Projet de STATISTIQUE LEXICALE (Licence Informatique)

Didier DE RYCK & Sébastien ROGERIEUX

Angers, 1997

Projet dirigé par Gilles HUNAULT

Table des matières

1 Description du projet **4**

2 La statistique lexicale **4**

3 Le programme **5**

3.1 Présentation de StatLex 5

3.2 Précautions d'emploi 6

3.3 Utilisation de StatLex sous Unix 6

3.3.1 Analyse d'un ou de plusieurs fichiers texte 7

3.3.2 Cumul de résultats 8

3.3.3 Mise à jour d'une base de données 9

3.3.4 Impression des résultats 9

3.4 Utilisation de StatLex par le Web 10

3.5 L'implémentation de StatLex 11

3.5.1 Présentation de Perl 11

3.6 Conception du programme 11

4 Définitions, conventions et cas particuliers **12**

4.1 Définition d'un mot 12

4.2 Unigrammes, bigrammes et trigrammes 12

4.3 Initiales et finales 13

4.4 Définition d'une phrase 14

4.5 Définition d'un paragraphe 14

4.6 Construction du dictionnaire des fréquences 14

4.6.1 Tirets et apostrophes 14

4.6.2 Majuscules et minuscules 15

5 Le corpus **15**

6 Les résultats **16**

7 Conclusion: réflexions sur le projet **17**

Annexe A. Liste des caractères alphabétiques reconnus par la norme ISO-Latin **18**

Annexe B. Exemple de sortie PostScript : analyse de *Candide* (Voltaire) **19**

Annexe C. Tables de fréquences d'unigrammes en allemand, anglais, espagnol, français et italien **31**

Annexe D. Bibliographie et adresses *http* **41**

1 Description du projet

Le projet présenté ici nous a été proposé par Mr Gilles HUNAULT. Sa réalisation a duré un mois, du 05 Mai au 06 Juin 1997. Ainsi que l'indique l'intitulé, ce projet aborde le domaine de la statistique lexicale, sans toutefois l'explorer en profondeur. Il s'agit ici d'effectuer des comptages lexicaux simples et de fournir des outils pour la mise à jour d'une base de données sur les caractéristiques lexicales d'une langue ou d'un corpus de textes. Les problèmes plus pointus, comme la lemmatisation, seront l'objet de projets ultérieurs venant se greffer sur la base que nous proposons ici. Dans cette optique, la modularité et l'extensibilité de nos programmes est un des points importants du projet.

Dans un premier temps, nous abordons brièvement ce qu'est la statistique lexicale. La section suivante 3. décrit le fonctionnement et l'utilisation de StatLex, le programme que nous avons développé dans le cadre du projet. La section 4. présente les définitions et les choix de conception auxquels nous avons été confronté lors de l'écriture de StatLex. Enfin, le corpus sur lequel nous avons travaillé et les résultats auxquels nous avons abouti font l'objet des sections suivantes.

2 La statistique lexicale

La statistique lexicale étudie avec les moyens des outils statistiques traditionnels les données textuelles.

Dans ce domaine, les travaux du professeur Charles MULLER sont considérés comme fondamentaux. Ses ouvrages ont valeur de référence , aussi ce sont ceux que nous avons consulté en premier, et notamment celui intitulé *Principes et Méthodes de Statistique Lexicale* (*cf* bibliographie). Cet ouvrage présente de manière simple et claire les différents aspects de la statistique lexicale ; on y trouve notamment la définition de tous les termes employés dans ce domaine et des remarques pertinentes sur les problèmes inhérents à l'analyse quantitative d'un texte.

Associer ainsi les chiffres et les lettres est une idée relativement neuve mais qui a déjà fait ses preuves. Les résultats de la statistique lexicale trouvent leur application aussi bien dans la gestion de réponses à des questions ouvertes, lors de sondages d'opinions, que dans l'aide à l'analyse de textes littéraires.

Sur ce dernier point, les travaux du professeur Etienne Brunet sont remarquables. Ils montrent comment l'informatique, associée aux outils statistiques,se révèle d'une aide indispensable pour l'analyse de grands corpus (toute l'œuvre de Proust, de Zola . . .). Aujourd'hui, ces recherches ont abouti à un logiciel, Hyperbase,[1] et à plusieurs dizaines de livres exploitant les données ainsi recueillies.

3 Le programme

Le programme que nous présentons ici, StatLex, a été conçu:
- pour répondre à toutes les demandes précisées dans l'intitulé du projet,
- pour afficher les résultats de manière claire et accessible.

Il y a deux façons d'utiliser StatLex: sous UNIX et par le Web.

3.1 Présentation de StatLex

StatLex analyse n'importe quel fichier texte et donne en sortie les informations suivantes:

- le nombre de signes, espaces et caractères de ponctuation compris,
- le nombre de caractères alphabétiques, signes diacritiques compris,
- le nombre de mots et leur longueur moyenne,
- le nombre de phrases et leur longueur moyenne,
- le nombre de paragraphes et leur longueur moyenne,
- la liste des mots et leurs fréquences, classés par ordre alphabétique, par ordre alphabétique inverse et par ordre de fréquence décroissante,
- la liste des unigrammes et leurs fréquences,

[1] Des explications sur l'utilisation et le fonctionnement de ce logiciel sont données sur le Web - *cf* Annexe D.

- la liste des bigrammes et leurs fréquences,
- la liste des trigrammes et leurs fréquences,

Dans les trois derniers cas, on distingue de l'ensemble des unigrammes, bigrammes et trigrammes les initiaux et les finaux, c'est-à-dire ceux qui commencent et qui terminent les mots (voir section 4.3).

L'ensemble de ces résultats, ainsi que quelques calculs statistiques simples, est présenté de manière claire et accessible grâce à une sortie PostScript.

3.2 Précautions d'emploi

Un certain nombre de contraintes sont à respecter si l'on veut utiliser correctement StatLex:

1) Le fichier analysé doit être un fichier texte ASCII "pur". Par exemple StatLex ne peut pas analyser un fichier Word.[2] La norme utilisée par StatLex pour le traitement des signes diacritiques (caractères accentués, caractères spéciaux ...) est le format ISO-Latin 88-59.[3] Les fichiers texte qui ne sont pas codés dans ce format sont acceptés par StatLex, mais il est probable que certains résultats seront erronés.

2) Si le fichier texte analysé possède un entête, ou si l'on désire que certaines lignes du fichier ne soient pas analysées, il faut le(s) signaler par un dièse (#) en début de ligne.

3.3 Utilisation de StatLex sous Unix

Sous UNIX, l'usage de StatLex est similaire à celui des commandes UNIX classiques: toutes les options sont rentrées par l'utilisateur sur la ligne de commande, le programme traitant ensuite les requêtes demandées sans poser de questions.

L'appel se fait de la façon suivante:

```
statlex [-bi xx][-tri xx][-dico xx][-ps][-maj l][-a] txt1 ...
txtn [a_rep]
```

[2]Ou alors il faut convertir ce(s) fichier(s) grâce à un programme spécial, ce qui pourrait d'ailleurs faire l'objet d'un projet ultérieur.

[3]*cf* Annexe A.

3.3.1 Analyse d'un ou de plusieurs fichiers texte

Supposons que l'on ait un fichier `voltaire.txt` à analyser.La commande à rentrer est:

`statlex voltaire.txt`

Une fois la commande lancée, une barre de progression s'affiche sur l'écran, indiquant l'avancée de l'analyse du fichier. Une fois celle-ci terminée, StatLex crée le répertoire `Voltaire` dans lequel il place les fichiers suivants:

`voltaire.uni` : fichier des unigrammes, classés par fréquence,

`voltaire.bi1` : fichier des bigrammes, classés par fréquence,

`voltaire.bi2` : fichier des bigrammes, classés par ordre alphabétique,

`voltaire.tr1` : fichier des trigrammes, classés par fréquence,

`voltaire.tr2` : fichier des trigrammes, classés par ordre alphabétique,

`voltaire.dc1` : fichier dictionnaire, mots classés par fréquence,

`voltaire.dc2` : fichier dictionnaire, mots classés par ordre alphabétique.

`voltaire.dc3` : fichier dictionnaire, mots classés par ordre inverse,

`voltaire.smo` : informations sur le nombre de signes par mots,

`voltaire.mph` : informations sur le nombre de mots par phrases,

`voltaire.mpa` : informations sur le nombre de mots par paragraphes,

`voltaire.ppa` : informations sur le nombre de phrases par paragraphes.

Maintenant, si l'utilisateur relance la commande

`statlex voltaire.txt`

le programme refusera de s'éxécuter car le texte en question a déjà été traité (à moins que l'utilisateur n'ait effacé le répertoire `Voltaire` entretemps...). Pour retraiter un fichier, il faut utiliser l'option `-t`:

`statlex -t voltaire.txt`

L'extension `.txt` pour les fichiers à analyser n'est pas obligatoire.

Le nom du répertoire où sont stockées les données est créé à partir du nom du fichier analysé, moins son extension, en lui ajoutant une majuscule. S'il n'y a pas d'extension et que le nom du fichier texte commence par une majuscule, par exemple `Rousseau`, le répertoire de données s'appellera `Rousseau_sl`.

Les droits sur le répertoire de données sont, par défaut, les suivants:

- lecture, écriture et éxécution pour le propriétaire du répertoire,
- lecture et éxécution seulement pour tous les autres.

Enfin, StatLex peut traiter plusieurs textes à la suite. Il suffit d'écrire:

`statlex [-t] texte1.txt texte2.txt ...`

3.3.2 Cumul de résultats

Maintenant que l'on a analysé quelques textes, on peut souhaiter cumuler les résultats obtenus dans un seul répertoire de résultats. Il faut alors utiliser l'option `-a` (a comme ajouter) :

`statlex -a Rep1 Rep2 ... Rep_d`

Rep1, Rep2,... doivent exister. Rep_d est le nom du répertoire dans lequel on désire placer les résultats cumulés ; s'il n'existe pas il est automatiquement créé, et s'il existe déjà toutes les données qui sont dedans seront écrasées. Il vaut donc mieux utiliser cette option avec prudence; si l'on désire mettre à jour une base de données il faut utiliser l'option `-maj`.

On peut aussi passer en arguments les noms des fichiers texte plutôt que les noms de leurs répertoires de données:

`statlex -a texte1.txt texte2.txt ... Rep_d`

Dans ce cas, si le texte n'a pas été analysé, l'analyse se fera automatiquement.Par exemple, si on veut cumuler les résultats trouvés en analysant `voltaire.txt` et `diderot.txt` dans le répertoire de données `Vol_Did` on peut écrire:

`statlex -a voltaire.txt diderot.txt Vol_Did` ou :

`statlex -a Voltaire diderot.txt Vol_Did`

3.3.3 Mise à jour d'une base de données

Dans le cadre plus linguistique du projet, l'option `-maj` a été rajoutée afin de permettre facilement la mise à jour de bases de données sur le français, l'allemand, l'anglais, l'espagnol et l'italien. Elle s'utilise ainsi:

`statlex -maj langue texte`

Par exemple, si on veut mettre à jour la base de données `Francais` en lui rajoutant les données recueillies en analysant `voltaire.txt` on écrit

`statlex -maj Francais voltaire.txt` ou

`statlex -maj Francais Voltaire`

Bien sûr, l'utilisateur est entièrement responsable de ce qu'il rajoute ...

3.3.4 Impression des résultats

Les résultats conservés dans les répertoires de données créés plus haut, peuvent être imprimés et consultés tels quels, mais il ne s'agit que des données brutes: hormis les fréquences, aucun résultat statistique n'est conservé. D'autre part, la présentation de ces fichiers n'est pas forcément très esthétique; c'est pour cela qu'il est prévu une sortie PostScript écrite en LaTeX qui inclut:

- les informations générales sur le texte: nom du texte, nombre de signes, de caractères alphabétiques, de mots, de phrases, de paragraphes,

- quelques calculs statistiques élémentaires: longueur moyenne d'un mot, d'une phrase, d'un paragraphe,

- les unigrammes, classés par fréquence et répartis en plusieurs catégories: minuscules, majuscules, signes diacritiques,

- les x_1 premiers bigrammes, classés par fréquence puis par ordre alphabétique,

- les x_2 premiers trigrammes, classés par fréquence puis par ordre alphabétique,

- les x_3 premiers mots, classés par fréquence puis par ordre alphabétique.

Le code LaTeX utilise l'extension *t1enc* qui reconnaît presque tous les caractères acceptés par la norme ISO-Latin, sauf le caractère de code ASCII

223 (ß) qui s'écrit avec cette extension SS, et le caractère de code ASCII 255 (ÿ) qui lui sera reconnu comme le ß(*cf* Annexe A.).

La sortie Postscript est activée par l'option `-ps`. C'est l'utilisateur qui choisit les x_i premiers bigrammes, trigrammes ou mots à afficher, en utilisant les options `-bi`, `-tri` et `-dico`. Par exemple, la commande

```
statlex -ps -bi 60 -tri 70 -dico 80 Voltaire
```

va générer le fichier `voltaire.tex` tel que décrit ci-dessus, ainsi que les fichiers `voltaire.dvi` et `voltaire.ps`. Si les options `-bi`, `-tri` ou `-dico` ne sont pas utilisées, StatLex utilise les valeurs par défaut suivantes: 20 bigrammes, 20 trigrammes et 30 mots.

On aurait aussi pu utiliser la commande

```
statlex -ps -bi 60 -tri 70 -dico 80 voltaire.txt
```

auquel cas le fichier texte `voltaire.txt` aurait été automatiquement analysé si ce n'avait pas déjà été fait.

Enfin, l'utilisateur n'est pas obligé de tout imprimer à chaque fois. Pour cela, il doit passer `no` en paramètre. Par exemple, la commande

```
statlex -ps -bi 200 -tri no -dico no voltaire.txt
```

donnera en sortie, en plus des informations générales sur le texte, la liste des 200 prmiers bigrammes sans imprimer les trigrammes et le dictionnaire.

3.4 Utilisation de StatLex par le Web

Une interface Web est proposée afin de permettre l'analyse à distance de fichiers textes. Elle se trouve sur

`http: //www.info.univ-angers.fr/~gh/cours_gH/statlexp.html`

L'utilisateur de cette interface donne le nom du texte qu'il désire analyser, puis il récupère un fichier `.zip` contenant les fichiers compressés `.uni`, `.bi1` ... décrits plus hauts. S'il le désire, il peut aussi visualiser sur son écran puis imprimer le fichier LaTeX précédemment décrit, en précisant lui-même le nombre de bigrammes, trigrammes et mots à afficher. En pratique, le fichier `.html` est produit à partir du fichier `.tex` en utilisant l'utilitaire `latex2html`.

Au stade actuel de développement, cette interface n'est pas opérationelle.

3.5 L'implémentation de StatLex

StatLex a été implémenté sous Unix avec Perl.

3.5.1 Présentation de Perl

Perl, ou *Practical Extraction and Report Language*, est très certainement le langage le plus adapté à notre projet puisqu'il est spécialement conçu pour, entre autres, traiter de longs fichiers et en extraire facilement toutes les informations utiles que l'on désire y receuillir.

Il serait malheureusement trop long de présenter Perl en détail ici; il faut surtout noter que Perl permet aussi bien d'écrire des petits scripts de dix lignes que des applications CGI sécurisées, ou même de développer en orienté objet. L'un des grands avantages de Perl est sa rapidité d'apprentissage: il n'est pas nécessaire de connaître toutes les subtilités de Perl pour écrire rapidement des applications utiles. Ce n'est que si l'on prend la peine de se plonger dedans que l'on découvre petit-à-petit la diversité et la richesse de ce langage.

Perl a été choisi parmi d'autres candidats potentiels pour développer StatLex parce que:

1) Comme il est dit plus haut, Perl est plus facile d'emploi que la plupart des autres langages évolués, comme C++ ou Java. Entre autres, on n'a pas à s'occuper d'allocations mémoire; on n'est même pas obligé d'initialiser les variables. Lorsqu'on ne dispose que de quatre semaines, c'est une donnée importante.

2) Par rapport à des langages comme Awk, Perl est plus puissant. Il est par exemple spécialement conçu pour écrire des applications CGI, et ce n'est pas le seul domaine où ce langage s'avère particulièrement utile. De plus, Perl (version 5) bénificie actuellement d'une grande popularité qui justifie pleinement son apprentissage.

3.6 Conception du programme

StatLex est divisé en plusieurs sous-routines :

- `traitement` : analyse le fichier texte,

- `donnees` : écrit les résultats trouvés par `traitement` dans les fichiers de données décrits plus hauts,
- `cumul` : cumule les résultats conservés dans deux ou plusieurs des répertoires de données créés par `donnees`
- `affiche` : à partir des données sauvées par `donnees`, sort un fichier PostScript. Cette sous-routine est divisée en plusieurs procédures (affichage des unigrammes, des bigrammes ...).

Le programme principal appelle ces sous-routines en fonction des options passées dans la ligne de commande.

Pour de plus amples détails, le fichier source de StatLex est disponible.

4 Définitions, conventions et cas particuliers

Certains cas particuliers se présentent inévitablement lors de l'analyse d'un texte; il faut donc faire un certain nombre de choix que nous explicitons ici.

4.1 Définition d'un mot

Nous définissons un *mot* comme une suite non nulle de caractères alphabétiques, commencée et terminée par un ou plusieurs caractères non alphabétiques. Par *caractère alphabétique*, nous entendons, en plus des 26 lettres de l'alphabet latin, tous les caractères accentués et autres signes diacritiques[4] *(ß, ñ, ...)*. A cette liste, les linguistes ajoutent aussi les tirets *(-)* et les apostrophes *(')*.

4.2 Unigrammes, bigrammes et trigrammes

Un *unigramme* correspond à un caractère alphabétique, un *bigramme*, ou *digramme* à deux caractères alphabétiques conjoints et un *trigramme* à trois caractères alphabétiques conjoints. Par exemple, dans le mot *avion* on trouve:

[4] Nous nous limitons aux signes diacritiques inclus dans le format ISO 88-59 - *cf* Annexe A

- 5 unigrammes: *a, v, i, o* et *n*,
- 4 bigrammes : *av, vi, io* et *on*,
- 3 trigrammes: *avi, vio* et *ion*.

StatLex recense tous les unigrammes, bigrammes et trigrammes du texte analysé et les classe par fréquence ou par ordre alphabétique.

Tous les caractères alphabétiques, excepté les tirets et les apostrophes, entrent dans la composition des bigrammes et des trigrammes.

Notons enfin que les deux caractères spéciaux *æ* et*œ* sont comptés par StatLex comme des unigrammes et non des bigrammes.

4.3 Initiales et finales

On appelle unigramme *initial* le premier unigramme rencontré dans un mot, et unigramme *final* le dernier unigramme rencontré dans un mot. La définition s'étend naturellement aux bigrammes et trigrammes. Par exemple, pour *bateau* on a:

- unigramme initial: *b*,
- unigramme final : *u*,
- bigramme initial : *ba*,
- bigramme final : *au*,
- trigramme initial: *bat*,
- trigramme final : *eau*.

C'est l'une particularités de StatLex de différencier initiales et finales.

4.4 Définition d'une phrase

Nous définissons une *phrase* comme une suite de caractères quelconques (alphabétiques ou non alphabétiques) terminée par un ou plusieurs des signes suivants, que nous appelons *indicateurs de fin de phrase*: le point (.), le point d'exclamation *(!)* et le point d'interrogation *(?)*.

Dans la plupart des cas, cette définition est satisfaisante, mais il existe certains cas particuliers plus ou moins rares:

- les *!* et les *?* ne concluent pas forcément une phrase: Pour lever cette ambigüité, il suffit de vérifier que le premier caractère alphabétique suivant un *!* ou un *?* est une majuscule,

- les *:* peuvent parfois servir d'indicateur de fin de phrase lorsque cette phrase termine un paragraphe et annonce le suivant; ce cas reste très rare,

- le cas le plus ennuyeux auquel nous ayons pensé est celui des abréviations, comme *S.N.C.F.*. On pourrait tester s'il y a un espace après chaque point, la règle en matière de typographie étant de laisser un espace après tout signe de ponctuation, mais ça ne lèverait pas les ambigüités du genre: *"... la S.A.R.L. Martin & Co ..."*.

4.5 Définition d'un paragraphe

Un *paragraphe* est un bloc de phrases à la suite les unes des autres, sans retour à la ligne à l'intérieur mm̂e du paragraphe. Dans la pratique, on appelle paragraphe une suite non nulle de phrases terminée par au moins deux retours à la ligne consécutifs, sauf le dernier.

4.6 Construction du dictionnaire des fréquences

4.6.1 Tirets et apostrophes

Les tirets de liaison et les apostrophes sont considérés comme des caractères alphabétiques, et, à ce titre, ils sont inclus dans le décompte des signes alphabétiques proposé par StatLex. Néanmoins, nous avons jugé préférable, lors de la constitution du dictionnaire, de ne pas compter les apostrophes comme des caractères alphabétiques. Effectivement, lorsqu'on analyse une séquence comme *l'avion*, nous nous rendons bien compte qu'il s'agit de la contraction de deux mots: *le* et *avion*. L'apostrophe joue bien ici le rôle d'une lettre: le *e*. Mais la compter comme telle reviendrait à distinguer dans le dictionnaire des fréquences les termes *avion* et *l'avion*. Nous avons préféré risquer de nous retrouver avec des occurences de termes comme *l*.

Il est bon de noter que dans d'autres langues, comme l'anglais, le problème devient encore plus complexe: comment traiter des cas comme *don't*, qu'il faudrait séparer en *do* et *not*, et des cas comme *Mike's bicycle*, ou le *'s* indique la possession? Ces questions dépassant largement le cadre du projet, nous avons préférer les laisser de côté.

Les tirets, par contre, sont pris en compte lors de la constitution du dictionnaire, lorsqu'ils sont immédiatement entourés de caractères alphabétiques. Par exemple, *tire-bouchon* est compté comme un seul mot. Cette convention nous semble nécessaire, même si elle conduit à accepter comme un seul mot des termes tels que *suis-je*.

4.6.2 Majuscules et minuscules

Lors de la création du dictionnaire on conserve les majuscules et les minuscules afin d'éviter certaines ambigüités, notamment sur les noms propres; sinon l'utilisateur du dictionnaire serait dans l'impossibilité de reconnaître le nom commun *pierre* du nom propre *Pierre*. Cependant, il faudra prendre garde que certains mots seront répertoriés sous deux noms différents selon qu'ils commencent ou non par une majuscule.

5 Le corpus

Dans le cadre linguistique du projet, nous avons réuni un ensemble de textes représentant les langues suivantes: allemand, anglais, espagnol, français et italien. Les analyses effectuées sur ces textes grâce aux outils décrits plus hauts doivent constituer la base de départ d'une base de données de référence sur les caractéristiques lexicales des cinq langues citées. Les données numériques devraient être, à plus ou moins court terme, accessibles sur le Web et remises à jour périodiquement par l'apport de nouveaux textes.

Tous les textes que nous avons trouvés proviennent de recherches sur le Web. Ces recherches n'ont pas forcément été aisées: si l'on trouve une pléthore de textes en anglais sur le Web, il est plus difficile (mais pas impossible) d'en trouver en espagnol ou en italien; en tout cas on a beaucoup moins de choix.

Nous avons visité plusieurs bibliothèques "virtuelles", comme ABU (*Association des Bibliophiles Universels*), ATHENA ou certains sites miroirs du projet Gütenberg. Mais à notre avis l'IPL ou *Internet Public Library* les surpasse tous, que ce soit par la quantité d'ouvrages disponibles que par la qualité de sa présentation. L'un des points que nous avons le plus apprécié est la classification des ouvrages par catégorie qui reprend la norme en vigueur dans toutes les bibliothèques "normales" (100 pour la philosophie, 200 pour la religion ...).

6 Les résultats

Un exemple de traitement, effectué sur le texte de Jules VERNES, *Le Tour du Monde en Quatre-Vingt Jours*, est donné dans l'Annexe B. Des tableaux d'unigrammes, issus d'analyses sur cinq textes de cinq langues différentes, sont présentés en Annexe C. Bien que sommaires, puisque les résultats proviennent d'un seul texte à chaque fois, ils permettent déjà de se faire une idée des répartitions des lettres de l'alphabet.

Les applications de ces tableaux de fréquence, aussi bien pour les unigrammes que pour les bigrammes et les trigrammes, en cryptographie et en compression de données sont bien connues, aussi nous ne nous attarderons pas dessus.

Notons que nous n'avons pas réussi à trouver, sur le Web, de telles tables de fréquence, malgré leur utilité en informatique. D'autre part, celles que l'on trouve dans la plupart des livres concernant la cryptographie et la compression de données ne sont pas forcément très fiables. Mettre en accès libre sur le Web les résultats numériques auxquels nous sommes parvenus ne semble donc pas une mauvaise idée, pour peu que ces résultats soient aisément accessibles.

L'analyse linguistique de ces données numériques n'est pas abordée ici. Nous laissons à d'autres, plus compétents et plus courageux que nous, le soin de se plonger dans ce vaste champ d'investigation.

7 Conclusion: réflexions sur le projet

Ce projet a été intéressant à plus d'un titre. Il a demandé un investissement important de notre part, notamment en travaux de recherche sur les à-côtés du projet, mais nous avons été largement payés en retour. Ainsi, nous sommes partis à la découverte de la statistique lexicale, domaine qui jusqu'alors nous était complètement inconnu. Nous avons aussi pu apprécier la richesse et la puissance du langage Perl, que nous ne connaissions pas auparavant. D'autre part, pour la première fois, nous avons été confronté à une exigence de qualité au niveau de la présentation, ce qui nous a conduit à découvrir LaTeX.

En conclusion, rappelons que ce projet est destiné à être complété par d'autres travaux, plus orientés vers la linguistique. On peut envisager, par exemple, la recherche de concordances dans le texte analysé, un traitement élaboré des exceptions et cas particuliers rencontrés lors de l'analyse du texte, voire tenter une lemmatisation du dictionnaire des fréquences, bref, ce n'est pas le travail qui manquera à nos successeurs.

Annexe A. Liste des caractères alphabétiques reconnus par la norme ISO-Latin

Code ASCII	Car .	Code ASCII	Car .	Code ASCII	Car .	Code ASCII	Car .
65	A	100	d	198	Æ	227	ã
66	B	101	e	199	Ç	228	ä
67	C	102	f	200	È	229	å
68	D	103	g	201	É	230	æ
69	E	104	h	202	Ê	231	ç
70	F	105	i	203	Ë	232	è
71	G	106	j	204	Ì	233	é
72	H	107	k	205	Í	234	ê
73	I	108	l	206	Î	235	ë
74	J	109	m	207	Ï	236	ì
75	K	110	n	208	Ð	237	í
76	L	111	o	209	Ñ	238	î
77	M	112	p	210	Ò	239	ï
78	N	113	q	211	Ó	240	ð
79	O	114	r	212	Ô	241	ñ
80	P	115	s	213	Õ	242	ò
81	Q	116	t	214	Ö	243	ó
82	R	117	u	215	Œ	244	ô
83	S	118	v	216	Ø	245	õ
84	T	119	w	217	Ù	246	ö
85	U	120	x	218	Ú	247	œ
86	V	121	y	219	Û	248	ø
87	W	122	z	220	Ü	249	ù
88	X	192	À	221	Ý	250	ú
89	Y	193	Á	222	Þ	251	û
90	Z	194	Â	223	ß	252	ü
97	a	195	Ã	224	à	253	ý
98	b	196	Ä	225	á	254	þ
99	c	197	Å	226	â	255	ÿ

Annexe B. Exemple de sortie PostScript : analyse de *Candide* (Voltaire)

Statistiques générales

Informations générales

Texte analysé	:	Voltaire.txt
Nombre de caractères	:	199794
Nombre de caractères alphabétiques (dont "-" et "’")	:	148762
Nombre de mots	:	33610
Nombre de phrases	:	1541
Nombre de paragraphes	:	61

Statistiques sur les mots

Longueur minimale	:	1	lettres
Longueur maximale	:	37	lettres
Longueur moyenne	:	4.37	lettres
Ecart type	:	2.61	lettres
σ/m	:	0.59	

Statistiques sur les phrases

Longueur minimale	:	1	mots
Longueur maximale	:	132	mots
Longueur moyenne	:	21.81	mots
Ecart type	:	18.25	mots
σ/m	:	0.83	

Statistiques sur les paragraphes

Longueur minimale	:	2	mots
Longueur minimale	:	0	phrases
Longueur maximale	:	3516	mots
Longueur maximale	:	157	phrases
Longueur moyenne	:	550.98	mots
Longueur moyenne	:	25.26	phrases
Ecart type	:	704.21	mots
Ecart type	:	32.76	phrases
σ/m	:	1.27	(mots)
σ/m	:	1.29	(phrases)

Unigrammes

Les pourcentages des initiales et des finales sont donnés par rapport au nombre total de mots (i.e. d'initiales et de finales), soit 33610 mots, et le pourcentage total par rapport au nombre total de caractères alphabétiques, soit 148762 caractères.

Caractères alphabétiques normaux

	init.	%	final.	%	freq. totale	%
e	2267	6.745	8940	26.599	20455	13.750
a	2080	6.189	1474	4.386	11303	7.598
s	1727	5.138	6455	19.206	11184	7.518
i	859	2.556	1304	3.880	10408	6.996
n	948	2.821	2581	7.679	10384	6.980
t	1034	3.076	4803	14.290	9891	6.649
u	767	2.282	1211	3.603	9476	6.370
r	770	2.291	1781	5.299	9031	6.071
o	481	1.431	122	0.363	7738	5.202
l	2864	8.521	1288	3.832	7444	5.004
d	3889	11.571	640	1.904	6081	4.088
m	1699	5.055	117	0.348	3943	2.651
p	2345	6.977	53	0.158	3599	2.419
c	1831	5.448	307	0.913	3503	2.355
v	1136	3.380	1	0.003	2518	1.693
q	1385	4.121	12	0.036	1855	1.247
g	317	0.943	17	0.051	1459	0.981
f	927	2.758	31	0.092	1395	0.938
b	661	1.967	5	0.015	1388	0.933
h	314	0.934	62	0.184	1107	0.744
j	691	2.056	137	0.408	806	0.542
x	.	.	468	1.392	584	0.393
y	171	0.509	161	0.479	466	0.313
z	1	0.003	259	0.771	350	0.235
k	1	0.003	3	0.009	19	0.013
w	.	.	.	.	.	.

Tab. 1: minuscules

	init.	%	final.	%	freq. totale	%
C	840	2.499	.	.	860	0.578
E	164	0.488	.	.	443	0.298
I	205	0.610	.	.	402	0.270
M	320	0.952	.	.	390	0.262
A	163	0.485	.	.	292	0.196
P	247	0.735	.	.	289	0.194
L	231	0.687	.	.	280	0.188
T	79	0.235	.	.	234	0.157
N	50	0.149	.	.	188	0.126
D	114	0.339	.	.	176	0.118
J	173	0.515	.	.	173	0.116
O	94	0.280	.	.	159	0.107
V	135	0.402	.	.	159	0.107
R	37	0.110	.	.	155	0.104
S	96	0.286	.	.	155	0.104
U	36	0.107	.	.	123	0.083
Q	79	0.235	.	.	85	0.057
H	40	0.119	.	.	80	0.054
B	64	0.190	.	.	71	0.048
G	22	0.065	.	.	54	0.036
F	38	0.113	.	.	41	0.028
X	1	0.003	.	.	14	0.009
W	12	0.036	.	.	12	0.008
Z	2	0.006	.	.	10	0.007
Y	2	0.006	.	.	7	0.005
K	.	.	.	.	.	.

Tab. 2: majuscules

	init.	%	final.	%	freq. totale	%
e+E	2431	7.233	8940	26.599	20898	14.048
a+A	2243	6.674	1474	4.386	11595	7.794
s+S	1823	5.424	6455	19.206	11339	7.622
i+I	1064	3.166	1304	3.880	10810	7.267
n+N	998	2.969	2581	7.679	10572	7.107
t+T	1113	3.312	4803	14.290	10125	6.806
u+U	803	2.389	1211	3.603	9599	6.453
r+R	807	2.401	1781	5.299	9186	6.175
o+O	575	1.711	122	0.363	7897	5.308
l+L	3095	9.209	1288	3.832	7724	5.192
d+D	4003	11.910	640	1.904	6257	4.206
m+M	2019	6.007	117	0.348	4333	2.913
p+P	2592	7.712	53	0.158	3888	2.614
c+C	2671	7.947	307	0.913	4363	2.933
v+V	1271	3.782	1	0.003	2677	1.800
q+Q	1464	4.356	12	0.036	1940	1.304
g+G	339	1.009	17	0.051	1513	1.017
f+F	965	2.871	31	0.092	1436	0.965
b+B	725	2.157	5	0.015	1459	0.981
h+H	354	1.053	62	0.184	1187	0.798
j+J	864	2.571	137	0.408	979	0.658
x+X	1	0.003	468	1.392	598	0.402
y+Y	173	0.515	161	0.479	473	0.318
z+Z	3	0.009	259	0.771	360	0.242
k+K	1	0.003	3	0.009	19	0.013
w+W	12	0.036	.	.	12	0.008

Tab. 3: minuscules + majuscules

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	freq. totale	%
é	577	1.717	616	1.833	2541	1708.097
à	619	1.842	705	2.098	686	461.139
è	.	.	.	.	572	384.507
ê	115	0.342	.	.	267	179.481
â	23	0.068	.	.	106	71.255
î	2	0.006	.	.	105	70.583
ô	3	0.009	7	0.021	101	67.894
û	.	.	1	0.003	75	50.416
ù	.	.	48	0.143	47	31.594
É	24	0.071	.	.	44	29.577
È	.	.	.	.	31	20.839
À	23	0.068	.	.	23	15.461
ï	.	.	1	0.003	14	9.411
Ô	7	0.021	.	.	9	6.050
Ê	1	0.003	.	.	3	2.017
ë	.	.	.	.	3	2.017
Î	.	.	.	.	1	0.672
Â	.	.	.	.	1	0.672

Caractères spéciaux (minuscules et majuscules)

	freq. totale	%
-	929	0.624
’	1882	1.265

	init.	%	final.	%	freq. totale	%
ç	1	0.003	.	.	80	0.054
Ç	.	.	.	.	1	0.001
æ	.	.	.	.	1	0.001
ñ	.	.	.	.	1	0.001

Bigrammes

Nombre de bigrammes différents trouvés dans le texte	:		33610
Nombre total de bigrammes trouvés dans le texte	:		112147
Nombre de bigrammes possibles avec les 26 lettres de l'alphabet	:	26*26 soit	676
Nombre de bigrammes possibles avec les 26 lettres de l'alphabet plus les 25 caractères accentués et les 8 caractères spéciaux (à prendre avec précaution)	:	59*59 soit	3481

Classement par ordre décroissant de fréquence des 20 bigrammes les plus fréquents:

	init.	%	final.	%	fréq. totale.	%
es	371	1.104	2264	6.736	2893	2.580
de	1956	5.820	1965	5.846	2756	2.457
le	1502	4.469	1441	4.287	2715	2.421
en	690	2.053	617	1.836	2607	2.325
re	331	0.985	1166	3.469	2320	2.069
ai	229	0.681	182	0.542	2298	2.049
ou	75	0.223	52	0.155	2269	2.023
on	281	0.836	861	2.562	2226	1.985
nt	.	.	1362	4.052	2009	1.791
an	76	0.226	25	0.074	1998	1.782
it	12	0.036	1448	4.308	1969	1.756
qu	1464	4.356	353	1.050	1928	1.719
ur	1	0.003	731	2.175	1541	1.374
is	15	0.045	746	2.220	1461	1.303
er	1	0.003	630	1.874	1358	1.211
et	917	2.728	1016	3.023	1319	1.176
il	784	2.333	743	2.211	1303	1.162
eu	84	0.250	125	0.372	1237	1.103
te	135	0.402	519	1.544	1229	1.096
nd	.	.	218	0.649	1209	1.078

Classement par ordre alphabétique des 20 bigrammes les plus fréquents:

	init.	%	final.	%	fréq. totale	%
ai	229	0.681	182	0.542	2298	2.049
an	76	0.226	25	0.074	1998	1.782
de	1956	5.820	1965	5.846	2756	2.457
en	690	2.053	617	1.836	2607	2.325
er	1	0.003	630	1.874	1358	1.211
es	371	1.104	2264	6.736	2893	2.580
et	917	2.728	1016	3.023	1319	1.176
eu	84	0.250	125	0.372	1237	1.103
il	784	2.333	743	2.211	1303	1.162
is	15	0.045	746	2.220	1461	1.303
it	12	0.036	1448	4.308	1969	1.756
le	1502	4.469	1441	4.287	2715	2.421
nd	.	.	218	0.649	1209	1.078
nt	.	.	1362	4.052	2009	1.791
on	281	0.836	861	2.562	2226	1.985
ou	75	0.223	52	0.155	2269	2.023
qu	1464	4.356	353	1.050	1928	1.719
re	331	0.985	1166	3.469	2320	2.069
te	135	0.402	519	1.544	1229	1.096
ur	1	0.003	731	2.175	1541	1.374

Trigrammes

Nombre de trigrammes différents trouvés dans le texte	:		2878
Nombre total de trigrammes trouvés dans le texte	:		80857
Nombre de trigrammes possibles avec les 26 lettres de l'alphabet	:	26*26*26 soit	17576
Nombre de trigrammes possibles avec les 26 lettres de l'alphabet plus les 25 caractères accentués et les 8 caractères spéciaux (à prendre avec précaution)	:	59*59*59 soit	205379

Classement par ordre décroissant de fréquence des 20 trigrammes les plus fréquents:

	init.	%	final.	%	fréq. totale	%
ent	91	0.271	831	2.472	1104	1.365
que	612	1.821	588	1.749	895	1.107
ait	16	0.048	794	2.362	844	1.044
les	548	1.630	722	2.148	736	0.910
ous	.	.	645	1.919	683	0.845
and	7	0.021	141	0.420	650	0.804
lle	.	.	476	1.416	646	0.799
eur	17	0.051	372	1.107	616	0.762
ais	11	0.033	374	1.113	616	0.762
tre	32	0.095	409	1.217	559	0.691
our	2	0.006	225	0.669	510	0.631
ant	7	0.021	300	0.893	496	0.613
ien	.	.	267	0.794	485	0.600
ndi	.	.	.	.	454	0.561
qui	372	1.107	354	1.053	445	0.550
vou	405	1.205	1	0.003	427	0.528
ide	.	.	552	1.642	418	0.517
tou	373	1.110	1	0.003	411	0.508
can	398	1.184	1	0.003	403	0.498
des	370	1.101	381	1.134	402	0.497

Classement par ordre alphabétique des 20 trigrammes les plus fréquents:

	init.	%	final.	%	fréq. totale	%
ais	11	0.033	374	1.113	616	0.762
ait	16	0.048	794	2.362	844	1.044
and	7	0.021	141	0.420	650	0.804
ant	7	0.021	300	0.893	496	0.613
can	398	1.184	1	0.003	403	0.498
des	370	1.101	381	1.134	402	0.497
ent	91	0.271	831	2.472	1104	1.365
eur	17	0.051	372	1.107	616	0.762
ide	.	.	552	1.642	418	0.517
ien	.	.	267	0.794	485	0.600
les	548	1.630	722	2.148	736	0.910
lle	.	.	476	1.416	646	0.799
ndi	.	.	.	.	454	0.561
our	2	0.006	225	0.669	510	0.631
ous	.	.	645	1.919	683	0.845
que	612	1.821	588	1.749	895	1.107
qui	372	1.107	354	1.053	445	0.550
tou	373	1.110	1	0.003	411	0.508
tre	32	0.095	409	1.217	559	0.691
vou	405	1.205	1	0.003	427	0.528

Dictionnaire des fréquences

Nombre de mots différents	:	5592
Nombre de mots en tout	:	33610

Classement par ordre décroissant de fréquence des 20 mots les plus fréquents

	fréq.	%
de	1269	3.776
et	876	2.606
le	670	1.993
la	652	1.940
à	615	1.830
les	494	1.470
un	488	1.452
que	461	1.372
il	448	1.333
d	385	1.145
l	381	1.134
en	361	1.074
Candide	350	1.041
des	345	1.026
qui	341	1.015
dans	313	0.931
qu	308	0.916
dit	277	0.824
je	273	0.812
ne	259	0.771

Classement par ordre alphabétique des 20 mots les plus fréquents

	total	%
Candide	350	1.041
d	385	1.145
dans	313	0.931
de	1269	3.776
des	345	1.026
dit	277	0.824
en	361	1.074
et	876	2.606
il	448	1.333
je	273	0.812
l	381	1.134
la	652	1.940
le	670	1.993
les	494	1.470
ne	259	0.771
qu	308	0.916
que	461	1.372
qui	341	1.015
un	488	1.452
à	615	1.830

Annexe C. Tables de fréquences d'unigrammes en allemand, anglais, espagnol, français et italien

Fréquence des unigrammes en francais, dans le texte "La Curée" de Zola (682K).

	init.	%	final.	%	total	%
e+E	8079	7.184	32062	28.510	76446	14.936
a+A	7465	6.638	5487	4.879	42523	8.308
s+S	8900	7.914	21334	18.970	40170	7.848
i+I	3380	3.006	3056	2.717	36689	7.168
t+T	4091	3.638	16829	14.965	36144	7.062
n+N	2298	2.043	7095	6.309	33561	6.557
r+R	3593	3.195	6067	5.395	32415	6.333
u+U	3351	2.980	4177	3.714	31324	6.120
l+L	12151	10.805	4543	4.040	32016	6.255
o+O	1652	1.469	53	0.047	23927	4.675
d+D	14498	12.892	2879	2.560	19904	3.889
c+C	7790	6.927	1188	1.056	15071	2.945
m+M	5479	4.872	416	0.370	14930	2.917
p+P	7778	6.916	221	0.197	12151	2.374
v+V	3056	2.717	12	0.011	8244	1.611
f+F	3648	3.244	81	0.072	5951	1.163
g+G	1597	1.420	84	0.075	5017	0.980
b+B	2746	2.442	6	0.005	5077	0.992
q+Q	3385	3.010	49	0.044	4924	0.962
h+H	1287	1.144	162	0.144	4513	0.882
x+X	7	0.006	1645	1.463	2284	0.446
j+J	1831	1.628	168	0.149	2152	0.420
y+Y	378	0.336	315	0.280	1124	0.220
z+Z	3	0.003	407	0.362	592	0.116
w+W	24	0.021	1	0.001	44	0.009
k+K	5	0.004	.	.	20	0.004

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	total	%
é	2163	1.923	1627	1.447	8611	1682.381
à	1928	1.714	2196	1.953	2214	432.562
è	.	.	.	.	1776	346.987
ê	172	0.153	.	.	1093	213.546
â	32	0.028	.	.	423	82.644
ô	4	0.004	.	.	339	66.232
û	.	.	29	0.026	277	54.119
ù	.	.	260	0.231	260	50.798
î	12	0.011	.	.	223	43.569
ï	.	.	.	.	25	4.884
ë	.	.	9	0.008	10	1.954
ü	.	.	.	.	1	0.195

Caractères spéciaux (minuscules et majuscules)

	total	%
-	2684	0.524
’	6302	1.231

	init.	%	final.	%	total	%
ç	65	0.058	.	.	363	0.071
Ç	20	0.018	.	.	20	0.004

Fréquence des unigrammes en italien, dans le texte “Dei Sepolcri” de Ugo Foscolo(11K).

	init.	%	final.	%	total	%
e+E	181	8.868	574	28.123	1158	12.780
a+A	188	9.211	340	16.659	908	10.021
i+I	151	7.398	368	18.030	903	9.966
o+O	69	3.381	287	14.062	727	8.023
l+L	186	9.113	176	8.623	651	7.185
r+R	45	2.205	69	3.381	619	6.831
n+N	72	3.528	78	3.822	567	6.258
t+T	83	4.067	3	0.147	501	5.529
s+S	164	8.035	4	0.196	376	4.150
d+D	197	9.652	39	1.911	368	4.061
c+C	158	7.741	1	0.049	314	3.465
u+U	45	2.205	20	0.980	278	3.068
m+M	88	4.312	2	0.098	278	3.068
p+P	138	6.761	.	.	241	2.660
g+G	59	2.891	.	.	185	2.042
v+V	71	3.479	1	0.049	174	1.920
f+F	80	3.920	.	.	107	1.181
h+H	4	0.196	3	0.147	86	0.949
b+B	20	0.980	.	.	73	0.806
z+Z	2	0.098	.	.	41	0.452
q+Q	24	1.176	.	.	31	0.342
j+J	.	.	.	.	1	0.011
k+K	.	.	.	.	.	.
w+W	.	.	.	.	.	.
x+X	.	.	.	.	.	.
y+Y	.	.	.	.	.	.

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	total	%
à	.	.	19	0.931	22	242.799
ò	.	.	17	0.833	18	198.654
é	.	.	15	0.735	16	176.581
í	.	.	7	0.343	13	143.472
è	8	0.392	9	0.441	12	132.436
î	.	.	.	.	11	121.399
ú	1	0.049	9	0.441	10	110.363
ü	.	.	.	.	2	22.073
ê	.	.	.	.	1	11.036
ä	.	.	.	.	1	11.036

Caractères spéciaux (minuscules et majuscules)

	total	%
-	232	2.560
’	136	1.501

	init.	%	final.	%	total	%

Fréquence des unigrammes en espagnol, dans le texte “Vida de Lazarillo” (123K, auteur inconnu).

	init.	%	final.	%	total	%
e+E	1900	9.459	3966	19.745	11199	13.550
a+A	1683	8.379	3712	18.481	10787	13.052
o+O	281	1.399	3751	18.675	8133	9.840
s+S	1194	5.944	2385	11.874	5616	6.795
n+N	630	3.137	1730	8.613	5456	6.601
r+R	271	1.349	1163	5.790	4754	5.752
l+L	1541	7.672	1099	5.471	4654	5.631
i+I	102	0.508	667	3.321	4599	5.564
d+D	2046	10.186	115	0.573	4064	4.917
u+U	283	1.409	137	0.682	3746	4.532
c+C	1605	7.991	.	.	3191	3.861
t+T	886	4.411	5	0.025	2977	3.602
m+M	1636	8.145	11	0.055	2984	3.610
p+P	1492	7.428	.	.	1887	2.283
y+Y	1394	6.940	1267	6.308	1619	1.959
b+B	335	1.668	.	.	1449	1.753
q+Q	1098	5.466	.	.	1413	1.710
h+H	690	3.435	7	0.035	1014	1.227
g+G	161	0.802	1	0.005	910	1.101
v+V	538	2.678	10	0.050	917	1.110
j+J	61	0.304	2	0.010	464	0.561
z+Z	6	0.030	55	0.274	352	0.426
f+F	251	1.250	2	0.010	351	0.425
x+X	.	.	1	0.005	18	0.022
k+K	.	.	.	.	.	.
w+W	.	.	.	.	.	.

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	total	%

Caractères spéciaux (minuscules et majuscules)

	total	%
-	95	0.115
'	.	.

	init.	%	final.	%	total	%

Fréquence des unigrammes en anglais, dans le texte “The Raven” d’Edgar Poe (7K).

	init.	%	final.	%	total	%
e+E	28	2.518	228	20.504	627	12.662
t+T	166	14.928	118	10.612	448	9.047
o+O	64	5.755	29	2.608	372	7.512
n+N	48	4.317	88	7.914	379	7.653
r+R	26	2.338	83	7.464	343	6.926
a+A	96	8.633	15	1.349	345	6.967
h+H	42	3.777	27	2.428	293	5.917
i+I	72	6.475	32	2.878	323	6.523
s+S	112	10.072	93	8.363	278	5.614
l+L	42	3.777	38	3.417	230	4.645
d+D	42	3.777	123	11.061	195	3.938
m+M	70	6.295	22	1.978	163	3.292
u+U	19	1.709	5	0.450	122	2.464
g+G	19	1.709	68	6.115	123	2.484
y+Y	9	0.809	86	7.734	104	2.100
f+F	51	4.586	29	2.608	93	1.878
p+P	23	2.068	4	0.360	97	1.959
b+B	67	6.025	.	.	95	1.918
w+W	56	5.036	14	1.259	79	1.595
c+C	29	2.608	1	0.090	74	1.494
v+V	9	0.809	.	.	68	1.373
k+K	1	0.090	9	0.809	31	0.626
q+Q	9	0.809	.	.	9	0.182
x+X	.	.	.	.	3	0.061
j+J	2	0.180	.	.	2	0.040
z+Z	.	.	.	.	.	.

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	total	%

Caractères spéciaux (minuscules et majuscules)

	total	%		
-	43	0.868		
'	.	. . .	13	0.263

	init.	%	final.	%	total	%

Frequence des unigrammes en allemand dans le texte "Wanderer" (279K, auteur inconnu).

	init.	%	final.	%	total	%
e+E	4967	9.927	8160	16.309	38390	17.622
n+N	2502	5.001	10483	20.952	22648	10.396
i+I	2770	5.536	1037	2.073	17070	7.836
r+R	1297	2.592	6196	12.384	14901	6.840
s+S	4427	8.848	4050	8.094	13936	6.397
t+T	1221	2.440	5950	11.892	12699	5.829
d+D	6402	12.795	2546	5.089	11492	5.275
h+H	1871	3.739	2378	4.753	11028	5.062
a+A	3345	6.685	665	1.329	10597	4.864
l+L	1221	2.440	1170	2.338	8531	3.916
u+U	2546	5.089	1035	2.069	8192	3.760
c+C	416	0.831	372	0.743	6945	3.188
g+G	2770	5.536	1279	2.556	6904	3.169
m+M	2134	4.265	1465	2.928	5354	2.458
o+O	566	1.131	733	1.465	4076	1.871
b+B	1325	2.648	362	0.724	4033	1.851
w+W	3037	6.070	175	0.350	4091	1.878
f+F	1148	2.294	483	0.965	3062	1.406
z+Z	1455	2.908	226	0.452	2520	1.157
k+K	953	1.905	394	0.787	2495	1.145
v+V	1464	2.926	125	0.250	1690	0.776
p+P	341	0.682	71	0.142	991	0.455
j+J	554	1.107	11	0.022	568	0.261
y+Y	11	0.022	12	0.024	77	0.035
x+X	4	0.008	5	0.010	35	0.016
q+Q	19	0.038	1	0.002	26	0.012

Caractères accentués (minuscules et majuscules)

	init.	%	final.	%	total	%
ü	310	0.620	82	0.164	1508	692.207
ä	111	0.222	68	0.136	1252	574.697
ö	42	0.084	28	0.056	671	308.004
Ü	78	0.156	.	.	78	35.804
Ä	12	0.024	.	.	12	5.508
Ö	5	0.010	.	.	5	2.295
è	.	.	.	.	1	0.459

Caractères spéciaux (minuscules et majuscules)

	total	%
-	1084	0.498
’	11	0.005

	init.	%	final.	%	total	%
SS	29	0.058	472	0.943	881	0.404

Annexe D. Bibliographie et adresses *http*

Ouvrages concernant la statistique lexicale :

- E. Brunet, *Le Vocabulaire Français de 1789 à nos Jours, d'après les données du Trésor de la Langue Française*, Slatkine-Champion, Genève, Paris, 1981.
- E. Brunet, *Le Vocabulaire de Proust*, t. I, II & III, Slatkine-Champion, Genève, Paris, 1983.
- L. Lebart & A. Salem, *Analyse statistique des données textuelles*, Dunod, Paris, 1988.
- Ch. Muller, *Principes et Méthodes de Statistique Lexicale*, Paris, Hachette, 1977.
- Ch. Muller, *Langue Française, Linguistique Quantitative, Informatique*, recueil d'articles, 1980-84, Slatkine-Champion, Genève, Paris, 1985.

Ouvrages sur Perl :

- L. Wall, T. Christiansen & L. Scwhartz, *Programming Perl, 2nd edition*, O'Reilly & Associates Inc, Sebastopol, CA, 1996, trad. française J. Zundel, *Programmation en Perl*, O'Reilly International Thomson, Paris, 1997.
- G. Johnson, *Introduction to Perl or, Learn Perl in Two Hours*, University of Missouri - Columbia, 1996.
- J. Vromans, *Perl Reference Guide*,

Adresses *http* :

- `http: //134.59.31.3/ brunet/pub/brunet.html` : page d'accueil du professeur E. Brunet.

- `http: //www.ciril.fr/INALF/` : page d'accueil de l'*Institut National de la Langue Française*, avec accès (limité) à Frantext et au TLF.
- `http: //www.ipl.org/` : page d'accueil de l'*Internet Public Library*.
- `http: //www.ipl.org/reading/books/` : index thématique des ouvrages en ligne.
- `http: //www.perl.com/perl/` : page d'accueil de Perl.
- `http: //www.perl.org/` : page d'accueil du Perl Institute.
- `http: //www.info.univ-angers.fr/~gh/cours_gH/statlexp.html` : page d'accueil de StatLex.

Et le mieux pour la fin :

- `http: //www.info.univ-angers.fr/aides/tutperl.html` : tuteur pour Perl écrit par Candice CHAUVE et Véronique COLIN.