JEAN-BENOIT GUINOT

LE TRAITEMENT DE TEXTES DE FLAUBERT

V bis

Ils passèrent commande à Dumouchel d'un ordinateur, identique à celui qu'ils avaient vu chez M. de Faverges. Si personne n'aimait la Littérature, ils allaient, eux, lui faire rendre gorge.

L'objet arriva, accompagné de quelques traités techniques, qu'ils écartèrent. Ils préférèrent se risquer tout de suite sur le Réseau, pour des navigations sans fin, où ils se perdaient.

– Sais-tu que je suis cité deux cents seize fois dans *le Trésor de la Langue Française Informatisé* ? demandait Pécuchet.

Bouvard ne l'était que deux fois, et dix-huit dans les exemples fournis par *Frantext*. Il en conçut un peu de rancune.

– Ton dictionnaire ne sait même pas traiter correctement les caractères accentués. – Il lui montrait *polysémie, métaphore*, transformés en *polysïf83mie, mïf83taphore*, empêchant ainsi toute recherche d'aboutir.

– Et Flaubert ? Combien de fois est-il cité ?

Le logiciel ouvrait une fenêtre spéciale pour préciser que *Flaubert est identique à Flaubert ,* et que *Flaubert est donc apparenté à Flaubert,* puis en restait là. Ce résultat leur parut bête.

Un jour, ils découvrirent le site de *l'Association des Bibliophiles Universels* et téléchargèrent le texte de *Madame Bovary.* – Nous allons pouvoir le disséquer, et sans doute le comprendre.

Ils demandèrent à *Word* une analyse statistique – il fallait d'abord vérifier l'orthographe, la grammaire, et refuser toutes les propositions saugrenues.

Il y avait 113 923 mots en 21 941 phrases ; la longueur de mot était de 4,7 (lettres, sans doute); la longueur de phrase de 5,2 (mots, sûrement) ; les substantifs monosémiques représentaient 70 % ; les *communs abstraits* 57,9 % ; les *communs concrets* 42,1 %.

– Le total des *communs* fait bien 100 % ! se réjouit Bouvard.

La moitié des mots étaient signifiants, dont 44,9 % de substantifs, 11,1 d'adjectifs, 28,8 de verbes et 15,2 d'adverbes. Les verbes étaient à 17,5 % au présent, 77,9 à l'imparfait ou au passé, 1,7 au futur et 3 au conditionnel.

Mais ce n'était pas tout ! La machine avait effectué d'elle-même une analyse qui, après quelques observations comme « la complexité sémantique est moyenne ; le style est enlevé ; les phrases ont une structuration grammaticale simplifiée ; le texte offre un niveau

d'abstraction assez élevé », donnait un coefficient de lisibilité de 88 % et déclarait le roman du niveau cours élémentaire.

– Essayons avec *Salammbô* !

– 100 760 mots et 18644 phrases ; 68,1 % de substantifs monosémiques ; 16,2 % de noms propres. – Il n'y en a que 12,1 dans *Bovary*, lâcha Bouvard – 8,7 % de présent, 88 % d'imparfait / passé ; 1,2 % de futur et 2% de conditionnel. – On voyait bien que le roman se passait dans l'antiquité ! – Le texte était également conseillé pour le cours élémentaire.

Ils rédigèrent un programme : il fallait effectuer des traitements ordonnés, mais d'abord se constituer un Corpus

Qu'était-ce qu'un corpus ? Etait-ce une simple collection ? Ou bien fallait-il choisir d'après des critères de représentativité ? Sinclair avait distingué entre les corpus spécialisés, parallèles, de suivi, de référence et les corpus comparables. Reinert constatait trois moments logiques dans l'élaboration d'un corpus : celui de l'hypothèse ou désir de voir – posture de témoin – ; celui de l'acte ou désir de prendre – posture de l'acteur – ; et celui du constat ou désir de comprendre – posture du patient -.

Bouvard voulait étudier l'ensemble de l'œuvre de Flaubert, mais Pécuchet les limita à celles qui avaient été publiées en volume – Tant pis pour les scénarios, tant pis pour les manuscrits, d'ailleurs, à l'exception d'une petite partie de *l'Education sentimentale*, ils n'avaient pas encore été numérisés.

Ils cherchèrent les textes disponibles en ligne. Tantôt l'on ne trouvait qu'une partie des œuvres, tantôt elles étaient en mode image – impossibles à exploiter –, tantôt assorties de tant de fautes de saisie que la décence aurait dû empêcher de les publier. Nulle part n'était indiquée l'édition de référence. Il y avait tous les formats.

On trouvait partout *Bovary*, souvent *Trois contes* et l'*Education*, parfois *Salammbô* ou le livre qui parlait d'eux.

Certaines bases, comme *Frantext* ou l'*ARTFL*, qui paraissaient les plus complètes, étaient *Réservées aux Institutions Universitaires* . Pécuchet attaqua les impôts. Bouvard se demandait s'il était si difficile de séparer les textes du domaine public, afin de les rendre interrogeables par tous.

Bibliopolis vendait un CD-ROM, mais presque un an après le changement, il était toujours affiché en ancienne monnaie. Un courriel envoyé pour s'enquérir du prix en Euros n'obtint pas de réponse.

Ils se contentèrent de ce qu'ils purent glaner. D'ailleurs le même problème se posait pour les versions imprimées. Il y avait les erreurs de copistes, les lectures hasardeuses de manuscrits, les multiples éditions, dont il aurait fallu pouvoir comparer automatiquement les différences.

Fallait-il lemmatiser ? D'après Bernard, la lemmatisation des textes – qui ramène toutes les formes à leur forme canonique, telle qu'elle figure dans le dictionnaire – était depuis longtemps un point d'achoppement théorique et pratique. Brunet avait résumé la querelle et conclu que, chaque fois qu'il était possible, il était préférable de disposer d'un double réseau de décomptes, en formes graphiques et en lemmes. Il ajoutait qu'il était de toutes façons difficile de trouver un standard pour les corpus lemmatisés, chacun proposant ses propres règles. Il s'était d'abord consolé en déclarant que la statistique aimait les grands nombres et ne répugnait pas à l'impureté, avant de se raviser quelque peu pour préférer les données plus pures et plus sûres.

Il est vrai que Kastberg avait obtenu des résultats similaires en comparant une version graphique et trois versions des mêmes textes lemmatisés selon des méthodes différentes. Seuls ces derniers offraient des possibilités d'analyses morphologique et syntaxique, les corpus graphiques permettant la comparaison des travaux de différents chercheurs sur un même ensemble d'œuvres. Peu importait à Bouvard le procédé. Il voulait s'instruire, descendre plus avant dans la connaissance.

Comment fallait-il encoder les textes ? Il y avait les bons vieux formats *TXT* ou *ASCII*, assimilables par n'importe quelle machine – sauf pour les caractères accentués –, le *RTF*, et aussi *l'HTML* – simple langage descriptif de mise en page. L'unification promise par *XML* – qui n'était qu'une grammaire descriptive devant être déclinée pour chacun de ses usages – semblait encore très lointaine. La *Text Encoding Initiative* avait défini des recommandations, mais dans le cadre strict du *XML*, qui empêchait l'utilisation simple de balises chevauchantes. Le centre Kolb-Proust s'en servait néanmoins pour publier certains de ses documents.

De quelle manière fallait-il encoder, manuellement ou automatiquement ? Le codage humain pouvait sembler le plus sûr, mais Fortier avait prouvé que dans la plupart des cas, une dizaine de codeurs ne suffisaient pas pour garantir un accord statistiquement significatif entre les résultats. En cela il rejoignait Brudigou et *alii*, qui définissaient les textes comme des données, de ce fait justiciables de processus d'objectivation et de formalisation, fruits d'intentions signifiantes de la part des acteurs et objets de parcours interprétatifs de la part

des analystes. Celles-ci se retrouvaient jusque dans le choix d'un logiciel de codage et de son paramétrage.

La principale source d'erreur provenait du processus de codage lui-même, et risquait de noyer toute possibilité de certitude dans une marée d'erreurs présente dans les données de base et renforcée par l'action des procédés statistiques.

Ce défaut de certitude les contrariait.

Et puis, il aurait fallu catégoriser les données pour pouvoir les étiqueter, de manière à séparer les verbes des noms, les présents des passés, les articles des pronoms. Labbé avait brillamment plaidé pour un étiquetage massif et rigoureux, permettant, en outre, de combiner les formes graphiques et lemmatisées. On pourrait enfin disposer de corpus de référence, où le zéro faute serait impérieux.

- Dire que les anglais disposent déjà d'un tel outil, riche de 100 millions de mots, se lamentait Bouvard.

Plus sceptique que Labbé sur la possibilité d'éviter toute erreur, Hug n'en développait pas moins une méthode pour désambigüiser automatiquement les homographes verbe/nom. Dister proposait le traitement des formes homographiques en analysant leur contexte au moyen de grammaires locales

– Sais-tu qu'un tiers des mots de la langue française est constitué d'homographes : Je porte une porte ; cette bête est bête.

Même si Véronis avait montré que la technique était encore loin d'être opérationnelle, il y avait des perspectives : selon Habert, *CorTecs* pouvait aider à la correction de textes catégorisés. Chaque mot pouvait porter plusieurs étiquettes – lemme, catégorie morpho-syntaxique, catégorie sémantique –. Le travail par concordances permettait de regrouper des contextes similaires et de propager aisément des corrections. On obtenait une plus grande cohérence.

Pour l'instant, ils découvraient le corpus public sur Balzac et Rabelais constitué à partir d'*Hyperbase*. Ils savaient que se préparait une version Flaubert. Ils rêvaient aux résultats qu'allaient bientôt produire le *Dictionnaire hiérarchique,* le *Tableau de distribution des fréquences,* l'*Etendue du vocabulaire et hapax,* le *Tableau de l'accroissement lexical,* le même en *ordre inverse*, la *Distance lexicale* et son *Analyse factorielle*, l'*Histogramme des valeurs propres.*

Ils téléchargèrent la version de démonstration. Sans se laisser rebuter par l'interface fruste, ils y introduisirent le texte du *Bal à Vaubyessard*, le logiciel générant des messages d'erreur en anglais quand on voulait lui faire absorber le texte complet du roman.

Ils apprirent tout d'abord que le texte comportait 4307 mots et 1370 formes, que la *prob.p* variait de 0,04597 à 0,38356 selon les paragraphes, tandis que la *prob.q* , elle, s'étendait de 0,61644 à 0,95403. Malheureusement, rien dans l'aide fournie ne permettait de comprendre ce que cela pouvait signifier. Ils cherchèrent les spécificités du vocabulaire – une longue liste d'excédents et une courte liste de déficits apparut, d'où ils retinrent simplement qu'il y avait 7 fois le mot *Marquis* et 4 fois le mot *Marquise*.

Cela veut dire, calculaient-ils, que ces deux mots ne représentaient 0,2554 % des mots du texte.

– Et pourtant la scène se déroule bien chez le Marquis, fit remarquer Bouvard.

Ils entrevoyaient dans ce genre de calculs des potentialités à la fois confuses et merveilleuses. Ils voulurent comparer avec le *Trésor de la Langue Française* – le bouton ne fonctionnait pas, sans doute avaient-ils omis quelque procédure à l'installation. Enfin, ils restèrent ébaudis devant ce qui était dénommé *Dictionnaire* : des colonnes de chiffres intitulées *réel*, *théorique, écart, réduit, hapax, réduit*, et dont la signification profonde – il y en avait forcément une – leur échappait totalement.

Il fallait essayer encore, avec un autre logiciel. La liste que leur avait envoyé Dumouchel en comportait vingt-trois, mais était vieille de trois ans.

– Pourquoi pas *Hyperpo* ? suggéra Bouvard, qui ne se rappelait plus où il l'avait vu cité, mais dont le nom lui évoquait certaines parties de Madame Bordin.

Hyperpo ne trouvait que 3 *Marquises* et 5 *Marquis*, encore en affichait-il qu'il ne comptait pas. On disposait cependant du contexte autour de chaque mot, et l'on pouvait en choisir la longueur.

– C'est ce qu'on appelle une concordance, annonça Pécuchet. Le précurseur en a été Carlut, qui publia – à l'aide des premiers ordinateurs, mais sur papier ! – les concordances des principaux romans de Flaubert

Il y avait justement *Le Concordeur* qui, lui, ne trouvait que 3634 mots dans *Vaubyessard*, et non plus 4307.

Et *Cordial Analyseur*, à qui on pouvait adresser un texte par courriel, pour recevoir en retour une version lemmatisée étiquetée, agrémentée de cinq pages de calculs statistiques. – *Vaubyessard* était cette fois crédité de 4209 mots.

– Cela m'ennuie, dit Bouvard un soir, toutes ces colonnes sans fin de mots et de chiffres.

– Eh bien justement, proposa Pécuchet, cherchons le mot *ennui*.

Le Bulletin électronique *Littérature et Ordinateur* leur avait appris l'existence *d'Intratext*, un site italien, tenu par les Frères Maristes, qui proposait rien moins que 4035 titres de 700 auteurs en 36 langues. Parmi quelques centaines de textes religieux, il y avait *Madame Bovary* – mais de Flaubert, rien que *Madame Bovary*.

D'abord, ils crurent avoir trouvé le Graal : on pouvait lire le texte, avec un lien vers les concordances des mots. On pouvait les rechercher par ordre alphabétique, et l'on obtenait fréquence et concordance. Des statistiques par lettres initiales donnaient la distribution par ordre alphabétique. On pouvait rechercher des fréquences en choisissant un nombre ou une plage de nombres, le résultat était un classement de mots par fréquence décroissante, avec toujours un lien hypertexte vers les concordances de chaque mot. On disposait de statistiques de distribution des mots par groupe de fréquence, et d'une liste de *hapax*.

– Un hapax est un vocable n'ayant qu'une seule occurrence dans un corpus donné, déclara Bouvard, qui avait vérifié dans le *TLFI* – Il y en avait 6715, qui formaient une liste qu'ils déclamaient le soir à haute voix : *adagio, addition, adjoindre, adjudicataire, fourchette, fourmi, fourneau, vitrée, vitriol, vitupérations, vivante.* Mélie écoutait, bouche bée, derrière la porte ; Gorgu les regardait par la fenêtre.

– Si l'on dressait la liste des *nullax*, c'est-à-dire des mots qui ne sont *pas* dans le texte ?

On pouvait aussi s'intéresser aux inversions et chercher des mots par leur dernière lettre ainsi que leur distribution par lettre finale. On pouvait choisir des longueurs de mots et trouver ainsi leur nombre, leur fréquence, leur répartition statistique par groupe de longueur.

Mais, curieusement, on ne pouvait accéder à la concordance de tous les mots : pas de lien pour *chambre* (70 occurrences), *chose* (111) et *choses* (51) ; *devoir* (12), *entre* (129), *pouvoir* (22) – des homophonies, observa Bouvard ; Pécuchet ne releva pas – *toujours* (122), *yeux* (131).

Ils apprirent cependant qu'il y avait dans le roman 13 202 mots et 120 334 occurrences – 6 411 de moins que n'en avait compté *Word*, en partant pourtant du même texte source, c'est beaucoup, remarqua Bouvard – L'occurrence par rapport aux mots était donc de 9,1 ; il y avait 71 287 occurrences de fonctions en 255 mots de fonction ; la longueur moyenne des occurrences était de 4,49 tandis que celle des mots était de 7,78.

– Comment les occurrences pouvaient-elles être plus courtes que les mots ? s'interrogeait Pécuchet.

Il y avait 7 fois le mot *ennui*, 5 fois *ennuyait* et *ennuyer*, 4 fois *ennuie*, 3 fois *ennuyé* et *ennuyeuse*, 1 fois *ennuierais*, *ennuis*, *ennuya*, *ennuyant*, *ennuyée* et *ennuyez*, soit 33 occurrences.

Ainsi morcelée à l'infini, l'œuvre en devenait inintelligible. Comme lorsqu'ils étudiaient la médecine, ils avaient démonté le cadavre, et se trouvaient embarrassés pour remettre en place les morceaux.

Ils se consolèrent avec la loi de Zipf, qu'ils appliquèrent à *Vaubyessard*, à *Salammbô*, et à eux-mêmes. Zipf avait montré qu'en classant les mots d'un texte par fréquence décroissante, on observait que la fréquence d'utilisation d'un mot était inversement proportionnelle à son rang. La fréquence du second mot le plus fréquent était ainsi la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc.

Mais si cette loi s'appliquait partout, qu'apportait-elle ? Et comment interpréter les subtiles variations de la courbe autour de l'axe idéal ?

– Cherchons plutôt les thèmes que les mots !

Mais qu'était-ce qu'un thème ?

Certains l'avaient défini comme ce dont parlait un texte. D'autres avaient préféré renoncer. Lemaire opposait le texte documentaire, univoque car tentant d'être clair et simple, au texte littéraire qui, toujours en train de tourner autour du pot, disait trente-six choses en même temps et souvent de manière confuse ou masquée. Celui-là était polysémique. Deux conjectures accompagnaient le propos : l'inscription d'un thème dans un texte était inversement proportionnelle à la littérarité de ce texte ; plus un thème était concret, plus son inscription dans le texte était précise.

– Oui, mais comment le vérifier puisque le même Lemaire souligne l'importance de tout ce que le texte fait comprendre de manière voulue, évidente pour tout lecteur, sans le « dire » ? Surtout qu'il concevait le thème comme une construction intellectuelle élaborée par le lecteur à partir d'éléments textuels récurrents. Au sens étymologique du terme, c'était une abstraction. Il était donc tout à fait possible que le thème ainsi construit ne corresponde à aucune expression précise du texte, autrement dit que le thème ne soit pas inscrit dans le texte. Rastier avait observé que, « alors que le mot *ennui* se rencontre quatre fois dans *Madame Bovary* – Tiens, il avait compté à la main et en avait manqué ! – les composants du thème apparaissent souvent, notamment à propos de Charles. » Mais Erlich ajoutait que « les formes sont statistiquement très saillantes chez Verlaine, alors qu'elles ne le sont pas du tout chez

Flaubert, qui se refuse effectivement à l'interrogation statistique ». Et Rastier de renchérir : «Les grandes œuvres s'avèrent beaucoup moins prévisibles que les autres, et les méthodes statistiques ordinaires s'y appliquent assez mal. Le mot *ennui*, par exemple, n'apparaît guère dans *Madame Bovary*, mais pullule chez des auteurs mineurs. Peut-être sont-ils mineurs parce qu'ils disent lourdement ce que les grands ne font que suggérer. »

On pouvait peut-être s'en sortir en relevant les thèmes à la main – Bonne idée pour des copistes ! –

C'était le parti pris de la *Base de Données d'Histoire Littéraire*. Mais on retombait sur les mêmes limites que dans le cas du codage.

La méthode d'Hubert de Phalèse tentait d'y pallier en distinguant quatre étapes : la compilation, la sélection, le classement et la rédaction. La compilation résultait du croisement par l'ordinateur d'une liste préalable au texte lui-même. La sélection consistait à choisir manuellement parmi les occurrences compilées précédemment celles qui étaient pertinentes et celles qui étaient remarquables. Le classement, enfin, était l'organisation de la matière sélectionnée. Le travail du chercheur progressait donc dans un va-et-vient constant entre les informations fournies par l'ordinateur et le texte sur lequel se poursuivait sa réflexion.

Cette interaction entre l'homme et la machine se retrouvait dans la méthode de Beust pour le coloriage des corpus par construction de classes sémantiques. L'ordinateur était considéré comme un compagnon de travail, un instrument de recherche et non un appareil magique qui fournirait des solutions toutes faites, des analyses clé en main.

D'autres en restaient à la puissance de la machine : Forest utilisait un classifieur de type réseau de neurones auto-associatifs sans supervision. Sjoblöm et Brunet tentaient le repérage grâce à une sélection automatique ou raisonnée de mots privilégiés.

Mais ce dernier reconnaissait que la lecture humaine était irremplaçable lorsqu'il s'agissait d'évaluer l'importance relative des mots, le relief thématique d'un texte et la signification profonde d'une œuvre, et ceci fondé tant que l'ordinateur n'aura pas acquis le sens de l'humour.

- Comment interpréter des tableaux de résultats sans connaître les algorithmes de départ, donc leurs présupposés ou leurs postulats théoriques, qui d'ailleurs nécessiteraient des connaissances mathématiques que nous n'avons pas ?

– C'est peut-être une blague ?

– Comme l'arboriculture !

– Et comme l'agronomie !

Ils passèrent au style, question autrement considérable et encore plus difficile. Le Style n'était-il pas la chimère de Flaubert, celle qu'il *poursuivit avec un labeur atroce, avec une opiniâtreté fanatique et dévouée* ?

Ils s'arrêtèrent bien vite à la définition de Gicquel qui voyait le style comme *la marque individuelle donnée au matériau dans lequel il se réalise*. Gicquel avait esquissé, par analogie avec la caractérologie et l'étude des parfums, un programme d'informatisation de l'étude stylistique. On se baserait sur un triple classement des caractéristiques esthétiques : selon leurs polarités, par classes polythétiques, et par appartenance à des niveaux de phénomènes linguistiques. La réalisation de ce vaste programme restait à entreprendre.

En attendant, Monière et Labbé exploraient la linguistique quantitative, à l'aide de l*'Indice de diversité*, mesurant la propension plus ou moins grande d'un auteur à diversifier l'expression ou à fuir la répétition et de l'*Indice de spécialisation,* mesurant l'adaptation de l'auteur au thème abordé ou à l'inverse sa propension à employer le même vocabulaire quel que soit le sujet traité.

Mais sur Flaubert, concrètement, rien.

A part la *Clé des Procédés littéraires*, qui donnait un extrait de l'*Education sentimentale*, et en détaillait les ingrédients : *clou d'or, description, discours rapporté, effet par évocation, endophasie, énumération, focalisation, fonction narrative, hypotaxe, hypotypose descriptive, netteté syntaxique, Pandore, personnage banal, psycho-récit, récit, récit circonstancié, référent, symbole, texture, toponyme*.

Il s'agissait bien évidemment d'un relevé manuel. Comment aurait-on pu automatiser le repérage de telles notions ?

– Sais- tu ce qu'a écrit un jour Remy de Gourmont ? : « Quand on a, avec beaucoup de peine, établi des catégories, il faut bien souvent se résigner à n'avoir rien à enfermer dans l'enclos : les jolies bêtes s'échappent et vont jouer dans la forêt voisine. C'est cependant une grande satisfaction pour l'esprit que l'établissement des catégories : on est rassuré ; on garde l'intime conviction que les troupeaux, fatigués de leur liberté, regagneront un jour ou l'autre, les délicieux bercails où le foin de la logique pend à toutes les crèches. »

Alors comment découvrir le sens du texte ?

Reinert avait défini, en suivant Peirce, la spécificité du sens comme trinitaire : sens comme sensation – Imaginaire – ; sens comme direction – Réel – ; et sens comme signification – Symbolique –. Il se proposait de le rechercher par le découpage, *volontairement arbitraire,* d'un texte en unités de contexte, afin d'en déceler les écarts, les trébuchements, les discontinuités. Pour cela il avait créé le logiciel *Alceste,* considéré non pas comme un instrument de validation mais comme une aide à la construction d' hypothèses, ou même plus simplement, une aide à la lecture. Les études menées ne portaient pas sur un objet particulier qui se trouverait enfoui dans les textes mais sur la façon dont un Sujet se constituait à travers son propre tressage, à travers ses ancrages, ses insistances, ses redites, ses échappements

N'ayant pu comprendre, ils n'en croyaient rien.

Ils continuèrent encore quelque temps leurs études, mais sans passion.

Bernard les acheva, qui déclarait qu'un texte littéraire était un texte qui ne pouvait *justement pas* être traité correctement par un automate, car l'être humain aspirait à travailler avec des concepts, alors que la machine ne pouvait manipuler que des formes.
Et Sinclair d'ajouter que l'ordinateur n'accédait jamais au niveau du sens. Les informations qu'il nous livrait pouvaient provoquer du sens, mais il fallait reconnaître aussi qu'elles pouvaient n'en provoquer aucun. Les effets produits par le texte ne pouvaient se réduire aux éléments textuels individuels. Le tout - quel qu'il soit - était plus que la somme des parties - sauf peut-être de la perspective de l'ordinateur.

– Après tout, Flaubert l'avait pressenti: « Je me suis rué sur la forme, sans presque songer à ce qu'elle disait. Définissez-moi la, faiseurs d' esthétiques, classez-la, étiquetez-la, essuyez bien le verre de vos lunettes, et dites moi pourquoi cela m'enchante. »

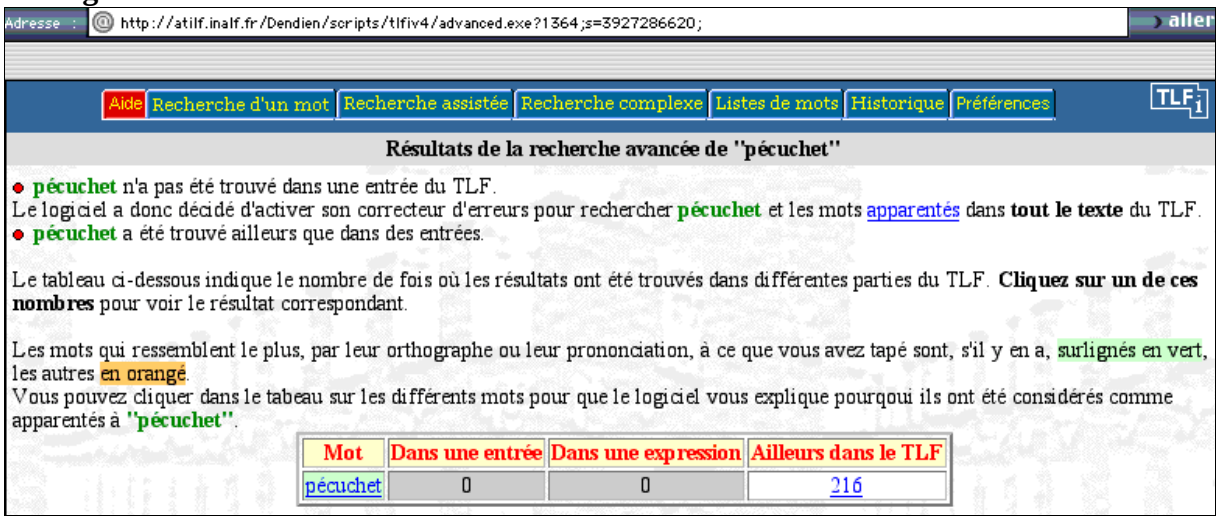
Pécuchet voyait tout en noir, peut-être à cause de sa jaunisse.

La Copie

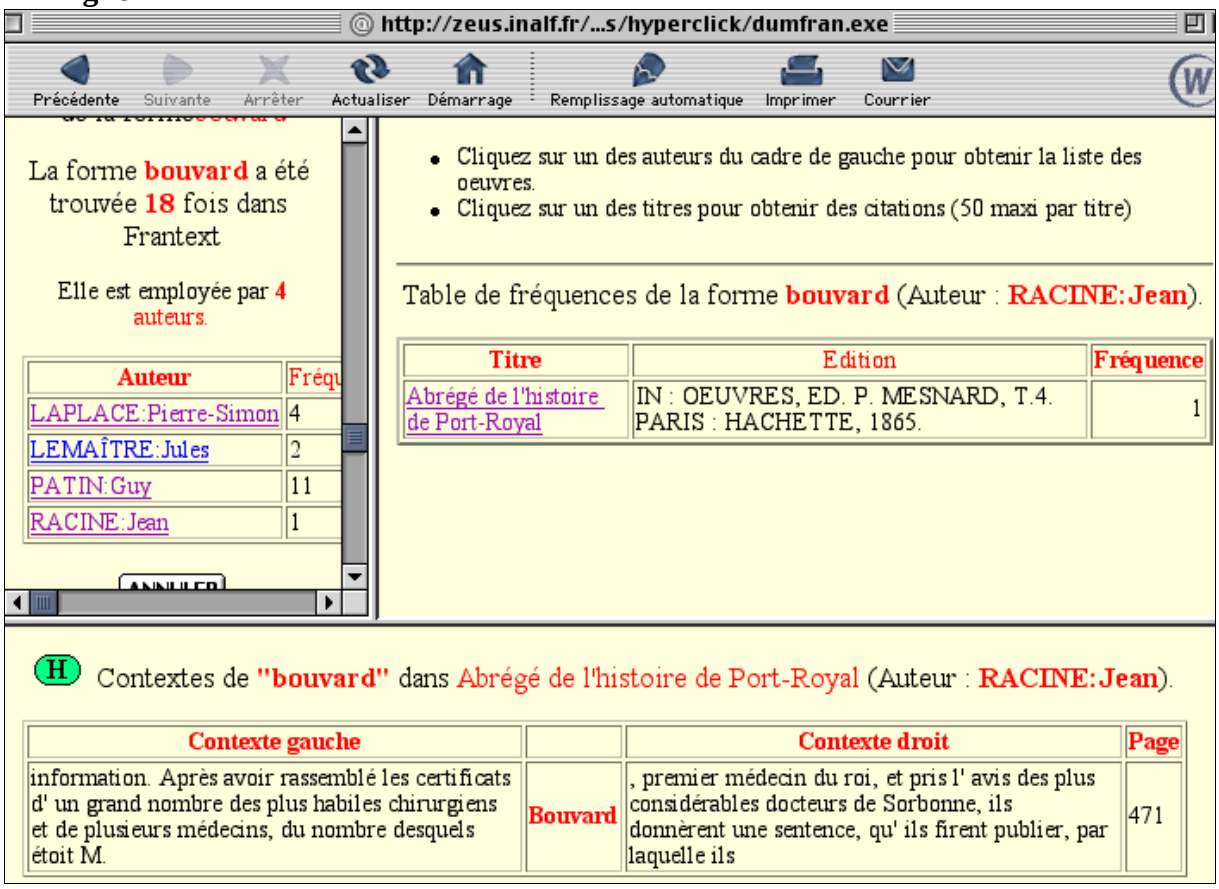
- **Ligne 12**

Trésor de la Langue Française Informatisé : http://atilf.inalf.fr/tlfv3.htm

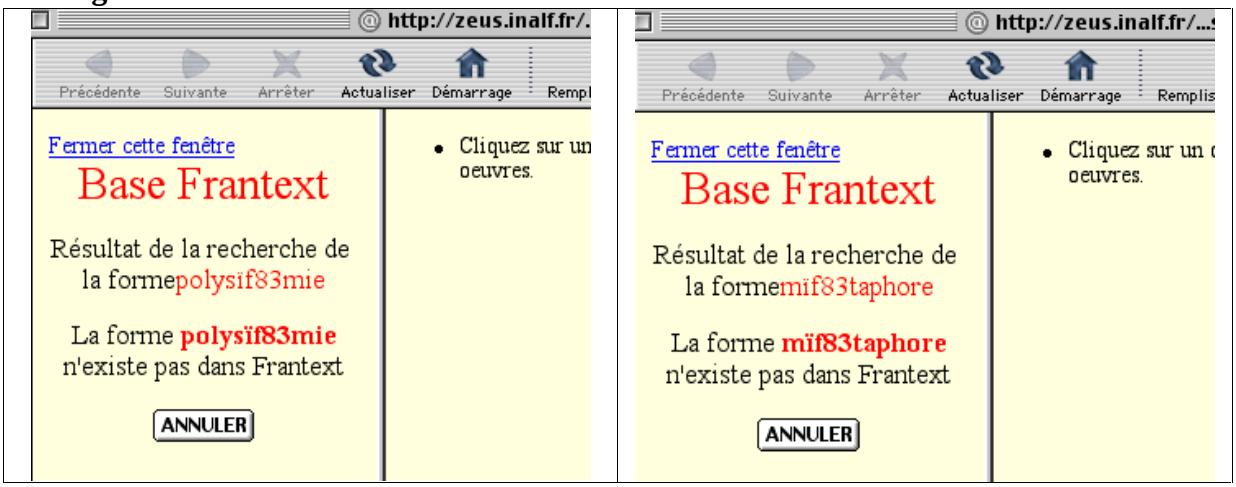
- **Ligne 12**



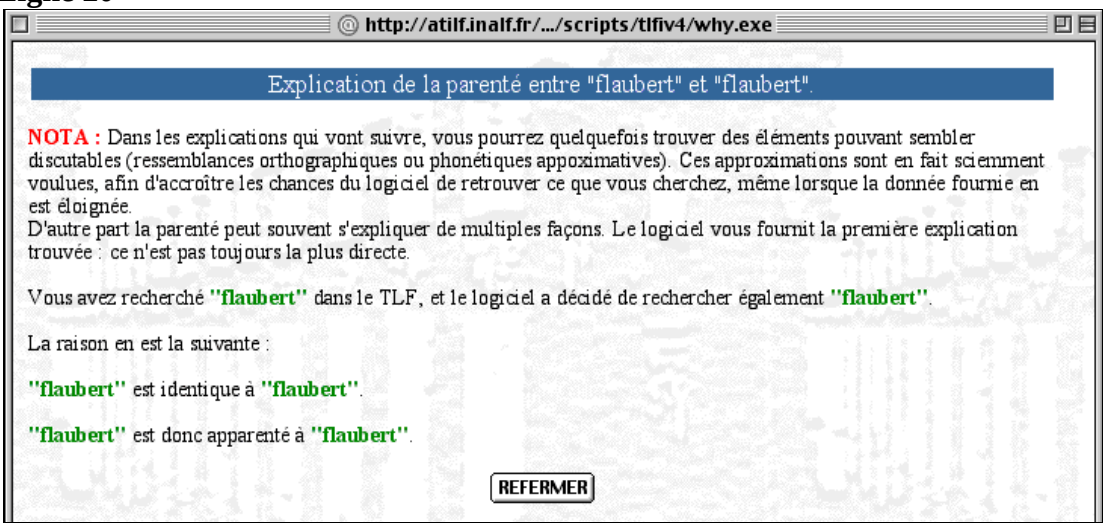
- **Ligne 14**



- **Ligne 16**



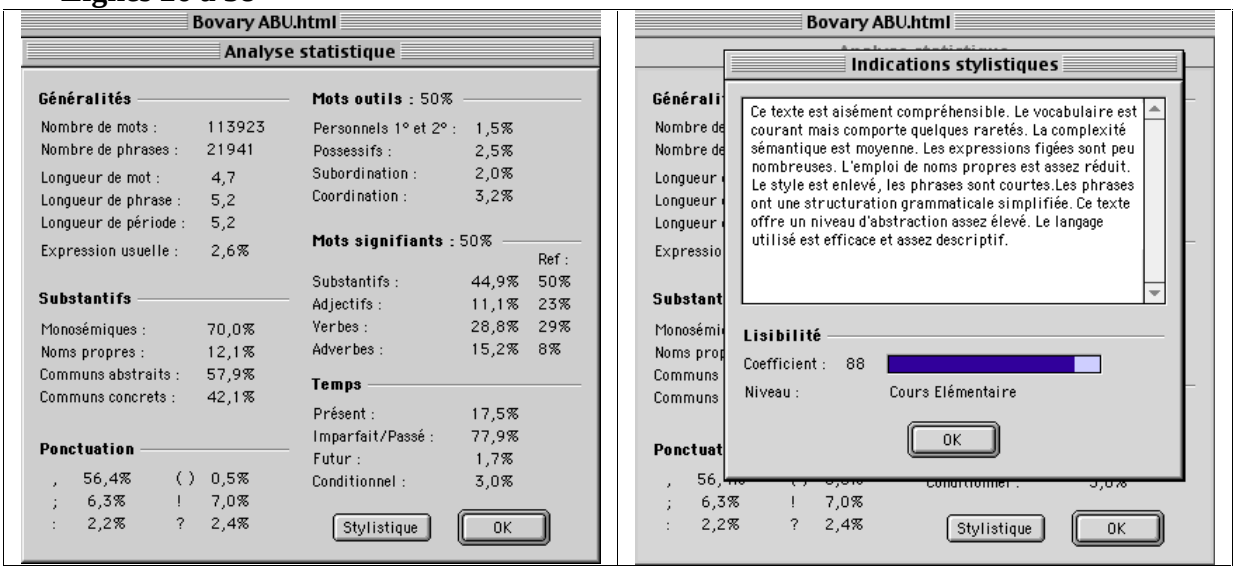
- **Ligne 20**



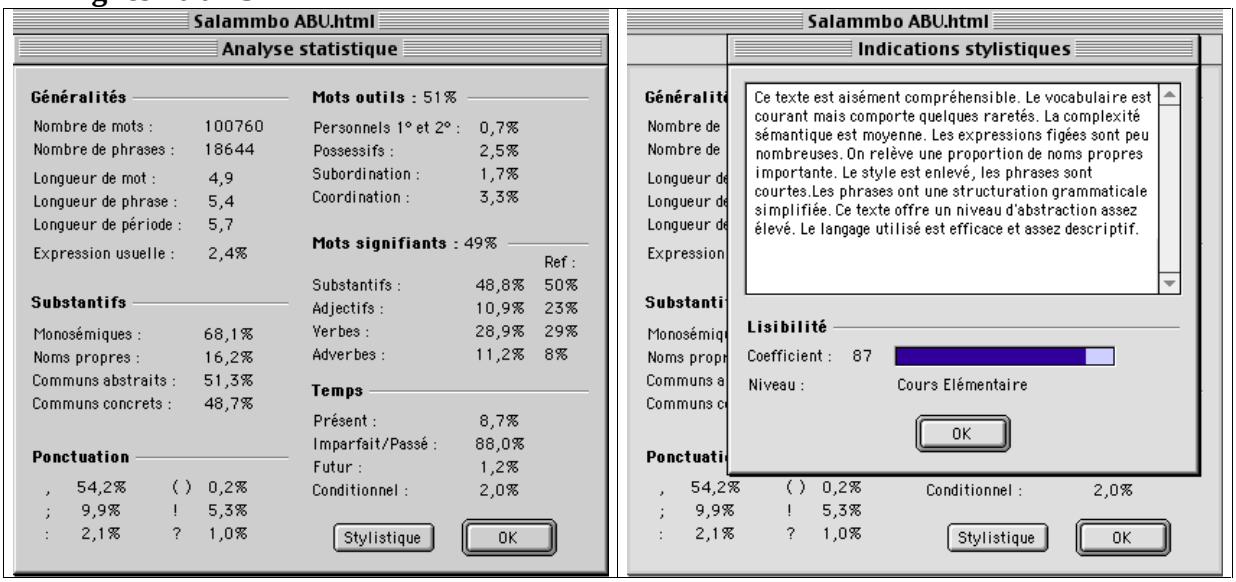
- **Ligne 22**

Association des Bibliophiles Universels : http://cedric.cnam.fr/ABU/

- **Lignes 26 à 38**



- **Lignes 40 à 43**



- **Ligne 48**

Typologie des corpus : http://www.rint.org/attrait/Documentation/Grilles/typocorpus.htm

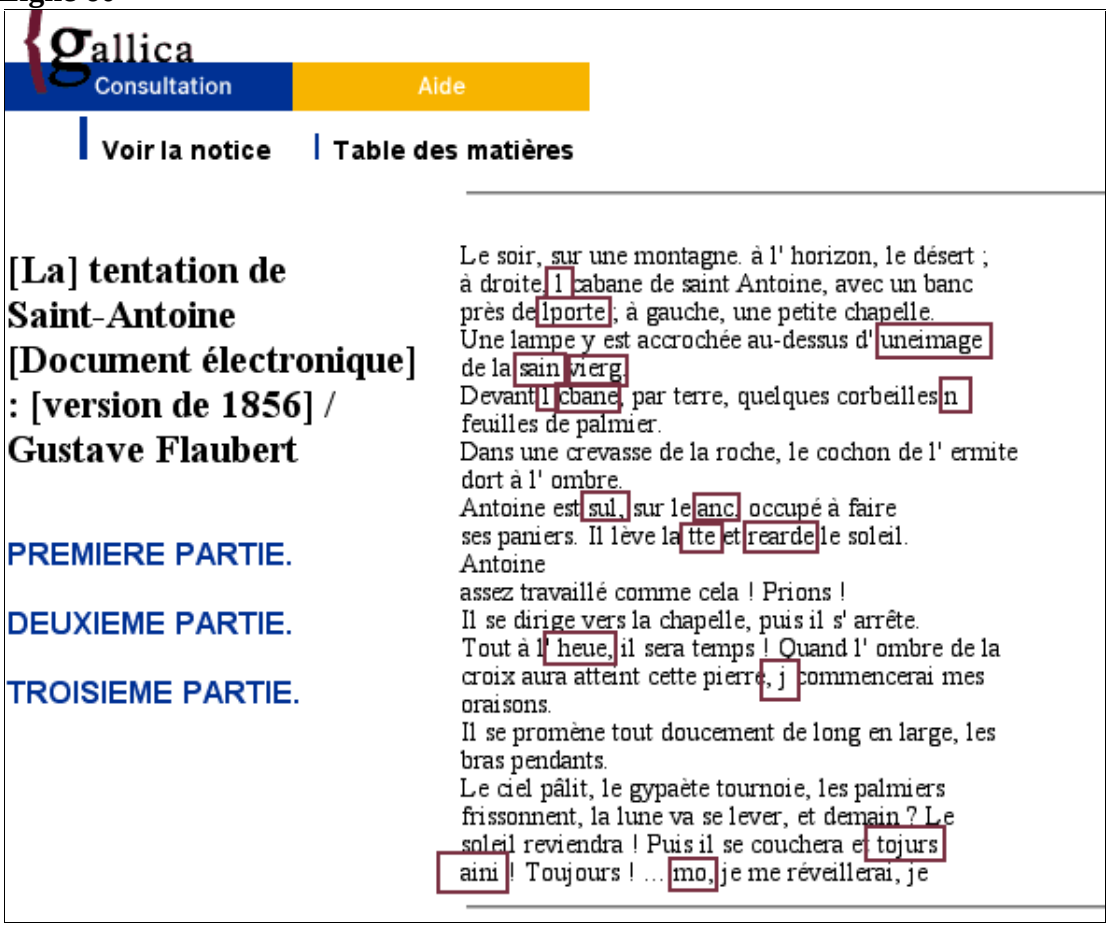
- **Ligne 49**

Max Reinert. Note sur la notion de corpus : http://www.printemps.uvsq.fr/Lettre09.htm

- **Ligne 55**

Les manuscrits de l'Education sentimentale : http://www.hull.ac.uk/hitm/

- **Ligne 60**



- **Ligne 62**

	Tentation	**Smarh**	**Souvenirs, Pensées**	**Première Education**	**Champs et grèves**	**Bovary**	**Salammbô**	**Education sentimentale**	**Bouvard et Pécuchet**	**3 contes**	**Corresp.**
ABU						**X**	**X**	**X**	**X**	coeur	
ARTFL	**49 56 74**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Bibliopolis						**X**		**X**		**X**	
Frantext	**49 56 74**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Gallica (texte)	**49 56 74**	**X**				**X**				**X**	**X**
Intratext						**X**					

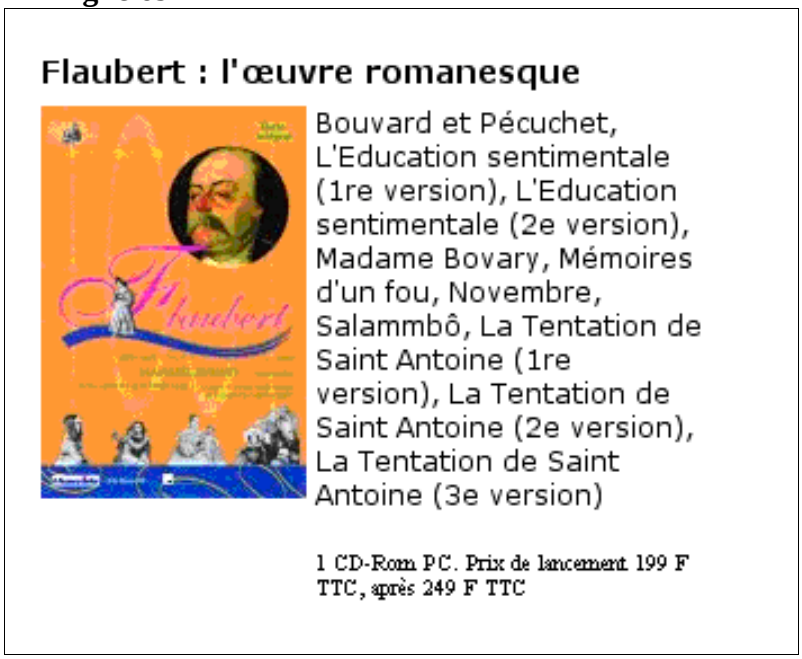
- **Ligne 64**

Frantext : http://zeus.inalf.cnrs.fr/frantext.htm
ARTFL : http://humanities.uchicago.edu/orgs/ARTFL/

- **Ligne 68**

Bibliopolis : http://athena.bibliopolis.fr:7900/

- **Ligne 69**



- **Ligne 75**

Michel Bernard. *Quelques remarques méthodologiques sur l'utilisation du TALN dans les ELAO.*
http://www.cavi.univ-paris3.fr/phalese/Cjouis/analyse.htm

- **Ligne 77**

Etienne Brunet. *Qui lemmatise dilemme attise.*
http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.PDF

- **Ligne 81**

Étienne Brunet. *Le lemme comme on l'aime.* JADT 2002.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/brunet.pdf

- **Ligne 84**

Margareta Kastberg Sjöblom. *Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus.* JADT 2002
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/kastberg.pdf

- **Ligne 95**

Text Encoding Initiative : http://www.tei-c.org/

- **Ligne 97**

Centre Kolb-Proust : http://www.library.uiuc.edu/kolbp/homeF.htm

- **Ligne 100**

Paul A. Fortier. *Le Codage des données textuelles.* JADT 2000
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/20/20.pdf

- **Ligne 102**

Mathieu Brudigou, Caroline Escoffier, Helka Folch, Saadi Lahlou, Dominique Le Roux, Patricia Morin-Andréani, Gérald Piat. *Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles.* JADT 2000
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/04/04.pdf

- **Ligne 115**

Dominique Labbé. *Analyse des données textuelles et Statistique lexicale*. 5e journées internationales d'analyse statistique des données textuelles. Mars 2002.
http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special1/spec1-texte1.htm

- **Ligne 118**

B. Habert, G. Illouz, P. Lafon, S. Fleury, H. Folch, S. Heiden, S. Prévost. *Profilage de textes : cadre de travail et expérience.* JADT 2000
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/56/56.pdf

- **Ligne 120**

Marc Hug. *Désambigüisation automatique d'homographes verbe / nom.* JADT 2002
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/hug.pdf

- **Ligne 121**

Anne Dister. *Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes.* JADT 2000.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/17/17.pdf

- **Ligne 126**

Jean Véronis. *Annotation automatique de corpus. Panorama et état de la technique*.
http://www.up.univ-mrs.fr/~veronis/pdf/2000hermes4.pdf

- **Ligne 127**

B. Habert, G. Illouz, P. Lafon, S. Fleury, H. Folch, S. Heiden, S. Prévost. *Profilage de textes : cadre de travail et expérience*. JADT 2000
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/56/56.pdf

- **Ligne 132**

Hyperbase Bazac : http://134.59.31.3/%7Ebrunet/BALZAC/BALZAC.htm

Hyperbase Rabelais : http://134.59.31.3/rabelais.html

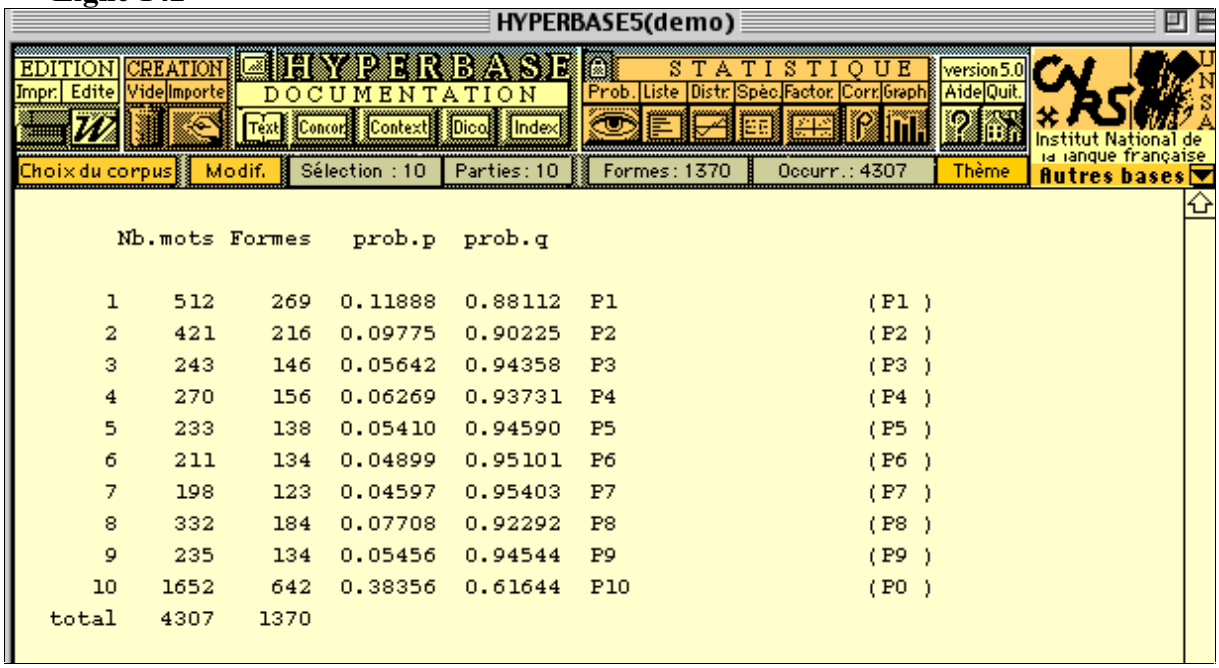
- **Ligne 133**

Hyperbase Flaubert (fichers PC de 25 Mo chacun) :
- Œuvres : http://www.unice.fr/ILF-CNRS/HYPERBAS/flaubert.exe
- Correspondance : http://www.unice.fr/ILF-CNRS/HYPERBAS/flaucorr.exe

- **Ligne 139**

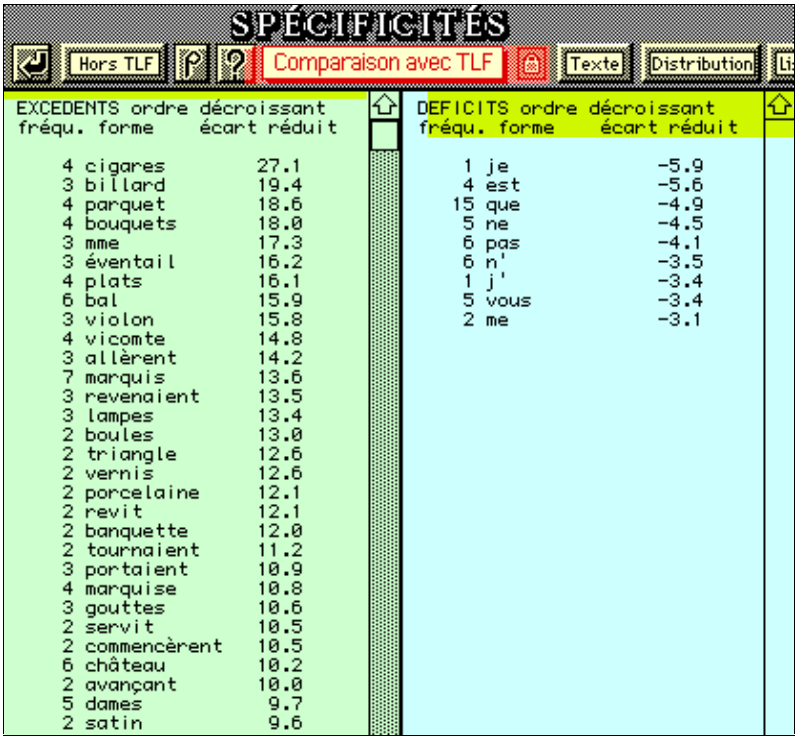
Hyperbase version de démonstration : http://ancilla.unice.fr/

- **Ligne 142**

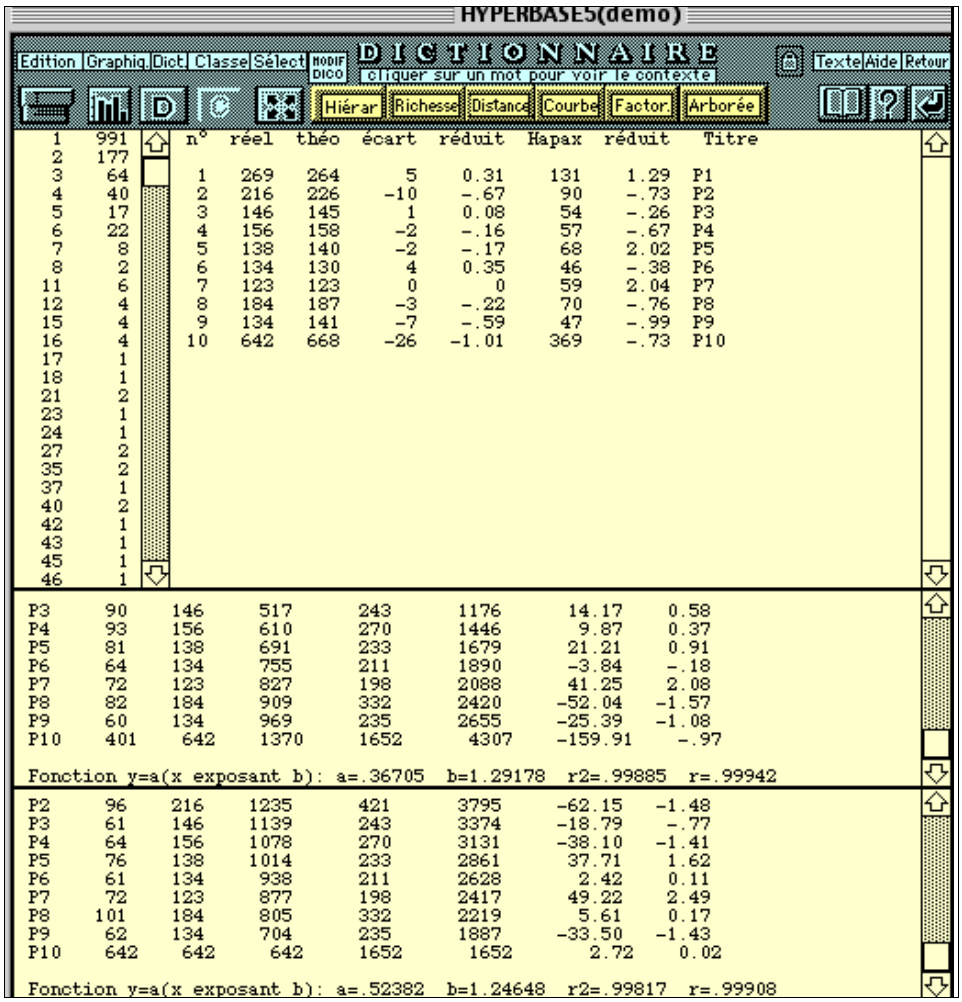


Nb.mots	Formes	prob.p	prob.q		
1	512	269	0.11888	0.88112	P1 (P1)
2	421	216	0.09775	0.90225	P2 (P2)
3	243	146	0.05642	0.94358	P3 (P3)
4	270	156	0.06269	0.93731	P4 (P4)
5	233	138	0.05410	0.94590	P5 (P5)
6	211	134	0.04899	0.95101	P6 (P6)
7	198	123	0.04597	0.95403	P7 (P7)
8	332	184	0.07708	0.92292	P8 (P8)
9	235	134	0.05456	0.94544	P9 (P9)
10	1652	642	0.38356	0.61644	P10 (P0)
total	4307	1370			

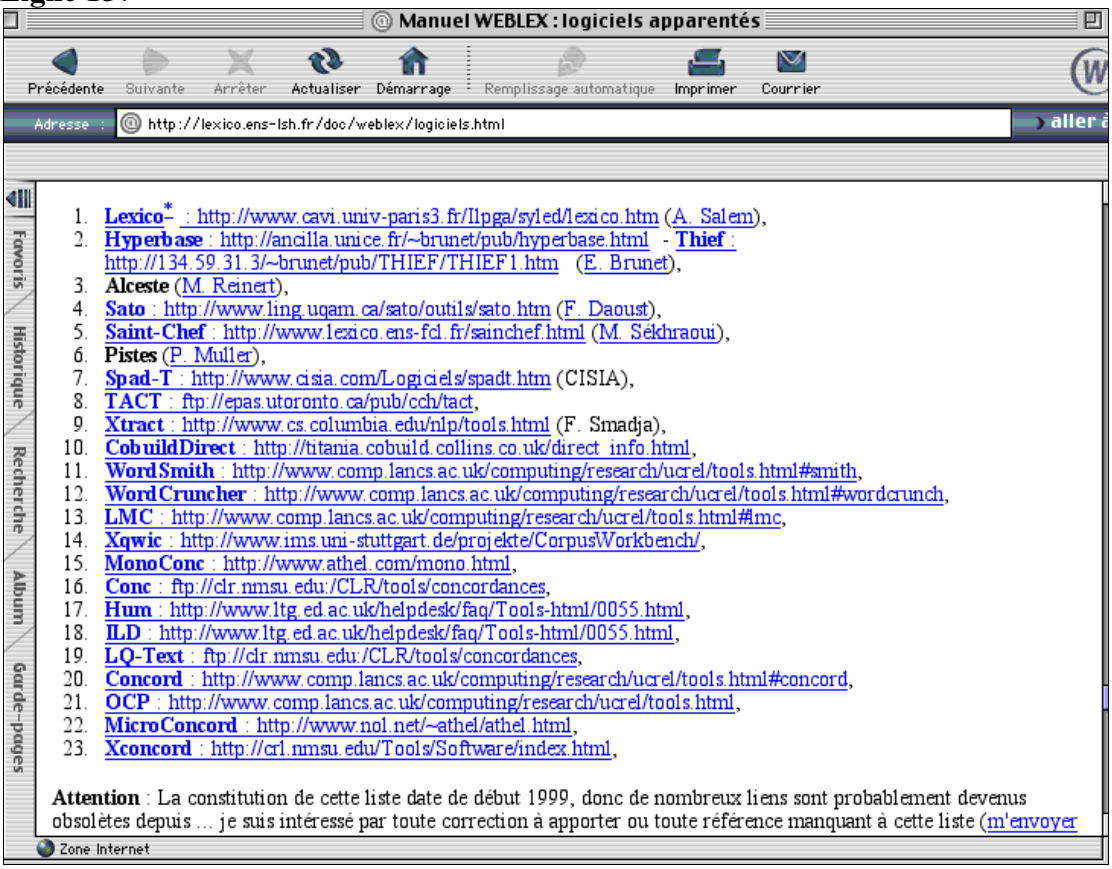
- **Ligne 145**



- **Ligne 154**



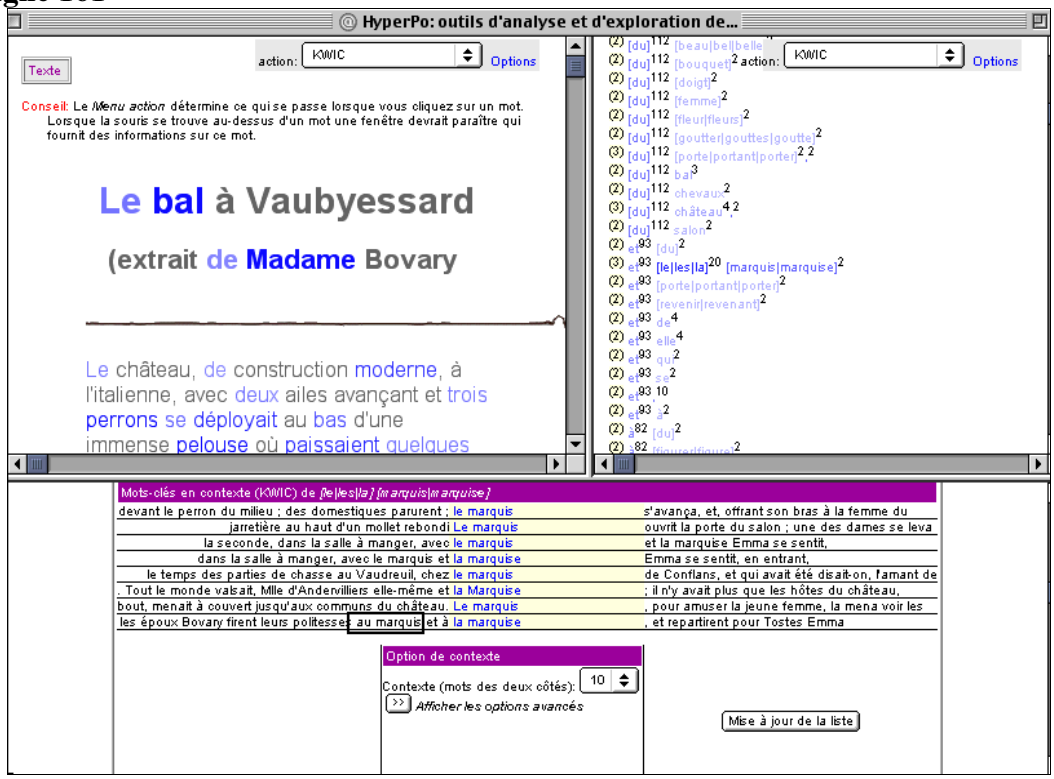
- **Ligne 157**



- **Ligne 159**

Hyperpo : http://qsilver.queensu.ca/QI/HyperPo/

- **Ligne 161**



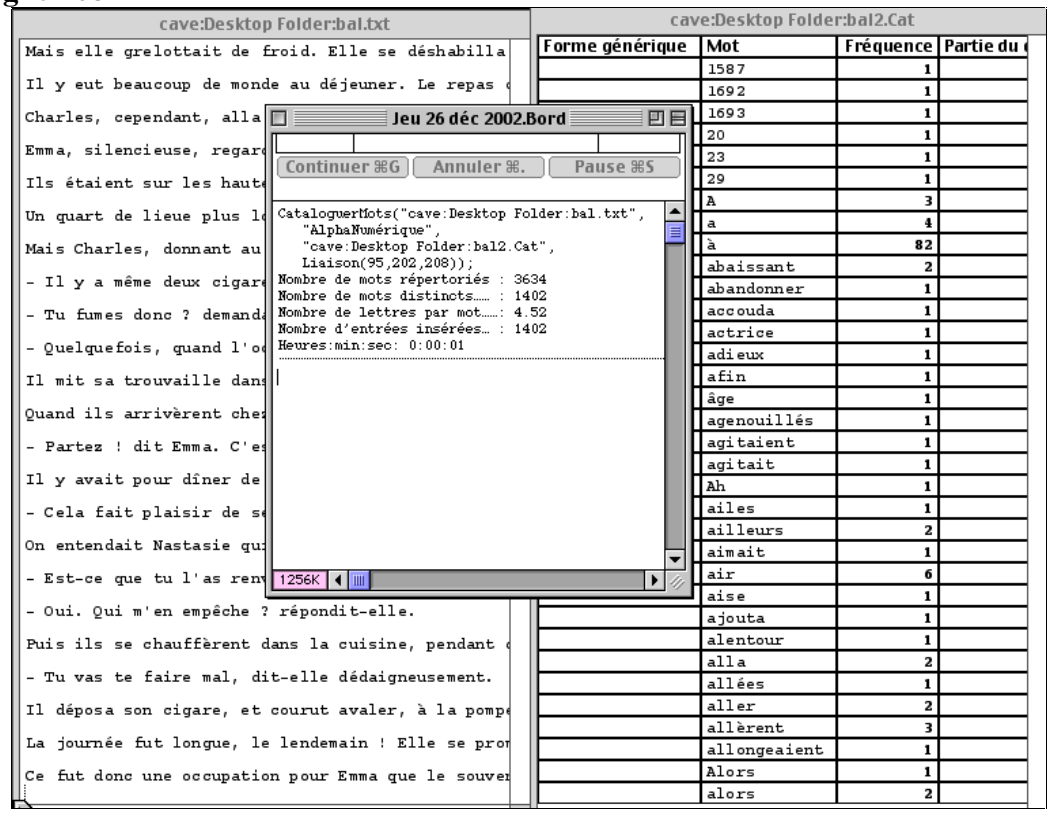
- **Ligne 164**

- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's Madame Bovary.* Garland. New York. 1978.
- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's Education sentimentale.* Garland. New York. 1978.
- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's Salammbô.* Garland. New York. 1979.
- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's Trois contes.* Garland. New York. 1979.
- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's La Tentation de Saint Antoine.* Garland. New York. 1979.
- Charles Carlut, Pierre Dubé et Raymond J. Dugan. *A concordance to Flaubert's Bouvard et Pécuchet.* Garland. New York. 1980.

- **Ligne 167**

Le Concordeur : http://omega.CRM.UMontreal.CA/~rand/CC_fr.html

- **Ligne 168**



- **Ligne 169**

Cordial Analyseur :
www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm

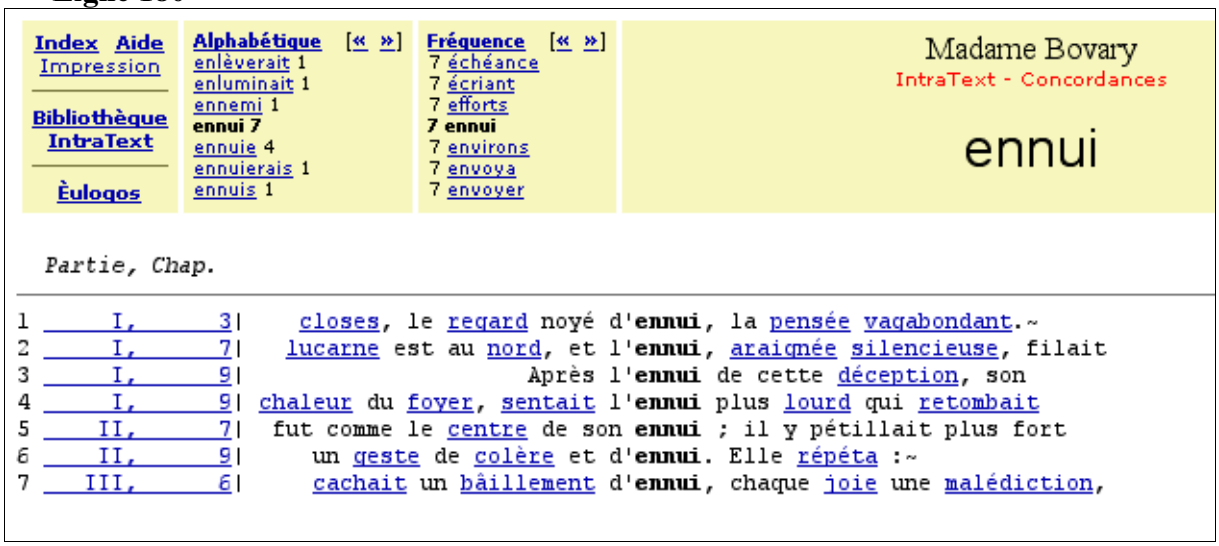
- **Ligne 174**

Abonnement à LITOR : envoyer un courrier à : admin-litor@univ-paris3.fr avec le message suivant : abonnement Nom Prénom

- **Ligne 174**

Intratext Bovary : http://www.intratext.com/IXT/FRA0023/_INDEX.HTM

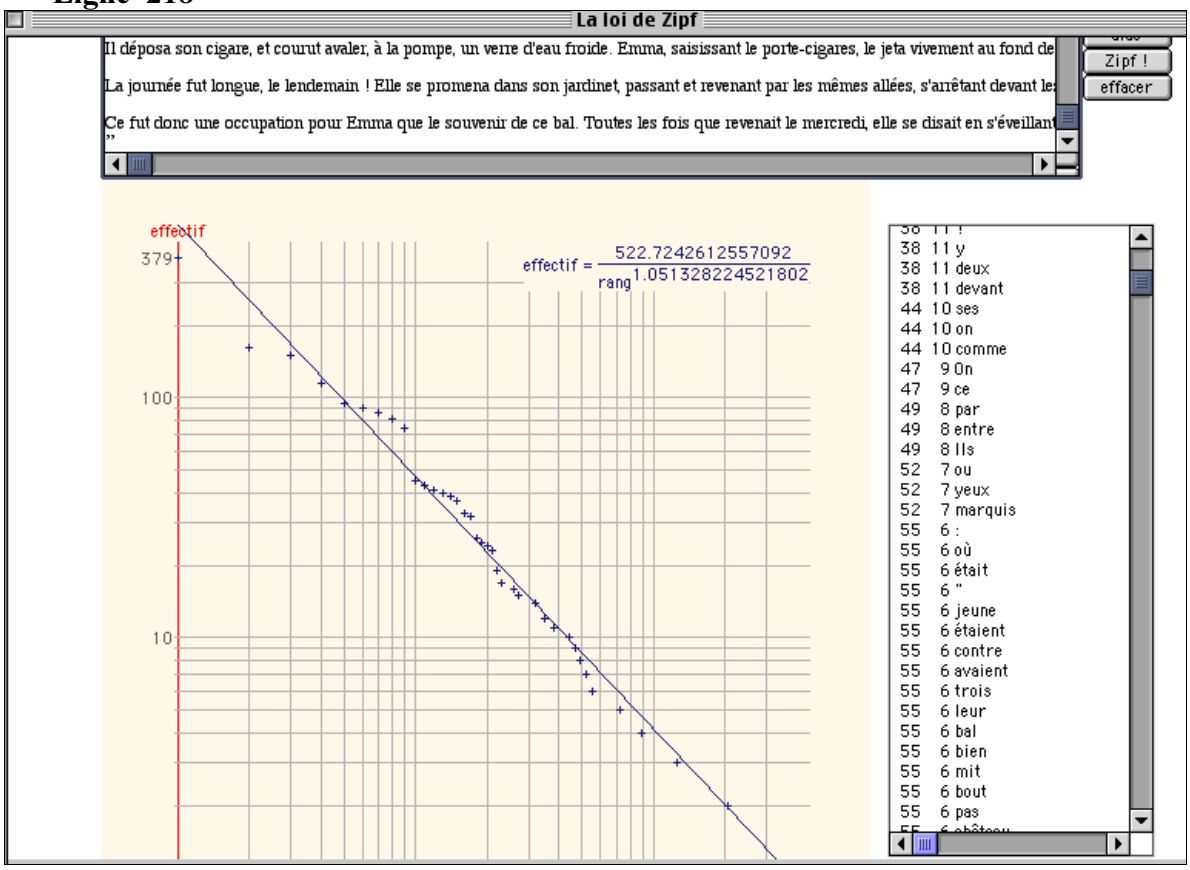
- **Ligne 180**



- **Ligne 212**

Loi de Zipf : http://users.info.unicaen.fr/~giguet/java/zipf.html

- **Ligne 218**



- **Ligne 225**

Michel Lemaire. *Le thème littéraire à l'épreuve de l'ordinateur.*
http://www.uottawa.ca/academic/arts/astrolabe/articles/art0022.htm/Epreuve.htm

- **Ligne 231**

Michel Lemaire. *Pratique de l'analyse thématique assistée par ordinateur.*
http://www.uottawa.ca/academic/arts/astrolabe/articles/art0023.htm/Pratique.htm

- **Ligne 237**

François Rastier. *La sémantique des thèmes ou le voyage sentimental* in *L'Analyse thématique des données textuelles.* Didier Erudition. 1995

- **Ligne 239**

David Erlich. *Une méthode d'analyse thématique. L'exemple de l'ennui et de l'ambition* in *L'Analyse thématique des données textuelles.* Didier Erudition. 1995

- **Ligne 241**

François Rastier. *L'Analyse thématique des données textuelles.* Didier Erudition. 1995

- **Ligne 249**

Base de Données d'Histoire Littéraire : http://michel.bernard.online.fr/bdhl/bdhl.php

- **Ligne 251**

Equipe Hubert de Phalèse : http://www.cavi.univ-paris3.fr/phalese/hubert1.htm

- **Ligne 262**

Dominic Forest, Jean-Guy Meunier. *La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques.* JADT 2000
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/63/63.pdf

- **Ligne 264**

Margareta Kastberg Sjöblom et Etienne Brunet. *La thématique. Essai de repérage automatique dans l'oeuvre d'un écrivain.* JADT 2000.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/90/90.pdf

- **Ligne 266**

Etienne Brunet. *Le vocabulaire de Zola.* Slatkine-Champion. 1985. Cité par Michel Lemaire. *Pratique de l'analyse thématique assistée par ordinateur.*
http://www.uottawa.ca/academic/arts/astrolabe/articles/art0023.htm/Pratique.htm

- **Ligne 271**

Soit un corpus, de taille N mots, comportant V vocables différents dont V_i de fréquence i (i variant de 1 à f) Pour un fragment de taille N' ($N'<N$), l'espérance mathématique du nombre de vocables différents contenus dans le fragment est, selon la formule de Muller (Muller 1977 et 1979) :

$$VCM(u) = V - \sum_{1}^{f} V_i Q_i(u) \quad avec \quad u = \frac{N'}{N} \quad et \quad Q_i(u) = (1 - u)^i$$

et selon le modèle de partition (Hubert-Labbé, 1988) :

$$VPA(u) = p.u.V + q \sum_{1}^{f} V_i Q_i(u) \quad avec \quad q = 1 - p$$

- **Ligne 280**

Lettre à Louise Colet. 15 août 1846.

- **Ligne 282**

Bernard Gicquel. *Stylistique littéraire et Informatique*. Artois Presses Université.1999.

- **Ligne 288**

Dominique Monière et Dominique Labbé. *Essai de stylistique quantitative : Duplessis, Bourassa et Lévesque*. JADT 2002.

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/moniere_labbe.pdf

- **Ligne 293**

La Clé des Procédés littéraires : http://www.cafe.umontreal.ca/cle/

Extrait de l'Education sentimentale : http://www.cafe.umontreal.ca/genres/e-edusen.html

- **Ligne 300**

Rémy de Gourmont. *Le problème du style.* Mercure de France. 1907.

- **Ligne 309**

Max Reinert. La tresse du sens et la méthode " Alceste " Application aux " *Rêveries du promeneur solitaire* ". JADT 2000.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/31/31.pdf

- **Ligne 323**

Michel Bernard. *Quelques remarques méthodologiques sur l'utilisation du TALN dans les ELAO.*
http://www.cavi.univ-paris3.fr/phalese/Cjouis/analyse.htm

- **Ligne 326**

Stéfan Sinclair. *Quelques obstacles dans le développement de l'analyse de texte informatisée.*
http://www.uottawa.ca/academic/arts/astrolabe/articles/art0021.htm

- **Ligne 332**

Gustave Flaubert. *Italie.* Œuvres de Jeunesse. Pléiade. Page 1111.