisi ne rend pas justice à la vertu discriminante d'un test dont la sensibilité est brutalisée par la thématique trop diverse de romans indépendants, appartenant à des époques et à des auteurs différents. Si le corpus est homogène – et cette qualité est généralement requise – la finesse du test échappe aux constatations triviales.

Relevé des cooccurrences et appréciation probabiliste

Il est facile de superposer deux distributions et de les distinguer par la couleur. Quant à les comparer plus précisément, la chose est délicate. Certes, s'il s'agit d'histogrammes, fondés sur la division du corpus en textes séparés, comme le nombre d'éléments est identique dans les deux distributions, la comparaison trouve un appui commode, soit qu'on se serve du coefficient de corrélation pour établir l'accord ou le désaccord des deux séries, soit qu'on établisse tout simplement le quotient des deux effectifs ou fréquences relevés pour chaque élément de la série. Mais, à la réflexion, que deux mots se trouvent avec des fréquences comparables dans le même texte ne prouve pas qu'ils ont des liens étroits. Il est possible qu'on les rencontre à des endroits différents du texte et jamais ensemble. La relation établie par le coefficient est au mieux un lien entre les textes, mais non entre les mots. Ce dernier lien ne peut être affirmé que si les deux mots apparaissent non pas dans les mêmes textes, mais dans les mêmes passages. Et par passage, il faut entendre une unité courte, qui peut être la phrase, le paragraphe, au maximum la page. C'est le paragraphe que nous avons choisi pour la fonction « thématique » de notre logiciel, c'est la même unité que nous considérons ici[12]. Seront retenus les paragraphes qui contiennent

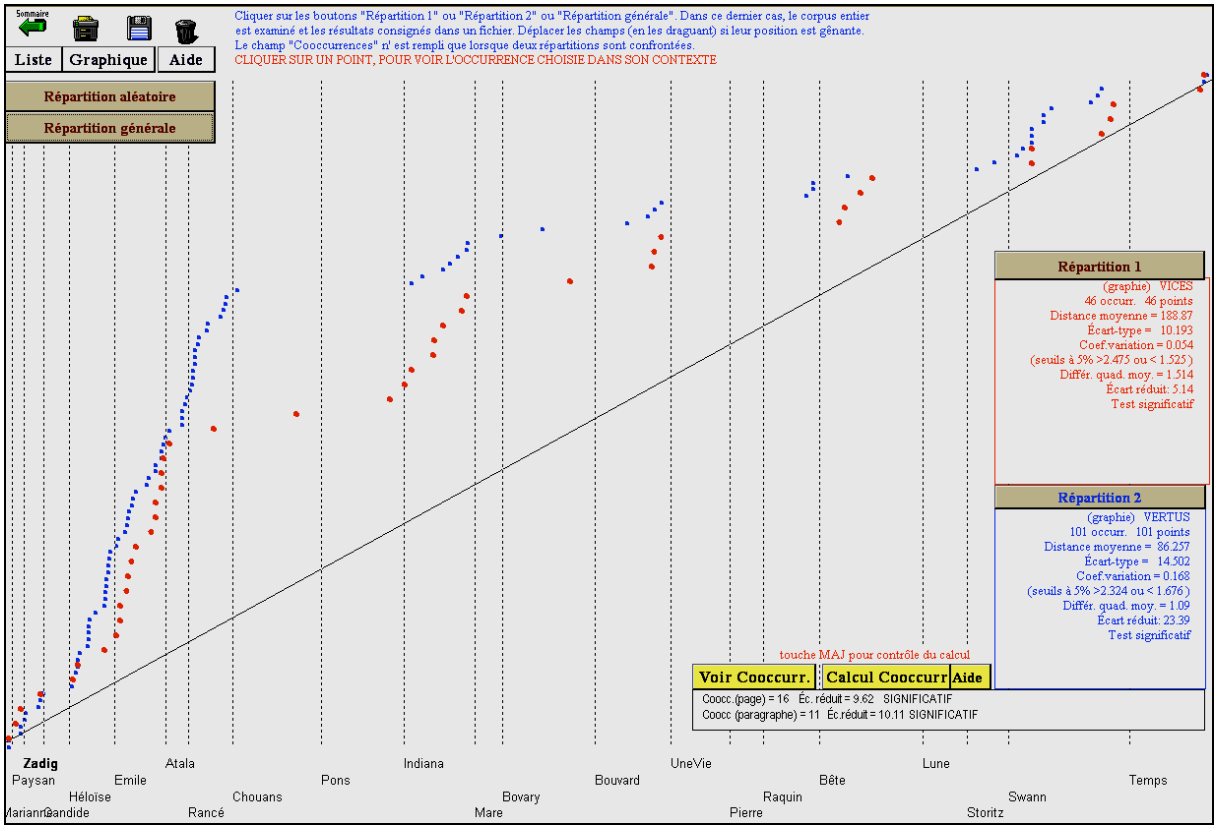
[12] C'est aussi la voie que recommande Pierre Lafon, après avoir exploré d'autres chemins, cf « Dépouillements… » p. 161.

à la fois les deux mots considérés. Mais comme cette contrainte peut être estimée trop forte, la cooccurrence est aussi évaluée au niveau plus large de la page. Ces deux méthodes sont proposées simultanément. Les passages où sont observées les cooccurrences étroites, dans le même paragraphe, peuvent en outre être montrés en clair à l'écran.

Reste à mesurer en probabilité la force de l'aimantation qui attire ou repousse les mots cooccurrents. Quand on met en regard deux distributions, la distance qui les sépare est d'autant plus faible que le nombre de cooccurrences est plus fort. Mais cela dépend aussi de la fréquence des deux mots et de la longueur du texte. La loi hypergéométrique est appelée ici à calculer la probabilité particulière attachée à la cooccurrence Le calcul est direct, selon la procédure exposée précédemment.

L'exemple qui suit illustre le couple inséparable *vices/vertus* et la liaison forte qui assez souvent rapproche bon gré mal gré les antonymes. Pour que la dispute s'instaure entre eux, il faut bien qu'ils soient en présence l'un de l'autre. Visuellement le parallélisme des deux courbes saute aux yeux. Impression confirmée par le calcul : dans un océan de deux millions de mots, les 46 occurrences de *vices* et les 101 occurrences de *vertus* ne devraient se rencontrer ni dans le même paragraphe, ni dans le même page (l'effectif attendu est proche de zéro). Or la cooccurrence est observée respectivement 11 et 16 fois.

Comparaison de deux distributions. Vices *et* vertus



Le même tropisme précipite l'un contre l'autre les deux partenaires au singulier : dans la bouche des moralistes le *vice* aime se frotter à la *vertu* (respectivement 11 et 15 fois). L'effet de style, même facile, est ainsi assuré. Encore faut-il que l'affrontement soit équilibré et que le singulier ne soit pas opposé au pluriel, ce qui rendrait l'antithèse boiteuse : le *vice* évite les *vertus* (une cooccurrence seulement) et la *vertu* les *vices* (un seul exemple également). Inversement les membres de la même famille n'ont rien qui les rapproche : *vice* et *vices* ne se rencontrent qu'une fois dans le même paragraphe, et de même la *vertu* ignore les *vertus*.*Comparaison avec d'autres modèles*

Deux modèles classiques peuvent être comparés au calcul hypergéométrique - auquel on a donné la préférence, parce qu'en proposant un effectif théorique, il autorise le calcul d'un écart et d'une probabilité. Les autres indices n'ont pas cet avantage, même si leurs résultats s'accordent parfaitement avec notre modèle.

Le premier indice, ou *Rapport de Vraisemblance (RV)* a été proposé par Dunning en 1993. Il s'appuie sur quatre paramètres:

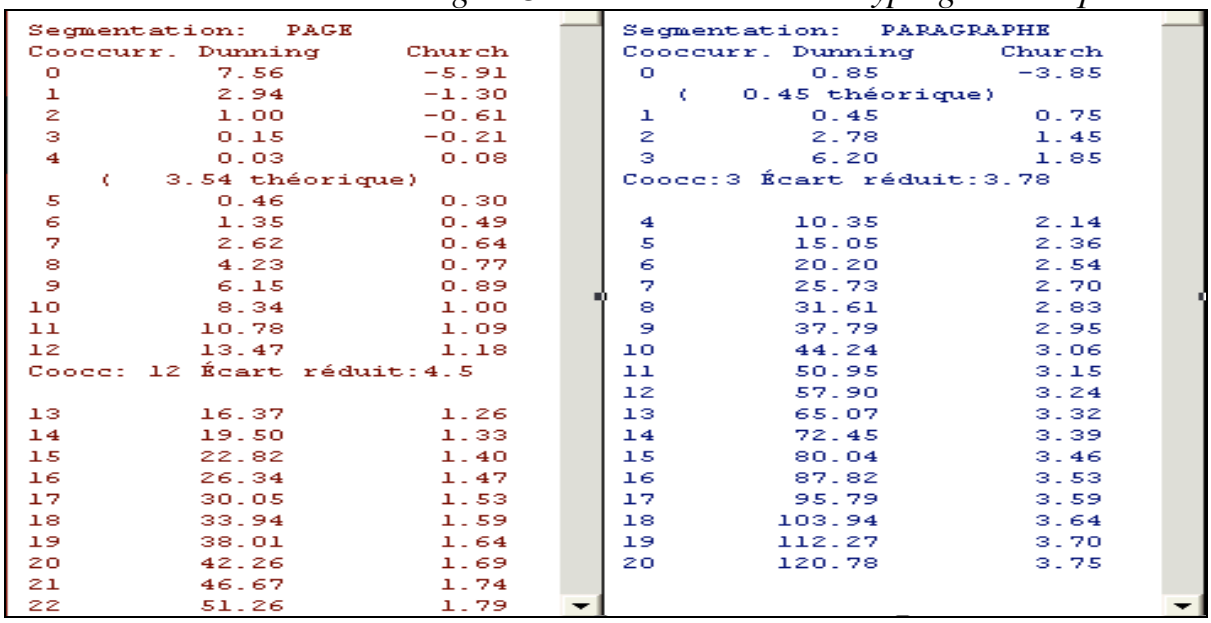
- *a* : nombre de cooccurrences des deux mots dans le champ exploré
- *b* : nombre d'occurrences du premier mot en l'absence du second
- *c* : nombre d'occurrences du second mot en l'absence du premier
- *d* : nombre d'occurrences des autres mots

$RV = -2 \log L = 2 (s1-s2)$

pour $s1 = a \log a + b \log b + c \log c + d \log d + (a+b+c+d)\log(a+b+c+d)$

$s2 = (a+c)\log(a+c) + (b+d)\log(b+d) + (a+b)\log(a+b) + (c+d)\log(c+d)$

Accord des indices de Dunning et Church avec le calcul hypergéométrique



Segmentation: PAGE		
Cooccurr.	Dunning	Church
0	7.56	-5.91
1	2.94	-1.30
2	1.00	-0.61
3	0.15	-0.21
4	0.03	0.08
(3.54 théorique)		
5	0.46	0.30
6	1.35	0.49
7	2.62	0.64
8	4.23	0.77
9	6.15	0.89
10	8.34	1.00
11	10.78	1.09
12	13.47	1.18
Coocc: 12 Écart réduit:4.5		
13	16.37	1.26
14	19.50	1.33
15	22.82	1.40
16	26.34	1.47
17	30.05	1.53
18	33.94	1.59
19	38.01	1.64
20	42.26	1.69
21	46.67	1.74
22	51.26	1.79

Segmentation: PARAGRAPHE		
Cooccurr.	Dunning	Church
0	0.85	-3.85
(0.45 théorique)		
1	0.45	0.75
2	2.78	1.45
3	6.20	1.85
Coocc:3 Écart réduit:3.78		
4	10.35	2.14
5	15.05	2.36
6	20.20	2.54
7	25.73	2.70
8	31.61	2.83
9	37.79	2.95
10	44.24	3.06
11	50.95	3.15
12	57.90	3.24
13	65.07	3.32
14	72.45	3.39
15	80.04	3.46
16	87.82	3.53
17	95.79	3.59
18	103.94	3.64
19	112.27	3.70
20	120.78	3.75

Le second indice est connu sous le nom *d'Information Mutuelle* (Church et Hovy, 1992). Là encore on a quatre données:

- *n* : nombre total de mots dans le champ exploré
- *n12* : nombre de cooccurrences, d'où *p12 = n12 / n*
- *n1*: fréquence du premier mot, d'où *p1 = n1 / n*
- *n2* : fréquence du second mot, *d'où* *p2 = n2 / n*

La formule est très simple: *IM = log (p12 / (p1*p2))*

Notre programme met en regard les trois indices et l'on vérifiera, dans le tableau ci-dessous, que l'effectif théorique établi par la loi hypergéométrique correspond toujours à la valeur minimale des deux autres indices.

CHAPITRE 11

Le menu SEGMENTS RÉPÉTÉS

1 - Pour traiter les segments répétés (bouton "Préparation"), on a procédé comme pour les mots simples, avec le même programme d'indexation (signé Anfosso). Mais la longueur admise pour les unités à indexer est étendue à 100 caractères (au lieu des 26 du programme normal). Dans un moule aussi large on colle les mots rencontrés entre deux signes de ponctuation en remplaçant les blancs intermédiaires par le "blanc souligné" (le caractère _). On a fixé la limite à 15 mots consécutifs, pouvant être enchaînés dans le même segment. Naturellement la fenêtre glissante enregistre tous les segments d'un ordre inférieur, jusqu'à la limite 2. Soit une séquence de 8 mots, notés de 1 à 8. Le programme enregistre tous les segments successifs de longueur 2 à 8 (on ne peut aller dans ce cas jusqu'à 15), soit au total: 1 segment de 8 mots (12345678), 2 de 7 (1234567 et 234567), 3 de 6 (123456, 234567, 345678), 4 de 5 (12345, 23456, 34567, 45678), 5 de 4 (1234, 2345, 3456, 4567, 5678), 6 de 3 (123, 234, 345, 456, 567 678) et 7 de 2 (12, 23,34,45, 56,67,78). Le "texte" ainsi généré devient très lourd. Là où le texte original avait 8 mots, son extension en produit 28, qui sont en outre beaucoup plus longs (de 2 à 8 fois plus). Le volume des données représente alors dix fois celui du texte original. Lors de cette première phase une seule limitation est imposée au segment: avoir plus de 12 caractères. Pour un corpus d'un million de mots, l'élaboration du "texte" exige 10 minutes, à quoi il faut ajouter une minute pour le tri et l'indexation. Il en résulte un énorme fichier (de 150 Mo pour le même corpus), qu'une seconde phase doit traiter.

2 - Cette seconde phase multiplie les critères de filtration pour épurer et condenser la masse des données, et faire en sorte que les millions de segments automatiques soient réduits à quelques milliers de segments significatifs. Le premier critère est la fréquence: tout segment qui n'est pas répété au moins x fois est éliminé (x a la valeur minimum 2 pour les segments de grande longueur; ailleurs x est fonction de la longueur L et aussi de la taille du corpus. En second lieu les segments formés uniquement de mots-outils ne sont pas jugés intéressants. Les mots-outils ne sont acceptés que si au moins deux mots de la combinaison sont ou risquent d'être des mots pleins, c'est-à-dire s'ils ont au moins 3 caractères (ou 4, voire , dans les grands corpus). En troisième lieu on évacue les segments contraints, c'est à dire ceux qui sont inclus dans un segment plus long avec la même fréquence. Ainsi si toutes les occurrences de "travailler

plus pour gagner" sont suivies de "plus", il n'est pas utile de garder ce segment provisoire qui n'apporte aucune information par rapport à la séquence plus longue. Mais il faudra le garder s'il a une fréquence supérieure, due à la variante "davantage". L'élimination des segments contraints est facile quand l'extension est à droite, comme dans l'exemple qui précède, puisque les deux segments se suivent dans le fichier. Il n'en va pas de même pour l'extension à gauche. Dans l'ordre alphabétique du fichier indexé, le segment éventuellement contraint, par exemple "plus pour gagner plus ", est éloigné du segment qui peut l'englober ("travailler plus pour gagner plus"). On ne s'étendra pas sur les procédures utilisées pour résoudre ces difficultés. Mais elles coûtent du temps.

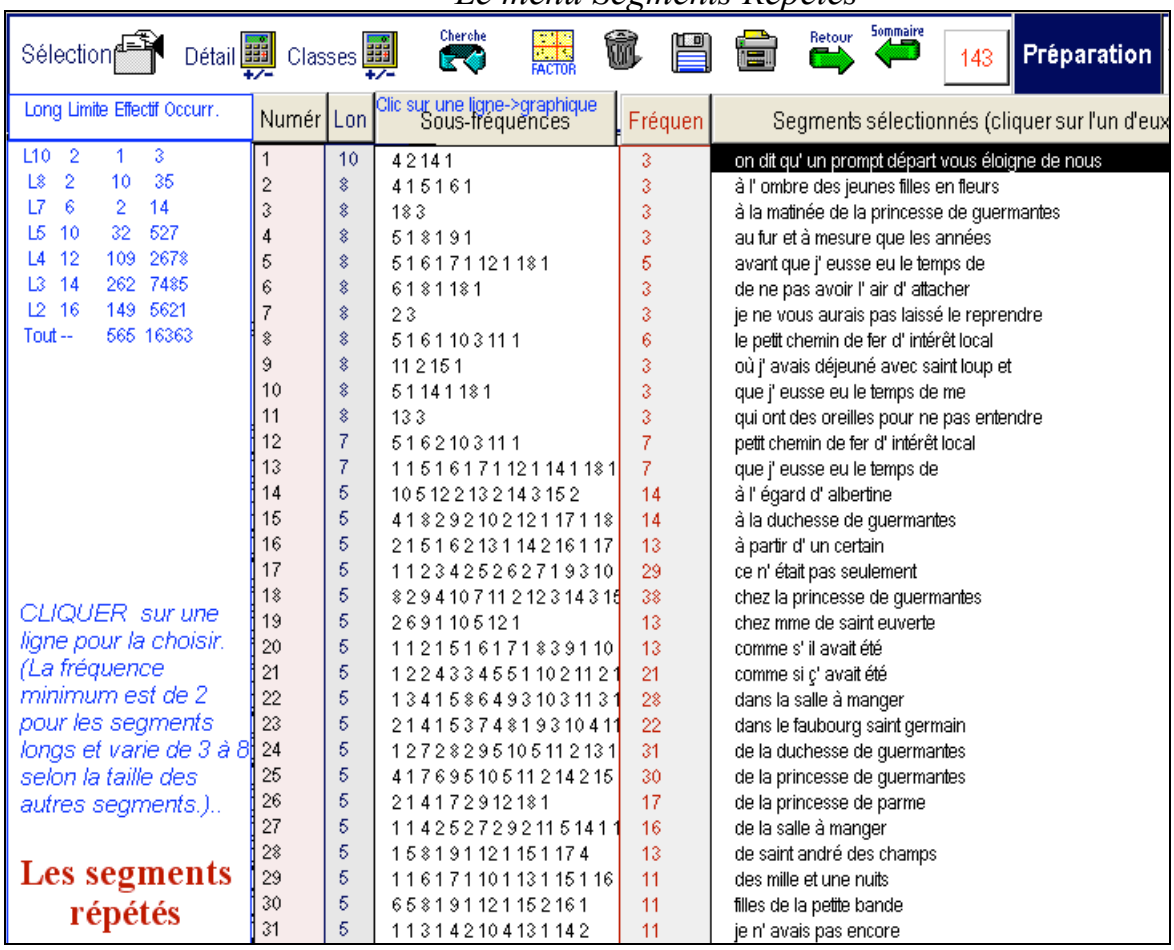
3 - À la fin du traitement, un récapitulatif est inscrit dans un champ en lettres bleues, où l'on peut lire le nombre de segments retenus et le nombre d'occurrences observées, et cela pour l'ensemble et pour chacune des classes de longueur, de 2 à 15. Pour disposer des données sur l'écran, cliquer sur une ligne du récapitulatif ou faire appel au programme "Sélection". Plusieurs critères sont proposés pour le choix: la classe de longueur, la fréquence minimum, la présence d'un mot simple particulier (bouton "cherche"). En corrigeant le tir selon la taille du corpus, on arrive à disposer d'un lot ni trop court, ni trop vaste. La bonne mesure semble être de 300 unités maximum. Au delà de cette limite, les segments, trop nombreux et trop longs, ne seraient guère déchiffrables dans les représentations graphiques. L'ordre alphabétique maintenu dans les listes permet de les explorer sans trop solliciter l'ascenseur.

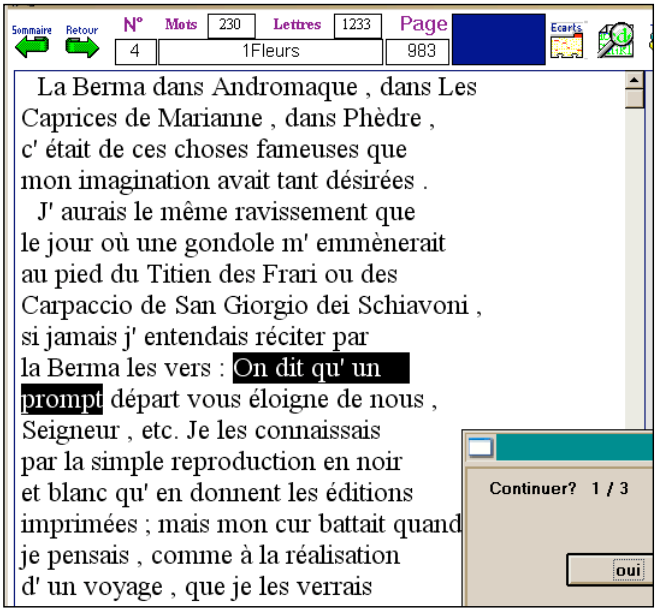
4 - Avant d'engager l'exploitation statistique, on peut retrouver dans le contexte du corpus chacune des occurrences d'un segment. Il suffit d'un clic sur le segment en question, comme certain vers de *Phèdre* répété trois fois dans la *Recherche du Temps perdu*.

5 - Le bouton "Segments communs" sert à détecter les emprunts ou segments identiques dans deux textes du corpus, choisis par l'utilisateur. Le champ des sous-fréquences est alors réduit aux deux fréquences observées dans les deux textes en question. Les écarts à la norme endogène sont signalés, s'ils dépassent le seuil: les trois valeurs indiquées concernent respectivement l'ensemble des deux textes comparés,puis le premier, puis le second. Retour au texte possible.

6 - Une fois que la sélection a été faite, le traitement statistique des segments relevés fait appel aux procédures classiques que propose la page "Liste". Les données segmentales sont transférées dans un tableau où les textes sont en colonne et les segments en ligne. La panoplie des programmes est disponible, pour la représentation des données, qu'il s'agisse des histogrammes, des analyses arborées ou des analyses factorielles. Il est plusieurs façons de remplir le tableau, soit en livrant le détail des segments ou leur cumul (bouton "Détail"), soit l'ensemble des classes de longueur (bouton "Classes"), soit le TLE ou tableau entier des segments (bouton "Factor").

Le menu Segments Répétés





7 - La colonne des sous fréquences est sensible au clic, qui déclenche le graphique du segment.

8 - Enfin le calcul des spécificités peut s'appliquer aux segments de la même façon qu'aux mot simples. Dès que la préparation est assurée, le calcul pour un texte donné peut être lancé si l'on sollicite le bouton "Spécificités". L'ordre des résultats suit la hiérarchie de l"écart réduit, du plus au moins significatif. Mais on peut préférer un autre classement (bouton "trier") Là aussi le retour au texte est assuré par un clic sur le segment souhaité.

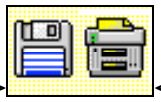
Voir Lebart et Salem, *Analyse statistique des données textuelles*, p. 145-178.

CHAPITRE 12

CONTRÔLER et IMPRIMER

CONTRÔLE DES OPÉRATIONS

Quelle que soit la recherche engagée, l'utilisateur assiste en témoin privilégié aux opérations en cours. Quand un programme a besoin de temps pour explorer le dictionnaire ou le corpus, l'utilisateur est tenu au courant de l'avancement par une information portée au bas de l'écran, où l'on peut suivre la progression dans l'alphabet ou dans la suite des textes du corpus. S'il s'agit d'une recherche dans le lexique ou dans le texte, chaque occurrence du mot cherché est montrée et encadrée dans la page même, en attendant qu'un clic de la souris invite à poursuivre le défilement. Inversement on peut toujours précipiter les choses et arrêter net l'opération en cours en appuyant sur la touche ALT. Si l'exemple relevé est jugé intéressant, la pression sur le bouton NOTES (en forme de disquette, comme ci-dessous) permet d'enregistrer dans le fichier des résultats EXTRAIT.txt la page portée à l'écran.

enregistrement -->  <-- impression

On a le choix entre l'ajout d'une note aux précédentes (le dernier enregistrement est montré pour vérification) et la remise à zéro du plan de travail. Le fichier ASCII où s'entassent ces notes peut être modifié et imprimé par tout éditeur ou traitement de texte, immédiatement ou ultérieurement. Un appel à l'éditeur de son choix est ménagé dans le menu principal (bouton EDITER). Les préférences de l'utilisateur sont préservées dans les réglages auxquels donne lieu une fois pour toutes le bouton INSTALLER.

Mais le fichier EXTRAIT.TXT qui intervient dans les pages-texte et les fonctions CONCORDANCE et CONTEXTE, n'est pas le seul apte à l'enregistrement des résultats. L'icône en forme de disquette est commune à presque tous les écrans, et voisine avec l'icône de l'imprimante. Dès qu'un résultat est acquis sur l'écran, l'utilisateur peut choisir entre deux sorties: une sortie immédiate mais éphémère sur papier, une autre plus souple et plus durable sur fichier. Ces fichiers, dont l'enregistrement est automatique (même si on ne s'en sert pas), portent des noms accordés à leur contenu: SPECIF.TXT pour les spécificités, DISTRIB.TXT pour les résultats relatifs à la structure lexicale, LISTEMOT.TXT pour les tableaux de fréquences. Ils sont ouverts par l'éditeur

de son choix dès que le bouton correspondant est sollicité. Lorsqu'il s'agit de graphique, une procédure particulière doit être appliquée: avant de sauvegarder l'écran sur un fichier, il faut en effet le copier dans le presse-papier, ce que Windows permet par la combinaison de touches ALT et F13 (la touche F13 porte généralement l'inscription IMP ECRAN ou PRINT SCREEN). Le bouton d'écriture propose alors de mettre en oeuvre soit un logiciel de dessin (comme PBRUSH), soit un traitement de texte (comme WORD). Dans l'écran vide et disponible, on collera alors le contenu du presse-papier.

Toutes les opérations sont sous le contrôle de l'utilisateur, même celles qui font appel à une application extérieure, par exemple un éditeur, ou un tableur. Même lors de ces excursions éloignées, le lien n'est pas rompu avec la base, qu'on retrouve au retour au même endroit et dans la même situation. Si l'on se sent perdu, un bouton en forme de flèche coudée et de couleur verte (placé généralement dans un coin, sous le nom de SOMMAIRE) permet toujours de revenir à la carte 1, qui sert de cabine de pilotage et qui permet la reprise du traitement normal. Un autre bouton symétrique du précédent (nommé BACK ou RETOUR) autorise les va-et-vient d'une page visitée à l'autre.

Les différents boutons de contrôle pour aller à la page suivante ou précédente ou pour gagner tel ou tel endroit stratégique, ont un graphisme et une désignation suffisamment explicites, pour qu'il ne soit pas nécessaire de s'y attarder. Pour éviter de se perdre deux boutons sont essentiels :

retour au menu principal---> 

<---retour à l'endroit d'où l'on vient

Pour quitter l'application, solliciter le bouton EXIT (ou QUITTER). Un adieu aimable apparaît alors à l'écran. À quelque endroit que l'on se trouve dans la pile, un départ précipité mais non incivil est possible quand on donne l'ordre d'urgence prévu par Windows.

Message d'adieu

LE RETOUR AU TEXTE

Le texte est toujours disponible pour vérifier les suggestions des chiffres ou les hypothèses du chercheur. Entre la démarche documentaire et les investigations statistiques la liaison a été renforcée par de nouvelles fonctions qui permettent d'accéder directement au texte dès qu'un résultat quantitatif suscite le soupçon, le doute, la curiosité ou la satisfaction. C'était déjà le cas

dans les versions antérieures en certaines occasions, un lien direct unissant chaque élément du dictionnaire des fréquences aux divers passages où un tel élément se trouve employé. Et de la même façon tout mot du texte était rattaché par un lien direct au dictionnaire. En outre la fonction thématique, dont le résultat est une liste ordonnée de corrélats, permettait un retour immédiat au contexte, chaque élément de la liste étant lié à tous ses emplois.

Mais ce souci de contrôle immédiat par le retour au texte a été poussé à l'extrême dans la présente version : dans toute liste de spécificités, toute représentation graphique et toute analyse factorielle, il suffit d'un clic (le plus souvent associé à une touche de contrôle) pour vérifier dans le texte même les indications fournies par les chiffres. S'il s'agit d'un histogramme, on cliquera sur un bâton représentant l'emploi, excédentaire ou déficitaire, d'un mot dans un texte pour voir défiler tous les emplois de ce mot dans le texte en question. Si l'on a affaire à une analyse factorielle, en désignant un mot avec la souris (touche CTRL enfoncée), on est renvoyé aux contextes de ce mot – ce qui permet souvent de comprendre pourquoi l'analyse lui a accordé la place qu'il occupe. Enfin, plus généralement, les listes auxquelles donnent accès les fonctions CONCORDANCE, CONTEXTE, SPECIFICITES, EVOLUTION, ASSOCIATIONS, TOPOLOGIE, SEGMENTS REPETES du menu principal sont pareillement sensibles au clic de la souris : tout mot ainsi isolé apparaît dans les différents contextes où il se trouve employé.

Ce luxe de contrôles immédiats nous a paru nécessaire. Car les résultats de la statistique deviennent vite abstraits et l'on peut craindre que leur enchaînement produise une sorte de vertige où la raison risque de perdre pied. En s'obligeant à contrôler, au moins partiellement ou ponctuellement, les résultats dès qu'ils apparaissent, on peut se prémunir contre l'emballement inconsidéré. Il arrive aussi que la statistique soit abusée par un artefact caché et insoupçonné ou par une influence si triviale qu'on n'y pense pas: le contact avec le texte concret réveille la conscience et dessille les yeux.

La recherche en sciences humaines est faite de ce va-et-vient incessant entre la machine et l'utilisateur, entre les hypothèses et les résultats, entre les traitements et les contrôles, entre la confiance et le soupçon.

LE CONTRÔLE DES DONNÉES INITIALES

HYPERBASE est à l'aise à une échelle qui est la sienne. Il risque d'être défaillant si le corpus est trop court, trop volumineux ou trop morcelé. Il n'y a guère de statistique possible dans les tout petits nombres et il n'y a rien à attendre d'HYPERBASE si le corpus n'a pas au moins 20 ou 50 pages. Inversement sa puissance a des limites: pour des raisons qui tiennent à l'environnement Toolbook, on ne peut guère aller au delà de 30 millions de mots. Enfin le nombre de textes confrontés dans le corpus est limité à 75. Cela peut rendre nécessaires des regroupements lorsque le logiciel est appliqué à des

enquêtes sociologiques, les unités traitées y étant souvent courtes et nombreuses.

Le moment le plus délicat est assurément celui de la préparation du texte. HYPERBASE n'a pas de grandes exigences pour le formatage des données. Mais il arrive souvent que, pressé d'exploiter un logiciel à peine déballé, l'utilisateur en fasse l'essai à l'aide du premier fichier qui lui tombe sous la main, sans se préoccuper aucunement du formatage. S'il s'agit d'un fichier HTML, PDF ou DOC, le résultat sera décevant. Plus perfidement les fichiers ASCII n'offrent pas les mêmes garanties s'ils viennent d'une machine Windows ou d'une machine Apple. HYPERBASE est né sur Macintosh il y a vingt ans mais il fonctionne désormais sous Windows (et uniquement sous Windows). C'est donc sous Windows qu'il faut lui préparer les textes. Faire cette préparation sur un Mac (où les retours de chariot et les lettres accentuées ont des codes différents) est dangereux, sauf si le fichier en enregistré au format RTF ou standard (suffixe .DOC). Dans ce cas le fichier peut être repris par Word sous Windows et converti en fichier ASCII (suffixe .TXT).

On déconseille fortement les tests extrêmes ou peu naturels. Hyperbase ne sait traiter que les alphabets latins. Il ignore les ressources de l'unicode et de XML. Il s'attend à trouver dans un texte des blancs, une ponctuation et des retours de chariot. Si rien de tout cela ne lui est fourni, il sera fort désappointé et le fera savoir.

LE RETOUR DE CHARIOT DANS LES CORPUS EN VERS

Les logiciels de traitement de textes n'utilisent souvent qu'un seul et même code, le retour de chariot, pour marquer la fin des vers et la fin des paragraphes, alors que le système Windows a prévu des symboles différents pour assurer ces deux fonctions distinctes. Les fins de paragraphes font appel à un signe composite (CRLF) alors que les fins de vers (ou de ligne) emploient un code simple (LF). Les effets d'une telle confusion à la saisie sont fâcheux pour certaines fonctions, et notamment pour la fonction CONTEXTE qui prend appui sur les fins de paragraphes et, les trouvant malencontreusement à chaque ligne, délimite des segments trop courts. Or il n'est pas toujours aisé de préparer les textes en évitant la confusion, car certains standards (notamment celui d'Apple) ignorent la distinction et ne disposent que du seul retour de chariot. On a donc proposé un programme (bouton OPTION VERS du menu principal) pour accomplir après coup ce toilettage. L'algorithme utilisé est simple: il conserve le retour de chariot qui suit une ponctuation forte et le remplace par le code LF dans les autres cas. Les données déjà inscrites et indexées dans la base n'en subissent aucune altération. Noter que parmi les versions lemmatisées qui seront exposées plus loin, il en est une (HYPERVER.EXE) qui traite les textes en vers et, ce qui est plus complexe, les corpus où prose et vers sont mêlés. Cela nécessite toutefois une désambiguïsation des codes de fins de ligne: ceux qui terminent un vers devront être précédés d'un blanc, les autres non.

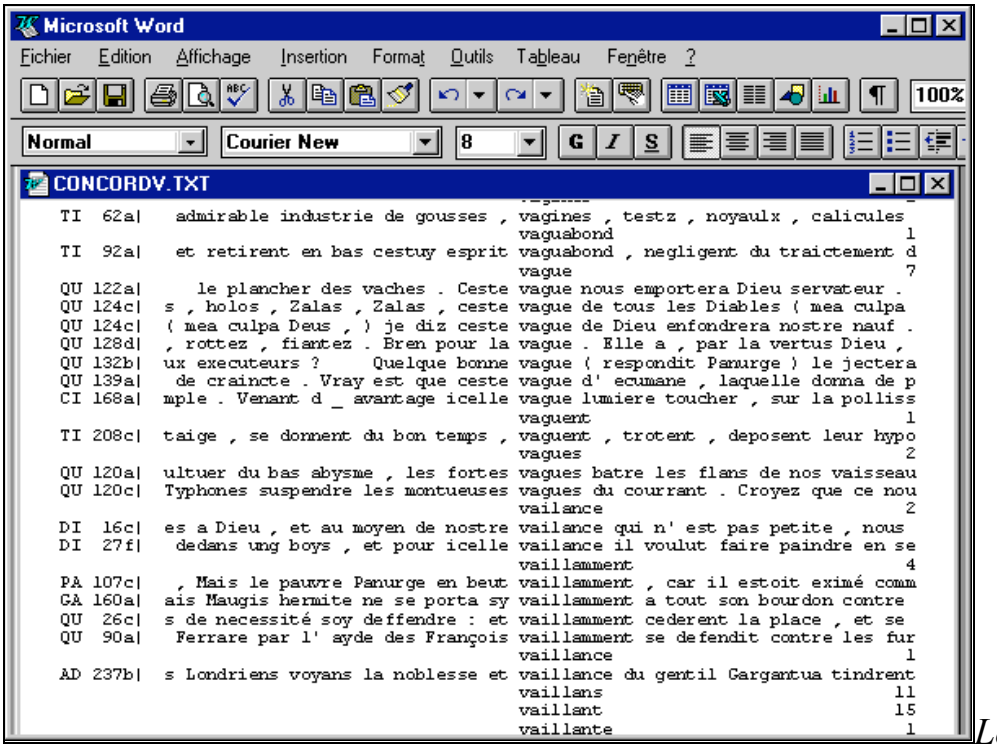
IMPRESSION DES RÉSULTATS

1 - Le symbole de l'imprimante laser apparaît sur toutes les cartes. Grâce à ce bouton, on peut envoyer à l'organe d'impression le contenu de l'écran et plus généralement le champ entier dont l'écran ne montre souvent qu'une fenêtre étroite. Cela peut aller jusqu'à l'intégralité de la concordance, de l'index ou du dictionnaire des fréquences. L'impression de la Concordance impose une police à espacement fixe. On a choisi la police *Courier* qui est la plus courante. Ce choix vaut aussi pour l'Index et le dictionnaire et pour la plupart des résultats. L'impression des textes et des contextes s'accommode mieux d'une police à espacement proportionnel.

Impression de l'Index

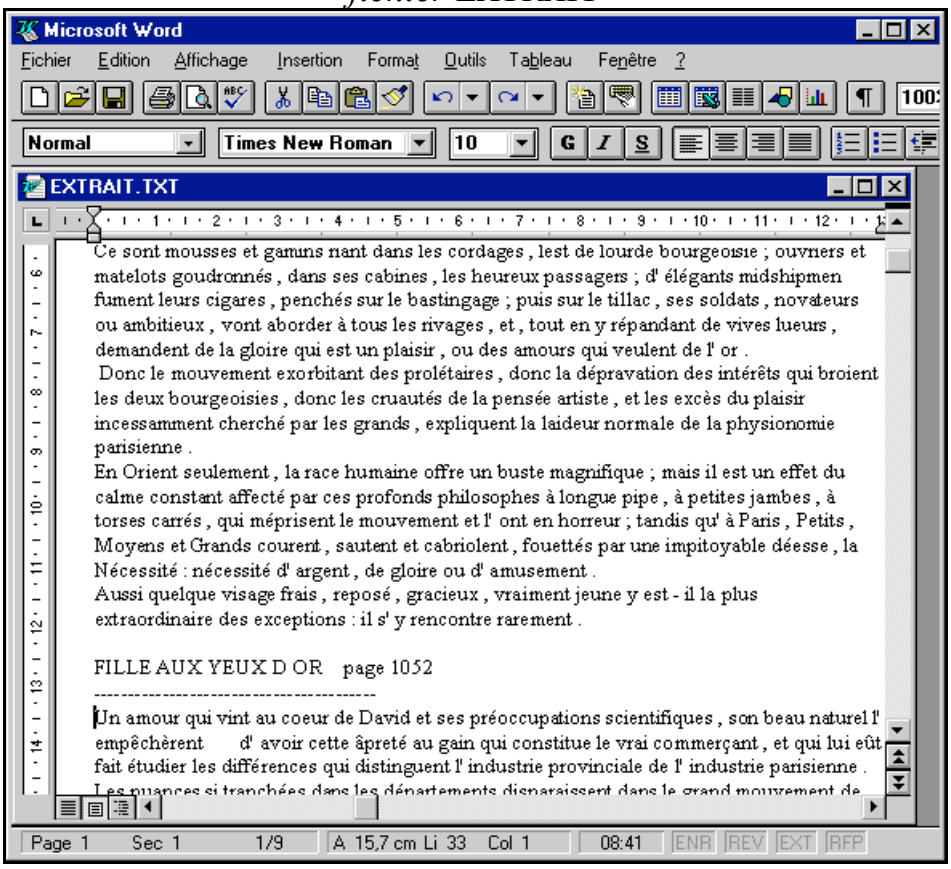
```
B                                   3
     GA   15c     TI  239a
     TI 239a
baaillans                           1
     CI   7d
baailloient                         1
     CI 113c
Babillebabou                        1
     QU 178a
Babillone                           1
     AD 178b
Babillonne                         11
     AD  261b     AD  263f
     AD  264b     AD  264b
     AD  264c     AD  264e
babouyneries                        1
     PA  82a
Babyloine                           1
     PA 148b
Babylone                            2
     GA 184d   CI 137d
Babylonien                          1
     TI  83a
bac                                 3
     TI  284d     TI  305c
     QU 183c
Bacabery                            1
     QU 118a
bacalarii                           1
     PA  51b
```

Impression de la Concordance générale

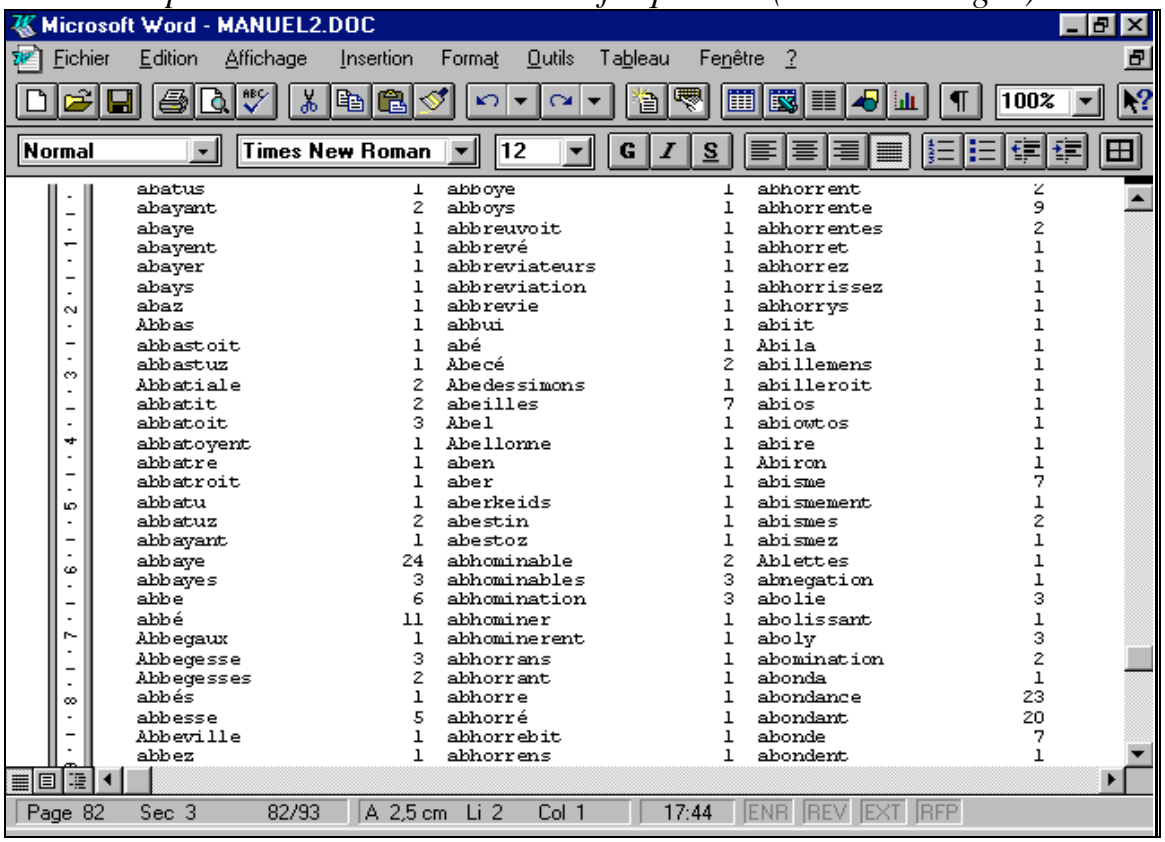


Le

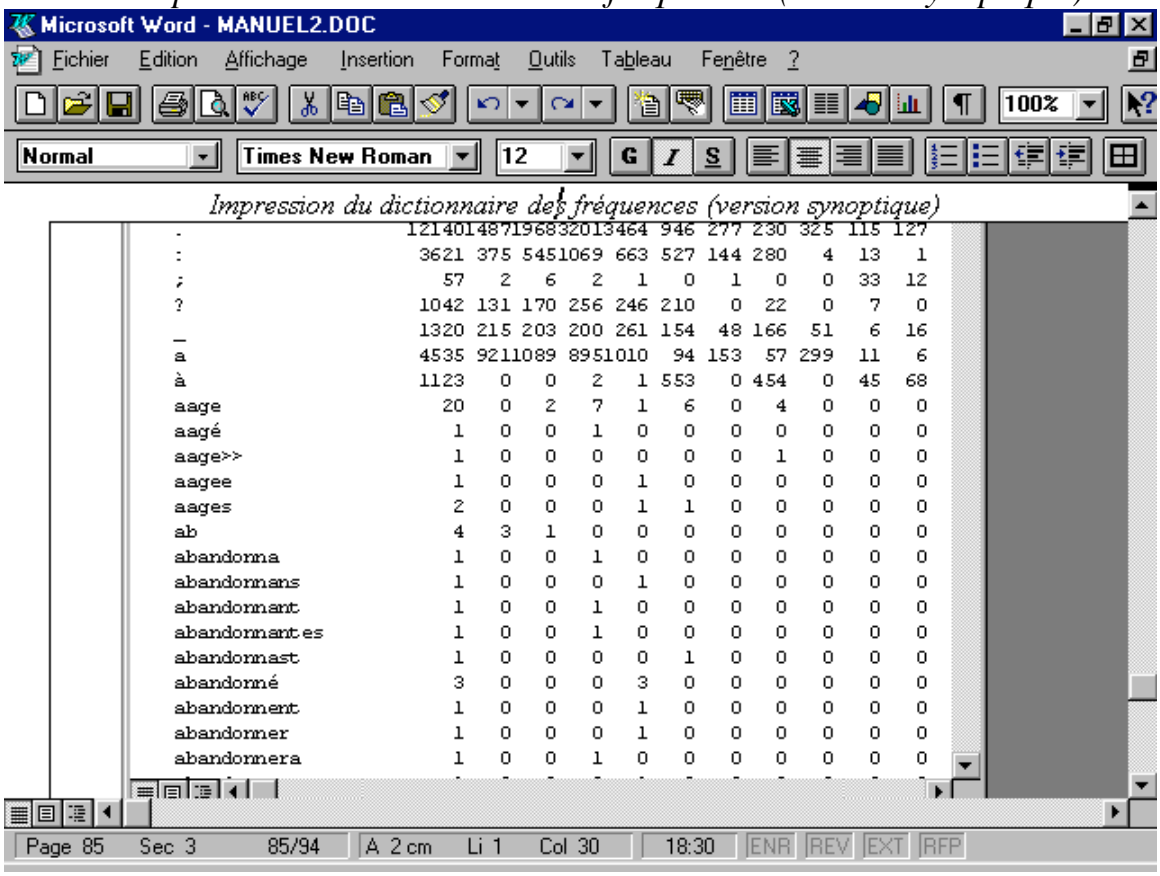
fichier EXTRAIT



Impression du dictionnaire des fréquences (version abrégée)



Impression du dictionnaire des fréquences (version synoptique)



.	12140	1487	1968	3201	3464	946	277	230	325	115	127
:	3621	375	545	1069	663	527	144	280	4	13	1
;	57	2	6	2	1	0	1	0	0	33	12
?	1042	131	170	256	246	210	0	22	0	7	0
_	1320	215	203	200	261	154	48	166	51	6	16
a	4535	921	1089	895	1010	94	153	57	299	11	6
à	1123	0	0	2	1	553	0	454	0	45	68
aage	20	0	2	7	1	6	0	4	0	0	0
aagé	1	0	0	1	0	0	0	0	0	0	0
aage>>	1	0	0	0	0	0	0	1	0	0	0
aagee	1	0	0	0	1	0	0	0	0	0	0
aages	2	0	0	0	1	1	0	0	0	0	0
ab	4	3	1	0	0	0	0	0	0	0	0
abandonna	1	0	0	1	0	0	0	0	0	0	0
abandonnans	1	0	0	0	1	0	0	0	0	0	0
abandonnant	1	0	0	1	0	0	0	0	0	0	0
abandonnantes	1	0	0	1	0	0	0	0	0	0	0
abandonnast	1	0	0	0	0	1	0	0	0	0	0
abandonné	3	0	0	0	3	0	0	0	0	0	0
abandonnent	1	0	0	0	1	0	0	0	0	0	0
abandonner	1	0	0	0	1	0	0	0	0	0	0
abandonnera	1	0	0	1	0	0	0	0	0	0	0

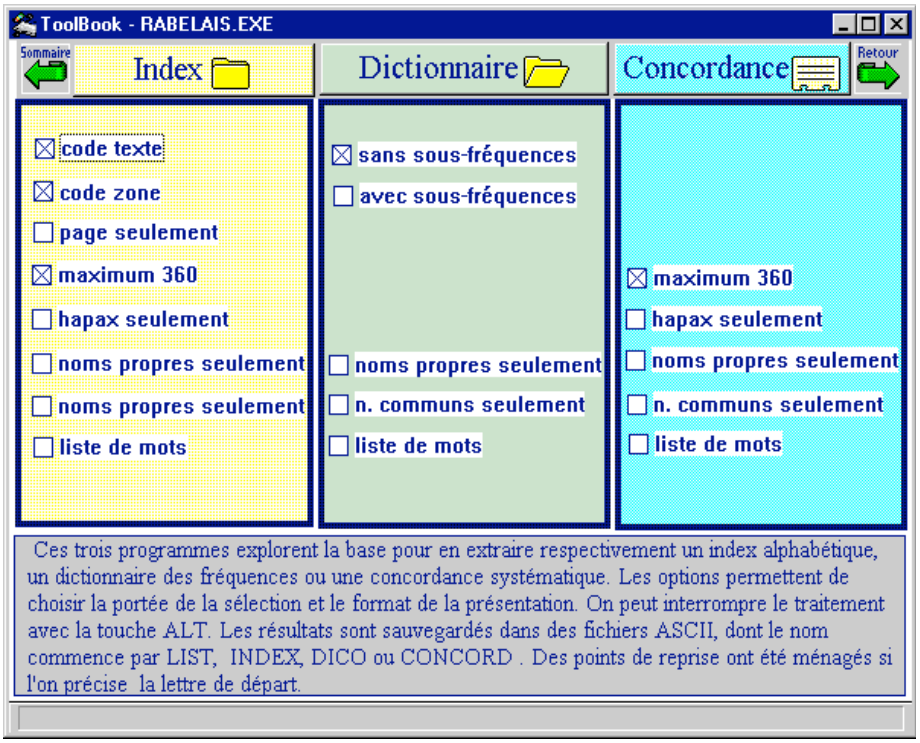
Le bouton EDITER apparaît en haut du menu principal. Il a pour fonction d'appeler un éditeur extérieur en lui fournissant le nom du fichier à éditer. Cela peut être l'un ou l'autre des fichiers dont le suffixe est en .TXT ou .XL ou .AFC. Leur nom les identifie aisément: ANALYSE, CONCORD, CONTEXTE, DICO, INDEX, LISTE, EXTRAIT, etc.. Comme certains de ces fichiers peuvent être volumineux, notamment les index et les concordances, un code alphabétique est ajouté qui précise quelle portion de l'alphabet est traitée dans le fichier. On peut modifier, éditer et imprimer après coup ces fichiers. Le choix de l'éditeur se fait soit par ce même bouton, soit, de façon plus générale, par les paramètres que propose le bouton INSTALLER.

La plupart de ces fichiers sont des entrepôts provisoires, qui gardent le même nom d'une séance à l'autre et d'une base à l'autre, en changeant de contenu. Les résultats y sont effacés dès que le programme correspondant en propose d'autres à leur place. Aussi convient-il soit d'imprimer les résultats dès qu'ils sont obtenus, soit de terminer la session pour exploiter ces résultats après la séance d'exploitation, en donnant aux fichiers un nouveau nom si l'on désire les conserver. Au moment où l'on abandonnera l'éditeur, le retour se fera sans problème dans HYPERBAS, et l'on retrouvera la page de départ.

Certains résultats sont préparés dès la constitution du corpus, le *Dictionnaire* notamment, dont le nom reprend le titre de la base en lui ajoutant le

suffixe.TXT. S'abstenir de changer le nom, l'emplacement et le contenu de ce fichier, qui sert au calcul des spécificités.

Les traitements lourds: index, dictionnaire, concordance



Mais on peut constituer un dictionnaire plus riche, où les sous-fréquences sont ordonnées dans une présentation synoptique. On aura recours à une page spéciale (voir ci-dessus) qui est commune aux traitements lourds et qui sert aussi aux index et aux concordances systématiques. On y accède, soit par le bouton IMPRIMER des pages-index, soit par le bouton TOUT de la page CONCORDANCE, et on a le choix entre trois programmes et diverses options de présentation ou de sélection.

CHAPITRE 13

Considérations techniques

PROTECTION

Normalement on protège les pages, les champs et les scripts, contre l'écriture ou l'effacement. L'item PROPRIÉTÉS du menu FICHIER de l'EXPLORATEUR Windows permet cette mesure de sécurité (en cochant l'option LECTURE SEULE). Cela n'empêche pas le déroulement normal des opérations qui se font en mémoire centrale et restent virtuelles, tout en donnant lieu à des résultats bien réels sur l'écran, et, plus substantiels encore, sur les fichiers et l'imprimante. Mais si l'on veut que la base conserve les modifications en cours, on enlèvera, provisoirement ou non, la protection.

À certains moments les champs - habituellement rebelles à l'écriture - laissent à l'utilisateur la liberté d'intervention. C'est notamment le cas des graphiques produits par l'analyse factorielle, qui ont souvent besoin d'être modifiés, au moins pour replacer les points doubles ou rendre plus explicite le nom des variables. Le point d'insertion apparaît alors dans le champ, ce qui autorise l'écriture.

On souhaite généralement que ces modifications de la base restent provisoires et, à moins de sauvegarder volontairement la base, les champs et les pages retrouveront leur virginité première lorsqu'on quittera la base. Cela peut se faire aussi expressément si on active le bouton VIDER du menu principal. Cela n'efface que les modifications faites pendant la séance, sans toucher aux données et résultats initiaux.

Dans l'utilisation normale la barre de menu n'apparaît pas au haut de l'écran. Cela permet de gagner de la place et aussi d'interposer un garde-fou contre les imprudences. L'utilisateur a cependant la possibilité de faire apparaître cette barre de menu, au moins pour régler certains paramètres de sauvegarde ou d'impression qui restent disponibles (les autres items restant grisés, c'est-à-dire interdits).

Le bouton rouge dont le graphisme évoque les panneaux de sens interdit est prévu à cet effet. Si on le sollicite l'interdiction tombe et le menu FILE apparaît en haut de l'écran. En le déroulant, on peut agir sur la mise en page des sorties imprimante et opérer des sauvegardes de la base - ce qui peut se faire aussi à l'aide des routines du système de Windows. Le bouton est alternatif: un nouveau clic rétablit l'ordre ancien et l'icône primitive.

Certains souhaiteront intervenir à l'intérieur même des scripts, c'est-à-dire des programmes. C'est un risque qu'on ne peut guère prendre si l'on maîtrise mal le logiciel et la programmation et qui est d'ailleurs interdit si l'on ne dispose que de la version runtime de Toolboook, la seule qu'on puisse distribuer sans entrave. L'auteur du présent logiciel est ouvert à toutes les suggestions et peut introduire de nouvelles fonctions dont on lui montrerait l'intérêt. Mais ouvrir le code à la curiosité publique, c'est ouvrir une boîte de Pandore et s'interdire la maîtrise des développements ultérieurs.

Si par exception le code est transmis à quelque utilisateur particulièrement intéressé, il devra veiller à faire une copie préventive, avant toute modification. Quand les précautions élémentaires ont été prises et que l'on dispose de la version complète de Toolbook, HYPERBASE s'ouvre très facilement pour peu qu'on tourne la clé, c'est-à-dire si l'on active la touche F3 en fournissant le mot de passe convenu. Attention! la protection est à plusieurs niveaux: il y a le niveau logique qui règle la modification des scripts et l'accès au mode *Author* et le niveau matériel qui règle de l'extérieur les droits d'intervention, soit que la base soit verrouillée (en lecture seule), soit qu'elle se trouve sur un support non-inscriptible (CD-ROM).

Au moment où la présente version 8.0 est livrée au public, le concepteur du logiciel est conscient de la fragilité des formats propriétaires et des contraintes qu'ils imposent à l'utilisateur et à l'auteur lui-même. L'évolution de la technologie et des systèmes d'exploitation prolonge sans fin la chaîne des adaptations et des développements. Vient un moment où le souffle vient à manquer à l'auteur d'un logiciel, surtout s'il a plus d'expérience que d'avenir. Ainsi la version Mac ne peut plus guère être maintenue, depuis que le nouveau système *MacOs X* a été généralisé. Comme ce système a abandonné

l'environnement *Hypercard*, pourtant créé par Apple et utilisé par notre version Mac d'Hyperbase, notre logiciel ne tourne plus sur Mac qu'en émulation *Classic*. Certes le code pourrait être repris et mis à jour dans un langage proche comme *Metacard* ou *Revolution*. Mais les garanties de pérennité resteraient insuffisantes, ce qui est le cas aussi de l'environnement *ToolBook* sur lequel repose la présente version Windows d'Hyperbase.

Comme Hyperbase s'adresse principalement à la recherche universitaire, où le commerce et l'argent n'ont pas encore éliminé la liberté et la gratuité, l'avenir d'Hyperbase est de se fondre dans une entreprise collective, fondée sur l'open source. D'autres logiciels comme *Lexico* ou *Weblex* seraient pareillement pris en compte pour offrir aux chercheurs les meilleures fonctions documentaires et statistiques.

MATERIEL REQUIS

Il est évidemment préférable de disposer d'une machine plus puissante, lorsqu'on procède à l'incorporation d'un texte. Les temps de préparation s'en trouvent considérablement réduits. Si ce texte est long et la mémoire étroite, le temps de traitement s'allongera, le programme étant paramétré de façon à se contenter de ce qui reste disponible.

Lors de l'exploitation, un ordinateur peu puissant peut suffire. Mais là encore si la pile est trop importante, certaines fonctions de recherche seront ralenties et la mémoire encombrée, et il vaut mieux utiliser une machine récente, et, si possible, un écran capable de restituer les couleurs dont le programme HYPERBAS est agrémenté (au moins 256 couleurs). On a résisté à la tentation d'entreposer dans la mémoire vive des tableaux trop importants, et par exemple le texte même. Le recours à la lecture des champs est systématique, les champs et les pages ayant été judicieusement segmentés pour raccourcir les temps d'attente.

De même cette version du logiciel HYPERBAS est assez peu gourmande en espace disque. Lorsqu'il s'agit de grands corpus la base occupe moins du double du fichier original. Ce rapport est moins favorable dans les petits corpus, parce qu'on a maintenu une part de la place vacante. Si la taille de la pile reste modérée, les fonctions de recherche disponibles ont une rapidité acceptable, d'autant que le traitement est facilité par les ressources de l'indexation. Dans les grands corpus, ces ressources jouent un plus grand rôle encore, car la taille des fichiers déconseille les techniques de recherche séquentielle d'autant que le CD-ROM où de tels corpus élisent domicile est un support de grande contenance mais de faible rapidité. La recherche indexée s'appuie sur des accès directs et des processus de Hashcoding qui limitent toute recherche à deux mouvements de la tête de lecture. Noter cependant que les accès directs sont moins nécessaires lorsqu'il s'agit de signes ou mots fréquents. Quoique l'indexation ait été

exhaustive dans la phase de création, on n'a pas cru bon de conserver ces index encombrants pour la virgule et les outils grammaticaux, qui sont dépourvus de références, mais leur fréquence même rend la recherche assez rapide, puisqu'on les trouve dans chaque page.

À titre d'exemple, un corpus de cinq millions de mots exigera plus d'une heure de traitement initial. Si le corpus est très volumineux, il sera préférable d'utiliser une machine rapide. Cette restriction ne vaut que pour la préparation de la base de données (qui n'a lieu qu'une fois et qui peut faire l'objet d'un emprunt extérieur et occasionnel), et non pour l'exploitation de cette base, qui n'a guère d'exigence.

Quant à la portée d'HYPERBASE, elle été poussée aussi loin qu'on l'a pu. D'une part les limites à la partition sont moins étroites. On avait prévu initialement 40 textes maximum, ce qui était suffisant, à notre sens, pour les recherches littéraires, linguistiques ou historiques. C'était là négliger certaines applications de la documentation automatique, de l'économie ou de la sociologie, à quoi notre logiciel a été appliqué et que nous n'avions pas envisagées au départ. Or de telles enquêtes sont faites d'unités plus courtes mais plus nombreuses, dont le nombre peut atteindre la centaine. La limite actuelle d'HYPERBAS est du double, soit 75 textes. Les corpus de grande taille peuvent y trouver place (un fichier de données de 30 millions de mots ou davantage peut être traité, comme cela a été tenté avec l'intégralité de la *Comédie humaine*, ou 300 numéros de la revue *Europe*).

UNE PRÉSENTATION PLUS LARGE ET PLUS CLAIRE

L'ancien format d'affichage 640x480 a été longtemps le format standard, compatible avec tous les écrans. HYPERBAS l'avait adopté. Mais cette contrainte n'a plus lieu d'être maintenue à l'époque actuelle où les écrans supportent des fenêtres plus larges, qui autorisent une lecture plus aisée et une présentation plus claire. HYPERBAS a donc fixé à 800x600 pixels la taille de la fenêtre qui lui est propre. Cependant il peut arriver qu'un réglage inadéquat de l'affichage rende l'écran trop étroit ou trop large et que le confort visuel de l'utilisateur en soit diminué. Le tableau de configuration de Windows permet de choisir la définition convenable. Si l'affichage en 480-640 est le seul disponible, Hyperbase s'en contentera et l'ajustement sera automatique. Le choix d'une présentation plus large ou plus étroite peut aussi être volontaire, si l'on sollicite le bouton *Option Écran* du menu principal.

STRUCTURE DE LA BASE

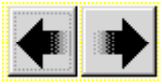
Il est un peu tard en conclusion pour décrire l'organisation interne de la base. Il est vrai que certains lecteurs commencent un livre par la fin. Cette structure est concentrée en un seul fichier et comprend trois types de pages:

- les pages-texte. Elles occupent les derniers rangs de la base, à partir de la page n° 1516. Leur nombre n'est pas limité.
- les pages-dictionnaire , qui occupent les premières places, de la page 11 à la page 1293. Leur nombre est fixe, même si les données ne remplissent pas la totalité de l'espace alloué (les pages-dictionnaire restées vides sont alors inaccessibles aux boutons de navigation).
- les pages-résultats, qui font tampon entre les deux zones, de la page 1294 à 1515 et qui assurent chacune un rôle spécifique:
  - graphiques
  - analyse factorielle et arborée
  - index (ou dictionnaires ou concordances alphabétiques)
  - listes de mots
  - vocabulaire spécifique du corpus et des textes qui le constituent
  - structure du vocabulaire
  - topologie.

A ce lot se rattachent les dix premières cartes, dont le rôle est essentiel pour le pilotage général, et particulièrement pour les fonctions qui intéressent la création d'une base nouvelle et l'exploitation systématique d'une base créée. C'est là qu'on trouve les deux principales fonctions de recherche documentaire CONCORDANCE et CONTEXTE.

CIRCULATION DANS LA BASE

La circulation n'a guère de sens interdit.

1 - D'une part à partir de chaque page d'un type donné, on peut aller à la suivante ou à la précédente du même groupe en utilisant les boutons flèche à droite ou flèche à gauche.  . C'est la circulation linéaire. L'adressage sélectif est aussi possible à l'intérieur du même groupe.

2 - D'autre part les relations sont aussi intergroupes. Car selon les méthodes de l'hypertexte, chaque mot (et donc la page-dictionnaire qui le contient) est relié aux pages où on le rencontre, de même que les formes d'une page donnée renvoient à l'emplacement qui est le leur dans les pages-dictionnaire, en sorte que le va-et-vient est incessant tout en restant ordonné: on revient toujours au point de départ. Le passage d'une zone à l'autre peut aussi se faire sans

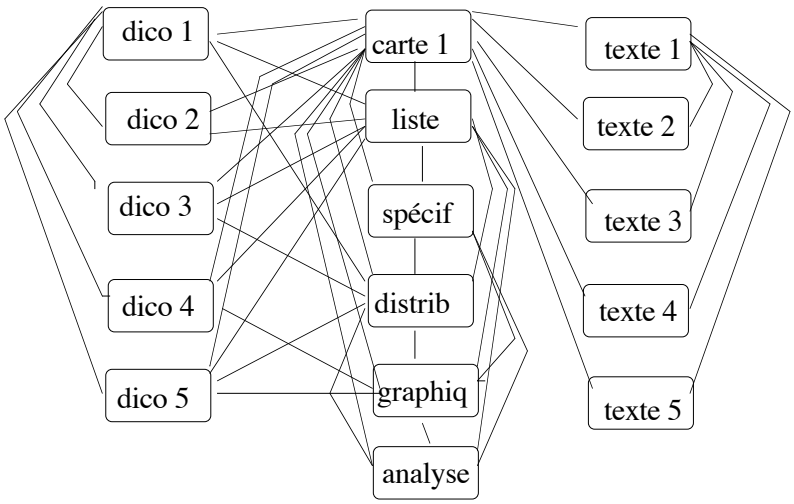
qu'aucune forme ne soit en cause, par la pression sur les boutons d'adressage direct.

3 - Enfin à partir des pages-résultats, on peut aller aux autres pages-résultats, et l'on peut aussi s'adresser au dictionnaire et au texte. Il arrive pourtant que par manque de place, tous les boutons de renvoi ne soient pas disponibles. En ce cas pour être conduit à la carte 1 où tous les aiguillages sont ouverts, il suffit de solliciter le bouton

SOMMAIRE 

qui est présent partout sous la même forme. Afin de prévenir la circulation désordonnée, on s'est appliqué à le rendre souvent seul visible et quasiment obligatoire, afin que l'utilisateur reconnaisse son chemin parmi les sentiers battus.

Circulation dans HYPERBAS



Les différents circuits possibles sont indiqués sur le graphique ci-dessous qui perd en lisibilité tout ce qu'il gagne en liberté. Il n'est pas très utile de s'interroger ici sur les chemins permis. Il vaut mieux porter son attention sur les circuits impossibles, dont on ne trouve quasiment aucun exemple. En sorte que la règle de circulation est d'une simplicité rabelaisienne: fais ce que voudras.

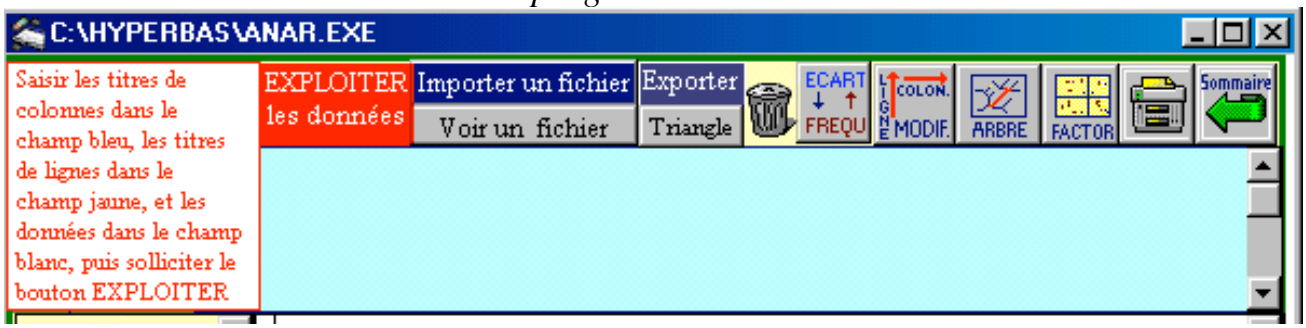
LES PROGRAMMES EXTERNES

Il y a peu de programmes externes, associés au présent logiciel, sinon le runtime TB40RUN.EXE dont la nécessité s'impose à toutes les applications fondées sur Toolbook. L'utilisateur aura recours aux utilitaires dont il a l'habitude, lorsqu'un fichier créé par HYPERBAS a besoin d'un tableur ou d'un éditeur.

LE PROGRAMME ANAR (ou HYPANAR)

Certains utilisateurs qui disposent de données quantitatives, issues ou non de corpus textuels, ont souhaité les soumettre aux mêmes méthodes qu'Hyperbase applique aux données textuelles. Avec ANAR (ANalyse factorielle et ARborée) on leur a donné le moyen de traiter de telles données sans qu'il soit nécessaire d'engager la chaîne des opérations qui conduisent à l'indexation et aux dénombrements et qu'HYPERBASE assure d'abord. Les programmes statistiques qu'Hyperbase propose en fin de parcours sont immédiatement accessibles, qu'il s'agisse d'histogrammes ou d'analyses multidimensionnelles. Ils ont cependant été adaptés pour prendre en compte des séries ou des tableaux extérieurs enregistrées dans des fichiers, au format "texte" ou "excel". La saisie directe des données est aussi possible, comme leur contrôle, leur correction ou leur modification. On a considéré que les données fournies en entrée seraient souvent des tableaux de contingence, nourris de fréquences brutes ou d'effectifs absolus. On a prévu dans ce cas le calcul classique de l'écart réduit., la fréquence théorique étant obtenue à partir du produit des deux totaux marginaux, de ligne et de colonne, pondéré par le total général. Mais on a pensé que les données pourraient être traitées telles quelles, s'il s'agit de distances, de pourcentages, de fréquences relatives, de données déjà pondérées ou même de données absolues, obtenues dans des conditions qui ne nécessitent aucune pondération (dans le cas de textes de longueur égale, par exemple).

Le programme ANAR



Qu'elles soient puisées dans un fichier ou saisies au clavier, les données sont d'abord contrôlées, et le traitement est suspendu s'il manque un élément du

tableau ou un nom de ligne ou de colonne. Le format d'enregistrement est peu exigeant ; il suffit qu'on utilise le blanc ou la tabulation pour séparer les éléments d'une même ligne, et le retour de chariot pour séparer les lignes. Si l'on retient l'option de calcul, on ne peut cette fois échapper à la loi normale car manquent ici les paramètres qu'exigerait le modèle hypergéométrique.

LE PROGRAMME *ANCORR* d'ANALYSE FACTORIELLE

(un programme semblable, écrit par Ludovic Lebart, est utilisé par la version lemmatisée d'Hyperbase; voir le chapitre consacré plus loin aux "corrélats")

Le programme d'analyse factorielle exige un complément d'explication. Cet ensemble de procédures multidimensionnelles, écrites en Fortran, est connu sous le nom de ADDAD. Voici l'adresse où l'on peut se procurer ce dernier programme qui a des versions implantées sur les gros systèmes et toutes les plates-formes, et dont les fonctionnalités sont beaucoup plus riches que ce que nous laissons entrevoir avec la seule analyse de correspondance:

ADDAD (Association pour le Développement et la Diffusion de l'Analyse des Données), 22 rue Charcot, 75013 Paris. Tél: 45 85 40 28. Responsable: J.P. Fénelon

Précisons que notre emprunt à ADDAD se borne au programme ANCORR.EXE qui est accompagné de trois fichiers: celui des données, appelé TABLEAU.AFC, celui des résultats (ANALYSE.AFC) et celui des paramètres (AFC.PAR). Ce dernier indique précisément comme paramètres les noms des fichiers TABLEAU et ANALYSE. Mais bien d'autres valeurs variables peuvent être transmises aux paramètres dont la variété n'est guère exploitée par HYPERBAS. HYPERBAS se contente d'indiquer le nombre de lignes (NI), le nombre de colonnes (NJ) et le nom des colonnes, derrière le mot-clé FLISTE.

Tous les autres paramètres sont fixés une fois pour toutes, selon le standard suivant:
(sont encadrés les paramètres pris en compte par HYPERBAS)

```
$RUN ANCORR
$L080
$F11=TABLEAU
$PRT=ANALYSE
$PAR=.
TITRE ANALYSE FACTORIELLE ;
PARAM NI = [34] NJ = [7] NF = 5 ;
OPTIONS IMPFI=1 IMPFJ=1 NGR=2 ;
GRAPHE X=1 Y=2 GI=1 GJ=1 ;
```

```
GRAPHE X=3 Y=4 GI=1 GJ=1 ;
FLISTE Argol Ténébreux Syrtes Forêt Lettrines
Ville Chemin;
(22X,A4,120F5.0) ;
$END
```

Mais rien n'empêche d'abandonner HYPERBASE pour modifier les paramètres avec un éditeur avant de lancer le programme ANCORR.EXE. Avec le même éditeur on récupérera les résultats dans le fichier ANALYSE. Voir l'exemple ci-dessous où la modification porte sur deux colonnes placées en éléments supplémentaires.

Modification des paramètres (encadrés)

```
$RUN ANCORR
$L080
$F11=TABLEAU
$PRT=ANALYSE
$PAR=.
TITRE ANALYSE FACTORIELLE ;
PARAM NI =34  NJ = 7  NF = 5  NJ2=2  ;
OPTIONS IMPFI=1 IMPFJ=1 NGR=2 ;
ORGAN  1  1  0  1  0  1  1  ;
GRAPHE X=1 Y=2 GI=1 GJ=3;
GRAPHE X=3 Y=4 GI=1 GJ=3 ;
FLISTE Argol Ténébreux Syrtes Forêt Lettrines
Ville Chemin ;
(22X,A4,120F5.0) ;
$END
```

Voici la liste de ces paramètres et leur signification:

```
$PAR=.      (maximum 80 caractères pour les lignes de paramètres)
TITRE               (maximum 72 caractères)
PARAM  NI= 22  NJ = 10 NF= 5  NI2= 4 NJ2= 3  ;
            NI >= nb de lignes réel
            NJ=  nb de colonnes
            NF = nb de facteurs <= min (NI-1, NJ-1)
            NI2= lignes en éléments supplémentaires (par déf.0)
            NJ2= colonnes en éléments supplément.  (par déf.0)
OPTIONS IOUT=1  IMPFI=1 IMPFJ =1 NGR=2;
            IOUT=0  défaut: pas de listing des données
            IMVP=0 défaut: pas d'impression des valeurs propres
            IMPI = 0 défaut  (même syntaxe pour IMPJ : colonnes)
                 = 1 lignes actives
```

= 2 lignes supplémentaires
= 3 toutes les lignes
NGR=0 défaut: nombre de graphiques
ORGAN 0 1 1 0 0 1 0 1 1 1 1 ;
ligne présente si NJ2 > 0 (1 pour les colonnes actives, 0 pour les supplémentaires)
GRAPHE X=1 Y=2 GI=1 GJ=1 NCHAR=2 ;
X: numéro du facteur en abscisse
Y: numéro du facteur en ordonnée
GI = 0 pas de projection des lignes
= 1 projection des lignes actives
= 2 projection des lignes supplém.
= 3 projection de toutes les lignes
GJ = 0 pas de projection des colonnes
= 1 projection des colonnes actives
= 2 projection des colonnes supplém.
= 3 projection de toutes les colonnes
NCHAR = 4 défaut (nb de caractères identificateurs)
OPT=3 défaut (1er point imprimé + légende)
= 1 (1er point imprimé)
= 2 (1er point imprimé , les autres dessous)
= 4 (graphique de densité)
NPAGE =1 défaut: nb de pages en largeur
CADRE =0 défaut
= 1 cadrage à 2,5 s
FLISTE Argol Ténébreux Syrtes Forêt ;
format libre pour le titre des colonnes
(A4,4X,6F8.0);
format FORTRAN pour la lecture des données

LES PROGRAMMES D'INDEXATION p114.EXE et p116.EXE

Le programme de tri et d'indexation , P114.EXE, a été réalisé par Jean-Pierre Anfosso et extrait de sa thèse. Il est écrit en langage C et intervient à plusieurs moments du traitement, pour trier ce qu'on lui propose: des graphies, des lemmes, des codes ou des séquences ou structures syntaxiques. Les signes de ponctuation sont traités comme des unités à part entière, s'ils sont précédés et suivis d'un séparateur. Seuls les blancs et les retours de chariot sont considérés comme des séparateurs Le programme est appelé au moment opportun. Pour assurer le synchronisme des opérations et poursuivre le traitement, l'utilisateur est invité à donner le signal de reprise, dès que l'indexation du corpus est

achevée, ce qui est signalé par la disparition d'un écran noir qui accompagne les programmes sous DOS.

Il suffit de lui proposer un fichier de données convenablement formaté, dans le même répertoire, et de lui donner le nom TEXTE.TXT. Le formatage consiste à séparer les textes du corpus par une ligne-jalon commençant et finissant par le symbole &&&, et à séparer les pages par un autre jalon commençant par le signe $ et indiquant le numéro de page (cela implique que le signe $ ne soit pas dans les données, non plus que le symbole &&&).

Le fichier TEXTE.TXT proposé au tri et à l'indexation.

```
&&&1,1,1&&&
"
$166451
 En entrant dans la chambre , Roubaud posa sur la table le pain d' une livre , le pâté et
la bouteille de vin blanc .
Mais , le matin , avant de descendre à son poste , la mère Victoire avait dû couvrir le feu
de son poêle , d' un tel poussier , que la chaleur était suffocante .
Et le sous - chef de gare , ayant ouvert une fenêtre , s' y accouda .
C' était impasse d' Amsterdam , dans la dernière maison de droite , une haute maison où la
Compagnie de l' Ouest logeait certains de ses employés .
La fenêtre , au cinquième , à l' angle du toit mansardé qui faisait retour , donnait sur la
gare , cette tranchée large trouant le quartier de l' Europe , tout un déroulement brusque
de l' horizon , que semblait agrandir encore , cet après - midi - là , un ciel gris du
milieu de février , d' un gris humide et tiède , traversé de soleil .
En face , sous ce poudroiement de rayons , les maisons de la rue de Rome se brouillaient ,
s' effaçaient , légères .

$186264
A gauche , les marquises des halles couvertes ouvraient leurs porches géants , aux vitrages
enfumés , celle des grandes lignes , immense , où l' oeil plongeait , et que les bâtiments
de la poste et de la bouillotterie séparaient des autres , plus petites , celles d'
Argenteuil , de Versailles et de la Ceinture ; tandis que le pont de l' Europe , à droite ,
coupait de son étoile de fer la tranchée , que l' on voyait reparaître et filer au - delà ,
jusqu' au tunnel des Batignolles .
```

Le résultat de l'indexation est enregistré dans le fichier GENERAL.TXT qui prend la forme ci-dessous, où chaque mot est suivi de sa fréquence et des numéros de page où on le rencontre. Suit une série de binomes, détaillant les sous-fréquences du mot dans les textes, le premier terme indiquant le numéro d'ordre du texte, le second terme la sous-fréquence dans ce texte.

Les noms de fichier d'entrée et de sortie étant toujours les mêmes, il n'y a pas lieu de préciser des paramètres lors de l'appel du programme. Il suffit de se positionner sous DOS dans le bon répertoire et de lancer l'invite de commande en précisant seulement le nom du programme: p114.

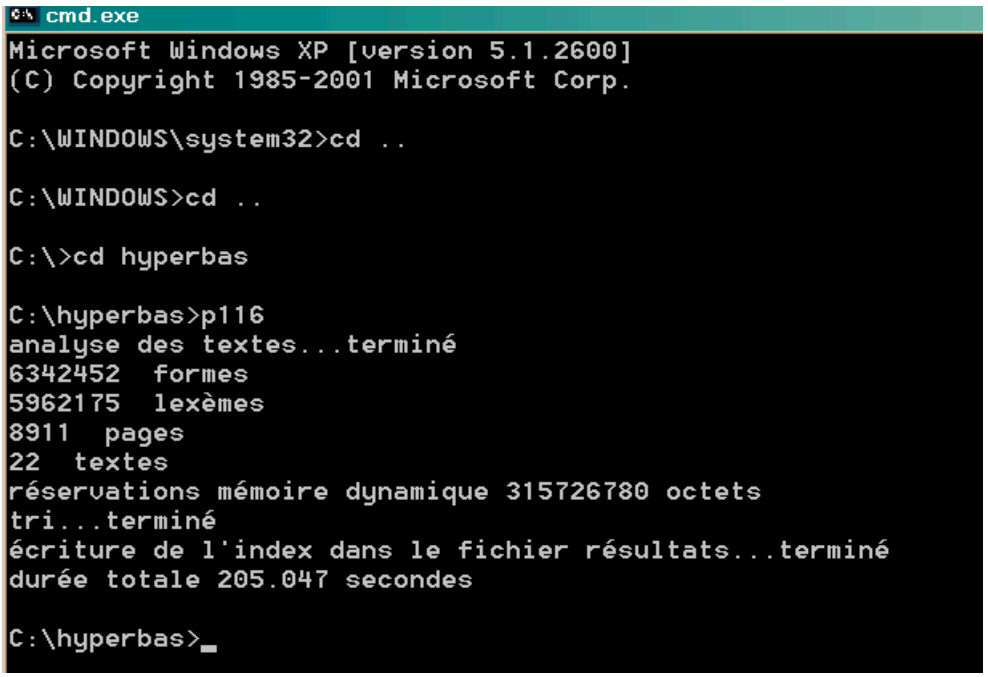
Le programme P116.EXE est une variante du précédent, la limite maximum du mot étant portée de 25 à 100 caractères. Cela est nécessaire pour trier les segments répétés, considérés comme des chaines de caractères englobant plusieurs mots consécutifs. En une telle circonstances non seulement les "mots" sont plus longs, mais aussi leur nombre est multiplié. Et en conséquence on demande au programme des performances extrêmes. On verra dans l'exemple ci-dessous que le programme P116.EXE s'acquitte

honorablement de sa tâche, puisqu'il se contente de 200 secondes pour indexer un fichier de 300 millions de caractères.

Le fichier GENERAL.TXT généré par le programme P114.EXE

```
bibles  2  276149 276247  , 2 1 3 1
bibliographie   1  275472  , 1 1
bibliographies  1  276460  , 3 1
bibliophile     2  276146 276149  , 2 2
bibliothèque    23  276107 276108 276108 276118 276138 276140 276144
276146 276147 276149 276220 276222 276222 276226 276417 276417 276417
276417 276417 276452 276453 276454 276460  , 2 14 3 9
bibliothèques   4  276146 276444 276444 276876  , 2 1 3 2 4 1
biblique        1  275791  , 1 1
bibliques       1  275499  , 1 1
bicyclette      1  276437  , 3 1
bien    923   , 1 127 2 313 3 238 4 245
bienfaisance    2  275581 275763  , 1 2
bienfaiteur     3  276034 276048 276191  , 2 3
bienfaitrice    1  275960  , 2 1
bienfaits       5  275787 275933 276035 276048 276049  , 1 1 2 4
bienheureuses   1  276113  , 2 1
biens   10  275499 275529 275556 275620 275620 275725 275726 275729
275800 276554  , 1 9 4 1
bientôt         18  275733 275736 275958 275984 275984 275985 276015
276018 276018 276022 276037 276051 276065 276077 276079 276092 276332
276337  , 1 2 2 14 3 2
bienveillance   3  275539 275752 275780  , 1 3
bienveillante   2  275612 275618  , 1 2
bienvenu        2  276927 276927  , 4 2
bienvenue       3  276033 276694 276785  , 2 1 4 2
bienvenus       1  276812  , 4 1
bière   4  276107 276139 276428 276429  , 2 2 3 2
bières  1  276428  , 3 1
```

Application du progr amme P116.EXE au tri des segments répétés

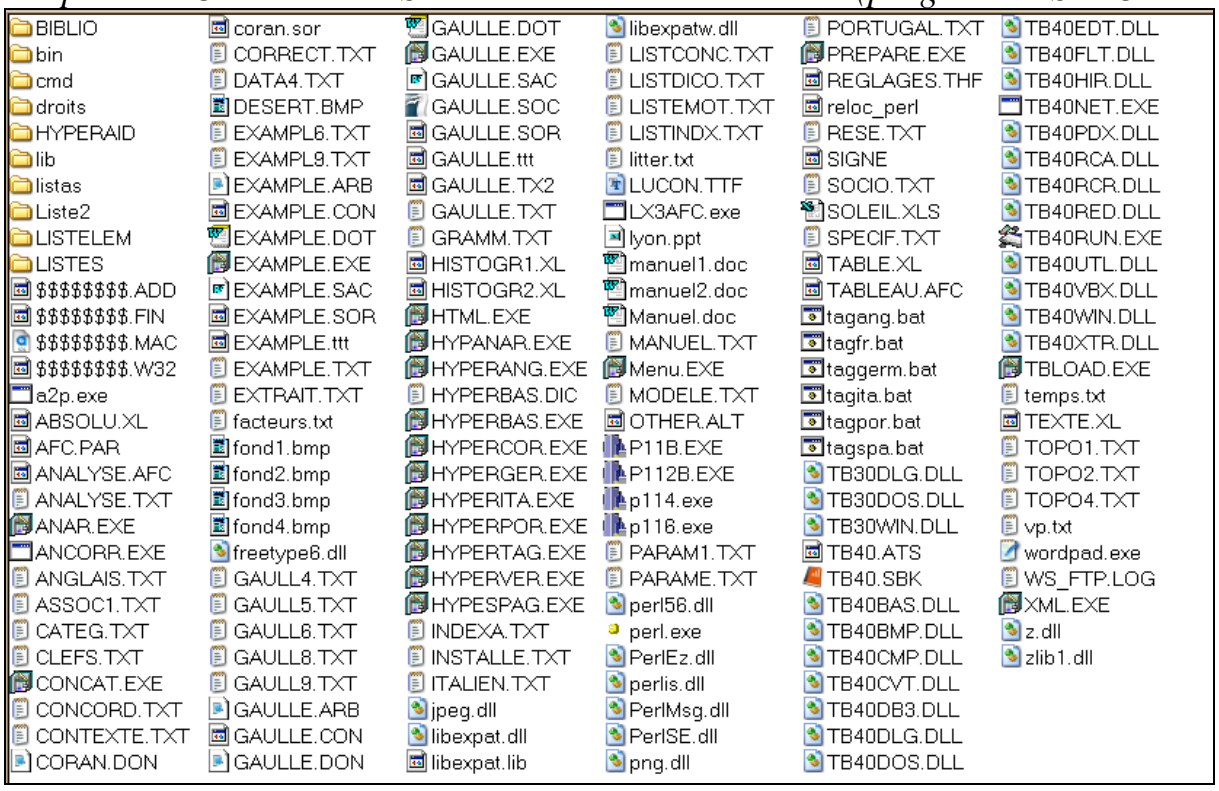


Le programme P114.EXE a la même efficacité: il ne lui faut que deux ou trois secondes pour indexer un corpus d'un million de mots.

LE CONTENU DU REPERTOIRE HYPERBAS

En réalité lorsqu'on explore le détail des programmes externes, copiés dans le répertoire C:/HYPERBAS/ au moment de l'installation (voir figure ci-dessous), on trouve bien d'autres fichiers, qui correspondent à des programmes utilisés par la version lemmatisée du logiciel (en particulier dans les répertoires BIN, CMD, et LIB).

Le répertoire C:/HYPERBAS/ au moment de l'installation (programme SETUP.EXE)



On peut y tranférer aussi, grâce au programme MENU.EXE, une quarantaine de bases gratuitement offertes à l'utilisateur et couvrant une grande part de la littérature française, de Rabelais à Proust. En dehors du modèle HYPERBAS.EXE la seule base qui soit fondée sur la version standard du logiciel est la base EXAMPLE, à laquelle nous avons emprunté la plupart des illustrations qui précèdent. Toutes les autres appartiennent au modèle lemmatisé, que nous nous proposons d'exposer maintenant.

Seconde partie

HYPERBASE pour WINDOWS

Versions lemmatisées

CHAPITRE 1.
LES LEMMATISEURS

LA LEMMATISATION

Le débat sur la lemmatisation a commencé il y a trente ans. À l'époque les chercheurs de Saint Cloud contestaient les recommandations de Charles Muller. La querelle est à ce jour apaisée mais la question reste en suspens. Aussi bien des travaux estimables ont été publiés qui suivent l'une ou l'autre option (et parfois les deux). Ceux qui s'en tiennent à la graphie sont sans doute les plus nombreux, non seulement parce que la préparation et le traitement y sont plus aisés, mais aussi parce que les résultats permettent plus facilement la comparaison, l'intervention humaine dans les données étant réduite au minimum. Ceux qui veulent traiter un produit raffiné et s'attachent au lemme s'échelonnent sur le long chemin qui va du réel à l'idéal. Deux obstacles principaux se dressent sur leur chemin, dont l'un tient aux traitement des expressions ou mots composés, l'autre aux homographes.

Même ceux qui s'abstiennent de lemmatiser ont des difficultés de définition. On a dénoncé depuis longtemps l'ambiguïté du mot mot. Mais il en va ainsi de la graphie. Il ne suffit pas de décréter que la graphie est l'espace imprimé entre deux séparateurs. Encore faut-il dresser la liste des séparateurs et prendre position sur les cas ambigus où un même caractère peut ou non jouer ce rôle de séparateur (par exemple le point, l'apostrophe et le trait d'union). Mais le lemme est plus difficile encore à délimiter. Il ne suffit pas d'invoquer la garantie d'un dictionnaire pour régler le problème de la nomenclature. Une fois réglés les regroupements qui réduisent les variations du genre, du nombre, du temps, du mode et de la personne, que fera-t-on des syntagmes ou expressions? Les dictionnaires les traitent à l'intérieur de l'article consacré à l'un de leurs constituants mais leur accordent rarement une entrée indépendante. Le Petit

Robert fait un sort particulier aux deux exemples canoniques qu'on cite toujours: le *chemin de fer* et la *pomme de terre*. Mais il refuse ce privilège à la *pomme d'Adam* et au *chemin de croix*. Et Alain Rey s'en explique franchement dans l'édition de 1977: « il nous a paru plus raisonnable de donner à l'ordre alphabétique ces vrais composés que sont *pomme de terre, chemin de fer* ou *point de vue*. De tels procédés ne peuvent être appliqués systématiquement dans un dictionnaire, sous peine d'augmenter excessivement le nombre des « mots » traités et le volume de l'ouvrage. Cependant, quand certaines expressions ou formes verbales, moins importantes, constituaient de véritables mots (on dit alors qu'elles sont lexicalisées), elles ont été présentées en capitales dans les articles où elles sont incluses, pour attirer l'attention sur leur autonomie. Ainsi : EXCÈS DE POUVOIR (à l'article *excès*), la locution À L'EXCEPTION DE (sous *exception*), la locution COMME IL FAUT, qui s'emploie comme un adverbe ou comme un adjectif (à *falloir*). » On pourrait certes relever tous ces syntagmes pourvus de la majuscule et les ajouter à la nomenclature. Mais il en est tant d'autres qui n'ont pas cette marque de reconnaissance et dont on ne sait s'il s'agit d'une suite de mots intangible ou seulement d'une suite de mots fréquente mais modifiable. On a affaire en réalité à un continuum qui va de l'expression figée et lexicalisée à la variation libre. La frontière entre l'une et l'autre est mouvante selon les époques, selon les écrivains (beaucoup prennent plaisir à casser ou modifier les expressions consacrées) et selon les dictionnaires. Et même à l'intérieur d'un même dictionnaire les décisions sont parfois peu cohérentes. Même les exemple relevés par A. Rey ne sont pas à l'abri de la contestation: il reste un peu de liberté dans les syntagmes les plus fermés: on dira parfois un *excès manifeste de pouvoir* au lieu d'un *excès de pouvoir manifeste*, et l'adjectif *notable* peut s'introduire dans l'expression *à l'exception de*. Et que dire de *comme il faut*, dont l'emploi adjectival ou adverbial ne peut être reconnu que dans le contexte? Et c'est ici que réside la principale difficulté de la lemmatisation. C'est le contexte qui fixe la nature et les limites de la chaîne graphique. Le recours aux dictionnaires ne résout pas le problème de la nomenclature.

Encore moins celui des homographes. S'il est difficile de regrouper ce qui doit l'être, il est plus délicat encore de séparer ce qui doit l'être. Ici aussi les dictionnaires électroniques sont utiles, mais insuffisants. Ils signalent l'ambiguïté d'une forme mais se gardent bien de décider laquelle des analyses possibles est exacte dans le cas précis qui leur est soumis. Même si le dictionnaire consulté est doté de pondérations ou de pourcentages, on ne peut toujours trancher en faveur du sens majoritaire ou, solution pire encore, donner raison à la première des analyses proposées. Ceux qui se dispensent de toute analyse et ne considèrent que les formes brutes peuvent invoquer là aussi la difficile standardisation dans cette opération de décantation, lors même qu'elle est menée dans le texte. Car la filtration peut être plus ou moins fine, et porter sur divers critères: la catégorie du mot, sa fonction et son sens, ce qui exige

qu'on sache reconnaître ces mêmes critères dans les mots voisins. Comme un simple codage grammatical ne suffit pas à distinguer *voler 1* (dérober) de *voler 2* (dans les airs), l'analyse ne peut se contenter de la seule syntaxe et doit s'engager (mais jusqu'où?) dans la voie sémantique et pragmatique. On court le risque de la babélisation, ce que reconnaît Benoît Habert quand il évoque « l'inévitable éparpillement des étiquetages »[13].

Dans les travaux de linguistique quantitative, on s'est contenté d'applaudir le courage de Gunnel Endwall, sans suivre son exemple, et la prudence a souvent choisi le même camp que la paresse. En s'abstenant de lemmatiser les données, on adoptait un profil bas, avouant l'impureté des données et faisant confiance à la statistique pour les dégager de l'entropie. Mais cette position attentiste peut-elle être indéfiniment prolongée? Les industries de la langue ont fait des progrès et des outils de plus en plus performants sont disponibles sur le marché. Rares sont les rédacteurs qui méprisent l'usage du correcteur d'orthographe. On lui pardonne ses bévues eu égard aux services qu'il rend pour signaler les fautes de frappe et les accords négligés. Or il n'y a pas de correction possible sans analyse préalable. Et la lemmatisation entre nécessairement dans le processus. Les concepteurs de logiciels statistiques ont suivi cette tendance, parfois à moindres frais. En s'appuyant sur la troncature, ils ont pu isoler le radical et soumettre au calcul des effectifs regroupés, où la dispersion des formes fléchies était neutralisée. Et notre HYPERBASE a tenté de suivre dans cette voie l'exemple de TROPES, d'ALCESTE et de SPHINX (pour s'en tenir au français).

LE LEMMATISEUR WINBRILL

Notre première tentative de vraie lemmatisation s'est révélée décevante et le prototype lemmatisé d'HYPERBASE n'a jamais été distribué. Il y avait une raison juridique à cela: cette version reposait sur le logiciel WINBRILL qui est certes gratuit mais dont la traduction française, fruit des efforts conjugués de deux chercheurs de l'INaLF, J. Lecomte et G. Souvay, ne nous appartenait pas. S'y ajoutait un embarras méthodologique: d'une part Winbrill n'opère qu'un étiquetage grammatical et l'on doit lui adjoindre des fonctions complémentaires pour accéder au lemme. D'autre part les codes qu'on y distingue sont peu classiques et peu précis. La classe des déterminants n'est pas détaillée ; celle des pronoms manque de clarté et celle des verbes ignore les modes, les temps et les personnes. Un autre logiciel de lemmatisation a été mis au point dans le même laboratoire et a servi à constituer la nouvelle version de FRANTEXT où la catégorie grammaticale s'ajoute à la panoplie des critères de sélection (une forme, un vocable, une expression, une cooccurrence, une liste, une alternative, ou toute combinaison de ces objets). Mais ce produit interne n'était pas disponible à l'extérieur.

---

[13] B. Habert, A. Nazarenko, A. Salem, *Les linguistiques de corpus*, Armand Colin, 1997, p. 23.

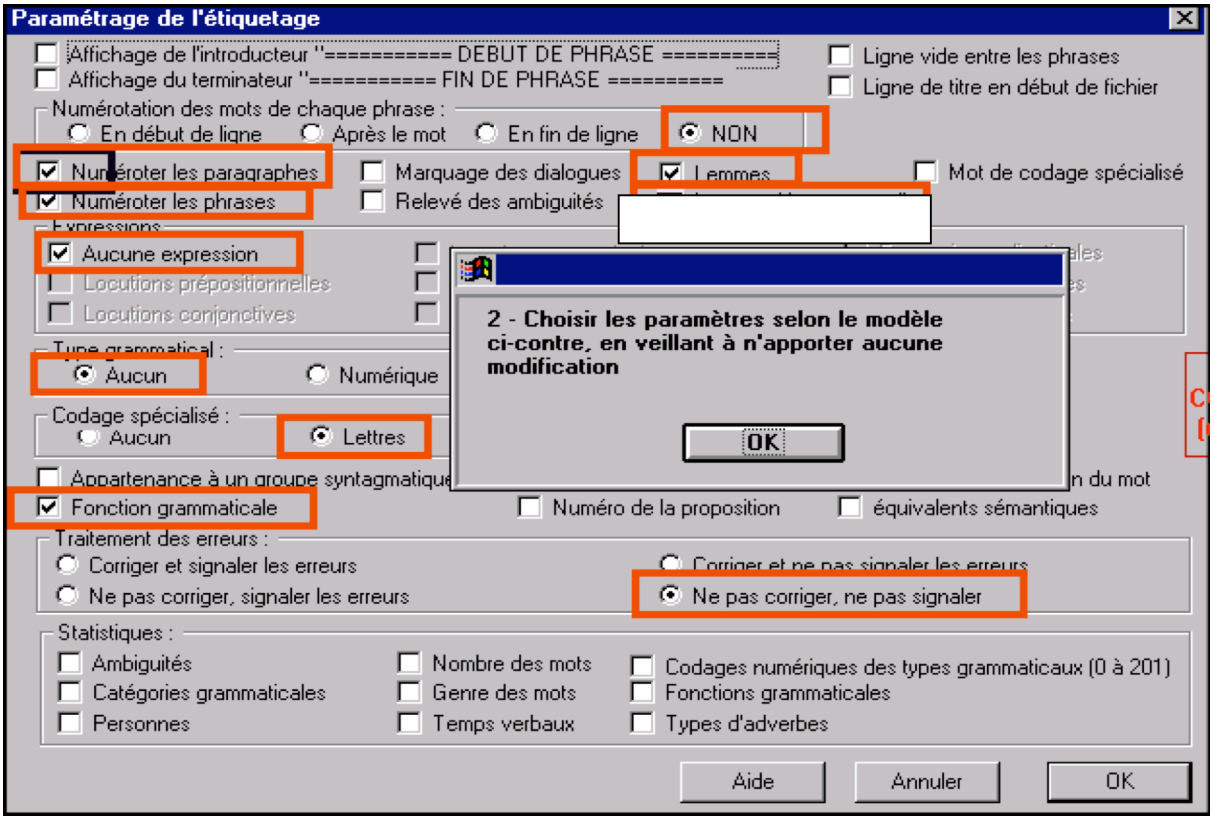
LE LEMMATISEUR CORDIAL (version HyperCor)

À qui donc s'adresser? On a songé d'abord à celui qui, sabre au clair, a maintenu sans faiblesse les exigences de la lemmatisation, à Dominique Labbé. Ses études lexicométriques, en particulier sur de Gaulle et Mitterrand, donnaient toutes les garanties souhaitables. Mais son logiciel, conçu pour une version ancienne du système Macintosh, exigeait une refonte préalable, rude tâche à laquelle ce chercheur a bien voulu s'atteler. En attendant que la nouvelle version soit disponible, un autre produit s'imposait, que beaucoup de gens utilisent sans le savoir et qui s'appellent Cordial. Le correcteur que Word Microsoft a parfois intégré à son traitement de texte est en effet emprunté à Cordial. Les nombreux prix glanés ici et là par ce logiciel s'accordent avec cette préférence enviée, qui en fait le correcteur le plus utilisé en France. Au reste les concepteurs de ce produit sont ouverts à la recherche universitaire et ont facilité l'expertise que mènent là-dessus François Rastier et son équipe. En particulier une version particulière du logiciel est destinée aux laboratoires spécialisés dans le traitement automatique de la langue, auxquels elle fournit un outil d'analyse et non plus seulement de correction. Cette version, anciennement dénommée « Cordial Université », est maintenant distribuée sous l'étiquette ANALYSEUR et correspond à la version 7 ou ultérieure du produit standard. Ce programme étant automatique n'est pas exempt d'erreurs. Mais il échappe à la fatigue, à l'inconstance, à la subjectivité et finalement au renoncement qui accompagnent les entreprises de désambiguïsation manuelle.

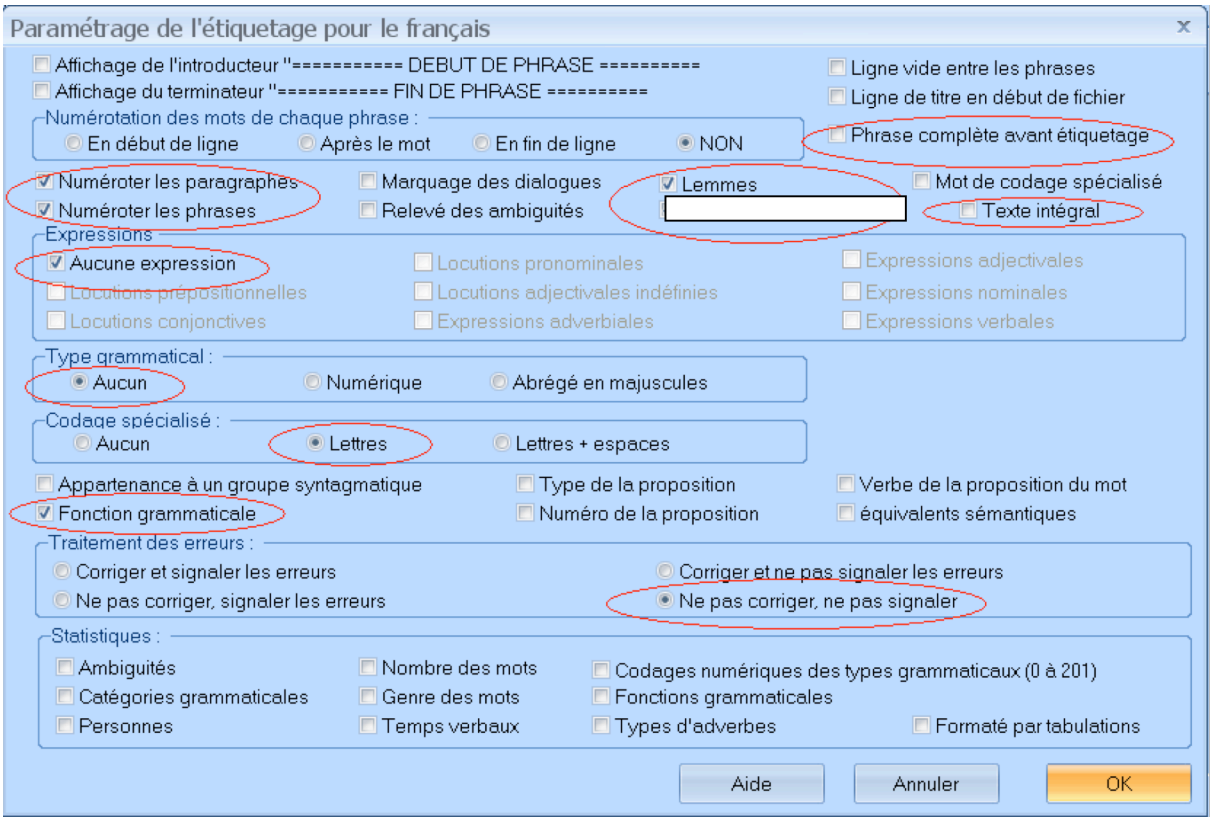
Comme le programme de Cordial est très rapide (quelques secondes suffisent pour un texte de 100 000 mots), nous n'avons eu aucun scrupule - ni aucun mérite - à lui proposer des corpus de grande dimension, celui de Balzac, qui compte 49 romans, celui de Hugo où 31 textes se trouvent rassemblés, et le même corpus romanesque qui nous a servi d'exemple dans la version standard .

Le corpus est soumis alors au programme de lemmatisation. Comme un gros corpus risque de dépasser les capacités de CORDIAL (l'échec est menaçant au delà d'un million de mots), les textes du corpus doivent être traités séparément, l'un après l'autre, en veillant à maintenir constants les paramètres selon les options qui sont cochées dans l'illustration ci-dessus. Le respect de ces paramètres est essentiel, faute de quoi la lecture du fichier serait fautive. En principe une fois qu'on a imposé les options adéquates (en rouge dans la figure ci-dessus), elles peuvent être enregistrées comme options par défaut. Cela est vrai des versions 7 et 8 de Cordial, mais ne l'est plus dans la version 14. À chaque lancement de Cordial, il faut donc redoubler l'attention et ne rien oublier des spécifications exigées. En particulier on veillera à cocher *Numéroter les paragraphes* et *Numéroter les phrases* et à désactiver *Phrase complète avant étiquetage* et *Texte intégral*, comme dans la copie d'écran ci-dessous. Il convient aussi de ne pas retenir toujours l'option *Lemmes fém->masculin*, car elle range la *femme* sous l'*homme* ce qui n'est pas toujours heureux.

Les paramètres à sélectionner dans le programme Cordial



Cordial version 14. Les options à choisir



Outre le code grammatical qu'il propose de trois façons différentes, Cordial ajoute de nombreux renseignements relatifs au traitement des expressions, à la fonction dans la phrase, à la place hiérarchique du mot dans l'arbre syntaxique, et même à la classe sémantique à laquelle le mot se rattache. Nous n'avons retenu que ce qui était strictement nécessaire à l'analyse, soit la moitié des possibilités offertes dans la figure 3, à savoir cinq champs remplis pour chaque mot : le numéro du paragraphe, le numéro de la phrase, la forme, le lemme, le code grammatical détaillé (codegram) et la fonction.

Un fichier lemmatisé par Cordial

N°	§	Phrase	Forme	lemme	ambig.	Typegra	CodeHexa	Codegram	Syntagme	Fonction	Num Sens	
==== DEBUT DE PHRASE ==												
1	1	1	Je	je		36	0xE480	Pp1.sn	1	S	1	
2	1	1	crois	croire	A3	101	-	Vmip1s	2	V	1	
3	1	1	que	que	A3	21	0x0000	Cs	-	-	2	
4	1	1	la	le	A3	15	0x6000	Da-fs-d	5|5	T	2	
5	1	1	langue	langue		26	0x6080	Ncfs	5|5	T	2	forme
6	1	1	est	être	A3	103	-	Vmip3s		6	2	
	*	*	*	*				*		*		

(L'astérisque désigne les champs retenus par Hyperbase, les deux derniers (Codegram et Fonction) étant agglutinés et combinés).

LE LEMMATISEUR TREETAGGER (versions HyperTag, HyperVer)

Le lemmatiseur *Cordial* ne s'applique qu'au français. D'autres lemmatiseurs étaient disponibles sur le marché national, comme *Winbrill*[14], *Sylex*[15], l'atelier de lexicométrie de Dominique Labbé ou encore *TreeTagger*. Si nous avons préféré *Cordial*, c'est que nous lui reconnaissons deux avantages :

1 - Il pousse plus loin que d'autres l'analyse, même si le nombre d'erreurs croît avec le niveau de profondeur. Certes ses résultats manquent de fiabilité s'ils portent sur la fonction des mots dans la phrase ou sur leur étiquette sémantique, mais la caractérisation purement grammaticale y est assez sûre, et, comme elle est très détaillée, elle permet une exploitation plus riche et plus variée.

2 – Il est largement répandu en France, ayant été choisi par Microsoft pour la correction orthographique du français.

---

[14] Une version éphémère d'Hyperbase, fondée sur le lemmmatiseur *Winbrill*, a vu le jour dans le passé. Elle a été abandonnée.

[15] *Sylex* est utilisé par le logiciel *Sphinx-Lexica*. En ce qui concerne *Alceste, Tropes* ou *Intex*, le traitement inclut des procédures internes de désambiguïsation, de regroupement et et de lemmatisation.

Mais lorsqu'on veut traiter des textes d'une autre langue, Cordial n'est d'aucun secours. Nous avons mené une enquête auprès des utilisateurs, au moins en ce qui concerne l'anglais. Il semble que TreeTagger remporte la majorité des suffrages, malgré quelques défauts :

1 – TreeTagger a besoin d'un apprentissage et ses résultats valent ce que vaut son apprentissage préalable. On n'est pas toujours sûr que les textes qu'on soumet à TreeTagger soient de même nature que ceux qui ont servi à sa mise au point.

2 – TreeTagger est de type statistique. Et les bévues du calcul ne sont pas rares. Il est vrai que l'approche par règles linguistiques n'est pas non plus une garantie contre les fautes d'analyse.

3 – Les sorties du traitement sont peu sophistiquées, chaque graphie étant accompagnée au maximum de deux informations seulement, le genre grammatical et le lemme. Encore ce lemme est-il porté inconnu (« unknown ») s'il n'existe pas dans le dictionnaire constitué au cours de l'apprentissage.

Mais cette rusticité peut aussi être portée à l'actif du logiciel, qui est immédiatement opérationnel. Nul besoin de paramétrer les sorties. Les options sont peu nombreuses et le plus généralement on utilise celles qui sont prévues par défaut, à savoir :

-token –lemma -sgml

à quoi nous proposons d'ajouter : *–no-unknown*[16]

Nul besoin de procéder soi-même à l'apprentissage. Des spécialistes de chaque langue y ont pourvu. On peut parfois choisir entre plusieurs apprentissages pour la même langue, comme c'est le cas pour l'italien. Pour le français il existe même une version spécialement adaptée à l'ancien français. On ne conseille pas de réaliser soi-même un apprentissage nouveau, ce qui est possible mais délicat. Mais on peut sans trop de difficulté corriger ou compléter le fichier lemmatisé, par exemple en proposant un lemme et un code là où le logiciel bute sur un mot « unknown ». Il faudra veiller cependant à respecter le format d'origine et ne pas bousculer l'ordre des tabulations et des retours de chariot, afin que chaque mot dispose de trois champs sur une même ligne.

Une précaution préalable doit être prise qui concerne le point et les signes de ponctuation. Si ces signes sont collés au mot qui précède, comme c'est la tradition typographique, le traitement de TreeTagger est souvent fautif. Prévenir les erreurs en mettant un blanc derrière mais aussi devant tous les signes de ponctuation relevés dans le fichier d'entrée (par exemple en utilisant une fonction de Word ou mieux en mettant à profit le programme PREPARE.EXE

[16] Cela n'est pas absolument nécessaire. Car, au moment de l'importation des données lemmatisées, Hyperbase élimine les mentions « unknown » et les remplace par un lemme identifié à la graphie.

qui accomplit cette tâche spécifique pour l'ensemble des fichiers qu'on veut traiter).

Les fichiers d'entrée sont au format "texte seulement". Si TreeTagger est sollicité à partir d'un Macintosh, on devra se méfier du code adopté par Apple pour la transcription des lettres accentuées, si de telles lettres existent dans la langue traitée. Comme les ressources linguistiques et notamment le dictionnaire machine sont au format PC, les fichiers à soumettre à TreeTagger doivent avoir le code PC, même si on travaille sur Macintosh. Word permet aisément de passer d'un code à l'autre.

Tous les fichiers à traiter doivent être réunis dans le répertoire où se trouve TreeTagger. Les fichiers de sortie seront nommés TEXTE1.CNR, TEXTE2.CNR, TEXTE3.CNR, etc, afin de faciliter le traitement automatique d'HYPERBASE. Contrairement à la version non lemmatisée d'Hyperbase, le corpus n'est pas constitué en un seul fichier, avec des jalons intérieurs pour séparer les textes. Comme Cordial, TreeTagger aurait été embarrassé pour traiter ces jalons qu'il aurait assimilés au texte même. Ajoutons que dans les très grands corpus, le traitement en un seul passage aurait fait difficulté (TreeTagger ainsi que Cordial ne peut guère dépasser un million de mots).

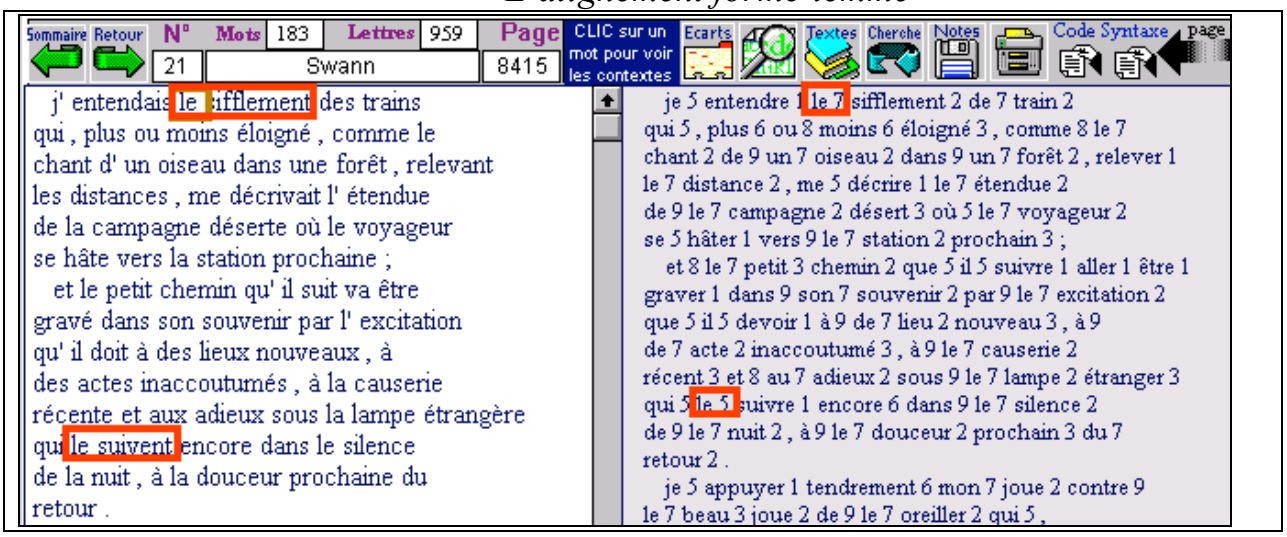
Au total, sans cacher notre préférence pour la lemmatisation plus fine et plus complète de Cordial, l'utilisateur peut être sensible à deux ou trois atouts de TreeTagger: d'abord sa gratuité et sa disponibilité, ensuite la simplicité et la constance du paramétrage, enfin son aptitude à traiter des langues différentes, au moins celles de l'Occident.

CHAPITRE 2.
LE TRAITEMENT DES TEXTES LEMMATISÉS

LES QUATRE NIVEAUX D'INDEXATION

Que la lemmatisation ait eu recours à TreeTagger ou à Cordial, Hyperbase prend en compte les trois principaux éléments d'un tel fichier et les distribue séquentiellement dans trois champs parallèles, voués respectivement aux formes, aux lemmes et aux codes. La figure 3 met en correspondance les formes et les lemmes d'une même page de Proust. On notera que les lemmes, dans la partie droite de l'écran, sont pourvus d'un indice numérique, afin de séparer les uns des autres les homographes. Ainsi *le 7* (dans *le sifflement*) distingue l'article du pronom codé 5 (dans *qui le suivent*). Ces codes simplifiés qui reproduisent la classification de Muller et de Labbé (1 verbe, 2 substantif, 3 adjectif, 4 numéral, 5 pronom, 6 adverbe, 7 déterminant, 8 conjonction, 9 préposition) n'appartiennent pas en propre à Cordial non plus qu'à TreeTagger, mais ont été dérivés de l'analyse complète fournie par le lemmatiseur.

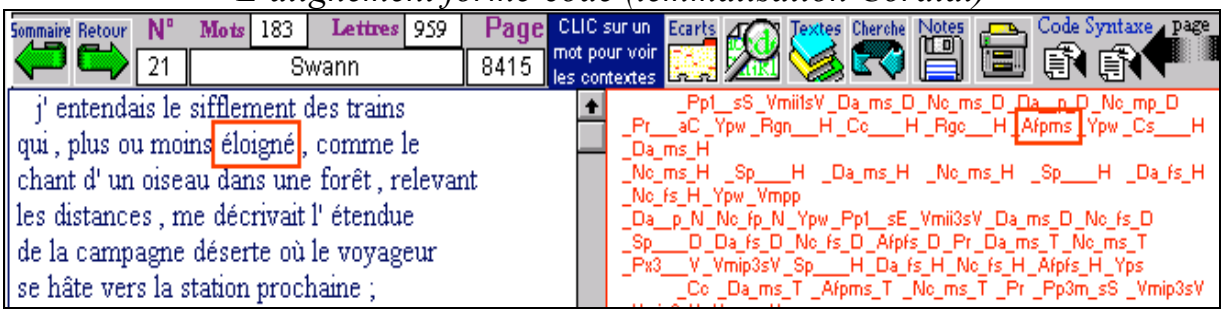
L'alignement forme-lemme



Cette analyse complète est rendue visible, quoique peu lisible, pour peu qu'on sollicite le bouton CODE situé à droite de la barre de menu. Là aussi l'alignement est rigoureux, en sorte que l'on sait précisément à quel mot correspond telle ou telle analyse. Ces trois champs sont sensibles au clic de la

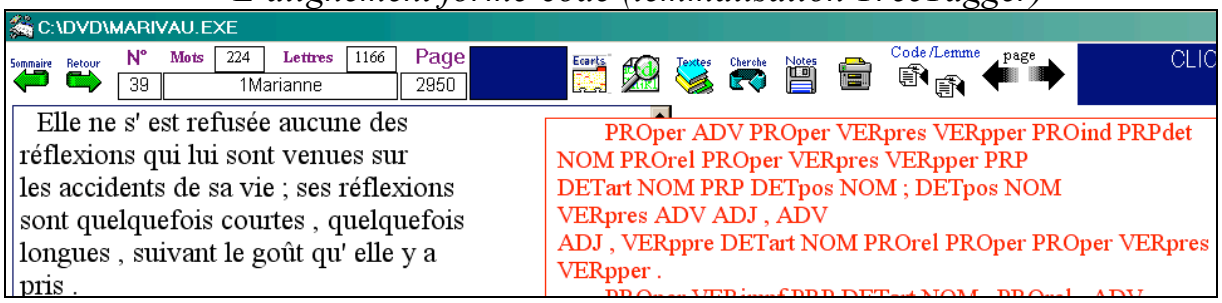
souris: tout objet que l'on désigne, qu'il s'agisse d'une forme, d'un lemme, d'un code ou d'une structure, renvoie aux autres occurrences où le même objet est rencontré, les relations hypertextuelles s'appliquant aux quatre champs. Mais ces relations lient aussi ces champs entre eux, en sorte qu'en cliquant sur un code grammatical dans le champ de droite (par exemple *_Afpms*, soit *adjectif qualificatif, positif, masculin, singulier*) on obtient successivement en vidéo inverse tous les adjectifs qui répondent à ce codage dans le champ de gauche.

L'alignement forme-code (lemmatisation Cordial)



Cependant cette possibilité offerte par *HyperCor* est désactivée dans les versions fondées sur TreeTagger (*HyperTag* et *HyperVer*). Car le nombre d'étiquettes y est beaucoup plus restreint (des dizaines au lieu de milliers) et le défilement des mots qui partagent le même code est sans intérêt quand leur nombre est très élevé.

L'alignement forme-code (lemmatisation TreeTagger)



L'indexation et toutes les opérations subséquentes sont alors répétées quatre fois, au niveau des structures syntaxiques (c'est à dire des séquences ordonnées des parties du discours), puis des codes grammaticaux, puis des lemmes, puis des formes. À l'issue de ce traitement, on obtient quatre index qui réagissent pareillement au clic de la souris. La forme ou le lemme ou le code ou la structure qu'on désigne montre le détail de ses occurrences, parmi lesquelles l'utilisateur fait son choix pour se référer au texte.

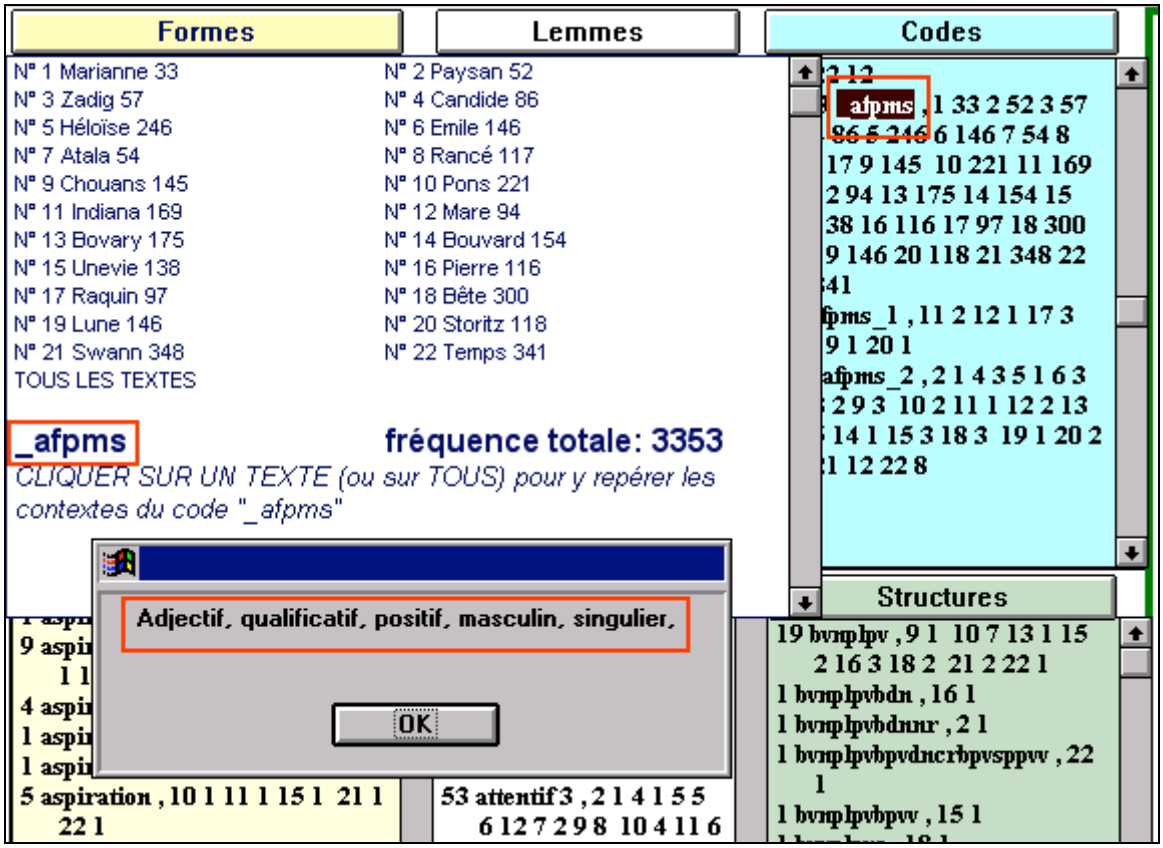
S'il s'agit d'un code grammatical, dont la signification peut être opaque, le décryptage est assuré et traduit en clair, comme dans l'exemple ci-dessous, relatif à l'adjectif qualificatif, positif au masculin singulier. Lorsqu'on veut, non plus reconnaître un code particulier, mais rassembler les mots qui ont reçu le même codage (par exemple pour constituer une concordance ou un

histogramme), on est renvoyé à une page spéciale qui dénombre toutes les combinaisons possibles.

A la différence de TreeTagger, Cordial pousse loin l'analyse, en relevant pour chaque mot la catégorie, la sous-catégorie, le genre, le nombre, la fonction et s'il s'agit d'un verbe le temps, le mode et la personne. Un clic dans une option provoque alternativement l'activation ou la désactivation correspondante. Certaines options sont impliquées ou exclues automatiquement, dès qu'une autre est choisie, de telle façon qu'il y ait toujours cohérence. Car il serait absurde de sélectionner le futur d'un substantif ou le féminin d'un verbe à l'infinitif. Chaque clic modifie le filtre dont l'affichage apparaît dans une fenêtre, en haut et à droite de l'écran, avec sa traduction en clair. Toute colonne non intéressée par la sélection est remplie par défaut par un joker, dont l'effet est d'admettre tout code qu'on rencontre à cet endroit. Ainsi dans l'exemple choisi la colonne 3 n'ayant pas été sélectionnée, tous les adjectifs seront retenus, quel que soit le degré, positif ou comparatif. De même le vide rencontré dans la colonne 7 laissera la sélection indifférente à la fonction dans la phrase.

Une fois que la sélection est faite, elle est transmise à la fonction appelante, qui délivre un contexte, une concordance, ou une liste , c'est à dire un tableau à deux dimensions dont chaque ligne dresse le profil d'une sélection grammaticale à travers le corpus et chaque colonne celui d'un texte du corpus à travers les codes grammaticaux.

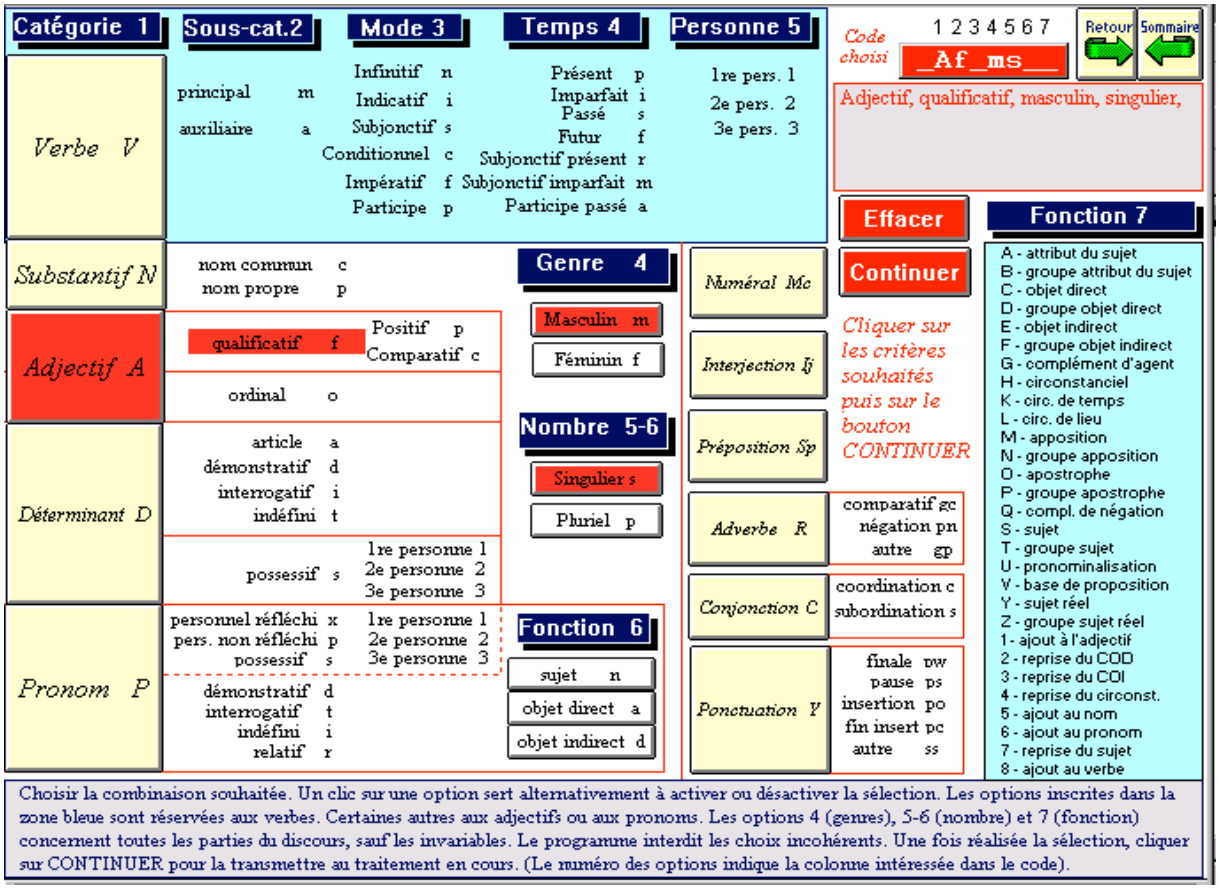
Les quatre index issus de Cordial



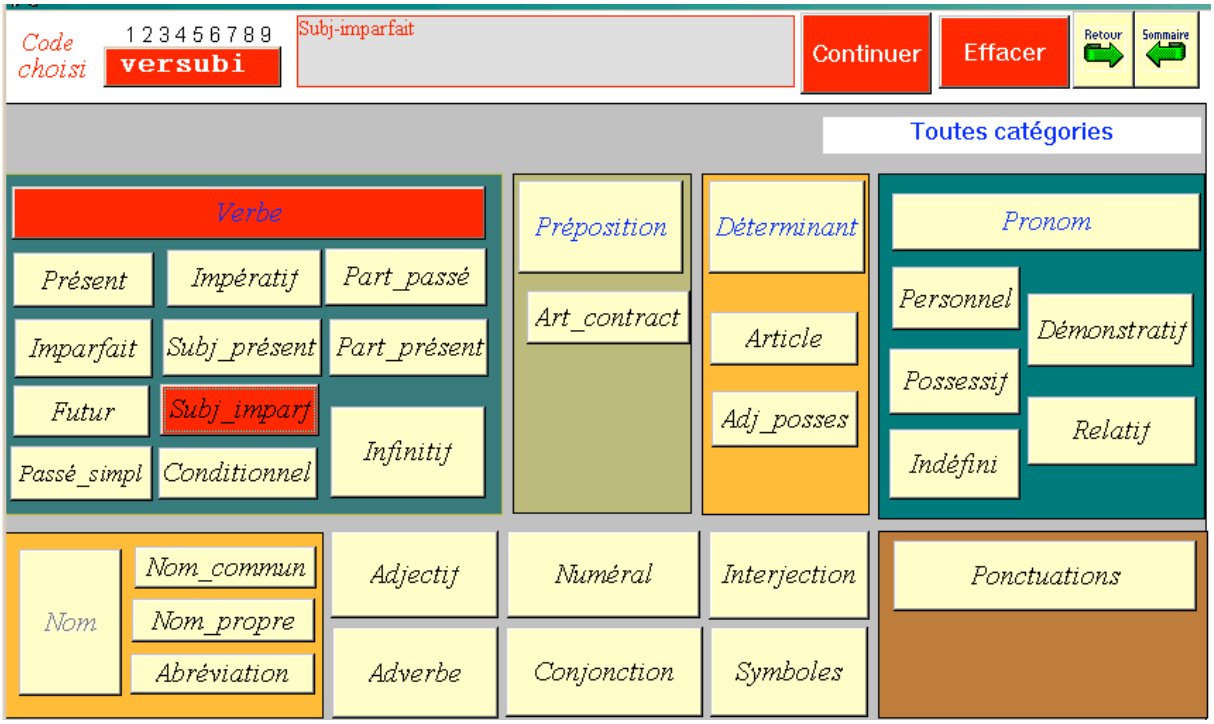
La sélection d'un code peut être faite *de visu*, soit à partir du texte, soit à partir du dictionnaire quand le code tombe sous les yeux, à portée de la souris (comme dans les écrans précédents). Mais elle peut se faire *in absentia* à partir d'un catalogue où toutes les combinaisons sont répertoriées. Celles-ci sont très riches si l'on se sert de Cordial, et bien moins variées si TreeTagger est mis en œuvre.

Choisir la combinaison souhaitée. Un clic sur une option sert alternativement à activer ou désactiver la sélection. Le programme interdit les choix incohérents. Une fois réalisée la sélection, cliquer sur CONTINUER pour la transmettre au traitement en cours. Les boutons bleus permettent d'atteindre d'un seul coup tous les éléments rangés sous leur bannière. Le bouton "Toutes Catégories" donne accès à l'ensemble des parties du discours, regroupées ou non.

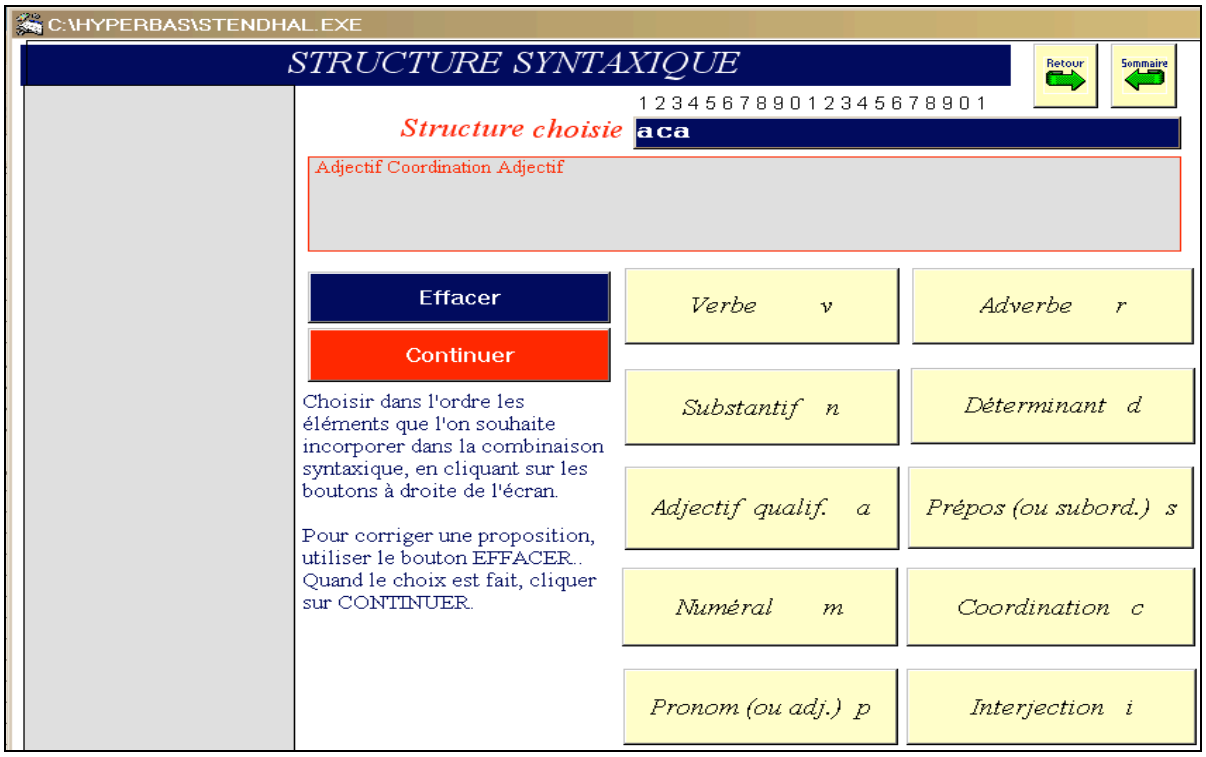
La table d'orientation grammaticale, selon Cordial. Le choix d'un code.



La table d'orientation grammaticale, selon TreeTagger. Le choix d'un code.



Le choix d'une structure syntaxique



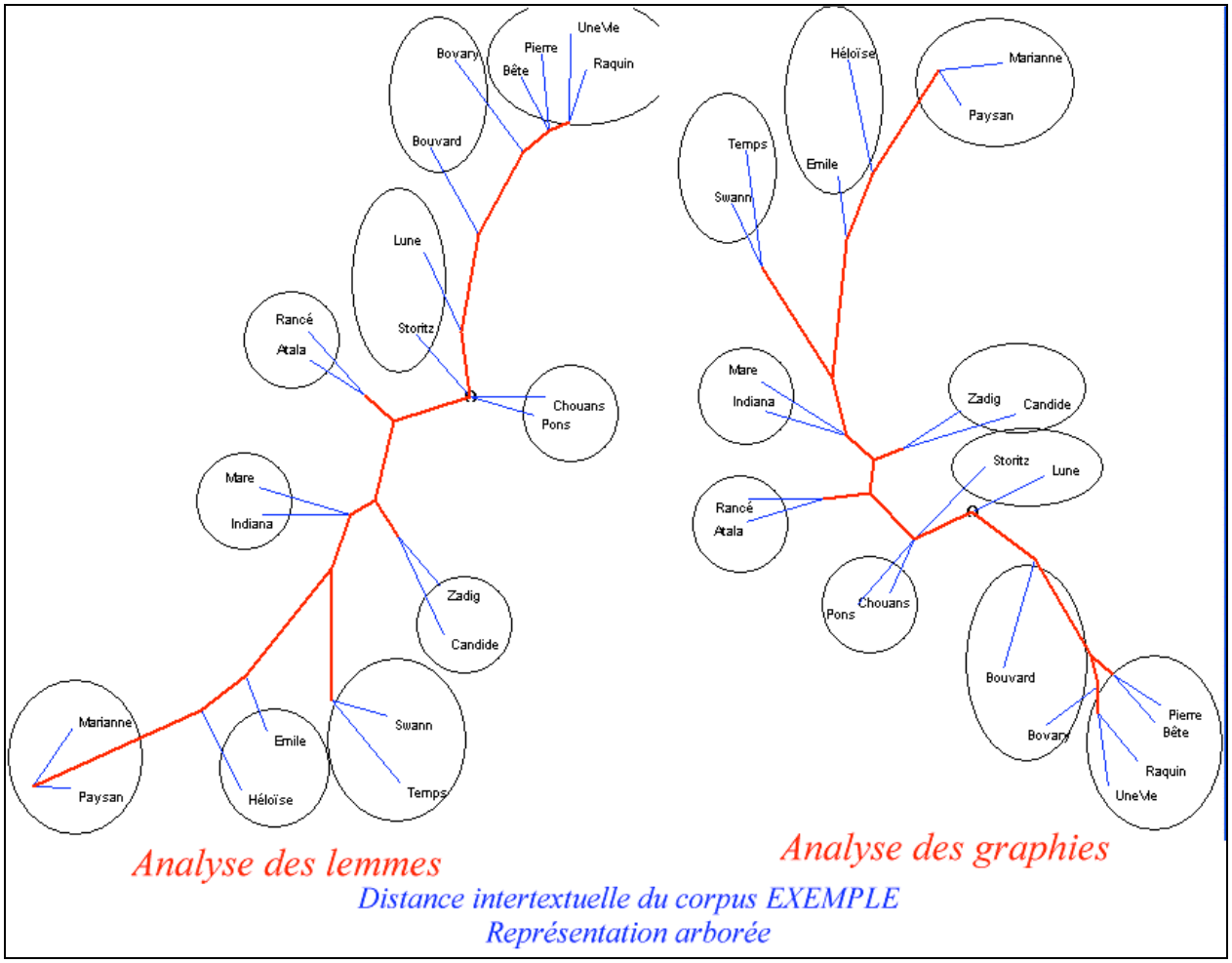
Le choix des structures syntaxiques se fait de la même façon en choisissant dans l'ordre les catégories qui forment la séquence souhaitée, par exemple adjectif + coordination + adjectif, comme dans l'exemple ci-dessus.

CHAPITRE 3.
EXTENSION DES RÉSULTATS
DANS LES TEXTES LEMMATISÉS

Tous les traitements de la version standard d'HYPERBASE sont repris dans les versions lemmatisées, qu'ils soient documentaires ou statistigues. Mais au lieu de s'appliquer seulement aux graphies, ils étendent leur portée aux lemmes, aux codes et aux structures.

LA DISTANCE INTERTEXTUELLE

Ainsi les 22 textes de notre corpus EXAMPLE se répartissent de la même façon lorsqu'est mesurée la distance entre leurs vocabulaires, lemmatisés ou non. La distance intertextuelle peut aussi être appréciée en dehors de toute influence thématique, en observant uniquement la distribution des codes grammaticaux ou des structures syntaxiques, indépendamment des mots auxquels ces codes ou structures sont attachés. Comme chacun des quatre niveaux d'observation peut donner lieu à un calcul fondé sur la fréquence (méthodes Muller et Labbé) ou sur la présence/absence (méthode Jaccard), on dispose en fin de compte de huit points de vue qui heureusement convergent. On préfèrera toutefois la méthode Muller s'il s'agit des codes car la variété y est limitée et les effectifs importants, et la méthode Jaccard pour la raison inverse s'il s'agit de structures.

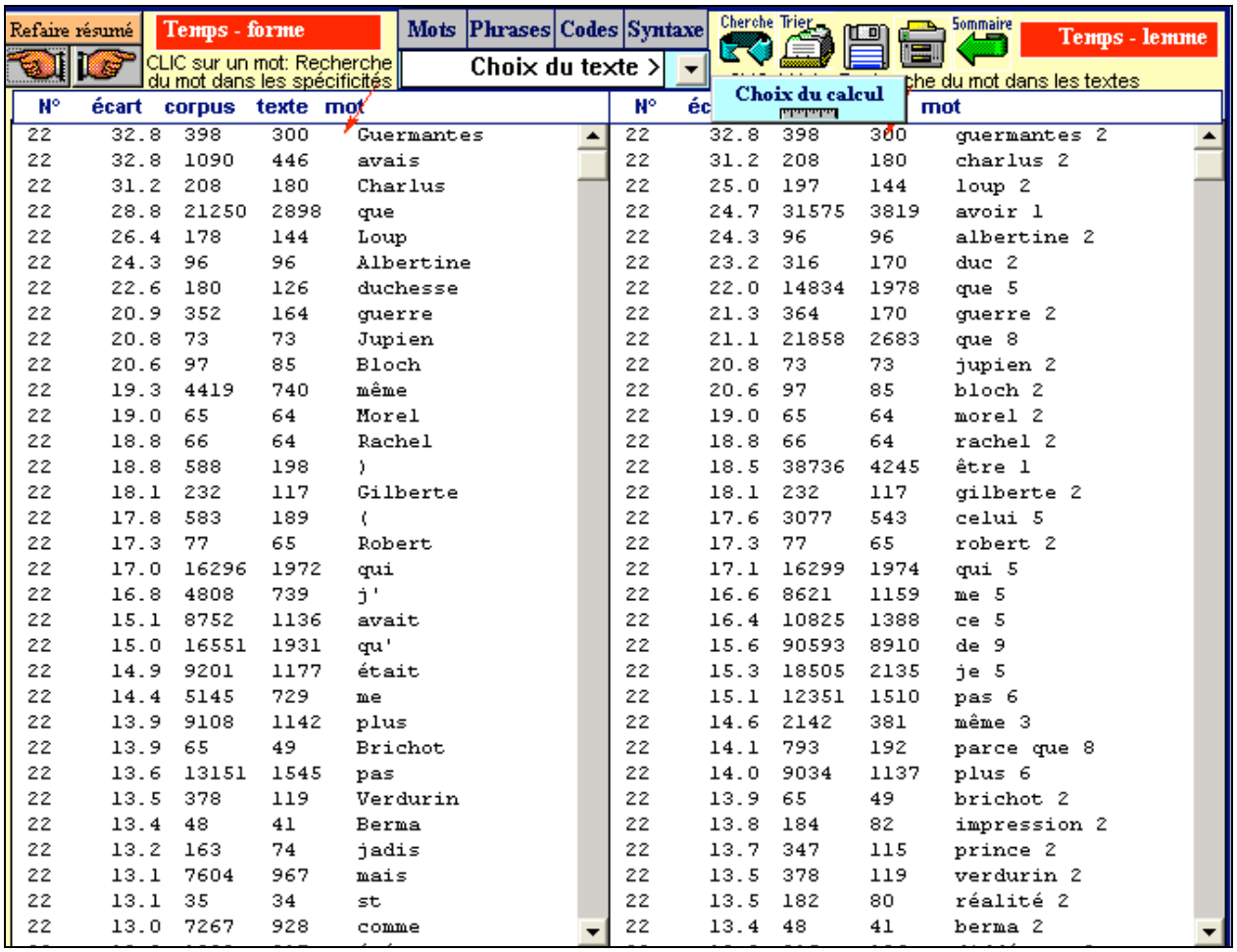


LES SPÉCIFICITÉS

De même les listes de spécificités obtenues pour chaque texte à partir des formes et des lemmes ont beaucoup d'éléments communs, quoique les informations qu'elles donnent pour les verbes soient beaucoup plus sûres si la lemmatisation a opéré le regroupement des formes fléchies. Les spécificités sont calculées non seulement pour les graphies et les lemmes, mais aussi pour les phrases caractéristiques, qu'elles soient constituées de formes ou de lemmes. Le calcul s'étend aussi aux codes et aux structures.

Le calcul des spécificités offre des variantes nouvelles, que le codage rend possible. S'il s'agit de lemmes, un tri peut s'exercer sur les parties du discours. Ces mêmes parties du discours peuvent même entrer dans le calcul. En isolant une catégorie et en neutralisant les autres, on peut calculer les spécificités, par exemple, d'un verbe en retenant pour référence les verbes du corpus, à l'exclusion de tout le reste. Le bouton *Choix du calcul* permet de réaliser de telles spécificités catégorielles. Il offre aussi la faculté de préférer la loi normale ou le calcul hypergéométrique. Enfin on a le loisir de repérer non seulement les spécificités positives (les excédents), mais aussi les spécificités négatives (les déficits).

Spécificités des graphies et des lemmes (et des phrases, des codes et des structures)



RÉFÉRENCE EXTÉRIEURE

S'il s'agit de textes en français, la comparaison se fait avec le corpus de FRANTEXT, comme on l'a expliqué page 79. Mais les fréquences n'y sont disponibles que pour les graphies, même si une grande partie du corpus a été lemmatisé. Au reste ces fréquences des lemmes n'auraient pas été très utiles pour la comparaison, car le programme qui les a créées n'est pas accessible au commun des mortels. Il a donc fallu constituer de toutes pièces un corpus lemmatisé de référence, en utilisant de façon constante le lemmatiseur dont l'usager d'Hyperbase est appelé à se servir. Comme TreeTagger est un produit courant et gratuit et qu'il a été utilisé dans la plupart de nos bases, nous avons pu constituer un dictionnaire des fréquences lemmatisées, suffisamment vaste et cohérent pour constituer une référence littéraire. Ce fichier, qui porte le nom de "modele2.txt", est établi sur une trentaine de monographies, de Corneille à Le Clézio. Gros de 60 millions d'occurrences, il offre, comme le fichier des graphies *modele.txt*, la possibilité de choisir une tranche de temps (siècle ou demi-siècle) pour adapter la comparaison au corpus à traiter. Au moment de la préparation, il y a donc deux confrontations successives à l'extérieur, l'une pour les graphies, l'autre pour les lemmes. Qu'on ne s'étonne pas si les résulats ne

coïncident pas absolument, même dans le cas où graphie et lemme sont identiques, comme pour le mot *temps*. La divergence peut venir des erreurs occasionnelles du lemmatiseur, mais elle est due plus souvent à la composition différente des deux corpus de référence ou au recouvrement imparfait des époques retenues.

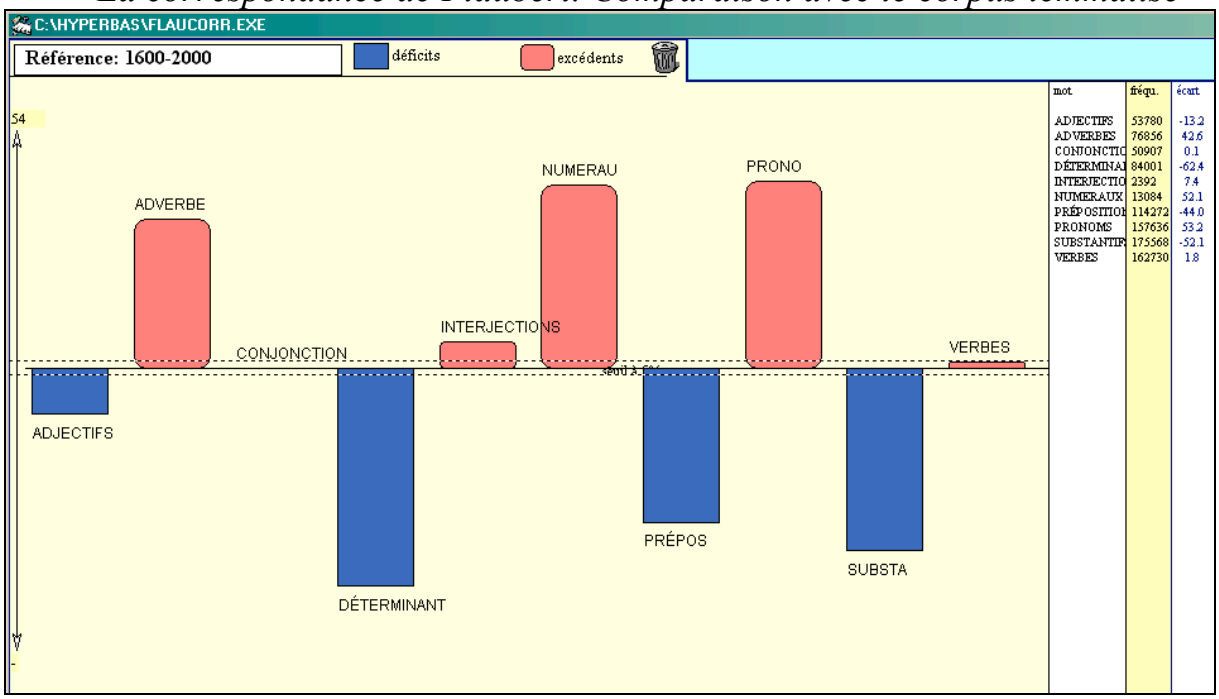
Comme Hyperbase s'accommode aussi du lemmatiseur Cordial et qu'il ne convient pas de mêler les résultats qui font appel à des lemmatiseurs différents, on a constitué aussi un fichier de référence avec des textes appartenant principalement au XXe siècle. A ce jour sa taille est de 40 millions d'occurrences. Il est réservé à la version HYPERCOR de notre logiciel.

Quant aux versions étrangères (HyperAng, HyperIta, HyperPor, HyperGer, et HypEspag) si les trois premières disposent d'une référence pour les graphies, aucune n'en propose pour les lemmes.

EXTENSION DE LA COMPARAISON EXTÉRIEURE AUX CODES

Les versions françaises (HyperTag, HyperVer et, dans un proche avenir, HyperCor) étendent aux codes grammaticaux les possibilités de comparaison extérieure (et aussi à d'autres mesures relatives à la longueur des mots ou aux classes de fréquence). Une fois qu'un tableau de graphies, de lemmes, de codes ou d'autres objets est établi dans la page LISTE, un bouton COMPARAISON EXTÉRIEURE est disponible qui donne lieu à un graphique analogue à celui qu'on présente ci-dessous. Noter qu'on peut mélanger les types d'objet et juxtaposer dans le même graphique des graphies, des lemmes et des codes ou parties du discours. Selon leur nature, les objets désignés font appel à trois dictionnaires différents (modele.txt, modele2.txt, modele3.txt).

La correspondance de Flaubert. Comparaison avec le corpus lemmatisé



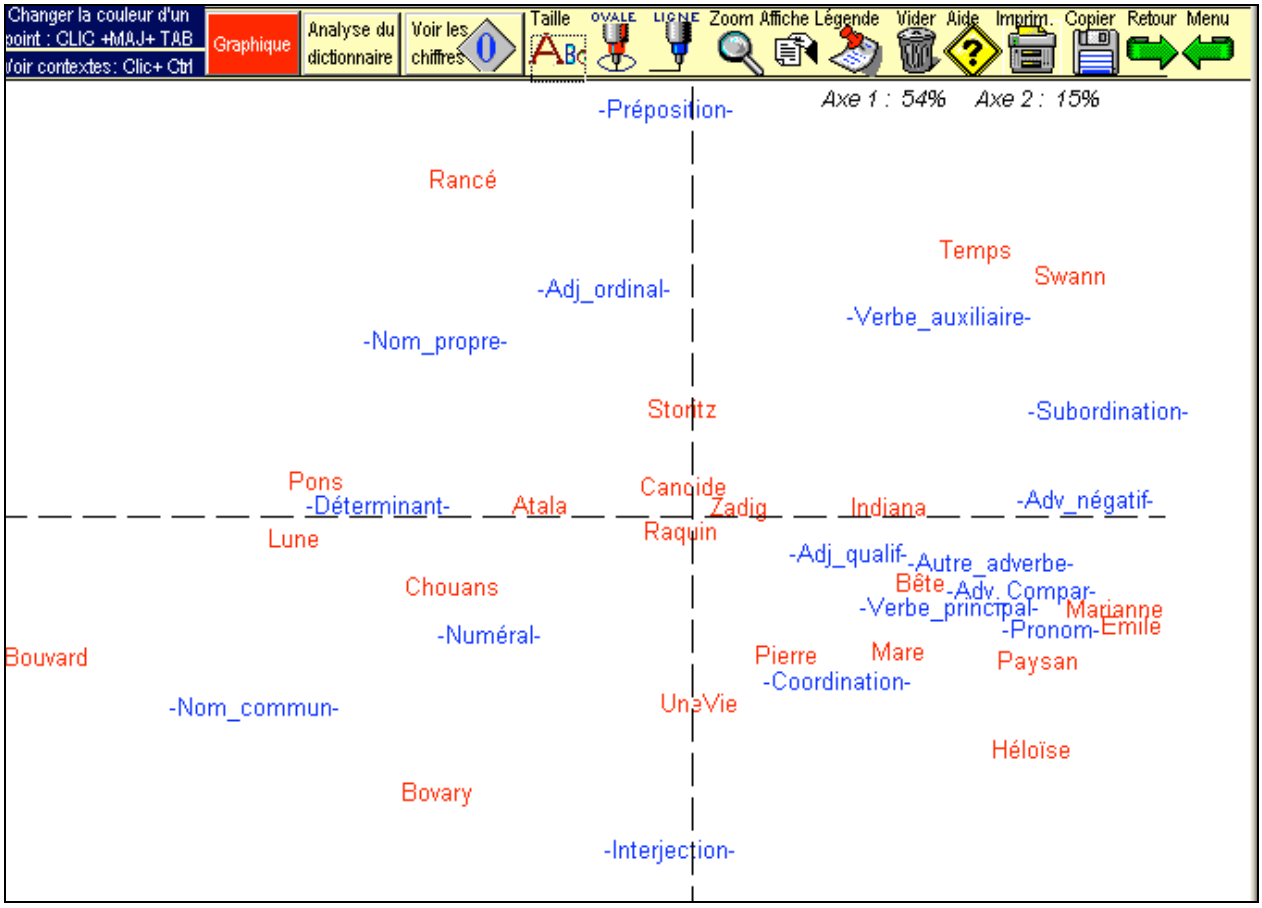
L'histogramme ci-dessus, consacré aux parties du discours dans la correspondance de Flaubert, est très différent de celui qu'on observe dans ses romans. Quand le premier privilégie les pronoms, les verbes et les adverbes, le second donne la faveur à la catégorie nominale. Ainsi peut-on comparer deux corpus en les superposant sur la même toile de fond.

ÉTUDE DIRECTE DES FAITS DE SYNTAXE ET DE STYLE

Dès qu'on aborde la syntaxe et les faits de style, l'étude des formes montre vite ses limites et ses faiblesses: l'approche biaisée des mots grammaticaux ne permet qu'une approximation timide, car beaucoup de mots-outils sont homographes et servent à plusieurs usages, comme les couteaux suisses. Sans lemmatisation préalable, bien des aspects du texte restent inaccessibles: le genre grammatical, le nombre, la fonction dans la phrase, le temps verbal, le mode , la personne, les parties du discours et leurs multiples combinaisons réalisées dans les syntagmes et les structures syntaxiques, des plus simples (bicodes ou tricodes) aux plus complexes, tout cela échappe aux formes brutes.

LES PARTIES DU DISCOURS

La table d'orientation grammaticale, représentée plus haut, offre mille possibilités de sélection. Le choix peut se rétrécir si l'on cumule les contraintes ou au contraire s'élargir si l'on fait appel aux boutons génériques (en bleu). Le plus général (*Catégorie 1* dans la version *Cordial*, *Toutes catégories* dans la version *TreeTagger*) regroupe tous les codes observés dans le corpus et les répartit dans des sous-ensembles correspondant aux parties du discours traditionnelles. Il en résulte un tableau de contingence où une dizaine de catégories se partagent tous les mots de chaque texte. D'un texte à l'autre les choix ne sont pas les mêmes et dans le corpus *Example* on observe ce qu'on a constaté dans beaucoup d'autres : l'opposition du verbe et du nom. L'analyse factorielle ci-dessous rend compte de ce clivage qui range dans le clan du verbe les pronoms, les adverbes et les subordonnants et dans l'autre camp les noms propres et communs, les déterminants, les numéraux et les prépositions. Habituellement les adjectifs rejoignent aussi la catégorie nominale, ce qui n'est pas le cas ici. Car l'adjectif est un élément moins stable, sensible aux modes, chéri par certains écrivains et chassé par d'autres. Or les textes et les écrivains prennent place aussi sur l'échiquier selon le goût qui les pousse dans un camp ou dans l'autre. Flaubert (quand il est romancier), Balzac, Chateaubriand et Verne sont du côté du nom, Rousseau, Marivaux, Proust et Sand prennent parti pour le verbe. Entre les deux, à cheval sur l'axe vertical, Voltaire hésite, comme aussi Maupassant et Zola.



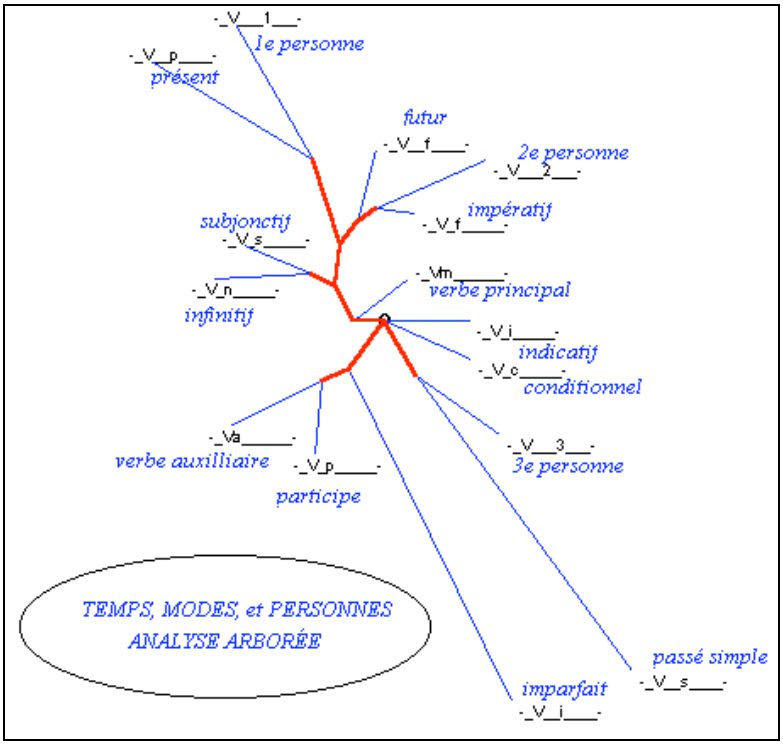
LE VERBE. TEMPS, MODES et PERSONNES (lemmatisation Cordial)

C'est le verbe qui gagne le plus à l'étiquetage: si l'on s'intéresse à son contenu sémantique, le regroupement des formes d'un même verbe offre un avantage indéniable, et si l'on étudie un temps verbal et qu'on regroupe les verbes qui partagent le même codage et le même temps, l'avantage est encore plus net. Et faute de pouvoir aborder les points de vue qu'on vient de passer en revue, c'est le système verbal qu'on voudrait étudier dans notre corpus, à titre d'exemple. Lorsqu'il s'agit de l'emploi des verbes, on a tout lieu de penser qu'un écrivain y porte attention. Le choix qu'il fait du passé ou du présent, de la première ou de la troisième personne, a des conséquences importantes pour la conduite du récit et une telle décision ne saurait être inconsciente. Le système verbal s'étageant sur plusieurs plans: le mode, le temps et la personne (sans compter le nombre, l'aspect et d'autres paramètres), on pourrait isoler successivement les trois plans principaux (les seuls que relève Cordial), ou bien les croiser et, par exemple, consacrer une ligne du tableau à la troisième personne du pluriel du présent de l'indicatif des verbes auxiliaires (croisement de 5 variables). Pour une première approche il nous a paru prudent de s'en tenir aux grandes divisions, sans croisement, mais sans exclusive. En étudiant ensemble, comme variables indépendantes, les modes, les temps et les personnes, on se donne le moyen de repérer lequel de ces trois paramètres est le plus discriminant, mais aussi quelle interaction s'exerce entre les uns et les autres.

Une première réponse est dans le graphique ci-dessous. On y voit que le mode n'est pas la pierre de touche qui puisse servir à classer les textes et les styles. Tous les modes restent groupés au centre du graphe, à peu de distance les uns des autres, à l'exception de l'impératif qui s'écarte vers le haut, ayant partie liée avec la deuxième personne, et du participe qui s'éloigne vers le bas en s'associant aux auxiliaires pour constituer les temps composés. Les personnes sont plus excentriques, et, comme elles sont trois, leur constellation prend la forme d'un y, la branche la plus longue étant le fait de la troisième, au bas du graphe, contre laquelle s'unissent les deux autres, au haut du graphe.

Mais la voix la plus forte appartient au temps; c'est elle qui impose sa loi au récit, en le sommant de choisir entre le présent et le passé. La tension la plus intense (sur le graphe la distance la plus longue) est en effet celle qui oppose le présent (en haut) à l'imparfait et au passé simple (en bas). Le futur accompagne le présent, tandis que les temps composés, principalement le passé composé, rejoignent l'imparfait. Les trois critères du verbes ne sont pas vraiment indépendantes. Si le mode maintient sa neutralité dans la partie engagée autour des temps (mis à part l'impératif et le participe), la personne exprime clairement ses préférences: la première pour le présent, la troisième pour le passé.

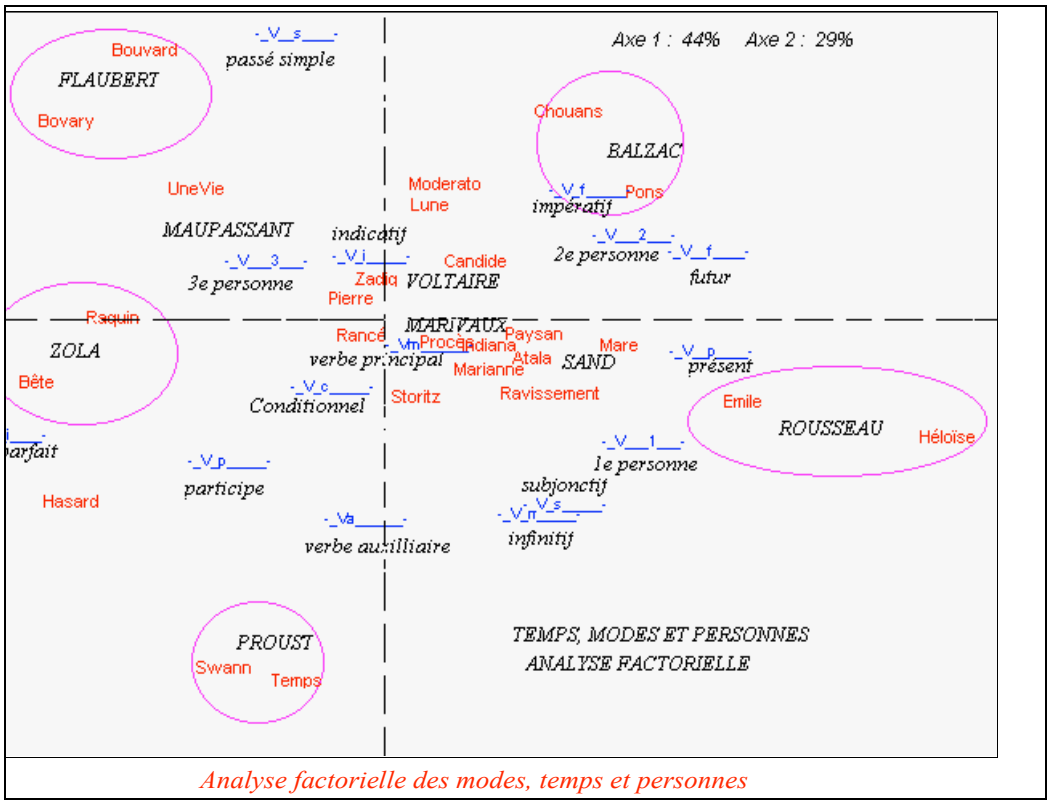
Analyse arborée des temps, des modes et des personnes



Les choix étant ainsi offerts, comment réagissent les textes? Comme nous l'avons observé à d'autres points de vue (distance lexicale, distance grammaticale, distance sémantique), ils vont par couples, les textes ayant des choix solidaires s'ils ont le même père. Cette fraternité est particulièrement étroite s'il s'agit de Marivaux, Rousseau, Voltaire, Sand, Balzac, Flaubert,

Maupassant, Zola ou Proust. Mais on distinguera, en les enveloppant dans un cercle sur le graphique ci-dessous, les couples qui affirment hautement leurs préférences et leurs exclusives communes, et ceux que le vote laisse indifférents et qui se rapprochent de l'origine des axes. Comme précédemment, Voltaire est parmi les abstentionnistes.

S'il n'y avait l'indécision de Verne, on pourrait admettre que la chronologie polarise les résultats: tous les écrivains antérieurs à 1850 se situent à droite, dans le présent, en compagnie des deux premières personnes. Sans doute la part du dialogue y est-elle plus importante. Mais ce n'est sans doute pas la seule raison. À gauche, dans la zone opposée, c'est, à partir de Flaubert, le règne de la troisième personne et du passé, particulièrement de l'imparfait que Proust admirait tant chez Flaubert. On peut être sensible à cette continuité stylistique qui prend naissance chez Flaubert et se maintient jusqu'au nouveau roman.



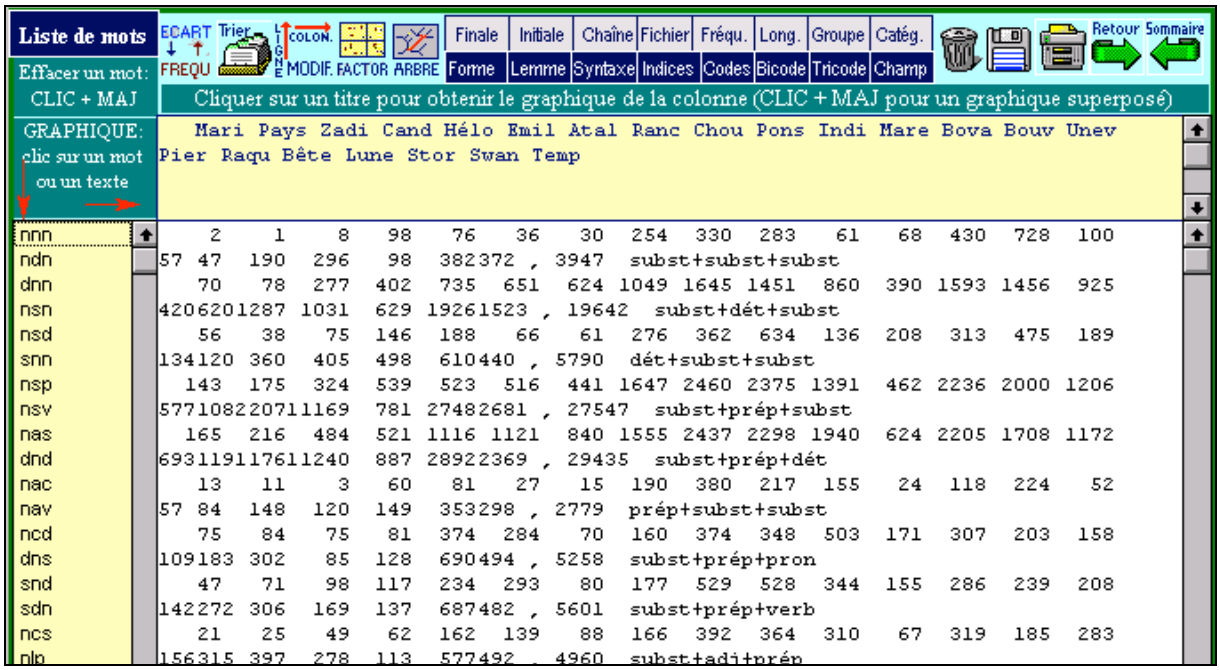
Analyse factorielle des modes, temps et personnes

LES STRUCTURES RÉCURRENTES. BICODES ET TRICODES

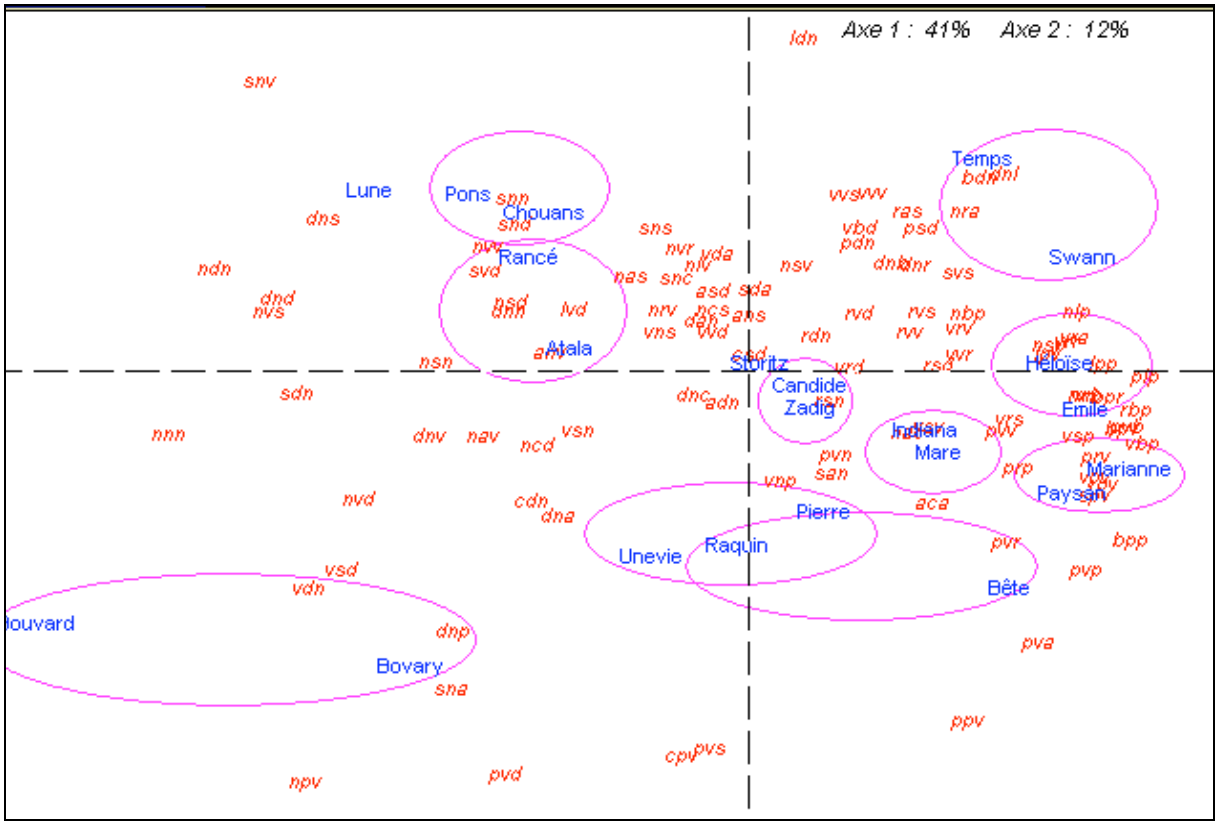
Il s'agit ici de structures syntaxiques limitées à deux éléments contigus (les bicodes) ou trois (les tricodes). Dans le premier cas la combinatoire reste limitée au carré du nombre de catégories grammaticales, soit 144 lignes dans le tableau. Le programme en fait un relevé exhaustif, ce qui permet par exemple de comparer la séquence *an* (adjectif antéposé au nom) et *na* (adjectif postposé) ou d'étudier l'ensemble des combinaisons dans une analyse multidimensionnelle. Dans le second cas, le programme ne retient que les séquences les plus fréquentes sur un total de plus d'un millier, dont certaines d'ailleurs ne sont pas

représentées, comme étant agrammaticales. Il est aisé de décrypter la composition des tricodes, si l'on connaît le code des éléments simples, à savoir : *n* = substantif, *v* = verbe, *a* = adjectif, *r* = adverbe, *s* = préposition, *p* = pronom, *d* = déterminant, *l* = relatif, *m* = numéral, *c* = coordination, *b* = subordination, *i* = interjection. Le relevé de quelques tricodes est présenté ci-dessous et leur analyse dans l'écran suivant.

La page LISTE. Relevé des "tricodes"



Analyse factorielle des tricodes

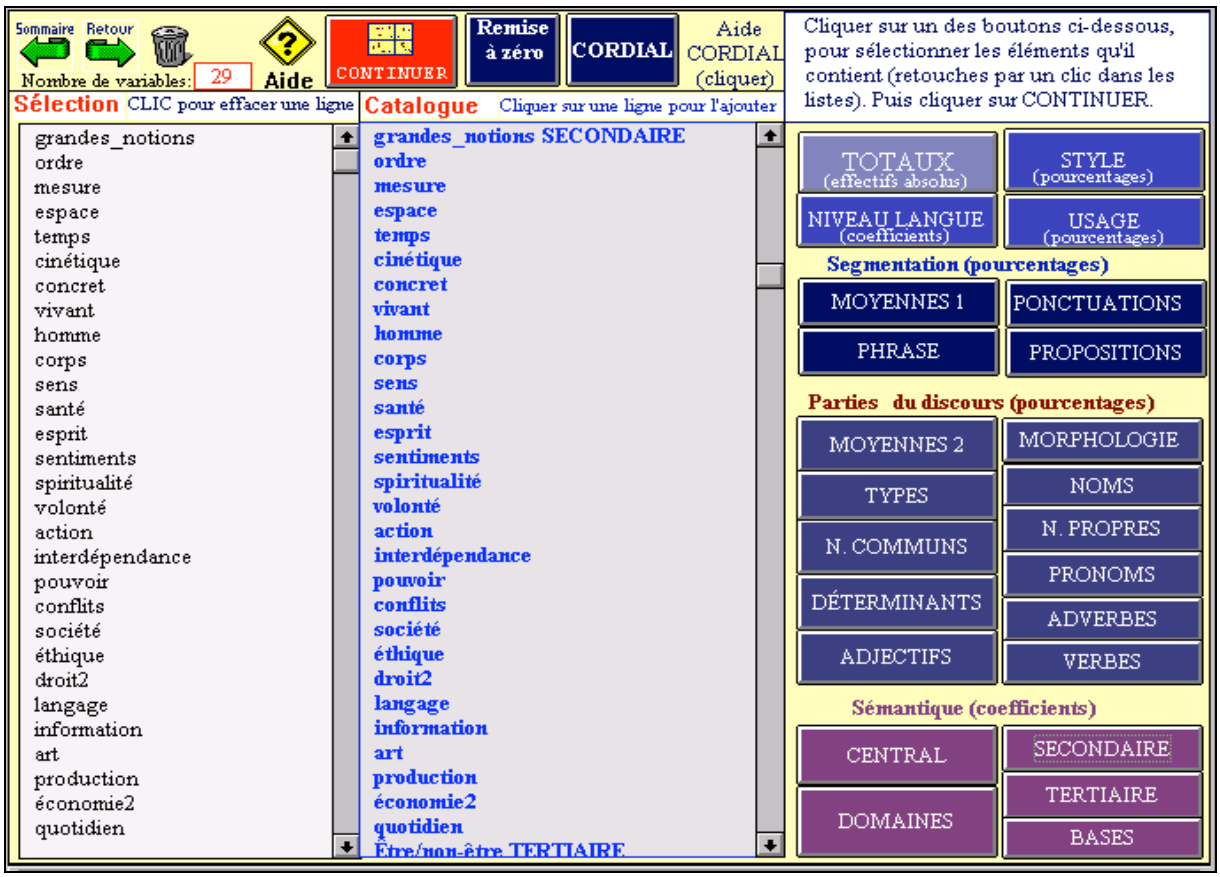


Le résultat de l'analyse n'est pas sans rappeler celui qu'on a obtenu précédemment à partir des catégories prises isolément ou assemblées deux à deux. S'il n'y a pas recouvrement intégral dans la figure ci-dessus (en particulier Flaubert n'a pas tout à fait les mêmes alliés sur la partie gauche), nombreux sont les textes qui font le même choix dans les différentes analyses. Et les textes d'un même auteur, ici comme ailleurs, sont toujours proches l'un de l'autre.

LA SÉMANTIQUE (lemmatisation Cordial)

Non content d'offrir une information grammaticale très riche, *Cordial* fournit encore des statistiques, dont certaines abordent le domaine de la sémantique et de la thématique. Il faut toutefois faire appel une nouvelle fois au programme ANALYSEUR, en sollicitant pour chaque texte l'item SAUVER LES STATISTIQUES COMPLÈTES DU TEXTE du menu SYNTAXE. Hyperbase lance le programme autant de fois nécessaire, en enregistrant les résultats au fur et à mesure (1345 indices pour chaque texte). Un regroupement est assuré en fin de parcours, afin de constituer un immense tableau qui comporte 1345 lignes et autant de colonnes que l'on compte de textes dans le corpus. Lorsque ces indices sont des pourcentages, Hyperbase les convertit en effectifs absolus puis en écarts réduits. Lorsqu'on a affaire à des coefficients, la transformation probabiliste est impossible et les données sont traitées directement par l'analyse factorielle.

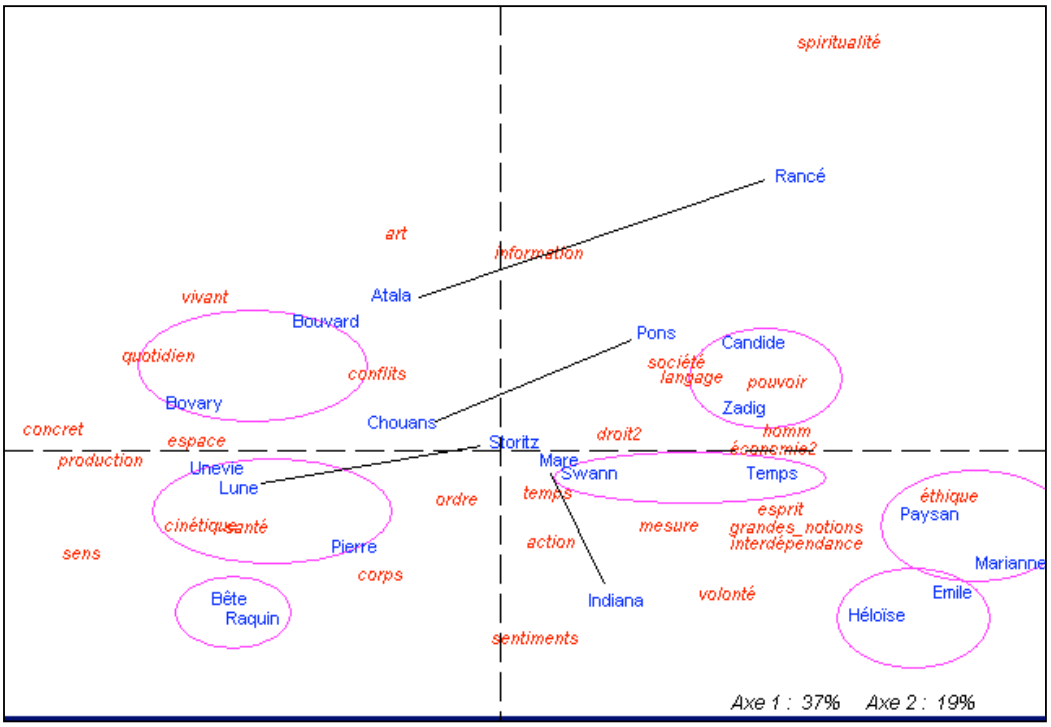
La page INDICES issue de CORDIAL



Une page spéciale (reproduite ci-dessus) donne accès à ces indices et en assure la gestion, par séries homogènes, de type grammatical, rythmique ou sémantique[17].

Le résultat de l'analyse thématique confirme ce que nous avait enseigné l'analyse des formes verbales. On observe la même dichotomie: d'un côté Flaubert, Maupassant et Zola mettent en scène le "milieu", les réalités concrètes dans lesquelles évoluent les personnages (*vivant, concret, quotidien, santé, corps, sens, mouvement, espace*). De l'autre l'attention est portée aux phénomènes psychologiques ou moraux qui touchent à l'âme ou à la société (*éthique, spiritualité, esprit, volonté, langage, société, pouvoir, droit*). Les auteurs du XVIIIe siècle se positionnent ici, mais aussi Sand et Proust, tandis que Chateaubriand, Balzac et Verne restent partagés.

Analyse factorielle des thèmes relevés par Cordial



En conclusion, une enquête sur la population des mots jouit de gros avantages si l'on a affaire à un état policé où les individus ont été recensés et possèdent une carte d'identité. C'est le cas des lemmes. L'étude prend l'aspect alors d'une recherche sociologique. En croisant la fonction, la catégorie, le temps , le genre, le nombre, etc., on peut suivre la même démarche que les autres sciences humaines, qui mettent en relation, à partir de leurs observations,

[17] La pose d'un code sémantique se heurte à une difficulté majeure: la polysémie. Les auteurs de Cordial, sensibles aux critiques faites à ce sujet, ont fait marche arrière dans la version 14. Les 1345 indices ont été maintenus pour des raisons de compatibilité avec les versions précédentes, mais un grand nombre de ces indices n'ont pas été évalués et présentent une valeur nulle.

la catégorie socioprofessionnelle, l'âge, le salaire, les opinions politiques, le niveau culturel, la mortalité, la fécondité, etc. Certains pourront regretter les formes brutes, dont la matérialité opaque pouvait receler quelque mystère, et renâcler devant un lemme blême, vidé de son sang, et réduit à un ensemble de traits abstraits. En réalité cette frustration, qui est bien réelle dans les sciences humaines (le sociologue qui dépouille une enquête n'a guère le moyen d'approcher les gens), n'a pas lieu lorsqu'on a affaire à des textes. Le texte est toujours présent dans sa matérialité originelle et tous les chemins de traverse sont des chemins de retour. En fin de compte les clignotants statistiques ramènent au texte en soulignant discrètement les faits saillants. On pouvait s'inquiéter de la pertinence des résultats, quand seules les formes graphiques étaient soumises au traitement. Mais lorsqu'on met en parallèle la forme, le lemme, l'analyse grammaticale, la structure syntaxique et le code sémantique, et qu'on obtient la convergence, on se rend compte que la voie empruntée jadis par les pionniers de la lemmatisation est appelée à devenir la voie royale dans l'étude des textes. Il reste seulement à élargir la route, à adoucir les angles, à supprimer les péages, à humaniser la technique, et...à convaincre les usagers.

CHAPITRE 4.
LES COOCCURRENCES. LA PROXÉMIE

Rappelons que, selon une problématique initiée par Pierre Lafon dans sa thèse, les séquences peuvent être opposées aux fréquences. En réalité l'étude des séquences et des cooccurrences est un large boulevard emprunté depuis longtemps par les chercheurs qui s'intéressent moins à la lexicométrie qu'au traitement automatique du langage. Dès que la perspective s'ouvre vers la documentation, le data mining, les systèmes-experts, la traduction, le résumé, on sort du cadre étroit où s'enferme la lexicométrie, condamnée à comparer les textes à l'intérieur d'un corpus et à manipuler des fréquences. Les textes sont certes des séquences, mais, au moins dans le domaine littéraire, leur empan est trop large pour que la cooccurrence de deux mots ou objets linguistiques ait un sens précis si elle est observée à longue distance (même si parfois le texte littéraire contient des rappels et des échos qui se répercutent de loin en loin). Les séquences impliquent une segmentation courte, qui est celle, non des textes, mais des phrases, des paragraphes, voire des pages ou des fenêtres, glissantes ou successives, de *n* mots. Le voisinage, immédiat ou proche, de deux mots, y prend une signification qui échappe au hasard, qu'il s'agisse d'une contrainte syntaxique, d'une aimantation sémantique, d'une convenance prosodique ou de quelque autre raison attachée à la situation ou à la langue.

Rien n'empêche d'ailleurs, une fois que les relevés ont été faits dans les séquences, de les projeter dans l'espace partitionné du corpus où l'on retrouve la division en textes. Les méthodes traditionnelles de la lexicométrie reprennent alors leurs droits et leur matériau de base, les fréquences, mais elles ne s'appliquent plus à des mots ou à des codes individuels mais à des combinaisons des uns et/ou des autres, à des profils, à des faits observés dans les séquences. Au fond il ne s'agirait là que d'une extension, d'un perfectionnement de la

lexicométrie, qu'on a attendu trop longtemps parce que les traitements préalables de désambiguïsation et de lemmatisation n'étaient guère disponibles, sinon de façon manuelle et limitée. Maintenant encore le traitement des séquences reste trop artisanal et un long chemin reste à faire pour accéder à la standardisation, même si on voit la trace à suivre qui est celle du codage XML et des filtres sophistiqués d'interrogation, à base de grammaires spécialisées et d'*expressions régulières*[18] .

Sans aller jusqu'à construire une telle grammaire, nous proposons ici une démarche exploratoire où cinq fonctions THEME, CORRÉLATS, ASSOCIATIONS, ALCESTE et TOPOLOGIE relèvent de la même approche, orientée vers l'étude des séquences plutôt que des fréquences. On y considère les mots (ou d'autres objets) dans leur environnement immédiat (paragraphes ou pages) en ignorant la partition en textes.

1 - La fonction TOPOLOGIE représente la distribution, aléatoire ou non, d'un ou de deux objets dans l'espace du corpus, et, le cas échéant, mesure la distance entre les deux distributions. Ce point, déjà développé dans la version standard d'HYPERBASE, ne sera pas repris ici. Mais on suppose acquis le calcul exposé, qui relève de la loi hypergéométrique.[19]

2 - La fonction THEME recense et assemble tous les passages où un mot (ou autre objet) est rencontré dans le corpus et oppose ces passages au reste du corpus. Il en résulte une liste de spécificités associée à l'objet de la recherche, graphie ou lemme. Ces mots associés au mot-pôle peuvent avoir entre eux des liaisons qui sont explorées, phrase après phrase dans le texte. Il en résulte un tableau de cooccurrences, représenté dans un graphe. Cette fonction, ne

---

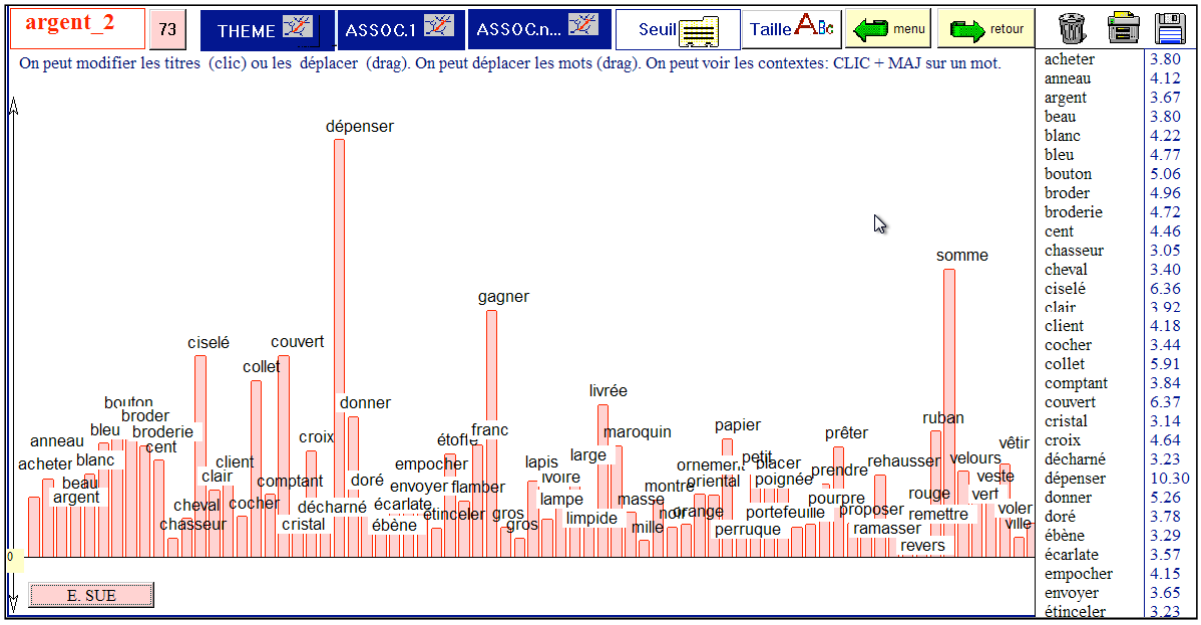
[18] L'exemple le plus achevé d'un traitement des séquences est celui de *Frantext*. Le logiciel de consultation *Stella* propose non seulement les opérateurs booléens, les jokers et les expressions régulières, mais aussi une grammaire évoluée qui mêle constantes et variables et rend le filtrage aussi fin et aussi souple que l'on veut. Dommage que les fonctions statistiques de *Frantext* ne soient pas à la hauteur des fonctions documentaires et ne s'appliquent pas à des relevés aussi finement établis.

[19] On avait expérimenté deux autres indices, *le Rapport de Vraisemblance* de Dunning et *l'Information mutuelle* de Church, tous les deux issus de la formule de Jaccard et utilisant les mêmes ingrédients - a : nombre de cooccurrences des deux mots dans le champ exploré (ici le paragraphe - b : nombre d'occurrences du premier mot en l'absence du second - c : nombre d'occurrences du second mot en l'absence du premier - d : nombre d'occurrences des autres mots. Or la convergence n'est pas observée dans toute l'échelle des valeurs On s'en est donc tenu à la méthode hypergéométrique qui n'est pas la plus économique mais qui reste la plus sûre. Il est en effet possible d'obtenir directement la probabilité de la coccurrence observée, grâce à une itération du calcul hypergéométrique pour cumuler les probabilités partielles de 1 à k (k étant la valeur observée). Voir le détail de ce calcul dans une communication de Serge Heiden *Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex*, publiée dans *JADT04, Le poids des mots*, p 578-588. Ajoutons que cet article apporte une contribution fondamentale au traitement des cooccurrences. On y trouvera en particulier sous le nom de lexicogramme une représentation graphique très séduisante des réseaux lexicaux.

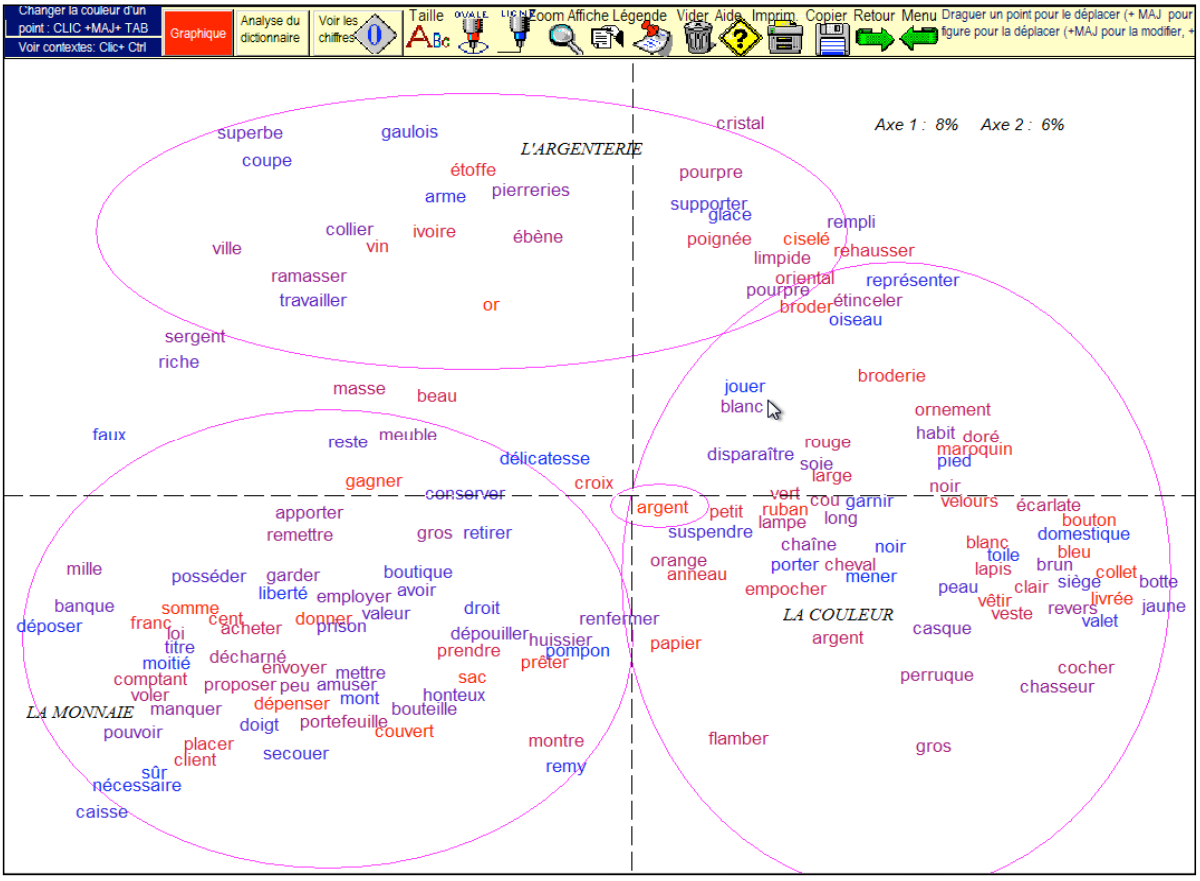
.

présupposant pas des données lemmatisées, est aussi traitée dans la version standard. Mais elle donne des résultats plus fins et plus solides si elle s'applique à des lemmes plutôt qu'à des graphies, comme le montre l'exemple du mot *argent* dans le corpus *Eugène Sue.* Livrons à la machine deux éléments : le sous-corpus bâti autour d'un mot par la fonction CONTEXTE (c'est le fichier DATA1.txt, quel que soit le nom de la base) et la liste des spécificités qui en est extraite par la fonction THEME. Le traitement va remplir le tableau des cooccurrences des mots de cette liste et le livrer sans autre procès au calcul factoriel. Le résultat, pour le même mot *argent* dans le même corpus, est dans les deux figures ci-dessous, la première étant un histogramme des spécificités qui entourent l'*argent*, l'autre une analyse factorielle des cooccurrents de *l'argent.*

Visiblement l'argent n'est pas rare dans les feuilletons d'Eugène Sue, surtout dans les *Mystères de Paris*. Il est vrai que c'est avec l'amour l'ingrédient principal de l'intrigue romanesque, de Balzac à Zola. Mais les 458 occurrences de l'argent ne brillent pas du même éclat dans tous les contextes. Tantôt il s'agit de la monnaie et cette acception ordinaire s'établit dans la partie basse et gauche de l'analyse factorielle. Tantôt il s'agit du métal et l'argenterie miroite au haut du graphique. Tantôt il s'agit de la couleur, qui illumine le flanc droit. On notera que c'est souvent le vêtement qu'on qualifie ainsi et que la transition entre le métal et la couleur passe par la broderie où le fil du métal donne à l'étoffe certaines de ses propriétés.



*Histogramme des corrélats de l'*argent *dans le corpus* Eugène Sue



*Analyse factorielle des mots attirés par l'*argent *dans le corpus* Eugène Sue

3 - La fonction CORRÉLATS regroupe les substantifs ou les mots sémantiques qui sont les plus fréquents dans le corpus et établit la carte synthétique de leurs cooccurrences (par une analyse factorielle de correspondance). Le programme commence par établir une liste de mots (au moins les substantifs, adjectifs ou verbes qui sont les plus fréquents) et enregistre toutes leurs rencontres, occasionnelles ou insistantes, dans le même paragraphe. Un lien est établi entre deux mots quand ils ont tendance à se donner rendez-vous. La "tendance" tient compte du nombre de cooccurrences (compte tenu de la fréquence respective des deux mots). Le registre est tenu dans un tableau carré où les mêmes éléments sont portés sur les lignes et les colonnes.

L'option en faveur du paragraphe ne permet pas d'échapper totalement aux contraintes syntaxiques mais l'élimination des mots fréquents et des mots-outils concourt à privilégier les relations sémantiques ou thématiques plutôt que les rapports de dépendance grammaticale. On notera que la division en textes est ignorée. La cohabitation à longue distance dans un même texte n'entre pas dans le calcul. Seule compte la proximité immédiate dans la même page, là où l'on a le plus de chances de relever les isotopies.

Le choix des termes est enclenché par le bouton *Prépare* de la page *Corrélats*. Compte tenu de l'étendue du corpus, le programme de sélection s'arrange pour retenir entre 300 et 400 candidats parmi les mots-pleins

(substantifs, adjectifs et verbes, ensemble ou séparément)[20]. Ensuite vient une phase, assez longue, d'exploration séquentielle du corpus. Dans chaque paragraphe on teste la présence ou l'absence des éléments de la liste, en notant les cooccurrences. Le tableau final est soumis à un programme d'analyse factorielle de correspondance, plus puissant que celui qu'Hyperbase utilisait jusqu'ici (ANCORR.EXE). Écrit en fortran par Ludovic Lebart dans les années 70, le programme LX2ACL.EXE n'est pas limité comme le précédent à 75 colonnes. On a fixé à 400 le seuil supérieur mais on pourrait doubler ce chiffre, n'était la nécessité de rendre lisible le résultat. Celui qu'on obtient pour notre corpus de Lamartine en sollicitant le bouton *Graphique,* est déjà suffisamment encombré (figure ci-dessous). L'interprétation en est pourtant assez claire : de la la droite à la gauche on passe du concret à l'abstrait. Quant au deuxième facteur, il paraît séparer ce qui relève de la société, en bas, et ce qui appartient à l'individu, en haut.

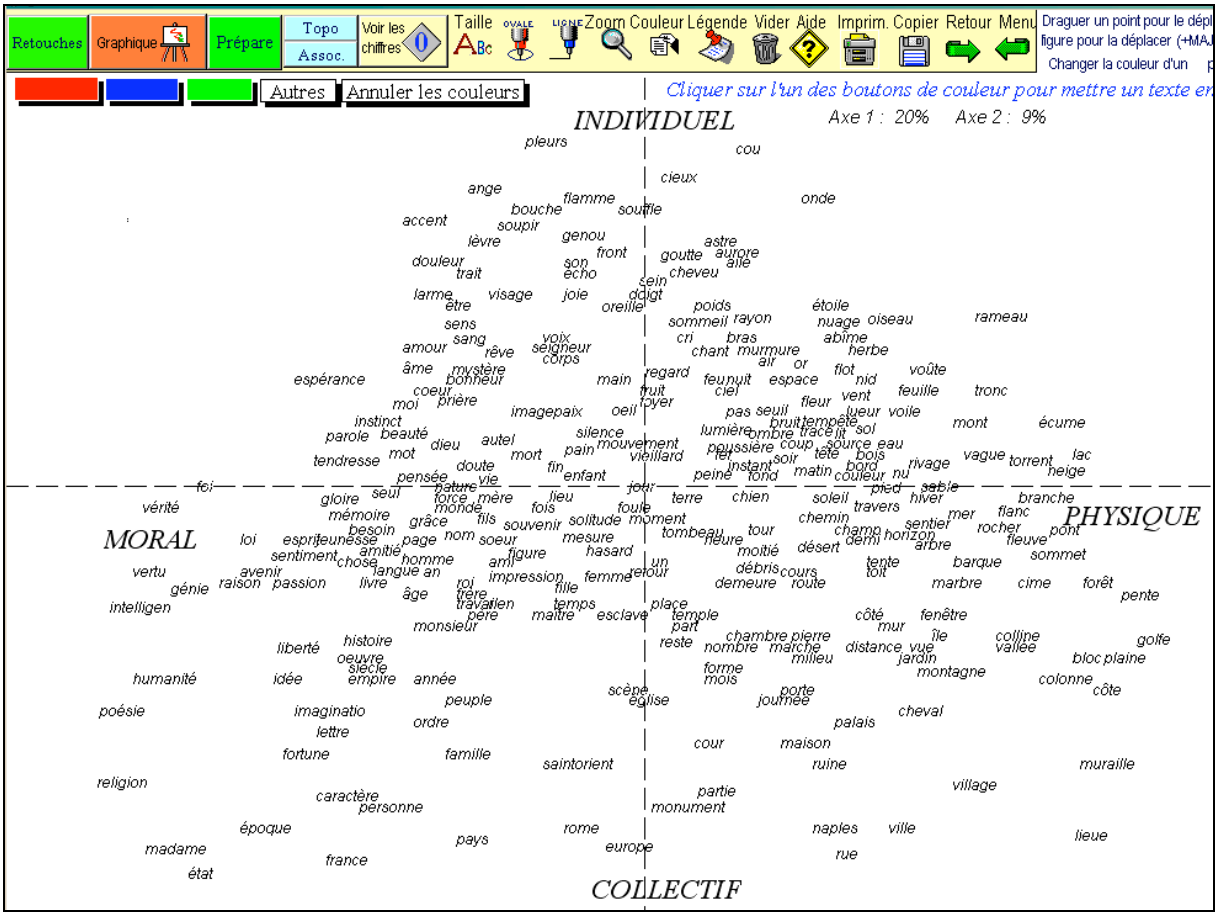


Figure 1 . Analyse factorielle des corrélats dans le corpus Lamartine

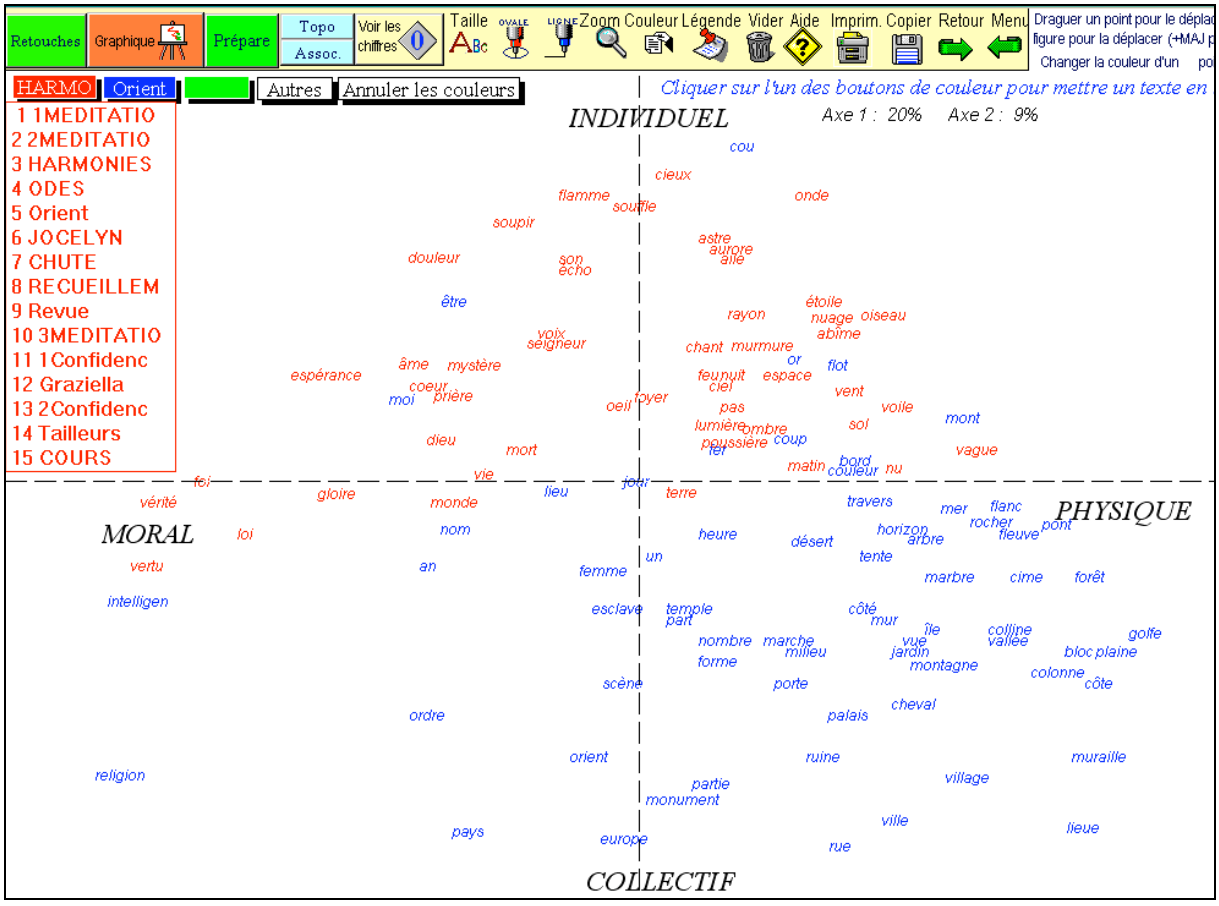
Chose curieuse, cette structure se retrouve dans des monographies plus cohérentes, comme celles de Flaubert, de Stendhal, de Zola, de Proust, de Gracq. Sans doute doit-on y voir non pas le partage des mêmes thèmes mais un reflet de

[20] La liste une fois établie reste modifiable. On peut y supprimer les indésirables. Même lorsque les calculs ont été exécutés, il est possible de les reprendre, en neutralisant soit un seul élément, soit une ligne entière (ou colonne) du tableau.

la composition romanesque, qui dispense dans le même roman des développements narratifs, descriptifs, dialogués ou réflexifs et qui s'impose d'un auteur à l'autre. Cette spécialisation de l'écriture, dictée par la situation où la plume intervient dans le cours de la rédaction, agit comme la loi du genre, mais à l'échelle de la microstructure. L'exploitation du tableau généralisé des cooccurrences ne se réduit pas à cette synthèse rapide. Il sert de base à des représentations moins synthétiques mais plus fines que nous exposerons plus loin.

Remarquons que les textes du corpus n'ont pas leur mot à dire dans cette carte des thèmes, générée par le vote individuel des paragraphes, libre de toute obédience. Et pourtant ces votes subissent sourdement l'influence du milieu, et les textes du corpus peuvent se reconnaître à tel ou tel endroit de la carte. En sollicitant l'un des boutons de couleur au haut de l'écran on peut mettre en relief les préférences d'un texte parmi les mots offerts à l'étalage. La couleur choisie pour ce texte est communiquée aux éléments qui appartiennent au vocabulaire spécifique de ce texte. Ainsi en est-il du rouge qui est attribué au recueil des *Harmonies* et qui colore la partie haute du graphique, et du bleu qui représente les choix *du Voyage en Orient*, circonscrits dans la moitié inférieure. Même si les divisions larges du corpus (les textes) n'entrent à aucun moment dans le calcul, une place précise leur est accordée, a posteriori, quand les jeux sont faits. C'est précisément la méthode choisie par le logiciel *Alceste* que nous nous proposons d'explorer.

Figure 1 bis



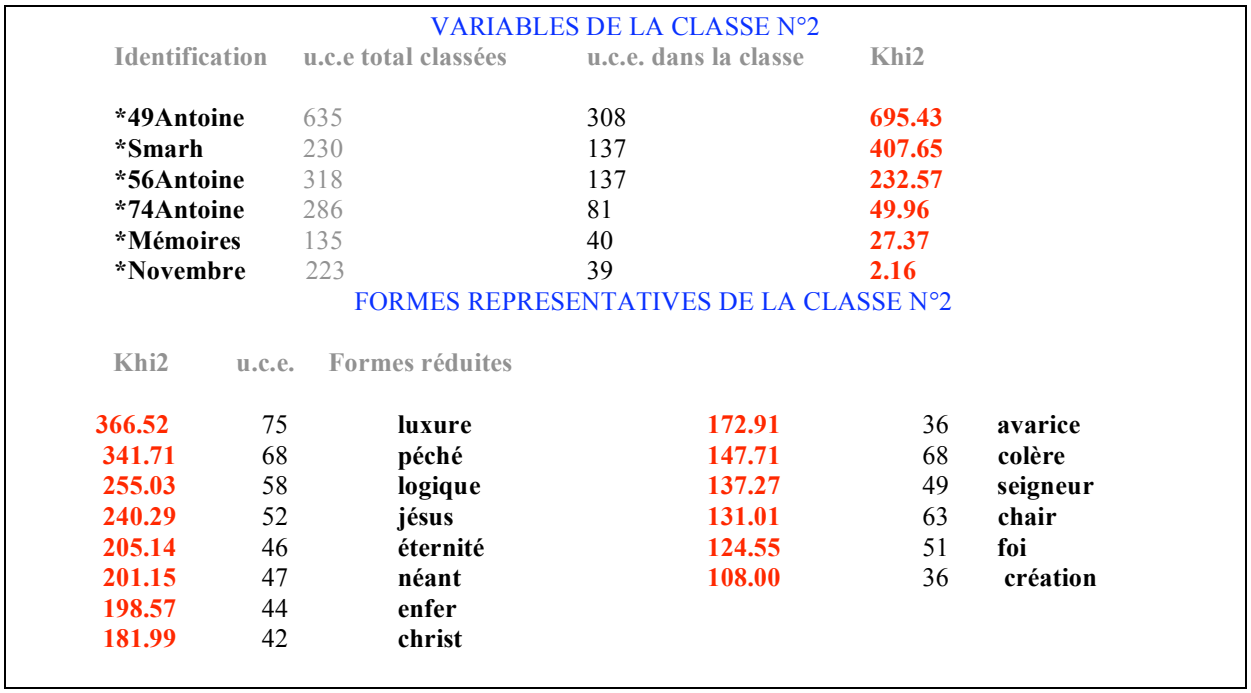
4 - La fonction ALCESTE établit un pont avec le logiciel ALCESTE. Elle lui fournit les données convenablement formatées, en lui transmettant la liste des substantifs les plus fréquents qu'on trouve associés dans un contexte étroit, paragraphe après paragraphe. Après traitement elle en reçoit les résultats sous forme de « classes ». La procédure CORRÉLATS qu'on vient d'exposer donne un avant-goût de ce que réalise *Alceste*. Le point de départ est le même : un réseau de mots associés. Mais dans *Alceste* la notion de cooccurrence est en principe plus étroite, puisqu'elle s'exerce dans des unités plus courtes de 2 ou 3 lignes, et non à l'échelle du paragraphe ou de la page, du moins lorsque les données fournies sont des textes suivis. De plus le calcul ne porte pas sur un échantillon de 400 mots-pleins, mais sur l'ensemble du vocabulaire. Enfin les résultats sont décantés par des filtres discriminants qui séparent les classes et les thèmes, au lieu que notre programme de *Corrélats* présente les alliances et les oppositions en une chaîne continue où les thèmes se succèdent en fondu-enchaîné.

Aussi bien avons-nous jeté un pont vers *Alceste*, sans pouvoir, hélas, fournir ce logiciel qui est un produit du commerce. Ceux qui le possèdent n'auront pas à se soucier de préparer les données. *Hyperbase* s'en charge si l'on actionne le bouton *Préparation des données*. Comme précédemment un seuil minimal et maximal est fixé selon la taille du corpus pour constituer un échantillon d'un millier de substantifs[21]. Et de la même façon chacune des pages est explorée et réduite à une suite d'une dizaine de mots, qui appartiennent à la liste préétablie et qui sont présents dans la page considérée. *Alceste* considèrera ces extraits comme des *unités de contexte élémentaires*, sur lesquelles s'exerce son algorithme quand l'ordre de *lancer Alceste* est donné. Précisons que le paramétrage est le plus simple et qu'il n'y a pas lieu de cocher la case relative à la lemmatisation, puisque les données sont déjà lemmatisées. Dès lors l'utilisateur quitte momentanément Hyperbase et peut à loisir recueillir et commenter les résultats produits par *Alceste,* comme nous l'avons fait pour les 5000 pages de l'oeuvre romanesque de Flaubert .

Huit classes ont été distinguées, auxquelles on doit donner un nom qui les résume au mieux, comme on fait pour les facteurs d'une analyse factorielle. Mais la liste des mots qui constitue une classe est suffisamment suggestive pour expliciter la classe (ou le thème), d'autant que le programme délivre une indication précieuse : les textes où le thème est exploité. Qu'il s'agisse des textes ou des mots, un Chi2 mesure l'appartenance plus ou moins étroite à la classe en question. Dans l'exemple de la figure 2, un extrait, même court, de la liste suffit à isoler les questions philosophiques et religieuses qui préoccupent

[21] On voit que la limite de 400 éléments a été reculée par rapport au programme *Corrélats*. Cela tient au fait qu'il n'est plus nécessaire de représenter les résultats dans une figure unique, où mille mots ne peuvent pas trouver place sans nuire à la lisibilité.

Flaubert dans ses premiers écrits et qui se maintiennent dans les trois versions de la *Tentation de Saint Antoine*.



VARIABLES DE LA CLASSE N°2

Identification	u.c.e total classées	u.c.e. dans la classe	Khi2
*49Antoine	635	308	695.43
*Smarh	230	137	407.65
*56Antoine	318	137	232.57
*74Antoine	286	81	49.96
*Mémoires	135	40	27.37
*Novembre	223	39	2.16

FORMES REPRESENTATIVES DE LA CLASSE N°2

Khi2	u.c.e.	Formes réduites	Khi2	u.c.e.	Formes réduites
366.52	75	luxure	172.91	36	avarice
341.71	68	péché	147.71	68	colère
255.03	58	logique	137.27	49	seigneur
240.29	52	jésus	131.01	63	chair
205.14	46	éternité	124.55	51	foi
201.15	47	néant	108.00	36	création
198.57	44	enfer			
181.99	42	christ			

Figure 2. Une classe isolée par Alceste *dans le corpus de Flaubert*

Les résultats sont distribués par *Alceste* dans une multitude de fichiers où l'utilisateur peut se référer en différé. Il en est un qu'*Hyperbase* rapatrie plus particulièrement : c'est le résumé de l'analyse, qui détaille le contenu des classes et dresse la carte des thèmes en y incorporant les « variables étoilées » c'est-à-dire le nom des textes du corpus. Certes les jalons textuels n'ont eu aucune influence sur les calculs, mais une fois que les classes ont été établies, les textes sont invités à choisir leur camp. C'est ce que montre l'analyse factorielle ci-dessus, où les huit classes occupent un espace particulier du graphique (avec un point de la même couleur au centre de gravité de cet espace).

Chaque mot du corpus peut y prendre place (on s'en est abstenu pour éviter l'encombrement) mais aussi les textes eux-mêmes qui prennent position selon leurs affinités avec les classes établies : dans la moitié supérieure la trilogie des romans modernes, à gauche , s'oppose à *Salammbô,* à droite ; dans la partie inférieure se retrouvent les textes autobiographiques des débuts de l'écrivain, à gauche, et , à droite, les tentations répétées de *Saint Antoine*.

L'analyse brute proposée dans les corrélats apparaît comme une ébauche affinée dans *Alceste*. La première n'est qu'un nuage de mots, sans autre repères que les points cardinaux. Dans la seconde des lignes de démarcation apparaissent, délimitant des constellations lexicales identifiables, tandis que la carte des textes, en surimpression, facilite l'interprétation. On aurait pu en rester là et s'en tenir à une relation client-fournisseur, *Hyperbase* préparant les données pour *Alceste* avant d'en recevoir les leçons.

Figure 3. Analyse factorielle faite par Alceste sur les substantifs de Flaubert

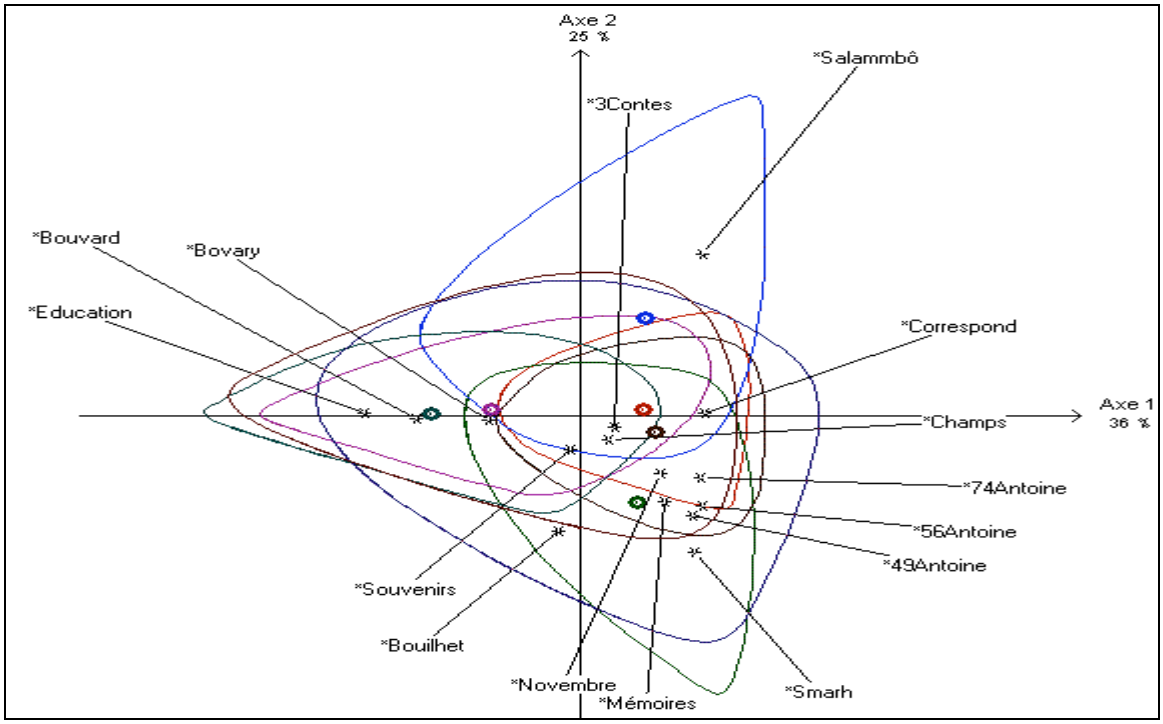
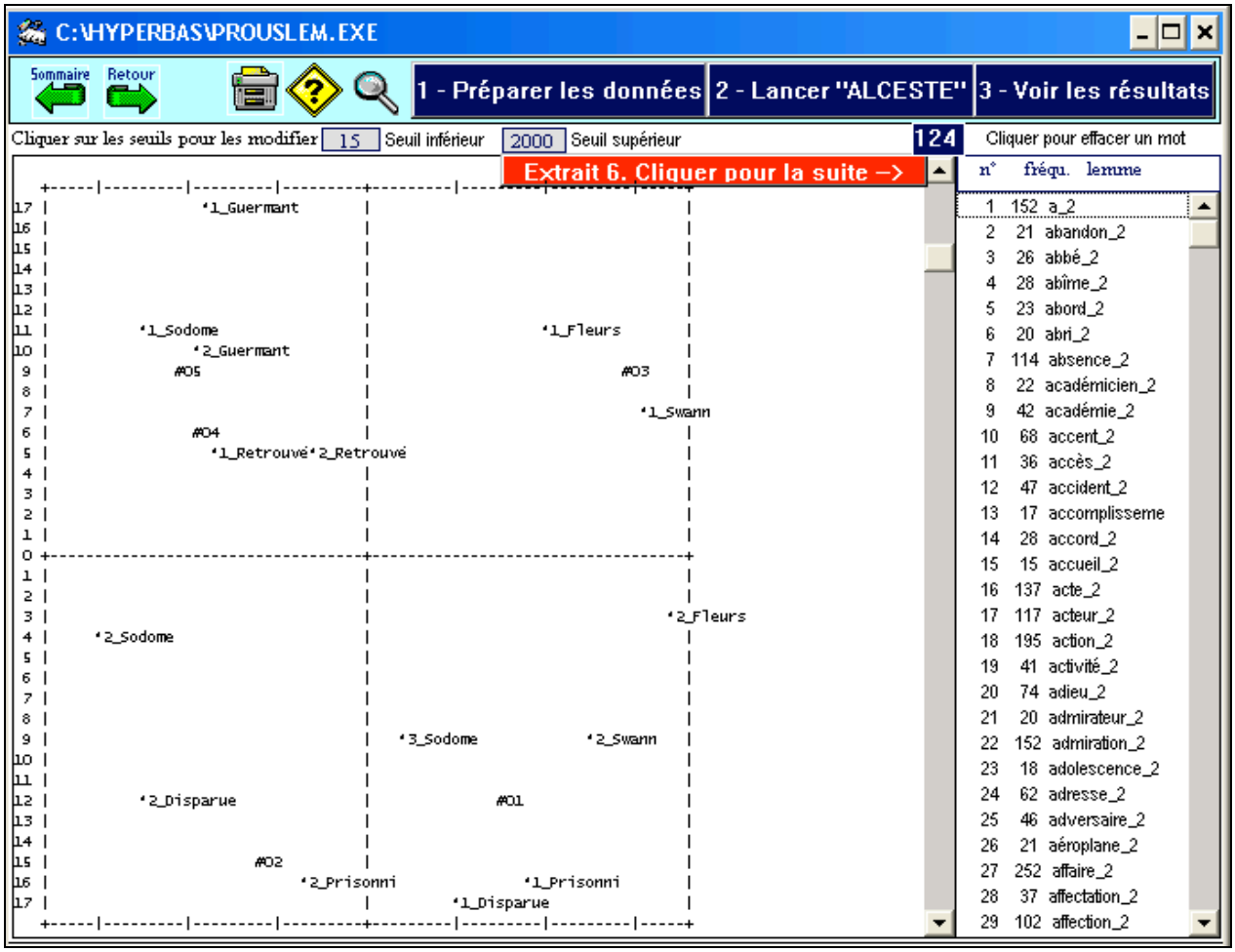


Fig.4. Analyse de l'oeuvre de Proust. Résumé d'Alceste rapatrié dans Hyperbase



5 - La fonction ASSOCIATIONS généralise la fonction THEME et étend le statut de pôle à un ensemble de plusieurs centaines de termes (ceux-là même qui ont été retenus comme "corrélats"). En s'appuyant sur la fréquence, une liste de mots pleins est d'abord constituée et donne lieu à un tableau carré de cooccurrences. Quand le tableau est rempli par un balayage complet du corpus, le détail des associations deux à deux est trié et analysé, et une représentation, sous forme de graphe, est proposée pour rendre compte des liens préférentiels qui tissent un réseau autour de chaque élément du tableau.

Si l'exploration entière du corpus est nécessaire pour offrir à *Alceste* les données particulières qu'il réclame, en revanche on n'a pas à renouveler ce long balayage, pour approfondir le réseau des associations, du moins si le tableau général des cooccurrences a déjà été constitué[22]. La recherche sur les associations s'appuie en effet sur le tableau des cooccurrences, dont la fonction CORRÉLATS a d'abord fourni une vue d'ensemble, sous forme d'analyse factorielle. La carte thématique du corpus y apparaît très claire, mais peut-être trop, car elle souligne assez trivialement les oppositions qui se font jour dans le lexique entre concret et abstrait, collectif et individuel, et qui se réalisent habituellement dans le discours romanesque. Il convient donc de répondre à des questions plus ciblées et de proposer des zooms sur des zones précises du vocabulaire.

Pour une base nouvelle, en supposant qu'on a déjà créé la liste des substantifs (ou verbes ou adjectifs) retenus et qu'on dispose du tableau général des cooccurrences, il faut procéder au calcul et au tri de tous les indices qui évaluent la distance entre les mots pris deux à deux. Ce rôle est joué par le calcul hypergéométrique, comme expliqué plus haut. Le seuil minimal de cet indice est établi par défaut à une valeur convenable vu la taille du corpus. Quand le calcul a été exécuté pour l'ensemble du tableau, le résultat n'en est pas un tableau de même taille où les éléments nuls seraient majoritaires et encombrants, mais une liste épurée qui détaille les associations privilégiées et abandonne les autres. C'est cette liste ordonnée qui est désormais consultée pour les recherches ultérieures.

Dans l'extrait qui en est livré dans le tableau 1 et qui est relatif à la *Recherche du temps perdu*, on ne s'arrêtera pas aux premières associations qui relèvent de la phraséologie et même du lexique et qu'une lemmatisation étendue aux mots composés aurait dû éliminer. Mais ces scories (*point (de) vue, maître (d') hôtel, chef (d')œuvre, œuvre (d')art*) n'entachent que la tête de liste, comme une écume mal dissoute. Dès que le coefficient échappe aux cooccurrences triviales et fixées dans la langue, des couples solides apparaissent, *défaut-qualité, père-mère, mensonge-vérité, ombre-soleil, musique-peinture*, dont le lien tient à la sémantique et à l'attirance magnétique que les mots opposés

[22] Si ce n'est pas le cas, la fonction ASSOCIATIONS déclenche le départ de l'exploration et lance le programme CORRÉLATS.

exercent l'un sur l'autre comme les pôles d'un aimant. Mais le plus souvent les couples se forment par le partage de goûts et de sèmes communs, par quelque raison métonymique, comme la relation de la partie au tout ou de la cause à l'effet, ou la proximité dans l'espace ou le temps.

LISTE HIÉRARCHIQUE		
test mot1 mot2		
74.52 loup saint	21.11 balbec plage	15.09 baron morel
62.35 pied valet	20.58 bord eau	14.64 salle table
50.71 hôtel maître	20.05 lumière soleil	14.62 mensonge vérité
50.66 chef oeuvre	20.00 phrase vinteuil	14.55 fenêtre soleil
49.44 point vue	18.44 mer soleil	14.42 duc duchesse
31.31 charlus morel	18.36 trait visage	14.36 écrivain talent
30.25 chambre valet	16.81 eau soleil	14.27 musique peinture
26.85 bord mer	16.80 ciel soleil	14.24 france roi
26.54 cottard docteur	16.69 mer plage	14.23 ombre soleil
24.23 art oeuvre	16.54 joue nez	14.14 mère père
23.98 défaut qualité	16.36 duc frère	13.95 oeil regard
23.97 avenir passé	16.28 lèvre sourire	13.90 joue oeil
23.85 gens monde	16.24 champs gilberte	13.75 fête invité
22.01 baron charlus	16.05 duchesse princesse	13.75 musicien vinteuil
21.72 maison maîtresse	15.90 elstir tableau	13.48 expression visage
21.46 musique vinteuil	15.76 elstir peinture	13.45 bouche regard
	15.54 bruit oreille	13.35 eau mer
	15.49 chambre lit	13.28 arbre soleil
	15.22 regard sourire	13.24 écrivain livre

Tableau 1. Les associations que Proust privilégie parmi les substantifs (extrait partiel)

1 – On peut tout d'abord isoler une ligne du tableau (en cherchant ce qu'il en reste dans la liste, quand tous les éléments nuls et non significatifs ont été expurgés) et la transposer dans un histogramme. Cette représentation simple est disponible, quoique réductrice. Elle a pourtant son intérêt si le mot représenté dans son environnement lexical est recherché de la même façon dans d'autres corpus. Les fréquences brutes du mot en question peuvent être semblables ou différentes dans ces corpus comparés, ce n'est pas là ce qui compte. On ne se soucie que de confronter leur entourage respectif, selon le principe « dis-moi qui tu fréquentes et je te dirai qui tu es ».

S'agissant de Proust, l'*amour,* qui fait l'objet de la figure 5, est lié à la *souffrance* et à la *jalousie*, c'est l'amour maladie. Chez Rousseau l'*amour* s'accorde avec la *vertu*, l'*amitié* et la *patrie*. Flaubert est plus sensible au *baiser*, à la *volupté* et aux satisfactions du *cœur*, tandis que Zola associe la *chair* et le *désir* à la *passion*.

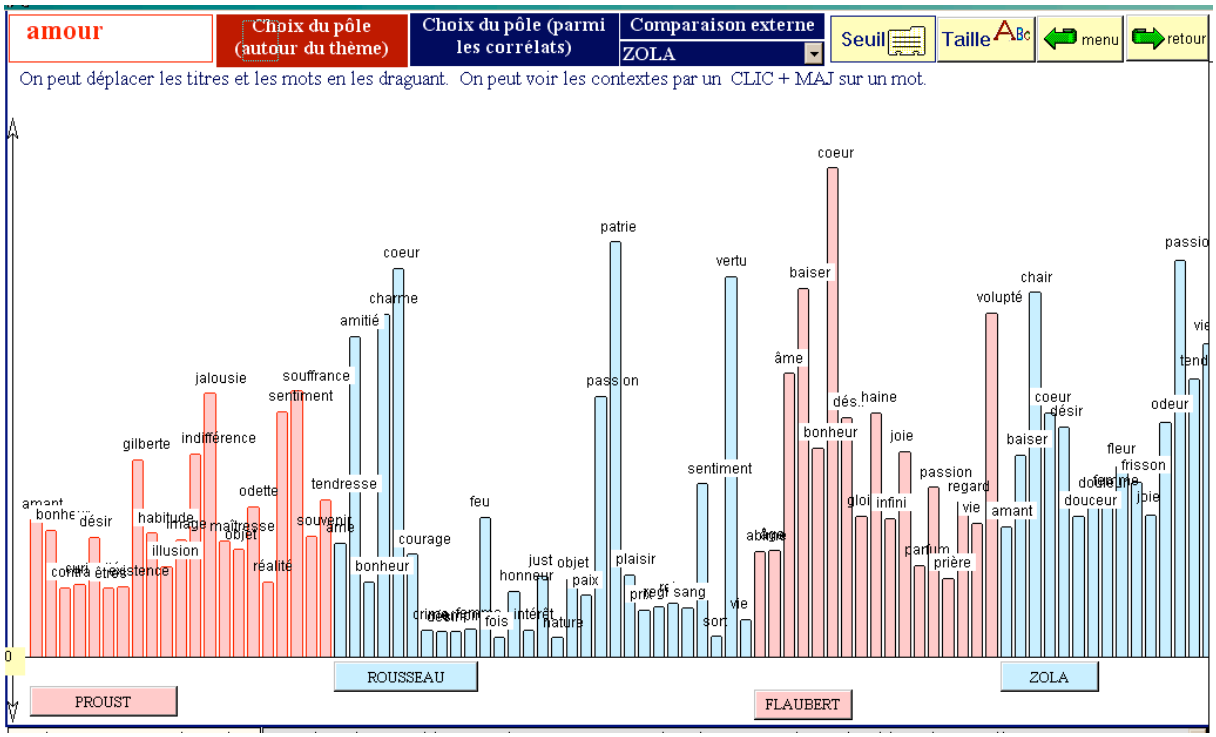


*Figure 5. La constellation lexicale autour de l'*amour *chez Proust, Rousseau, Flaubert et Zola*

2 – Entre la vue lointaine de l'analyse factorielle (figure 1) et le détail myope de l'histogramme (figure 5), il y a place pour un échelon intermédiaire : comme précédemment on commence par s'attacher à un mot parmi les 400 disponibles. Une fois que l'hameçon est accroché, on tire sur le fil et on sort de l'eau non seulement les mots-amis qui sont liés au mot-pôle, mais aussi ceux qui sont proches de ces proches. L'enquête qui s'étend donc aux amis des amis vise à dessiner un réseau complexe autour du pôle[23].

Nous prendrons le mot *mémoire* à l'intérieur de la *Recherche du temps perdu*. Les liens représentés dans le graphe 6 sont en rouge s'ils concernent le mot-pôle (ce sont ceux qui ont donné matière à l'histogramme 3) , ils sont en bleu s'ils concernent les mots liés au pôle et en noir dans les autres cas. Les mots eux-mêmes sont différenciés par la couleur: le rouge est réservé aux noeuds fréquentés, le noir aux noeuds isolés (moins de 5 liaisons). La force des liaisons influe sur l'épaisseur des traits et la taille des caractères. Si les relations collatérales (en noir) encombrent sans profit le graphique, on peut les faire disparaître (ou les rétablir) en faisant appel au bouton SIMPLIFIER/ENRICHIR.

[23] Des raisons pratiques nous ont dissuadé d'approfondir encore le champ exploré et d'envisager un troisième niveau. À chaque étage le champ s'élargit en effet comme le carré du précédent et on aurait vite atteint les limites d'un tableau. pourtant gros de 16000 éléments. En outre la polysémie qu'on peut rencontrer dans le mot-pôle et à chaque étage du réseau produit beaucoup de dispersion et le nuage des points s'effiloche au gré des courants et diversions polysémiques.

Rappelons que le calcul du graphe arborescent et de la position des noeuds et des arcs est assuré par le logiciel libre *GRAPHVIZ* (licence GNU)[24].

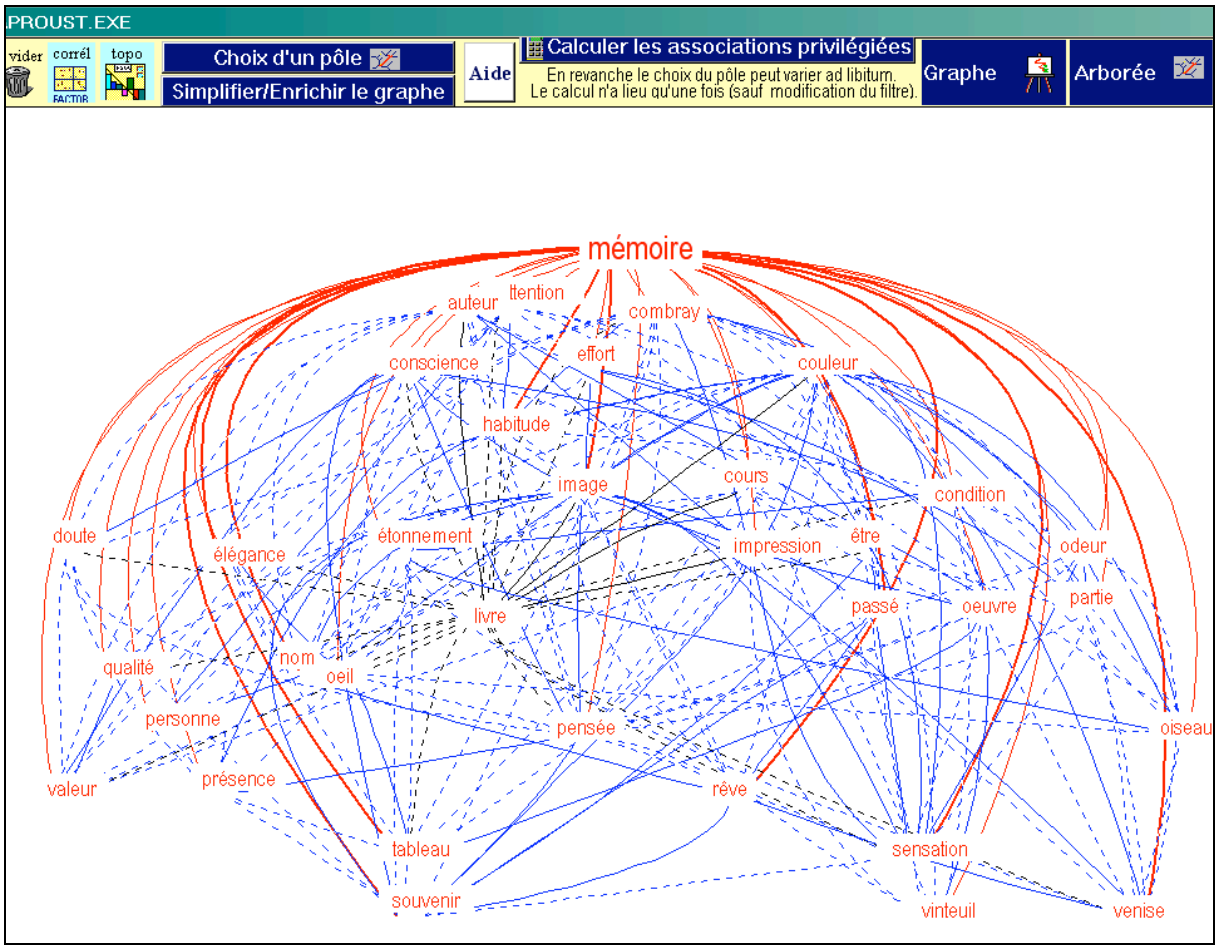


Figure 6. Le graphe de la mémoire *dans la* Recherche du temps perdu

En réalité le logiciel *GRAPHVIZ* ignore les poids et les pondérations et ne veut connaître pour chaque élément du tableau des cooccurrences qu'une information grossière du type présence/absence, comme si l'on circulait dans un réseau binaire avec des portes ouvertes ou fermées mais non entrebâillées ou grandes ouvertes. Comme pour chaque arc nous connaissons la force d'attraction calculée par l'hypergéométrique, nous avons pu réintroduire cette information en épaississant les traits ou en grossissant les caractères. Mais on aurait aimé que le dessin du graphe soit ordonné en tenant compte non seulement de l'existence d'un lien mais aussi de la mesure de son intensité. Ces sortes de graphe sont vite illisibles et on regrette qu'un élément de clarté ait été négligé.

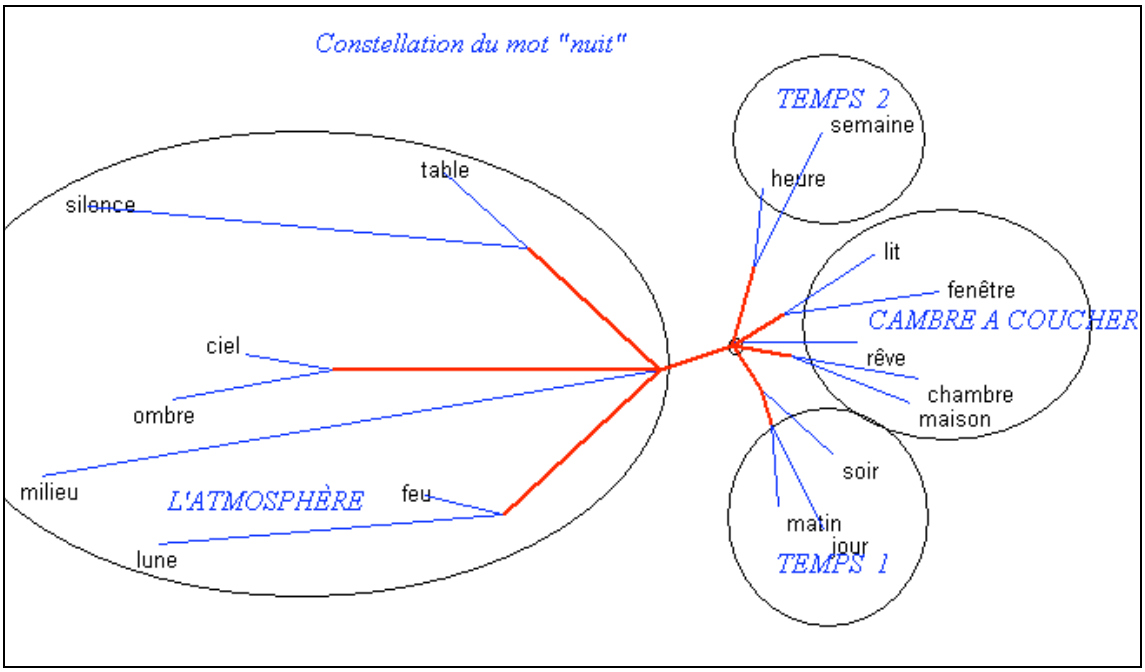
[24] Nous n'avons utilisé GRAPGVIZ que pour le calcul de la position des arcs et des points. Quant au dessin des arcs, nous avons aménagé leur courbure pour faciliter l'analyse. Il est possible d'accentuer ou d'atténuer cette courbure avec la souris et de déplacer légèrement un point lorsqu'il se produit un recouvrement gênant.

3 – On a constitué un tableau carré où les valeurs de proximité se substituent à la mesure brute des cooccurrences. Les effectifs absolus sont en effet trop dépendants de la fréquence des mots. Dans l'approche qui précède, c'est ce tableau entier qui est exploré, les ramifications pouvant aller loin quand elles se communiquent de voisin à voisin. Mais on peut fixer une barrière à cette propagation, en constituant un sous-tableau, carré comme le grand, et qui porte en marge des lignes et des colonnes la liste des mots directement liés au pôle. Un tel tableau contient les cooccurrences pondérées des uns avec les autres, en excluant précisément le pôle et en neutralisant les liens de chacun avec ce pôle. En somme une séance à huis clos, où les gens en relation avec l'intéressé sont invités à porter leur témoignage en son absence.

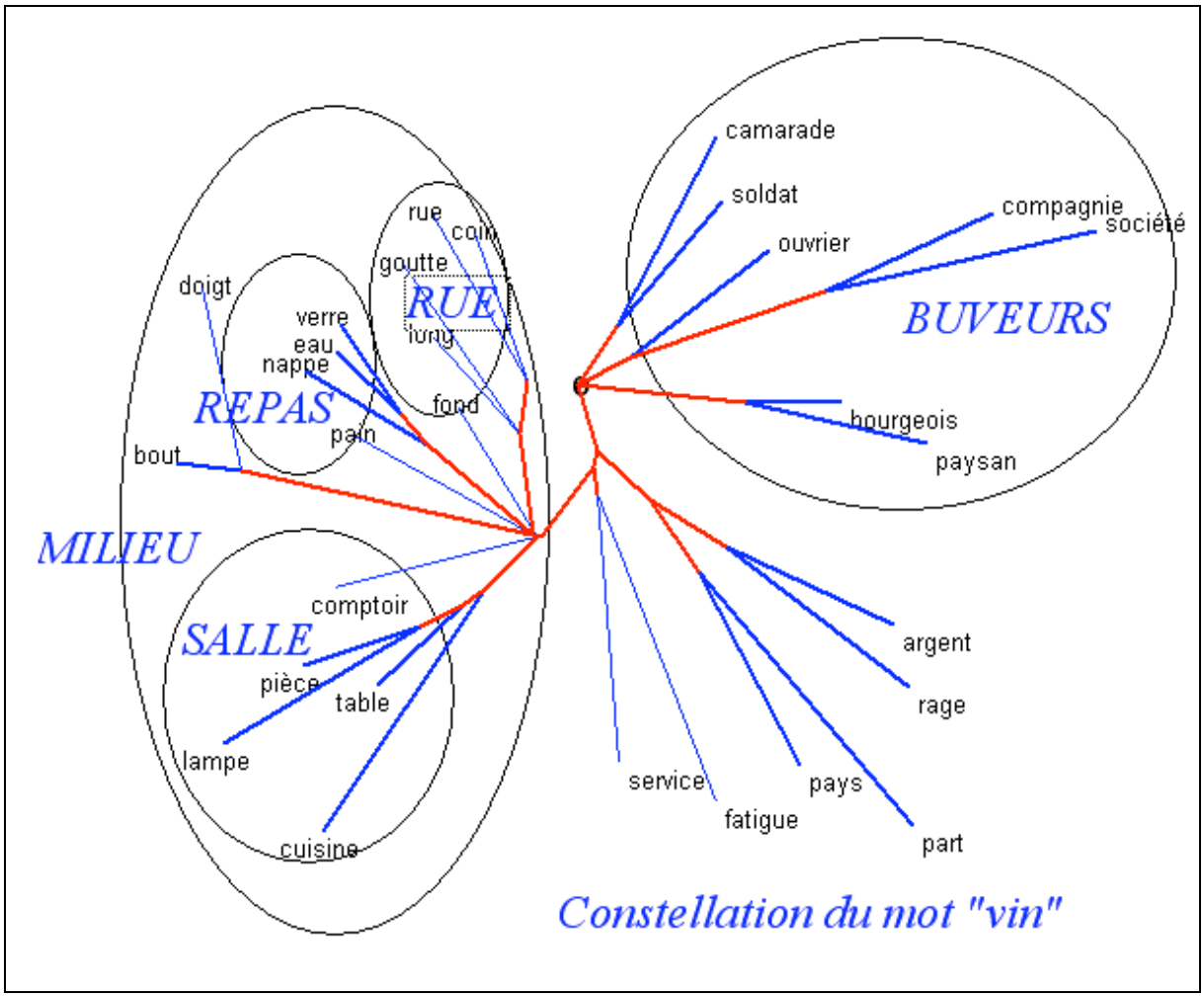
Ce sous-tableau est alors soumis aux méthodes habituelles: analyse factorielle de correspondance et analyse arborée. Deux boutons sont disponibles à cet effet et s'appliquent au mot qui a été choisi pour pôle. Nous éclairerons cette procédure avec le mot « nuit », emprunté à la base EXEMPLEM. Le graphe obtenu est encore plus complexe que celui de la mémoire. Des zébrures dans tous les sens y traversent l'espace et font penser à un feu d'artifice nocturne, au point qu'on a renoncé à le montrer. Tout se simplifie pourtant dans l'analyse arborée qui en garde la trace et en souligne la structure : la nuit peut être considérée d'abord sous l'aspect temporel ; elle voisine alors avec les autres unités de temps, celles qui sont à sa mesure, comme le *jour*, le *soir* et le *matin* et celles qui sont à une échelle différente comme l'*heure* et la *semaine*. La nuit s'exprime aussi dans l'espace : elle évoque d'une part le lieu clos de la chambre à coucher (*maison, chambre, lit, fenêtre*) et d'autre part l'atmosphère nocturne à ciel ouvert (*ciel, lune, feu, milieu, ombre, silence*).

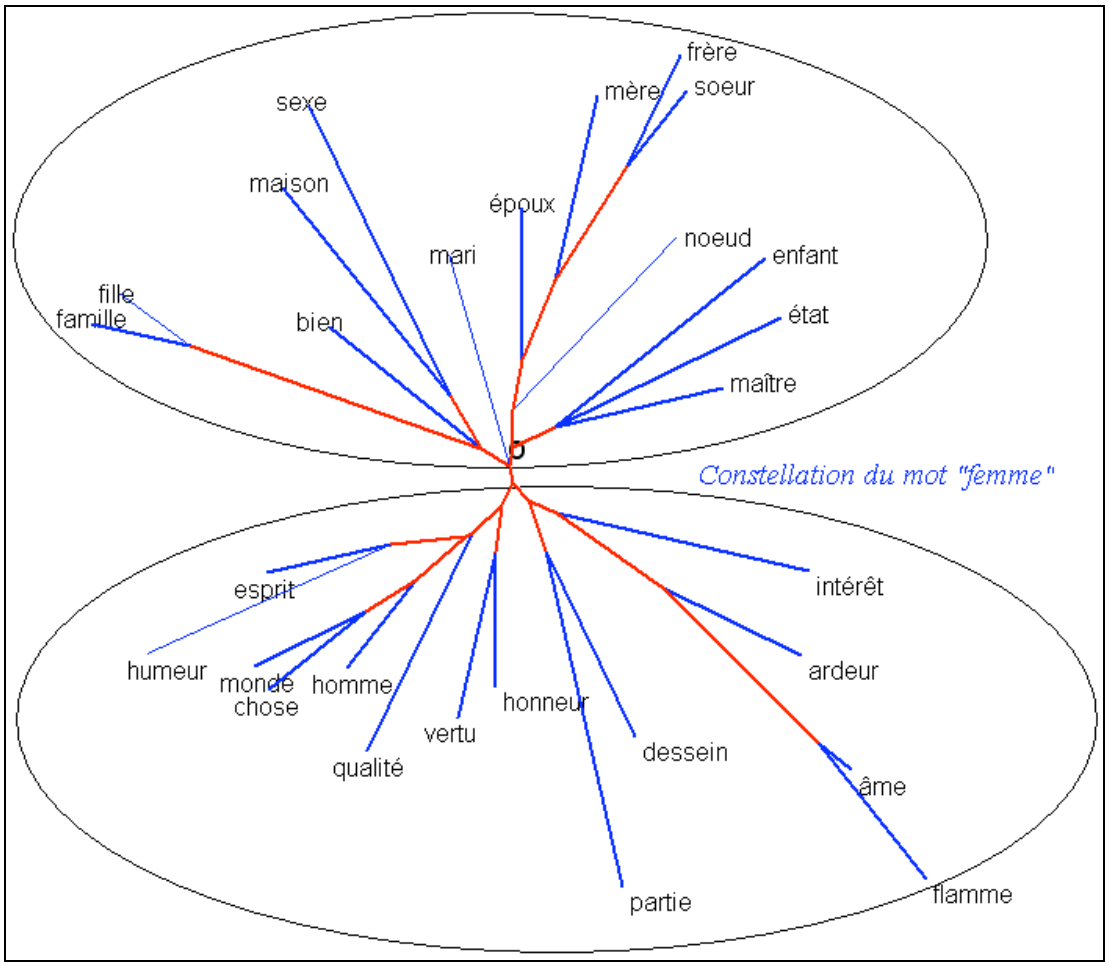
Certes ce schéma n'a pas la même précision qu'obtient Bruno Gaume, avec des méthodes semblables, dans la représentation graphique des verbes du *Grand Robert*. Mais nous partons ici non d'un dictionnaire mais de textes littéraires, où les mots voguent en liberté sans s'enfermer dans des définitions circulaires[25].

[25] Gaume B. (2006) *La proxémie : vers un modèle de sémantique lexicale pour un Traitement automatique des langues à ergonomie cognitive*, http://www.limsi.fr/Individu/habert//04-05/inex.html. B Gaume a mis au point un logiciel graphique, qui semble grandement supérieur à G*raphviz*. On aimerait que sa diffusion soit rendue possible.



L'analyse arborée, s'appuyant sur une distance chiffrée, a un pouvoir discriminant remarquable. Dans la constellation qui se forme autour d'un mot elle aide à distinguer les amas de mots qui s'attirent mutuellement, comme le groupe des buveurs autour du *vin*, ou le cercle de la famille autour de la *femme*.





En conclusion, on n'est pas certain que toutes ces opérations perpétrées sur les séquences et les cooccurrences puissent faire découvrir le sens des mots, sauf dans certains cas où la polysémie peut être facilement décantée, comme dans le cas de l'*argent*. Les champs sémantiques qu'on parcourt en arpenteur sont d'autant mieux délimités qu'ils appartiennent à des séries relativement fermées comme les liens de la parenté ou les parties du corps humain. Mais le bornage des champs est le plus souvent approximatif et les contestations sans solution. En outre les collocations ne sont pas nécessairement des connotations et la phraséologie ordinaire joue dans beaucoup de cooccurrences un rôle majeur qu'on peut trouver gênant, sauf si c'est là ce qu'on cherche. Mais si le sens d'un mot est, comme on l'a dit, la somme de ses emplois, on ne peut éviter de recenser ces emplois, en espérant ramasser dans le filet électronique quelques-unes des isotopies, dont le rayonnement parcourt les textes littéraires sans être facilement observable. Quel dommage que la physique ait tant d'outils pour déceler les multiples rayonnements dans le ciel étoilé au-dessus de nos têtes et qu'on en ait si peu pour les rayonnements textuels et ceux de la loi morale au fond de nos coeurs.

CHAPITRE 5.
AUTRES VERSIONS LEMMATISÉES

La version standard d'Hyperbase n'est pas réservée au français. Elle peut traiter pareillement des données en langue étrangère, pourvu que l'alphabet latin soit utilisé. Seule la référence à FRANTEXT est propre au français pour la comparaison extérieure. Mais un test est fait sur les données qui propose une autre référence s'il est constaté que le texte n'est pas en français: le *British National Corpus* pour l'anglais, le journal *Publico* pour le portugais, etc. Mais lorsqu'intervient un lemmatiseur, les caractères spécifiques de la langue doivent être prises en compte, ce qui rend nécessaire un toilettage des données préalable à l'opération de lemmatisation et une adaptation du logiciel aux résultats de celle-ci. On propose donc pour chaque langue une version particulière d'Hyperbase, qui sous une approche commune se plie à la diversité des codes grammaticaux. Afin d'éviter la dispersion, on a choisi le même lemmatiseur qui nous avait servi pour le français: TreeTagger.

TÉLÉCHARGEMENT de TreeTagger

TreeTagger est un logiciel libre. Tout un chacun peut le télécharger en s'adressant à son créateur, Helmut Schmitt, de l'université de Stuttgart, à l'adresse :
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
Voici ce que propose le site en question (début mai 2006):

The TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Spanish, Bulgarian, Greek, Portuguese and old French texts and is easily adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Sample output:

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

The tagger is described in the following two papers:

- "Probabilistic Part-of-Speech Tagging Using Decision Trees" (gzipped Postscript, pdf)
- "Improvements in Part-of-Speech Tagging with an Application to German" (gzipped Postscript, pdf)

Executable code for Sparc workstations, Linux and Windows PCs and Macs as well as parameter files for English, German, Italian, Spanish, Bulgarian, French and old French can be downloaded via the links below.

The French and the Italian parameter files are provided by Achim Stein.

The second Italian parameter files was provided by Marco Baroni.

The English parameter file was trained on the PENN treebank and uses the English morphological database created by Karp, Schabes, Zaidel and Egedi.

The Spanish parameter file was trained on the Spanish CRATER corpus and uses the Spanish lexicon of the CALLHOME corpus of the LDC.

The Bulgarian parameter file was created by Julien Nioche on the Bulgarian Treebank. It uses a UTF-8 encoding.

At the end of this page, you will also find a link to Pablo Gamallo's website where you can download a Linux parameter file for Portuguese.

This software is freely available for research, education and evaluation.

Please read the license terms, before you download the software! By downloading the software, you agree to the terms stated there.

The following steps are necessary to install the TreeTagger (see below for the Windows version):

1. Download the tagger package for your system (Sparc-Solaris, PC-Linux, Mac OS-X).
2. Download the tagging scripts into the same directory.
3. Download the parameter files for your system (Sparc-Solaris, PC, Mac).
4. Download the installation script install-tagger.sh.
5. Open a terminal window and run the installation script in the directory where you have downloaded the files:
   sh install-tagger.sh

6. Make a test, e.g.
   echo 'Hello world!' | cmd/tree-tagger-english
   or
   echo 'Das ist ein Test.' | cmd/tagger-chunker-german

If you have difficulties with the installation, have a look at the installation hints (kindly provided by Joachim Wagner).

Parameter files for Sparc-Solaris and Mac OS-X (Latin1 character set)

- English parameter file (3045 kByte, gzip compressed)
- German parameter file (7012 kByte, gzip compressed)
- small German parameter file (2415 kByte, gzip compressed)
- French parameter file (2375 kByte, gzip compressed)
- Italian parameter file (5484 kByte, gzip compressed)
- Marco Baroni's Italian parameter file (1972 kByte, gzip compressed)
- Spanish parameter file (918 kByte, gzip compressed)
- Bulgarian parameter file (603 kByte, gzip compressed)
- German chunker parameter file (52 kByte, gzip compressed)
  Note: The German tagger parameter file is needed, as well.
- English chunker parameter file (82 kByte, gzip compressed)
  Note: The English tagger parameter file is needed, as well.

Parameter files for PC (Linux and Windows, Latin1 character set)

- English parameter file (2945 kByte, gzip compressed)
- German parameter file (6642 kByte, gzip compressed)
- small German parameter file (2340 kByte, gzip compressed)
- French parameter file (2336 kByte, gzip compressed, information about this file)
- Italian parameter file (3238 kByte, gzip compressed, information about this file)
- Marco Baroni's Italian parameter file (1963 kByte, gzip compressed)
- Spanish parameter file (899 kByte, gzip compressed)
- Bulgarian parameter file (579 kByte, gzip compressed)
- German chunker parameter file (52 kByte, gzip compressed)
  Note: The German tagger parameter file is needed, as well.
- English chunker parameter file (82 kByte, gzip compressed)
  Note: The English tagger parameter file is needed, as well.

A Windows version of the TreeTagger is also available. The parameter files have to be downloaded separately.

Tagsets

Here is some information about the tagsets used in the parameter files:

- English (Penn-Treebank tagset)
  The tagset used by the TreeTagger is a refinement of this tagset where the second letter of the verb part-of-speech tags distinguishes between "be" verbs (B), "have" verbs (H) and other verbs (V).
- German (in German)
- French (in French)
- Italian
- Marco Baroni's Italian tagset
- Spanish
- Bulgarian
- Pablo Gamallos web page where you can download a parameter file for Portuguese.
- Achim Stein's web page on French and old French POS tagging with the TreeTagger
- Italian Online Tagger at the University of Odense
- Python Wrapper for the TreeTagger (developed by Laurent Pointal)

LEMMATISATION À DISTANCE.

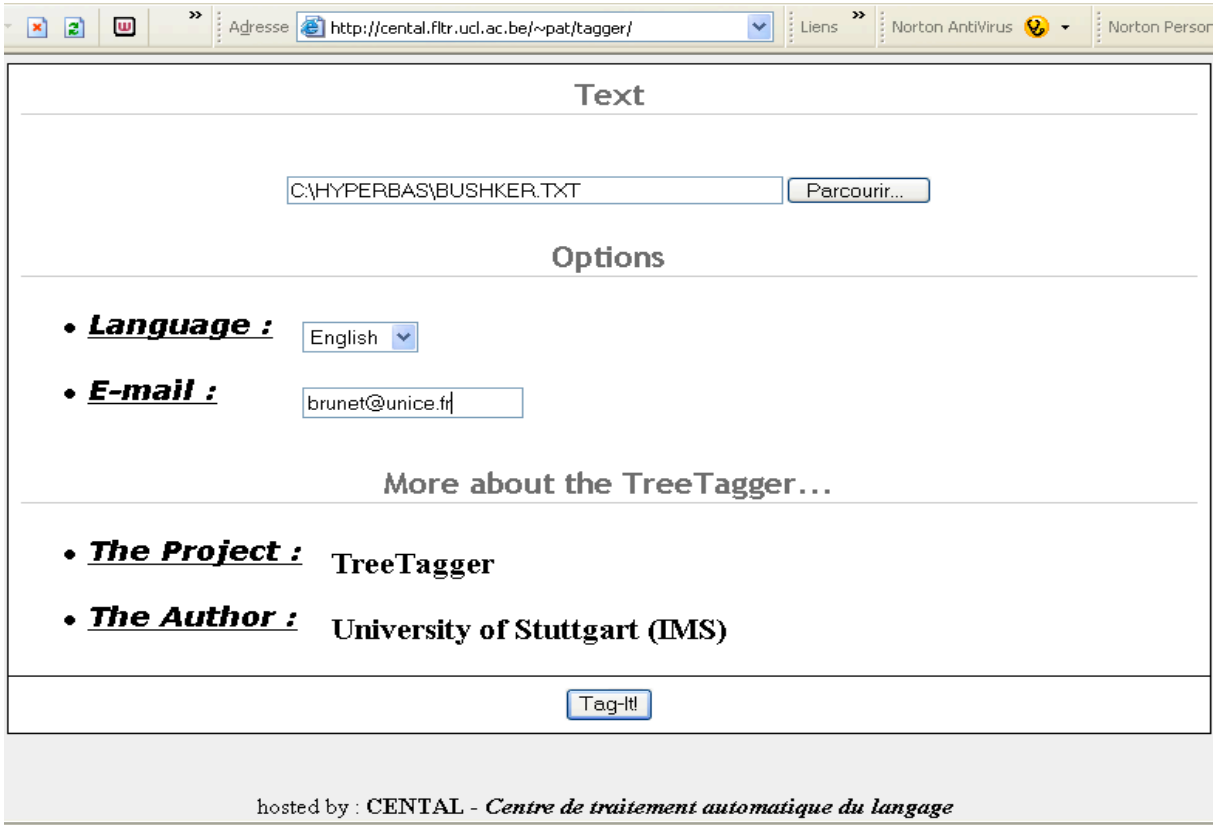
TreeTagger fonctionnant en client/serveur

Logiciel libre, TreeTagger est disponible pour différents systèmes, différentes langues et différents apprentissages. Beaucoup de collaborateurs ont œuvré à son développement et contribué à sa diffusion. Cela ne va pas sans quelques particularités dans la mise en œuvre du logiciel et son application à un cas précis. D'une version à l'autre on note des variantes dans les scripts qui formatent le texte, gèrent l'adresse des fichiers et et coordonnent les ressources. Et le néophyte peu habitué à l'environnement Unix peut parfois être embarrassé pour installer et lancer le lemmatiseur. Or si le temps ou les compétences lui manquent, l'utilisateur peut s'adresser à un serveur qui propose la lemmatisation à distance du texte de son choix. Le centre de recherche Cental de l'université de Louvain assure gratuitement ce service, à l'adresse :

http://cental.fltr.ucl.ac.be/~pat/tagger/

Il suffit de préciser le fichier à lemm atiser, la langue à utiliser (anglais, allemand, français ou italien) et l'adresse email où le fichier lemmatisé sera envoyé. L'opération est instantanée dès qu'on clique sur le bouton *Tag-It*. Il ne reste plus qu'à consulter son courrier électronique pour récupérer le fichier de sortie.

Dialogue émis par le serveur



LA VERSION ANGLAISE : HYPERANG.EXE

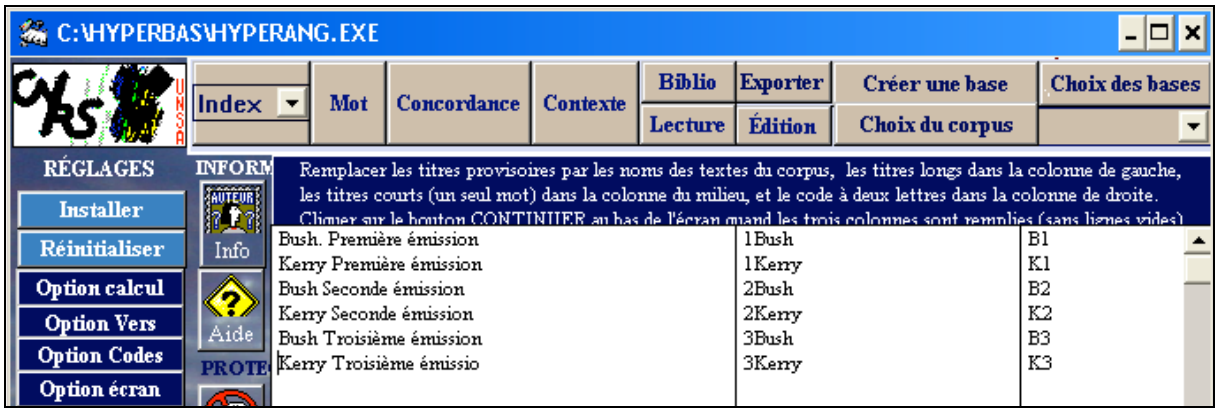
La base HyperAng.exe est vide, comme HyperCor.exe, HyperGer, HyperIta, et HypEspag. Ce sont des modèles, pourvues de fonctions et dépourvues de données. Dès qu'on veut créer une base nouvelle, on doit en faire un double auquel on donne un autre nom. Veiller à limiter à huit lettres la longueur de ce nom car un vieil héritage du DOS empêche le système de montrer les caractères surnuméraires et cela peut entraîner des confusions et une mauvaise reconnaissance des fichiers. Ce nom donné à la base est repris par d'autres fichiers qui lui sont associés et qui ont des suffixes différents, par exemple .txt, .tx2, .cor, .cnr, .don, .sor, .sta.

Le menu CREATION

Une fois faite la duplication, la base nouvelle s'ouvre sur le menu de création (voir ci-dessus). On suppose qu'à ce moment les fichiers lemmatisés par TreeTagger sont disponibles dans le même répertoire et qu'ils sont désignés par les noms Texte1.cnr, Texte2.cnr, Texte3.cnr, etc.. Le bouton CREATION déclenche les opérations, la première consistant à nommer les différents textes du corpus. Trois champs parallèles sont ouverts qu'on doit remplir convenablement, celui de gauche avec des noms développés, celui du milieu avec des noms courts, en un seul mot, et celui de droite avec des codes de deux

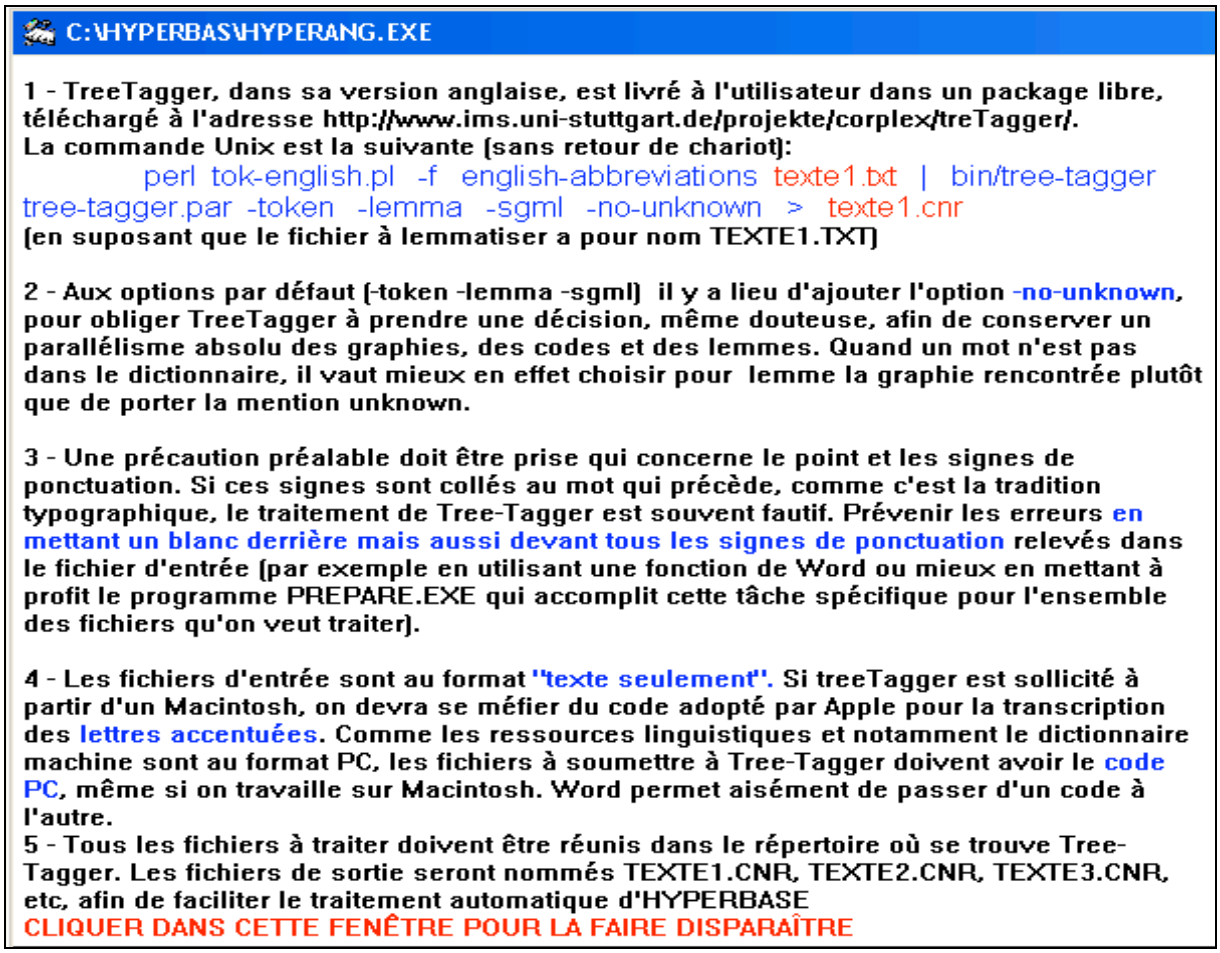
caractères. Éviter les retours de charriot inutiles : les trois champs doivent avoir strictement le même nombre de lignes, sans quoi on s'attire un rappel à l'ordre.

Le choix des titres



Des explication succinctes sont délivrées avant d'engager la chaîne des traitements, afin que l'utilisateur sache ce dont le logiciel a besoin. Si de plus amples informations sont nécessaires, en particulier sur la lemmatisation, une aide plus circonstanciée est disponible, reproduite ci-dessous).

La lemmatisation. Aide complémentaire



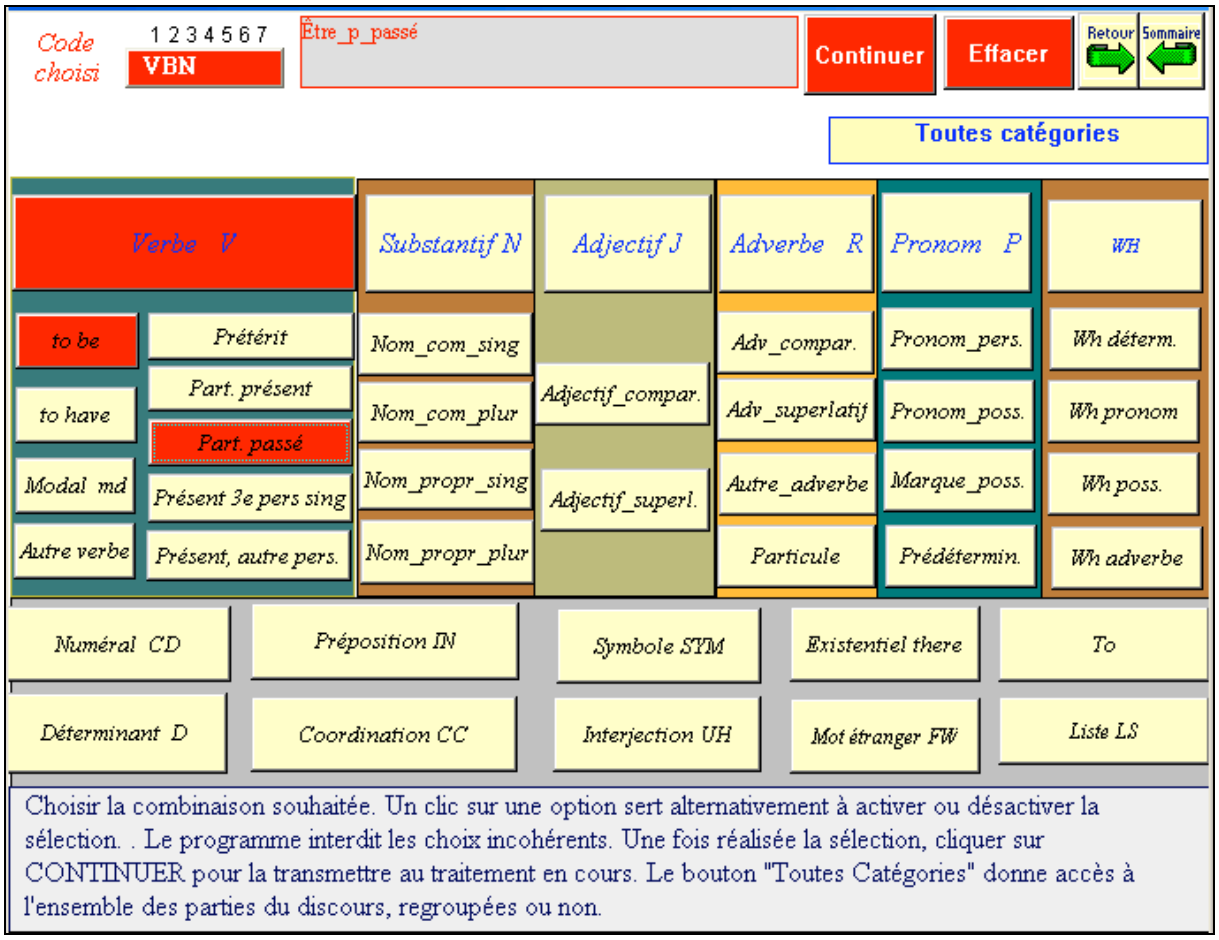
Le traitement de données lemmatisées est nettement plus long que celui des seules graphies. Car toutes les phases (de tri, d'indexation, de calcul de spécificités, etc.) sont répétées quatre fois, an niveau des structures syntaxiques, puis des codes grammaticaux, puis des lemmes et enfin des graphies. La présence de l'utilisateur est requise pendant toute la durée de l'opération, car certains dialogues sont proposés qui sollicitent son choix. En particulier l'emplacement des fichiers ou des programmes extérieurs doit être confirmé (il est sage de les entreposer dans le répertoire C:/HYPERBAS). La taille de la base croît en proportion : il n'est pas rare qu'elle dépasse 100 millions d'octets pour une monographie d'écrivain. Si cela obère quelque peu la phase de préparation, l'exploitation de la base n'en est pas ralentie, les recherches étant fondées sur l'adressage indexé. Nous ne décrirons pas le détail des opérations puisque ce sont les mêmes, mais répétées, que celles de la version standard. Cependant le transfert des données lemmatisées est plus lourd que celui des seules graphies et l'utilisateur est invité à prendre patience durant cette phase un peu lente, dont le rythme n'est que de 1000 lignes à la seconde.

Quand la réalisation est achevée et que la base est exploitable, les fonctions disponibles sont semblables à celles qu'on a décrites précédemment pour les bases réalisées avec *Cordial*. Nous renvoyons donc le lecteur au développement précédent. Deux particularités sont cependant à souligner : d'une part TreTagger ignorant tout codage sémantique, les fonctions correspondantes disparaissent. D'autre part le codage grammatical de l'anglais étant différent de celui du français, le menu de l'écran « grammatical » a été adapté aux choix imposés par Treetagger. Ces choix sont les suivants :

This section contains a list of tags in alphabetical order and the parts of speech corresponding to them.

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NP	Proper noun, singular
15.	NPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PP	Personal pronoun
19.	PP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Le menu grammatical



Quoique les codes retenus ne soient pas superposables à ceux que *Cordial* propose pour le français, les parties du discours sont regroupées de la même façon, à savoir : 1 verbe, 2 substantif, 3 adjectif, 4 numéral, 5 pronom, 6 adverbe, 7 déterminant, 8 conjonction, 9 préposition. Ces codes numériques sont ajoutés aux lemmes pour lever les ambiguïtés. Ils ne sont pas expressément fournis pas *TreeTagger* (non plus que par *Cordial*). Mais ils sont dérivés de l'analyse du lemmatiseur. Noter que dans la représentation des structures syntaxiques, ces mêmes parties du discours sont marquées par un code alphabétique, afin de faciliter le décryptage visuel (*v* = verbe, *n* = substantif, *a* = adjectif, *m* = numéral, *p* = pronom, *r* = adverbe, *d* = déterminant, *c* = conjonction, *s* = préposition). Ainsi la chaine *danv* traduit la séquence *déterminant + adjectif + nom + verbe.*

LA VERSION ALLEMANDE : HYPERGER.EXE

Les procédures qu'on vient d'exposer pour l'anglais valent pour l'allemand et les autres langues disponibles. Mais la version à dupliquer est ici *Hyperger.exe*. Le lancement de TreeTagger se fera avec les paramètres prévus pour l'allemand et invoqués par le script *tree-tagger-german* .
La commande UNIX est la suivante:

sh tree-tagger-german texte1.txt > texte1.cnr

Aux options retenues dans ce script : -token -lemma -sgml il conviendrait d'ajouter l'option -no-unknown . Veiller aussi, sur Mac, à changer GAWK en AWK dans ce même document car le système MacOs X ne connaît que ce nom AWK. .Les fichiers d'entrée (dans l'exemple TEXTE1.txt) sont au format "texte seulement". Les sorties ont un nom et un suffixe imposés: TEXTE1.CNR, TEXTE2.CNR,etc... Avant de faire appel à TreeTagger il est prudent de formater les données en faisant appel au programme PREPARE.EXE.

Ceux qui ne disposent pas de TreeTagger peuvent obtenir gratuitement et immédiatement la lemmatisation de leurs fichiers en s'adressant au site:
http://cental.fltr.ucl.ac.be/~pat/tagger/

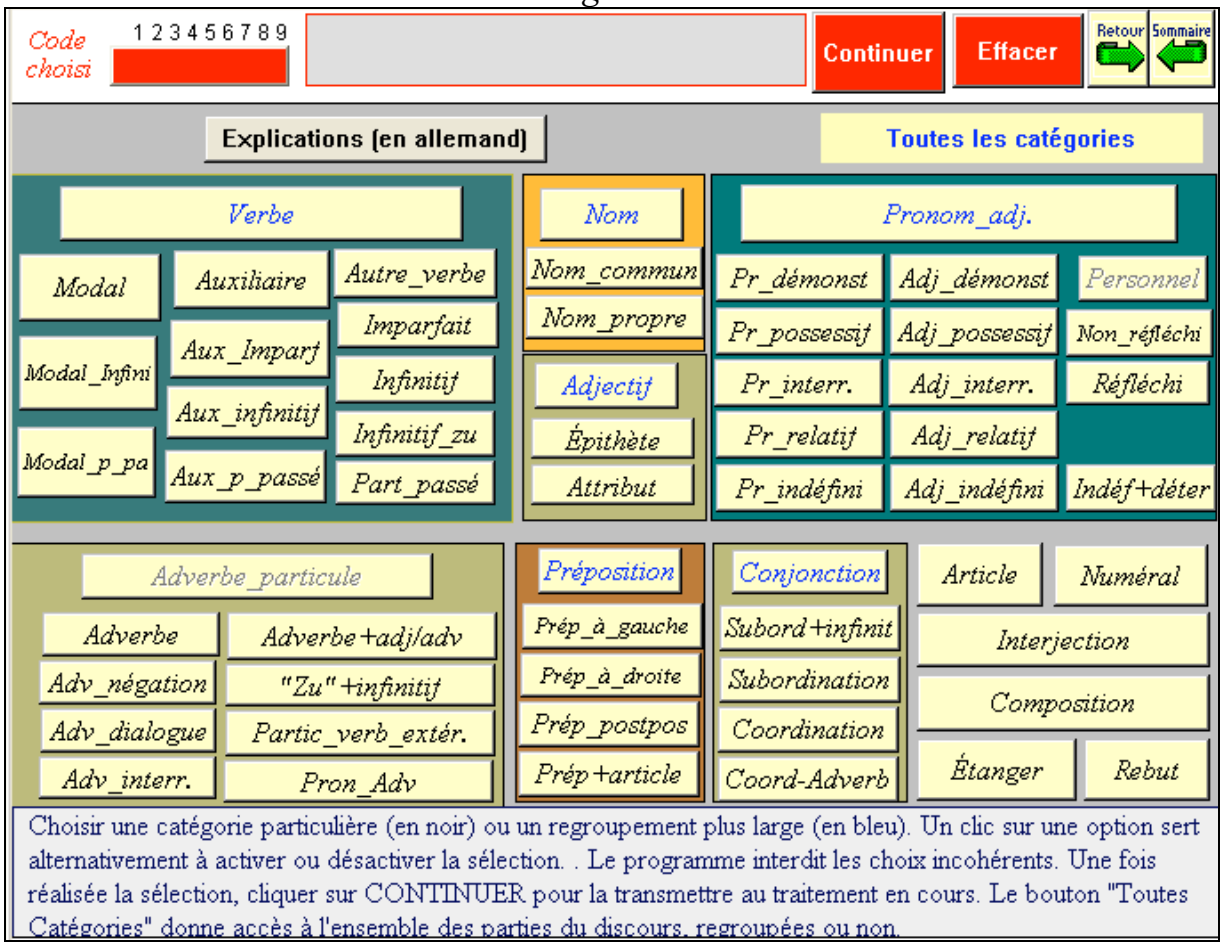
Le codage grammatical s'établit comme suit :

POS =	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	*um [zu leben],* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit,* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*
NN	normales Nomen	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	Relativpronomen substituierend	*[der Hund,] der*
PRELAT	attribuierend	*[der Mann ,] dessen [Hund]*

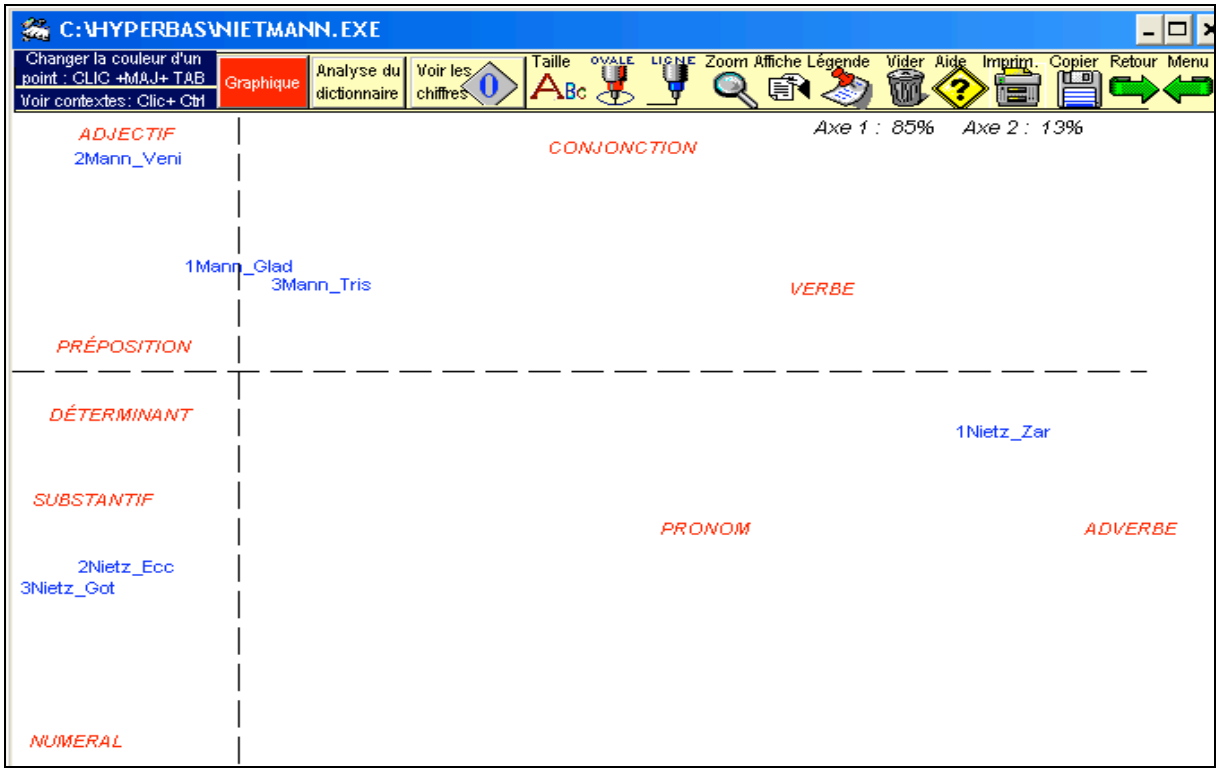
POS =	Beschreibung	Beispiele
	Relativpronomen	
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe], wessen [Hut]*
PWAV	adverbiales Interrogativ– oder Relativpronomen	*warum, wo, wann, worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	"zu" vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten], zu [schnell]*
TRUNC	Kompositions–Erstglied	*An– [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	– []()

Hyperbase traduit ce codage dans un menu approprié (voir ci-dessous). Les boutons représentés en bleu permettent des regroupement catégoriels. Quant à la présentation récapitulative des parties du discours, on y accède avec le bouton « Toutes les catégories » qui en établit la liste. Le tableau peut donner lieu à diverses analyses, arborées ou factorielles, à l'image de l'exemple ci-dessous emprunté à un corpus d'illustration où trois romans de Thomas Mann étaient confrontés à trois œuvres de Nietzsche, dont *Zarathoustra*. On voit que ce dernier texte se sépare non seulement des romans de Th. Mann mais aussi des autres textes de Nietzsche, et que l'utilisation des parties du discours y est radicalement différente, le verbe et ses acolytes (adverbes,et pronoms) étant privilégiés.

Le menu grammatical



Analyse *factorielle des parties du discours dans un corpus Nietzsche-Mann*



LA VERSION ESPAGNOLE : HYPESPAG.EXE

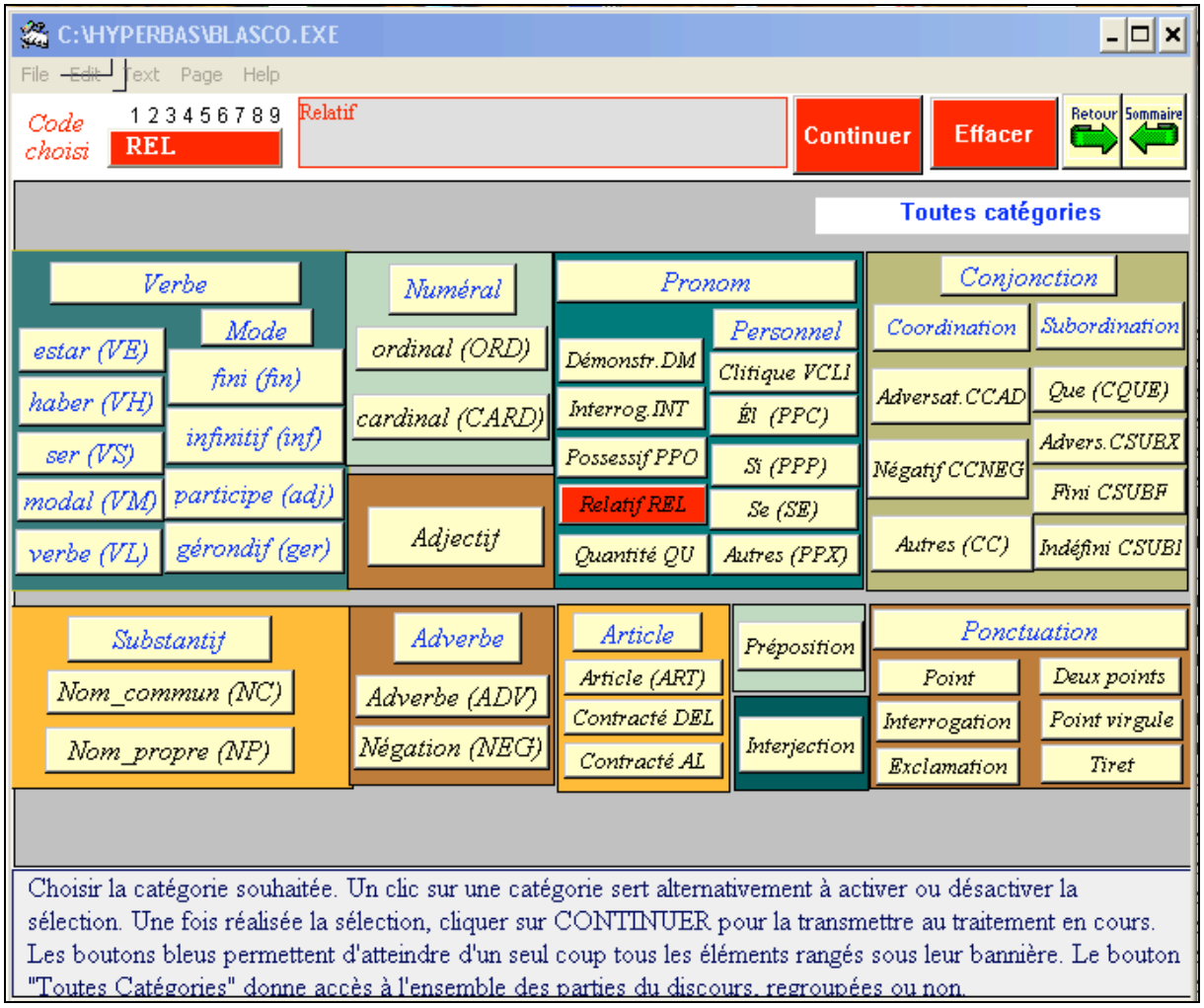
On prie le lecteur de se reporter aux pages qui précèdent pour les procédures de lancement et d'exploitation. On se contentera d'expliciter les particularités du codage grammatical telles qu'on les trouve pour l'espagnol dans *TreeTagger* .

```
FS    Full stop punctuation marks
SYM   Symbols
ART    Articles (un, las, la, unas)
ADJ   Adjectives (mayores, mayor)
ADV   Adverbs (muy, demasiado, cå«omo)
ALFP  Plural letter of the alphabet (As/Aes, bes)
ALFS   Singular letter of the alphabet (A, b)
CARD  Cardinals
CC    Coordinating conjunction (y, o)
CCAD  Adversative coordinating conjunction (pero)
CCNEG Negative coordinating conjunction (ni)
CODE  Alphanumeric code
CQUE   que (as conjunction)
CSUBF  Subordinating conjunction that introduces finite clauses (apenas)
CSUBI Subordinating conjunction that introduces infinite clauses (al)
CSUBX  Subordinating conjunction underspecified for subord-type (aunque)
DM     Demonstrative pronouns (å«esas, å«ese, esta)
FO     Formula
INT    Interrogative pronouns (quiå«enes, cuå«antas, cuå«anto)
ITJN   Interjection (oh, ja)
NC     Common nouns (mesas, mesa, libro, ordenador)
NP     Proper nouns
NEG   Negation
ORD   Ordinals (primer, primeras, primera)
PAL    Portmanteau word formed by a and el
PDEL  Portmanteau word formed by de and el
PE    Foreign word
PNC   Unclassified word
PPX    Clitics and personal pronouns (nos, me, nosotras, te)
PPO    Possessive pronouns (tuyas, tuya)
PPO   Possessive pronouns (tuyas, tuya)
PREP  Preposition
PREP   Negative preposition (sin)
QU    Quantifiers (sendas, cada)
REL    Relative pronouns (cuyas, cuyo)
SE     Se (as particle)
VEfin  Verb estar. Finite
Veger Verb estar. Gerund
Veinf Verb estar. Infinitive
VE    Verb estar. Past participle
VHfin  Verb haber. Finite
VHger  Verb haber. Gerund
Vhinf Verb haber. Infinitive
```

```
VHadj  Verb haber. Past participle
VLfin  Lexical verb. Finite
Vlger Lexical verb. Gerund
VLinf  Lexical verb. Infinitive
Vladj Lexical verb. Past participle
Vmfin Modal verb. Finite
Vmger Modal verb. Gerund
Vminf Modal verb. Infinitive
VM     Modal verb. Past participle
Vsfin Verb ser. Finite
Vsger Verb ser. Gerund
VSinf  Verb ser. Infinitive
VS     Verb ser. Past participle
```

Cette table des codes est ainsi interprétée dans le menu grammatical d'HypEspag :

Le menu grammatical dans la version espagnole



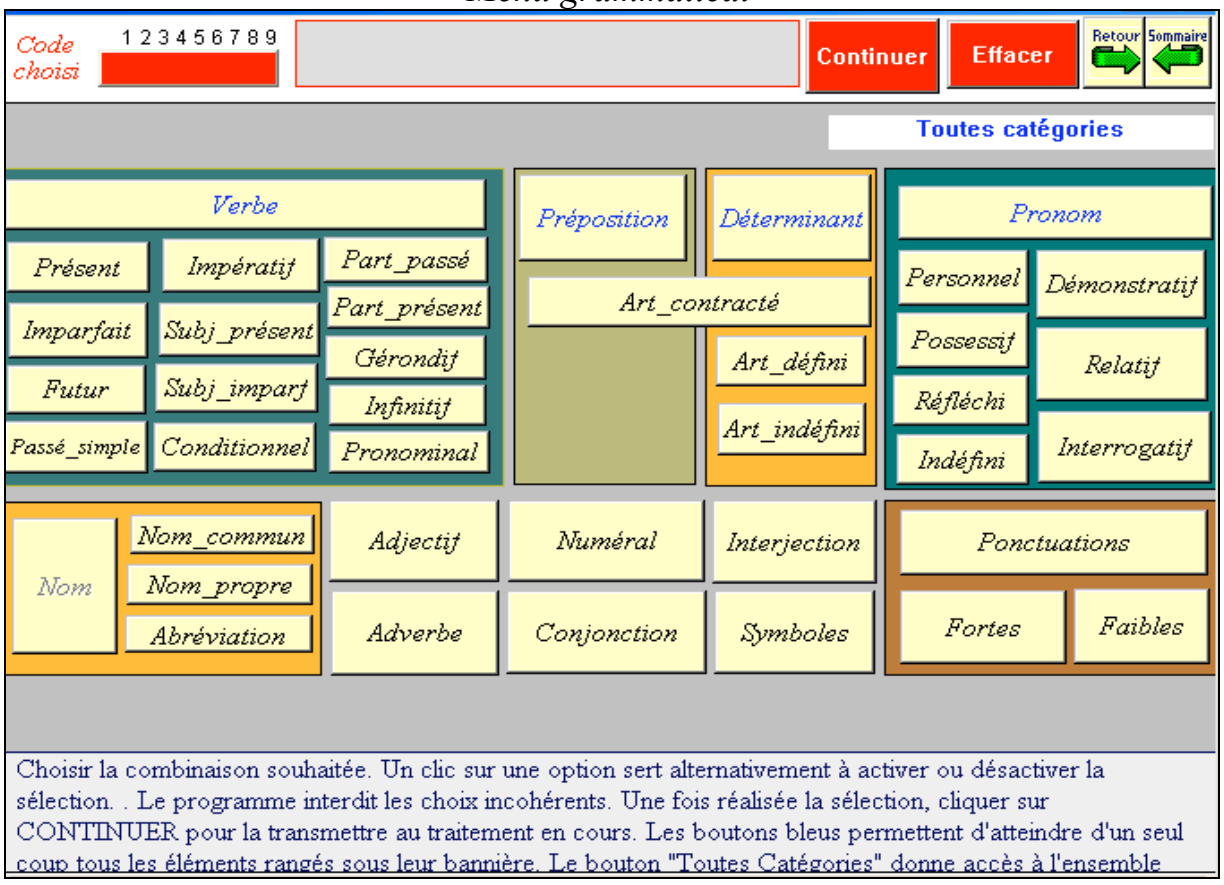
LA VERSION ITALIENNE : HyperIta.EXE

Table des codes

(Copyright Prof. Achim Stein, University of Stuttgart)

ABR abbreviation
ADJ adjective
ADV adverb
CON conjunction
DET:def definite article
DET:indef indefinite article
INT interjection
NOM noun
NPR name
NUM numeral
PON punctuation
PRE preposition
PRE:det preposition+article
PRO pronoun
PRO:demo demonstrative pronoun
PRO:indef indefinite pronoun
PRO:inter interrogative pronoun
PRO:pers personal pronoun
PRO:poss possessive pronoun
PRO:refl reflexive pronoun
PRO:rela relative pronoun
SENT sentence marker
SYM symbol
VER:cimp verb conjunctive imperfect
VER:cond verb conditional
VER:cpre verb conjunctive present
VER:futu verb future tense
VER:geru verb gerund
VER:impe verb imperative
VER:impf verb imperfect
VER:infi verb infinitive
VER:pper verb participle perfect
VER:ppre verb participle present
VER:pres verb present
VER:refl:infi verb reflexive infinitive
VER:remo verb simple past

Menu grammatical

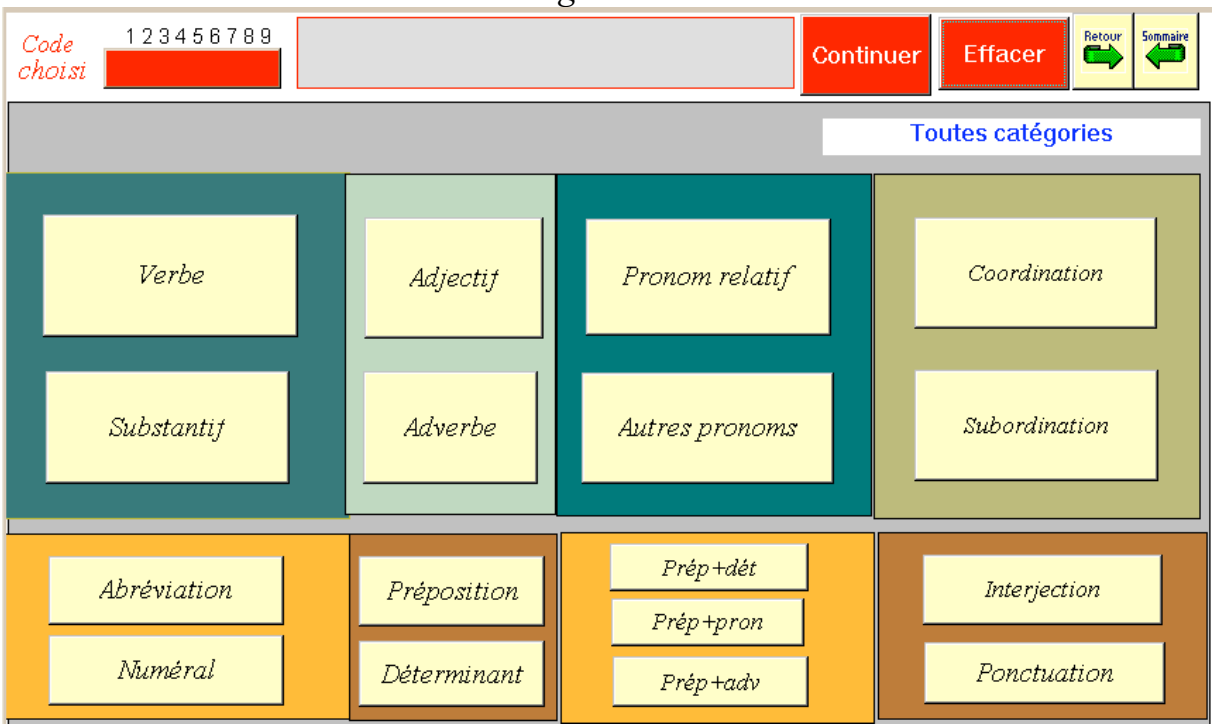


LA VERSION PORTUGAISE : HyperPor.EXE

Table des codes et fréquence de chacun dans un texte juridique

4158 adj
1210 adv
1751 card
2253 coor
3433 det
12 etrg
6 inj
11444 nom
852 prel
3049 prep
1069 pron
1 prpadv
3817 prpdet
227 prpp
56 sub
5441 verb

Menu grammatical



CHAPITRE 6.
LES MONOGRAPHIES

On distinguera les modèles, qui ne contiennent que les programmes, sans les données, et les bases exploitables, pourvues de données et riches de résultats. On vient de passer en revue les premières, soit HYPERBAS.EXE, HYPERCOR.EXE, HYPERTAG.EXE et HYPERVER.EXE pour le français et HYPERANG.EXE, HYPERGER.EXE , HYPERITA.EXE, HYPERPOR.EXE et HYPESPAG.EXE pour les langues étrangères. Reste à situer les secondes.

Ces bases, réunies sur la surface du DVD, ont été constituées à l'aide de la présente version 8 d'Hyperbase. Elles sont lemmatisées (par Cordial ou TreeTagger), sauf la base EXAMPLE.EXE qui nous a servi à illustrer la version standard. Cette base EXAMPLE a été prévue pour faciliter l'apprentissage du logiciel. Mais son jeu de données (22 oeuvres romanesques de Marivaux à Proust) est resté limité, afin de ne pas encombrer le disque dur de l'utilisateur, où doivent trouver place aussi des programmes DLL, des fichiers d'aide et quelques listes bibliographiques ou lexicologiques. Pour la maîtrise des bases lemmatisées on a fourni également un exemple, tantôt GAULLE.EXE, tantôt PROUST.EXE, suivant que l'utilisation attendue était d'ordre historique, sociologique ou littéraire.

Le programme de choix: MENU.EXE

Un clic sur l'un des auteurs du programme MENU.EXE (ci-dessus) convoque la base correspondante sur le disque dur, dans le répertoire C:/HYPERBAS/ (en cas d'absence cette base y est automatiquement tranférée à partir du DVD, en même temps que les fichiers associés qui partagent le même nom).

Le menu MENU.EXE est commode pour passer d'une base à l'autre. Mais on peut aussi atteindre chaque base directement, soit en cliquant sur une base éxécutable du DVD ayant le suffixe.EXE (il y a alors tranfert sur C:/HYPERBAS/), soit en sollicitant une base du disque dur, si le transfert a déjà eu lieu. Dans tous les cas c'est la base du répertoire C:/HYPERBAS/ qui est mise en oeuvre et qui seule permet l'écriture.

Certains textes ont été saisis par nos soins. Mais la plupart viennent des sources anciennes ou modernes. Parmi les premières on citera *FRANTEXT* (principalement le théâtre classique). Parmi les fournisseurs actuels, les plus sollicités ont été *Gallica*, le *Projet Gutenberg*, *Poesies.net* et *Wikipedia*, et dans une moindre mesure *ABU*. Mais il existe aussi des sites spécialisés, qui distribuent généreusement le texte même des oeuvres de l'écrivain considéré

(par exemple la totalité des textes de Balzac vient d'une collaboration avec le Professeur Kazuo Kiriu)..

Afin d'unifier les conventions de saisie et de permettre les comparaisons, nous avons souvent filtré les textes reçus. En particulier les ponctuations ont bénéficié d'un traitement commun. De même à l'intérieur d'une même monographie les options de lemmatisation restent constantes et l'on n'a pas mêlé les résultats de lemmatiseurs différents.

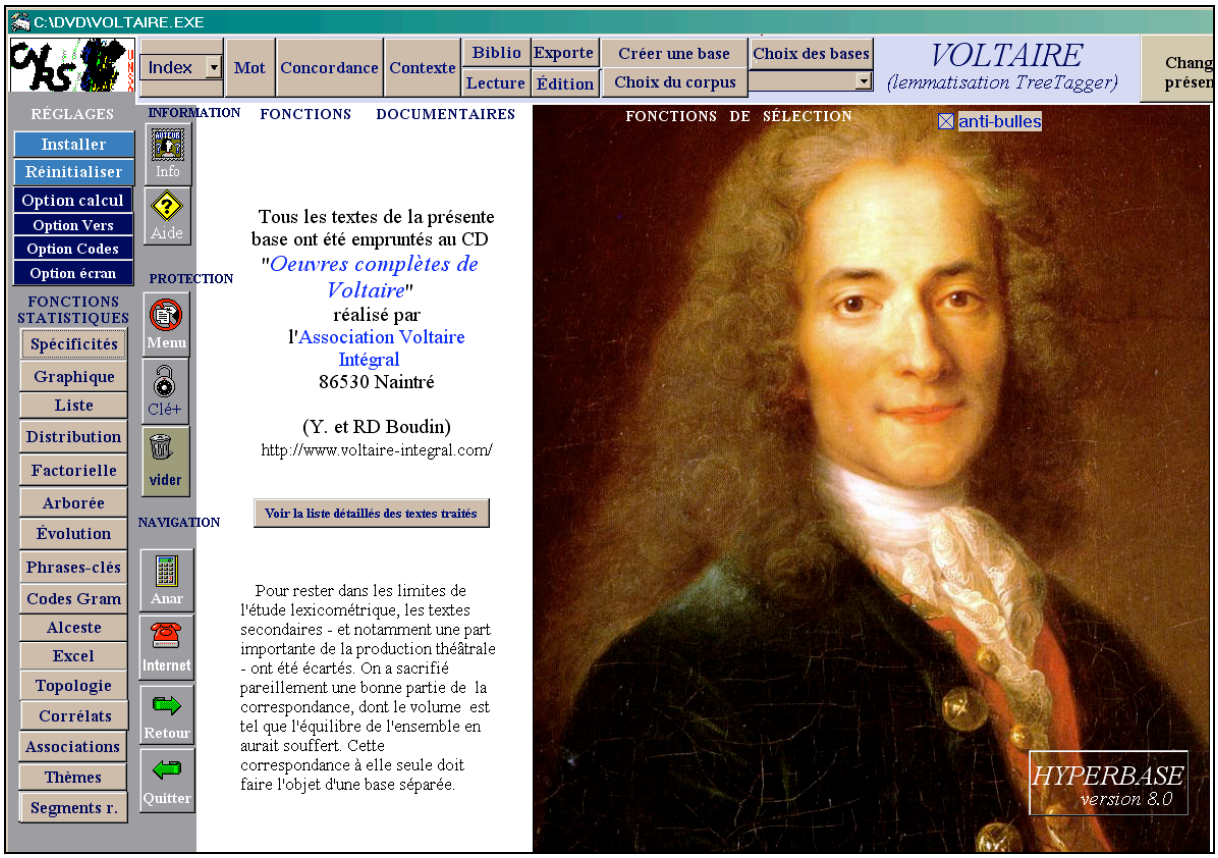
Certaines bases ont été constituées anciennement à un moment où les textes numérisés étaient rares. Il y a alors des lacunes regrettables dans le corpus. En puisant dans les ressources d'Internet, on se propose d'en faire ultérieurement une refonte. Mais l'exhaustivité n'est pas nécessairement requise quand on poursuit une étude lexicométrique, la statistique étant une science apte à combler les trous.Il arrive même que l'abondance et la multiplicité des textes soient un obstacle à leur exploitation. Ainsi alors que nous disposions des oeuvres complètes de Voltaire, patiemment réunies par Y. et R.D. Boudin, nous avons décidé d'écarter bon nombre de ses pièces et la masse écrasante de sa correspondance.

Les bases livrées sur ce DVD ne concernent que les textes du domaine public. Elles n'intéressent que la littérature, de Rabelais à Proust. Le XVIe siècle y est peu représenté, à cause de l'inconstance de l'orthographe et de l'inaptitude des lemmatiseurs à traiter les textes plus anciens (Rabelais et Montaigne figurent ici dans une transcription moderne). Quant à l'époque contemporaine, qui intéresse vivement les chercheurs, elle a fait l'objet d'une exploitation intense, comme en témoigne la dernière rangée de noms, qui va de Breton à Le Clézio. Malheureusement le copyright s'oppose à une divulgation incontrôlée. Ces bases peuvent cependant être communiquées, à titre personnel et dans un but de recherche, si les garanties de confidentialité sont données.

L'ensemble des bases rassemblées ici représente plus d'un millier de textes et quelque 60 millions de mots. Il y en a autant et même davantage dans deux bases particulières, accessibles au bas de l'écran et issues de Frantext. L'une, grosse de 117 millions de mots, rend compte de l'évolution de la littérature, à travers quinze tranches chronologiques découpées dans la production littéraire de cinq siècles. L'autre (55 millions de mots) envisage le même ensemble en s'intéressant individuellement aux auteurs, et plus précisément aux 70 écrivains les mieux représentés dans Frantext. Prendre garde que ces deux bases sont uniquement statistiques et que le le texte en est absent. Toutes les autres bases permettent le retour au texte et l'exploration documentaire aussi bien que l'exploitation statistique.

On a représenté ci-dessous le menu principal de la base Voltaire qui comprend 52 textes et plus d'un million de mots.

La base Voltaire.



Celle de Rousseau compte 1,5 million d'occurrences. Le texte en a été emprunté au site ATHENA (http://un2sg4.unige.ch/rousseau/rousseau.html), créé par Pierre Perroud.

Celle de Maupassant (voir ci-dessous) est plus importante encore (1,7 million d'occurrences) et compte 38 sous-ensembles, dont 24 recueils de contes, 8 romans, le reste étant constitué de chroniques, de récits de voyage, de correspondance et d'oeuvres en vers. Les textes ont été réunis par Thierry Selva et sont disponibles sur la toile à l'adresse:

http://lib.univ-fcomte.fr/PEOPLE/selva/maupassant/

Les données sur La Fontaine viennent du site http://www.lafontaine.net créé par Marc Bassetti. Quant au théâtre classique, la source remonte loin dans le passé, à la saisie faite dans les années 60 par l'équipe de Besançon et dirigée par Bernard Quemada. Charles Bernet, de l'Institut National de la langue française, en a assuré le contrôle, la correction et le codage.

CHAPITRE 6.
LES BASES TRANSVERSALES

La plupart des bases proposées sont des monographies, qui sont consacrées à un auteur unique et tendent à la représentativité, sinon à l'exhaustivité. Mais il en est d'autres qui visent un objet plus composite, où la variable isolée peut être le genre, l'auteur, le temps ou le lieu.

LES GENRES LITTÉRAIRES

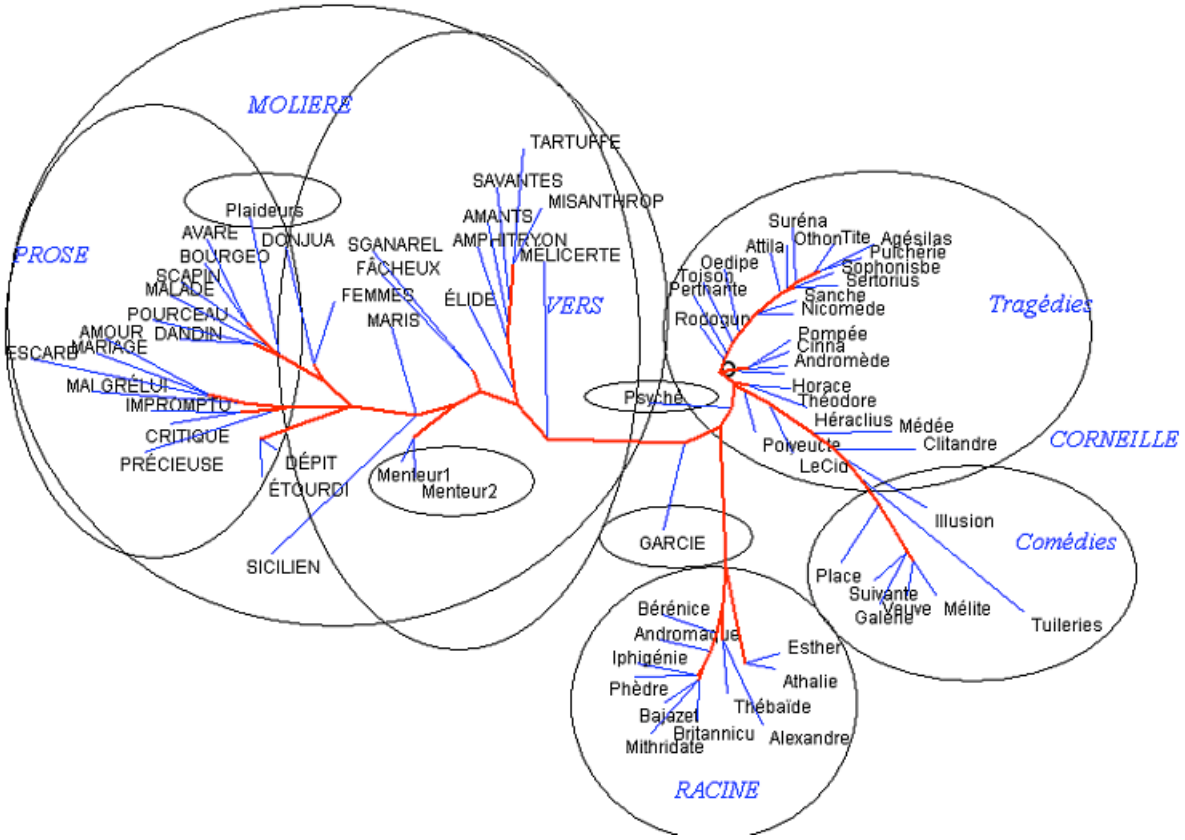
La statistique est plus sûre quand s'exerce la loi des grands nombres et quand le calcul porte sur des corpus plus étendus, là où précisément la mémoire humaine rencontre des limites. Parmi les trois ou quatre variables qui peuvent ainsi être étudiées, la première porte sur le genre. En s'appuyant sur les données de l'Institut National de la langue française, on pourrait constituer une base romanesque, une base de poésie, une base de théâtre, etc... Afin de donner un aperçu de cette exploitation des genres, on a constitué en corpus les textes du théâtre classique (soit 75 au total, en cumulant les pièces de Corneille, Racine et Molière.

L'analyse arborée, fondée sur la connexion lexicale de Muller et représentée ci-dessous en est une illustration. Elle analyse la distance qui sépare chaque texte de tous les autres, sans que l'analyse prenne en compte la signature, le genre ou l'époque. Or au bout du traitementb les 75 pièces apparaissent classées par auteurs, Racine en bas à droite, Molière à gauche, et Corneille au centre. Et pour chaque auteur la chronologie et le genre permettent un sous-classement. Chez Racine les premières pièces et les dernières se détachent du reste. Chez Corneille les comédies, se distinguent des tragédies. Chez Molière les pièces en vers précèdent les pièces en prose. Si l'on est aveugle aux auteurs, le paysage est encore lisible : à droite c'est la tragédie, à gauche la comédie. À droite c'est le domaine du vers, à gauche celui de la prose. Et si l'on croise genre et versification, on a une progression très claire : tragédie en vers, puis comédie en vers, puis comédie en prose.

Il y a pourtant trois ou quatre points où l'harmonie des genres et des signatures ne règne plus. Qu'arrive-t-il lorsqu'il y a conflit ? C'est le genre qui prévaut. Ce qui ne va pas à l'encontre des prérogatives de l'écrivain, puisque

c'est lui qui choisit librement le genre, même s'il n'a pas toute liberté pour en modifier les lois. Les points de divergence sont très localisés dans le graphique : au centre, près des tragédies on trouve une pièce de Molière, *Don Garcie de Navarre ou le Prince jaloux*. Quoique le sous-titre y soit ambigu ("comédie héroïque"), il s'agit d'une pièce sérieuse où Molière jouait un rôle tragique. Ce fut un échec, pour l'acteur comme pour l'auteur, et Molière se le tint pour dit. À l'opposé, parmi les comédies de Molière, on relève une pièce étrangère, les *Plaideurs*, qui est de Racine, et la seule comédie que Racine ait écrite. Le genre lui a dicté sa place, sans contestation. Enfin reste à considérer le cas des deux *Menteur*. Ils prennent place parmi les premières comédies de Molière et surtout celles qui sont écrites en vers. Là encore le genre a parlé. S'y ajoutent les contraintes de la versification, et - plus faiblement - un rapprochement chronologique, les dernières comédies de Corneille n'étant guère antérieures aux premières de Molière.

Analyse arborée du théâtre classique, à partir de la connexion lexicale



On restera sur ce constat, qui n'est pas le premier où l'on constate la force du genre. Il y a vingt ans Muller avait proposé une expérience de laboratoire qui consistait à mêler les écrivains et les genres. On avait relevé une liste de mots grammaticaux dans des œuvres poétiques, théâtrales et romanesques de trois écrivains de la même école et l'analyse était conviée à un problème d'attribution. Elle en trouva trois : un poète qui avait écrit les *Contemplations*, les *Méditations*

et les *Nuits*, et pareillement un romancier et un dramaturge. Le genre s'était interposé devant l'écrivain[26].

L'examen statistique, s'il est pratiqué de bonne foi et avec de bons outils, comme la connexion lexicale de Muller, s'inscrit donc en faux contre la thèse de Pierre Louÿs, bien mal épaulée par Labbé. Au reste dans ces questions historiques, si la statistique peut fournir des indices et même des présomptions, elle ne peut produire des preuves au même titre que la philologie et l'histoire littéraire. Muller il y a quarante ans avait prévu et prévenu ces imprudences dans la conclusion prémonitoire de sa thèse : "Nous avions déclaré d'emblée que cette œuvre ne pose pas de problèmes philologiques importants, et que notre étude ne promettait ni révélations ni solutions inédites. Ne serait-elle pas de nature, plutôt, à mettre en garde ceux qui, en l'absence de renseignements historiques, attendent de la statistique lexicale des certitudes en matière de datation et d'attribution ?" [27]

LES ÉPOQUES. LA BASE *CHRONO*

Quand on veut suivre l'évolution d'un mot à travers le temps, *Frantext* offre une assise très solide pour ce type de recherche historique. Mais l'abondance des textes peut nuire à leur homogénéité. Trop de textes, trop de genres s'y trouvent mêlés, et l'évolution, si on la constate, doit pouvoir être isolée des influences parasites. On a rendu le corpus plus homogène en écartant les textes techniques pour ne conserver que la littérature. Cela représente encore une masse énorme, de 117 millions d'occurrences. Le corpus constitué est celui qui était disponible sur le réseau en 1996, au moment où cette base chronologique a été réalisée, sous le nom de *THIEF* (*Tools for Helping Interrogation and Exploitation of Frantext*). La présente base est extraite de *Thief*, dont elle a repris les fonctions *off line* (mais la consultation *on line* de *Frantext* est ménagée par le bouton INTERNET).

Précisons que cette base transversale et synthétique, pour des raisons de copyright, est dénuée de texte et que les recherches documentaires n'y ont pas cours. Les informations qu'elle donne sont seulement comparatives et quantitatives.

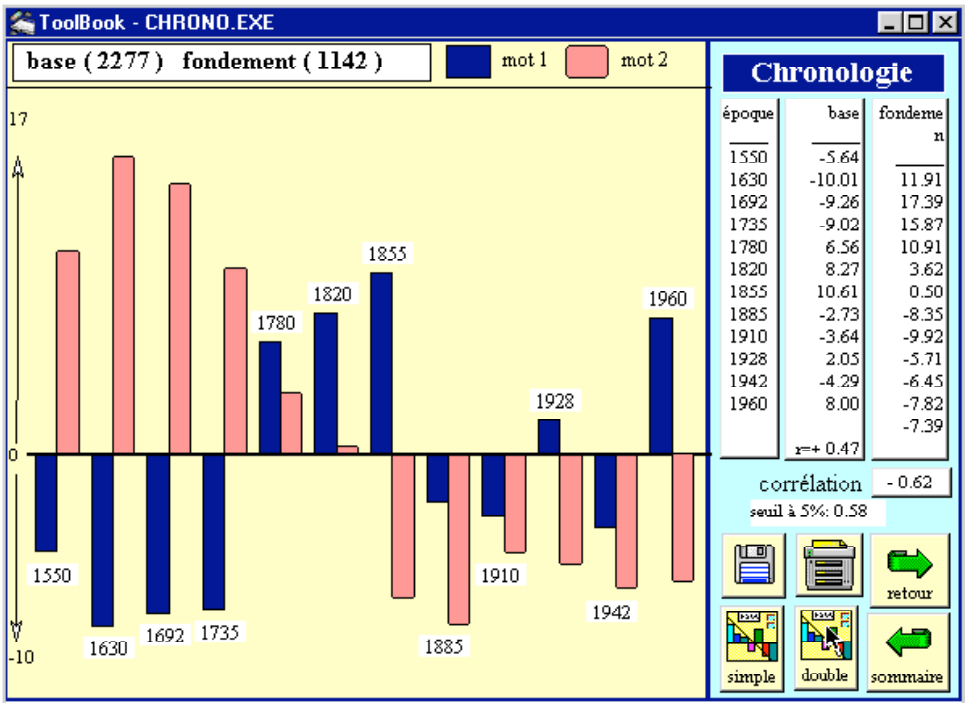
La figure ci-dessous illustre ce qu'on peut attendre de cette base quand on croise deux mots, ici les mots *base* et *fondement*, dont la fortune historique s'oriente inversement.

---

[26] Brunet É & Muller Ch. (1988). "La statistique résout-elle les problèmes d'attribution?", *Strumenti critici* III,3, p.367-387.

[27] Muller Ch. (1967*). Étude de statistique lexicale*, op.cit. p.263.

Évolution opposée des mots base *et* fondement



LES ÉCRIVAINS. LA BASE *AUTEURS*

On compte des centaines d'écrivains dans *Frantext*. Là encore c'est à *Frantext* qu'il faut s'adresser si l'on veut avoir une vue d'ensemble du paysage littéraire, région par région, auteur par auteur. Comme on ne traite ici que des données quantitatives, le copyright n'est pas embarrassant et les écrivains modernes ont été pris en compte, jusqu'à Gracq. On a ainsi réuni 70 écrivains sans dépasser le 17e siècle (le premier de la liste est Honoré d'Urfé). Car la base *Chrono* nous a averti que l'instabilité de l'orthographe au 16e siècle rendait les comparaisons délicates, d'autant que dans Frantext les textes anciens ne sont pas souvent modernisés.

La liste des auteurs retenus est affichée dans le menu principal de la base (voir ci-dessous). Un clic sur l'un d'entre eux fait apparaître la fiche bibliographique qui le concerne, à l'image de celle de Giono, qui est livrée ici et qui fait regretter un choix trop mesquin pour un auteur qui ne l'est pas.

La base Auteurs *(56 millions d'occurrences)*

La présente base est purement quantitative. Elle rend compte de l'usage comparé des mots chez les grands écrivains de notre littérature.Pour avoir accès aux textes qui ont servi à constituer cette base, consulter FRANTEXT.
LISTE des textes de Giono :

K629 GIONO.J/COLLINE/1929
PROSE,ROMAN
PARIS : GRASSET, 1929.

K628 GIONO.J/UN DE BAUMUGNES/1929
PROSE,ROMAN
PARIS : GRASSET, 1929.

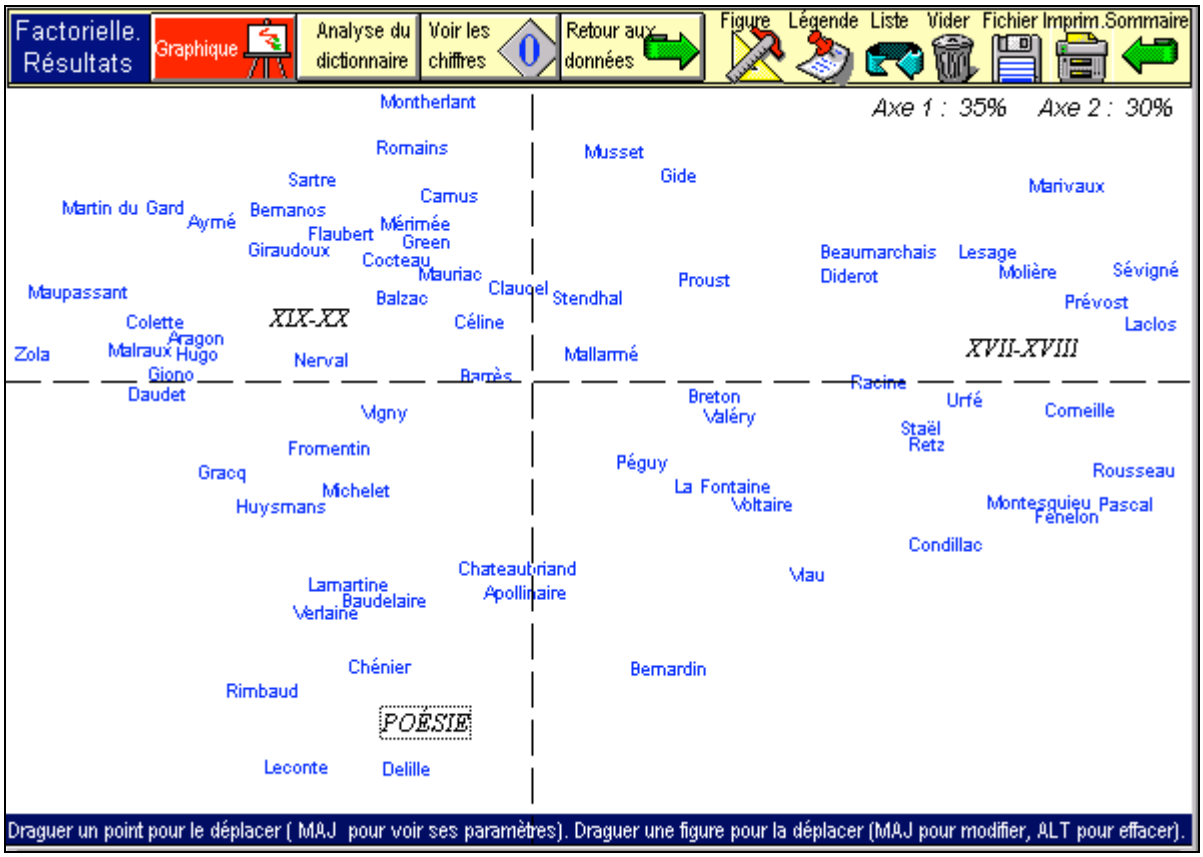
K633 GIONO.J/REGAIN/1930
PROSE,ROMAN
PARIS : GRASSET, 1937.

K632 GIONO.J/LE GRAND TROUPEAU/1931
PROSE,ROMAN
PARIS : GALLIMARD, 1931.

Les textes de Giono présents dans la base

Au total cette base *Auteurs* enveloppe un corpus considérable de 56 millions de mots (et 236 000 formes différentes). Au risque de la déflorer, on en illustrera la richesse en dressant la carte des écrivains selon la distance lexicale où chacun s'établit au regard des autres. Ce calcul de "connexion lexicale" est le même que celui du théâtre classique représenté plus haut. Son interprétation,

alors que tous les mots sont entrés dans le calcul, offre la même lisibilité: c'est le temps qui parcourt l'espace de droite à gauche (du XVIIe au XXe siècle) et c'est le genre qui oppose le haut (la prose) et le bas (la poésie).



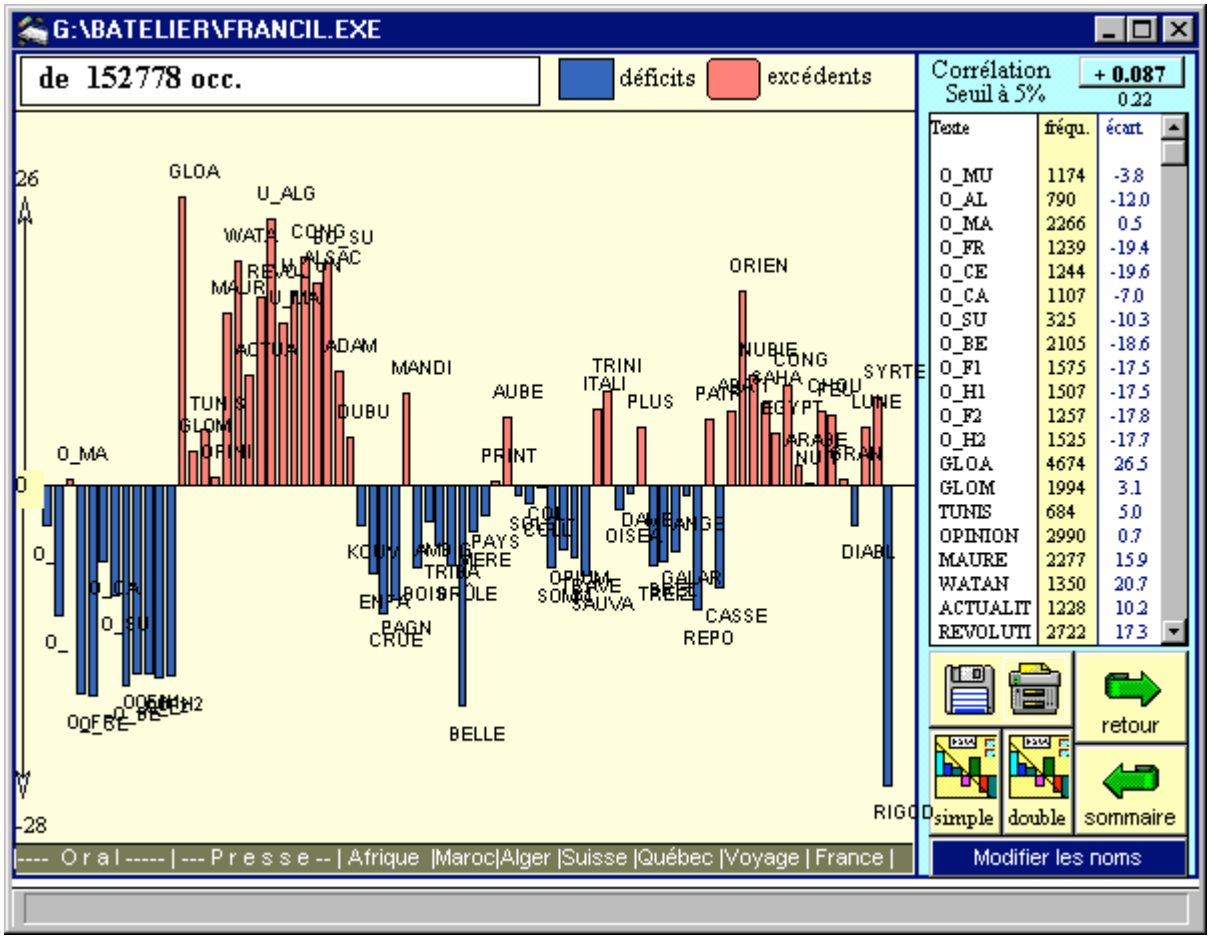
Analyse factorielle de la distance lexicale dans la base Auteurs

LIEUX ET MILIEUX. LA BASE *FRANCIL*

Une dernière base enfin est proposée pour donner réponse aux questions que l'on peut se poser à propos des diverses variétés du français et des variables mises en jeu dans son exercice. Ces variables divergent autant que les populations qui partagent l'usage du français et que divers facteurs peuvent opposer: l'espace géographique, le temps historique, les conditions sociologiques, l'environnement économique, politique et culturel, sans compter le tempérament, les goûts et les choix personnels des écrivains. S'y superposent les variables proprement linguistiques qui opposent l'oral à l'écrit, l'information à la fiction, l'utilitaire au littéraire. Pour tenter de maîtriser toutes ces variables et en dénouer les fils invisibles et entremêlés, on a étendu le champ de l'enquête à l'ensemble de la francophonie. Les données ont été empruntées aux observatoires du français établis au Québec (*Québétext* et *Catiq*), en Belgique (*Valibel* et *Beltext*), en Suisse (*Suistext*), en France et en Afrique (*Gars* et

Frantext). Certaines données relatives à la presse ou à l'oral ont été recueillies expressément en vue de ce programme *Uref/Aupelf*. Au total le corpus recouvre 4,5 millions d'occurrences, où le français parlé est confronté à l'écrit, le littéraire à l'utilitaire, et le nord au sud.

Pour illustrer la complexité des interactions qui animent le langage, on représente ci-dessous la courbe obtenue pour une articulation syntaxique du discours, la préposition *de*.



La distribution de la préposition de *dans la base* Francil

Cette préposition (et ses variantes contractées) a partie liée avec le discours informatif et accumule ses emplois dans la presse, mais aussi là où l'information prend la forme adoucie de la description, dans les récits de voyage. Ici la variable géographique est atténuée mais on peut déceler ses effets quand le genre est constant, dans le roman (partie droite du graphique): les écrivains suisses ou français utilisent plus volontiers cet outil grammatical que ceux de l'Afrique, du Maghreb ou du Québec. On pourrait comparer cette distribution à celle de *que* (tous emplois réunis: subordonnant, relatif, interrogatif, adverbial). L'histogramme, inversement, invite à voir dans *que* une caractéristique de l'oralité, mais il semble aussi que cette structure soit plus fréquente dans l'usage maghrébin et africain, même là où l'oral n'est pas en cause. Ces deux témoignages choisis parmi des milliers d'autres montrent que les variables en jeu

peuvent s'ajouter, se neutraliser, ou rester indépendantes. Comme la mer où s'exerce la force conjuguée ou contrariée des vents, des courants, des marées et des obstacles, le langage obéit à la mécanique des fluides, et la statistique a fort à faire pour en rendre compte.

Troisième partie

HYPERBASE pour Macintosh

version standard 5.0

On a expliqué précédemment que le développement du logiciel Hyperbase est rendu difficile sur le standard Apple, maintenant que le système MacOs X ne reconnaît plus l'environnement Hypercard, pourtant créé par Apple. Cela n'empêche pas Hyperbase de fonctionner sur les machines Apple si elles utilisent la version 9 du système, ou même si elles tournent sous MacOs X, pourvu que l'émulation Classic soit disponible.

Mais on a renoncé à développer la dernière version 5.0 qui date de 2003. Le manuel correspondant n'a donc pas subi de modification et on n'a pas jugé utile, par économie, de le reproduire de nouveau sur papier. Ceux qui en auraient besoin le trouveront sur le cédérom.

Appendice

Lemmatisation TreeTagger sous Windows 7

Le lemmatiseur TreeTagger est très largement répandu et librement téléchargeable. Outre ces avantages il offre une grande simplicité d'emploi, une fois installé. Les fichiers résultats qu'il produit sont facilement exploitables, une ligne étant consacrée à chaque mot traité, avec trois champs : la graphie, le lemme et le code grammatical. Et surtout fondé sur un apprentissage de nature statistique, et non sur des règles linguistiques propres à chaque langue, il est apte à traiter la plupart des langues occidentales. C'est pourquoi nous avons généralisé son emploi dans Hyperbase, au moins pour le français, l'anglais, l'espagnol, l'allemand et l'italien. D'autres lemmatiseurs ne sont proposés que pour le français (Cordial) et le latin (LASLA).

Malheureusement TreeTagger créé initialement par Helmut Schmitt (Université de Stuttgart) n'a pas suivi instantanément l'évolution des systèmes Windows et se trouve actuellement incompatible avec Vista et System 7. Il y a donc lieu d'utiliser le system XP pour préparer les données et les soumettre à la lemmatisation de Treetagger. On a envisagé et testé trois façons de contourner l'obstacle :

Première procédure

Confier provisoirement à une machine XP la création d'une nouvelle base. Une fois que la base est constituée, la transférer sur une machine Vista ou Windows 7, dans le dossier C:\HYPERBAS\. Prendre garde à déporter non seulement la base elle-même (avec suffixe .TBK), mais aussi les index ou fichiers qui sont associés et qui portent le même nom avec d'autres suffixes, notamment .TXT et .TX2. Naturellement cela suppose que l'installation d'Hyperbase soit faite également sur la machine XP, et sur la machine Vista ou Windows 7.
Le programme d'installation n'a pas de limitation et peut être lancé sur des systèmes différents autant de fois que nécessaire.

Deuxième procédure

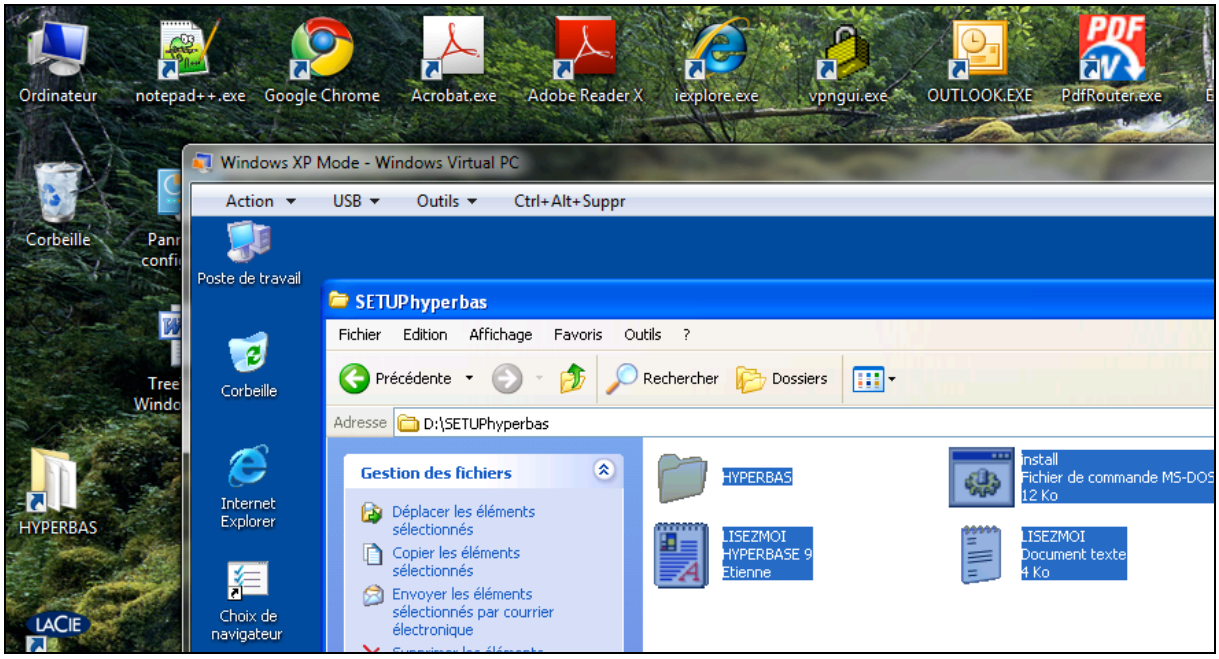
On peut se contenter de confier à une machine extérieure sous XP la phase seule du traitement où s'opèrent la lemmatisation et le recours à TreeTagger. Dans cette étape préalable, les fichiers de données (avec suffixe .TXT) sont traités successivement par TreeTagger en donnant naissance à des fichiers lemmatisés

portant le même nom (avec suffixe .CNR). Ces fichiers CNR doivent alors être transmis au répertoire C:\HYPERBAS, où Windows 7 (ou Vista) les retrouvera quand la création de la base sera lancée. La bifurcation dans le traitement se fait grâce à un dialogue où l'utilisateur est invité à choisir entre deux options de lemmatisation :
"Traiter des fichiers lemmatisés"
"Lemmatiser à la volée"
On choisira la première option, en désignant les différents fichiers CNR qu'il faut prendre en compte.
Cette procédure apparemment plus légère est en réalité plus délicate et plus fragile. Car HYPERBASE n'est plus invité à contrôler les paramètres de TreeTagger, ni pour la segmentation en mots ni pour le sort réservé aux mots inconnus. La segmentation dépend du statut des ponctutations, de l'apostrophe et généralement des délimiteurs, ce statut variant d'une langue à l'autre. Treetagger impose un prétraitement du texte, généralement confié à un automate écrit en langage PERL. D'autre part il laisse la possibilité de laisser les lemmes inconnus du modèle sous l'étiquette UNKNOWN, ce qui n'est pas la meilleure solution. Il vaut mieux associer un lemme identique à la graphie quand la graphie n'est pas reconnue, par exemple dans le cas des noms propres. En pareille situation HYPERBASE corrige la proposition, s'il trouve des UNKNOWN intempestifs. Quand Hyperbase est le maître d'œuvre il dirige aussi les automates de prétraitement. Mais il reste impuissant si la segmentation a été réalisée sans pertinence et sans son concours.

Troisième procédure

Windows a prévu un moyen de pourvoir à l'incompatibilité que certains logiciels opposent, soit à Vista, soit à Windows 7, soit aux systèmes 64 bits. Apple, au moment du passage à OS X, avait proposé avec CLASSIC un moyen provisoire d'émuler l'ancien système 9 et d'utiliser les logiciels non encore convertis. De la même façon Windows 7 propose de faire un pas en arrière pour accompagner ceux qui sont en retard. Pour ce faire, il faut avoir recours à l'émulateur, qui se trouve caché et disponible sous le nom « Windows XP mode ». Lancer DEMARRER au bas de l'écran à gauche, puis TOUS LES PROGRAMMES, activer l'ascenseur de la liste et cliquer sur WINDOWS VISUAL PC et enfin sur WINDOWS XP MODE. La machine virtuelle XP apparaît alors en surimpression sur l'écran de Windows 7. Reste à installer HYPERBASE sur cette machine virtuelle, à partir du DVD original. Il est inutile de transférer la totalité des bases et l'option 1 de l'installateur suffit (SETUPmin.exe).

Installation d'Hyperbase sur une machine virtuelle en mode XP



Une fois Hyperbase réinstallé dans ce nouvel environnement, commencer par introduire les fichiers de données dans le répertoire C:\HYPERBAS\ de la machine XP. L'échange de fichiers entre les deux environnements peut se faire via le presse-papier, voire même par le glisser-déposer.
On doit alors choisir et lancer la version qui convient aux données : HYPERang pour l'anglais, HYPERtag, pour le français, HYPEspag pour l'espagnol, HYPERpor pour le portugais, HYPERita pour l'italien. Il n'y a pas lieu de lancer HYPERBAS qui ne fait pas appel à TreeTagger et qui fonctionne sans problème sous Windows 7, non plus que HYPERCOR, qui utilise le lemmatiseur Cordial, déjà adapté à Windows 7.

Le déroulement du programme n'appelle pas de commentaires particuliers. On choisira bien entendu l'option « lemmatiser à la volée » quand l'invitation sera faite de procéder à la lemmatisation.

Lorsque le processus parvient à son terme et que la base est réalisée, son exploitation peut alors être délocalisée et revenir dans le giron de Windows 7. Comme dans la première procédure précédemment évoquée, on doit rapatrier la base et ses fichiers associés dans le répertoire C:\HYPERBAS\ de Windows 7, ce qui se fait facilement via le presse-papier pour par glisser-déposer.

Cette troisième procédure n'est pas la plus simple à réaliser, mais c'est la seule qui soit autonome et ne fasse pas appel à l'extérieur. Il reste à espérer qu'elle soit provisoire et que les auteurs ou responsables de TreeTagger procèderont sans tarder à la mise à jour de leur logiciel.

TABLE DES MATIÈRES

PREMIÈRE PARTIE — HYPERBASE standard pour Windows