Tutoriels pour l'analyse textométrique

[Tutoriels]

André Salem
salem@msh-paris.fr

Résumé : Ces *tutoriels* devrait permettre à l'utilisateur débutant de ***Lexico3*** (et de *mkAlign*) de se familiariser avec les différentes fonctionnalités du logiciel, à partir de corpus de recherche concrets et, au delà de cette prise en main, d'entrevoir quelques-unes des possibilités offertes par l'approche textométrique des corpus de textes.

Complétant la documentation disponible sur ***Lexico3*** :

- ***Manuel d'utilisation*** ;
- ***User's Manual***, traduction anglaise du même manuel ;
- ***Les 10 premiers pas avec Lexico3***, manuel de prise en main ;
- http://www.cavi.univ-paris3.fr/lexico3www site web *de* ***Lexico3***[1],

et sur *mkAlign* :

- ***Manuel d'utilisation en ligne*** :
  http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

ces ***Tutoriels*** devrait permettre à l'utilisateur débutant, au delà d'une simple prise en main, de se familiariser avec les différentes fonctionnalités de ces logiciels, à partir d'un corpus de recherche concret et d'entrevoir quelques-unes des possibilités offertes par l'approche textométrique des corpus de textes.

- Le corpus ***Père Duchesne*** choisi dans les deux premiers tutoriels pour servir de base à cette exploration guidée est le même que celui utilisé dans les brochures précédentes. Ce corpus à fait l'objet de plusieurs études de caractère pluridisciplinaire dont on trouvera les références dans la dernière section. La ressource textuelle ***duchn.txt*** qui sert de support à ce tutoriel est diffusée en tant que corpus d'essai sur toutes les versions du logiciel ***Lexico***. Accessible sur le CD-Rom ***Lexico3***, elle est installée automatiquement dans le dossier Lexico3 créé lors de l'installation du logiciel. Elle peut également être téléchargée directement depuis le site du logiciel.
- Le corpus ***Investiture Obama*** utilisé dans le troisième tutoriel est disponible en ligne sur le site de *mkAlign*.

---

[1] Le logiciel, la documentation et les ressources textuelles (parmi lesquelles la ressource ***duchn.txt***) utilisées dans le présent manuel peuvent être téléchargées depuis ce site.

On a tenté, dans ce qui suit, de trouver un compromis acceptable entre la nécessité de présenter les principales fonctionnalités du logiciel que le lecteur pourra utiliser dans d'autres entreprises textométriques et le compte-rendu d'une recherche qui nous a conduit à agencer l'utilisation de ces méthodes en fonction des objectifs fixés au départ de l'étude, des résultats que nous avons obtenus, mais aussi des perspectives de recherche qui se sont ouvertes à cette occasion. Dans chaque cas, nous nous sommes efforcés de faire en sorte que le lecteur dispose des informations suffisantes pour reproduire par ses propres moyens les fonctionnalités décrites. Ces informations sont rassemblées, à chaque étape, en fin de paragraphe dans un encart annoncé par la séquence `=== Lexico3 === ou === mkAlign ===`

On se reportera aux manuels d'utilisation pour une description plus détaillée de chacune des fonctionnalités**.**

Le ***Tutoriel n°1****, Exploration du corpus Père Duchesne,* devrait permettre à l'utilisateur de se familiariser avec les notions de *ressources numériques textuelles, de corpus textométriques,* de dépouillement d'un corpus en *unités textuelles*, de partition d'un corpus textométrique et d'acquérir quelques notions sur les principales méthodes textométriques qui permettent d'explorer ces corpus de textes.

Le ***Tutoriel n°2****, Séries textuelles chronologiques,* est consacré à l'étude d'un type de corpus particulier que l'on rencontre très souvent dans le domaine textométrique, qui est celui des corpus rassemblant une série de textes produits au cours du temps par un même émetteur. L'étude de ces corpus obéit à des règles particulières que l'on s'est efforcé de décrire dans ce tutoriel.

Le ***Tutoriel n°3****, Investiture Obama,* est consacré à l'étude d'un corpus aligné avec *mkAlign*.

.

Tutoriel n°1:

Exploration du corpus *Père Duchesne*

Corpus, unités textuelles, partitions, méthodes textométriques

[Duchesne1]

Apprendre à :

- Construire une ressource textométrique
- Introduire des jalons textuels
- Choisir des unités d'analyse textométrique
- Utiliser les outils textométriques de base
- Conduire une exploration textométrique

1 Le corpus *Père Duchesne*

Le corpus ***Père Duchesne*** que l'on considère ici est constitué de 96 livraisons d'un journal édité par Jacques-René Hébert (1757-1794), parues entre juillet 1793 et mars 1794, durant la Révolution française, dans une période de luttes particulièrement âpres entres différentesfactions politiques. Du fait de sa reproduction et de son acheminement systématique en direction des armées, ce journal a connu une diffusion exceptionnelle pour l'époque qui lui permet de prétendre au titre de *premier media de masse de l'époque moderne*. Le corpus a été réuni dans le cadre d'une étude plus large portant sur la presse jacobine de l'époque et a donné lieu, depuis, à de nombreuses publications[2]. On peut voir sur la figure 1 une reproduction de la première et de la dernière page d'un des exemplaires du *Père Duchesne*, feuille imprimée, pliée en quatre, vendue à la fois par abonnement et à la criée dans les rues de Paris.

1.1 Etablissement de la version numérique du corpus

Lors de la saisie initiale sous forme numérique de cette ressource textuelle, quelques normalisations orthographiques mineures ont été effectuées à l'époque par les chercheurs qui ont transcrit le corpus sous forme numérique. Ainsi, les terminaisons en *oit* ont toutes été ramenées à l'orthographe moderne en *ait* (ex : *foutoit* est devenu *foutait*). Les enrichissements

[2] Des recherches sur ce corpus ont été réalisées dans le cadre de l'équipe *Révolution française* de laboratoire de l'ENS de St-Cloud [Guilhaumou, 19xx], [Salem, 1993].

textuels (italiques, gras, etc.) ont été négligés. Les majuscules du texte ont été remplacées par le signe ***** suivi de la minuscule correspondante (ex : *Paris* -> **paris*)[3].

1.2 Balisage du corpus

Afin de permettre la comparaison entre les différents textes réunis en un même corpus, on a introduit des *jalons textuels* ou *balises* servant à délimiter des *parties*. Dans cette version de Lexico3, les *balises* qui permettent d'introduire les partitions sont du type[4] :

<type=contenu>

Chaque ***type*** particulier de balise (partie située avant le signe « = ») permet de définir une partition du corpus. Pour un type fixé, si on ignore tous les autres types, les différents *contenus* (partie située après le signe « = ») correspondent à autant de parties différentes .dans le corpus. Ainsi, par exemple, la sélection de la clé ***numero*** (<numero= xx>) permet de découper le corpus en 96 parties correspondant chacune à une des 96 livraisons qui constituent le corpus.

Les balises introduites dans le corpus ***Duchn.txt*** sont :

- <*Epg=x*> qui permettent de localiser chacune des pages à l'intérieur d'un même numéro ;
- <*numero=x*> qui permettent de délimiter chacune des 96 livraisons du corpus ;
- <*mois=x*> qui permettent d'opérer un regroupement des livraisons parues à l'intérieur de chaque période d'un mois. Ces périodes sont notées (M1, M2, …, M8) ;
- <*quinzaine=xx*> qui permettent d'opérer un regroupement de ces mêmes livraisons par quinzaines.
- <*semaine =xxxx*> qui permettent d'opérer un regroupement de ces mêmes livraisons par semaines.

2 Zones textuelles

Pour pouvoir s'appuyer sur une division du texte en paragraphes, on a fait précéder chacun des paragraphes par le caractère « **§ »**[5].

Il est également possible de réaliser un découpage correspondant approximativement à un découpage en phrase en fournissant aux outils qui assurent un tel découpage une liste de caractères délimiteurs de phrases (par exemple : « . ?! »)

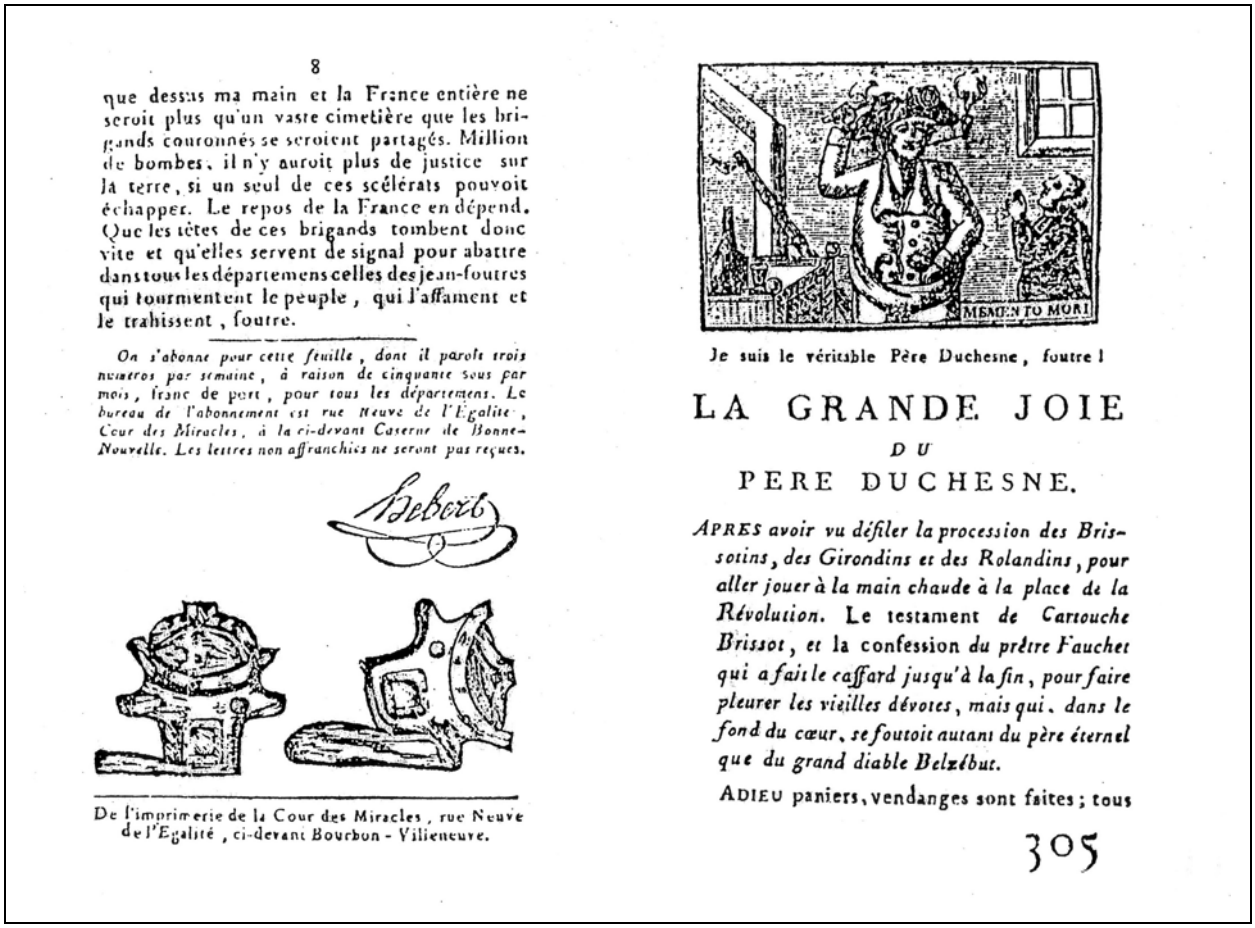
Comme on va le voir dans les sections qui suivent, les découpages en partitions constituent avec les systèmes de découpage en sections un dispositif articulé qui permet de renvoyer les constats textométriques à des zones textuelles délimités avec une précision que l'on peut faire varier.

---

[3] Cette technique permet de différer la décision de savoir si les formes qui ne diffèrent que par une majuscule initiale doivent être décomptées séparément. Lors des segmentations ultérieures de la ressource on aura le choix entre deux options : a) on considère que le caractère * est un caractère délimiteur et les formes **abc* et *abc* seront alors considérées comme deux occurrences d'un même type (*abc*) ; b) on décide que le caractère * n'est pas un délimiteur et les formes **abc* et *abc* seront alors considérées comme des occurrences de deux types différents.

[4] Le système de balisage du texte décrit dans ce paragraphe a été élaboré avant .l'apparition de normes plus consensuelles dans la communauté des études textuelles réalisées avec l'aide de l'ordinateur. Les prochaines versions du logiciel prennent en compte les formats d'entrée des textes construits à partir de la norme XML (EXtensible Mark Up Langage). Les fonctionnalités textométriques de ces différentes formes de balisage restent cependant très voisines.

[5] Ce remplacement peut être effectué de manière générique à l'aide d'un logiciel de traitement de texte en remplaçant le caractère « retour-chariot » par la séquence « retour-chariot » suivi de « § ». Avec le logiciel Word, par exemple on utilisera les commmandes : Chercher : ^p Remplacer par :^p §.



8

que dessus ma main et la France entière ne seroit plus qu'un vaste cimetière que les brigands couronnés se seroient partagés. Million de bombes, il n'y auroit plus de justice sur la terre, si un seul de ces scélérats pouvoit échapper. Le repos de la France en dépend. Que les têtes de ces brigands tombent donc vite et qu'elles servent de signal pour abattre dans tous les départemens celles des jean-foutres qui tourmentent le peuple, qui l'affament et le trahissent, foutre.

On s'abonne pour cette feuille, dont il paroît trois numéros par semaine, à raison de cinquante sous par mois, franc de port, pour tous les départemens. Le bureau de l'abonnement est rue Neuve de l'Egalité, Cour des Miracles, à la ci-devant Caserne de Bonne-Nouvelle. Les lettres non affranchies ne seront pas reçues.

Hebert

De l'imprimerie de la Cour des Miracles, rue Neuve de l'Egalité, ci-devant Bourbon-Villeneuve.

MEMENTO MORI

Je suis le véritable Père Duchesne, foutre !

LA GRANDE JOIE
DU
PERE DUCHESNE.

APRES avoir vu défiler la procession des Brissotins, des Girondins et des Rolandins, pour aller jouer à la main chaude à la place de la Révolution. Le testament de *Cartouche Brissot, et* la confession *du prêtre Fauchet qui a fait le cafard jusqu'à la fin, pour faire pleurer les vieilles dévotes, mais qui, dans le fond du cœur, se foutoit autant du père éternel que du grand diable Belzébut.*

ADIEU paniers, vendanges sont faites; tous

305

Figure 1a :
Fac simile de l'édition originale du numéro 305 du ***Père Duchesne*** (1793)

```
<numero=305><Epg=1>
§ la grande joie, du *père *duchesne après avoir vu défiler la procession des *brissotins,
des *girondins et des rolandins, pour aller jouer à la main chaude à la place de la
révolution. le testament de *cartouche,*brissot , et la confession du prêtre *fauchet qui a
fait le cafard jusqu'à la fin,pour faire pleurer les vieilles dévotes, mais qui,dans le
fond du coeur, se foutait autant du père éternel que du grand diable *belzébut.
§ adieu paniers,vendanges sont faites ; tous <Epg=2> les châteaux en *espagne, que vous
avez bâtis, infâmes *brissotins, s'en vont tous en fumée. non, foutre, non la république
que vous aviez vendue aux brigands couronnés ne sera point déchirée. le roi *georges
*dandin, et *pitt, porte-esprit, ont tiré leur poudre aux oiseaux. nous serons républicains
malgré toutes les guinées de l'*angleterre, et tout l'or de l'*autriche, et de l'*espagne
.partout nos affaires prennent la meilleure tournure. Les brigands de la *vendée, sont
dispersés et leurs cadavres engraissent la terre qu'ils ont souillée par leurs crimes; ce
qu'il en reste est cerné de toutes parts et va bientôt tomber sous les coups des généreux
défenseurs de la république; tandis, foutre, que l'armée du nord partout victorieuse est
aux trousses des gros talons et des pieds plats que commande *cobourg ; tandis que *mons,
ouvre ses portes au brave *jourdan ,*brissot, et sa clique marchent à l'échafaud.
```

Figure 1b :
Extrait de l'édition numérisée du numéro 305 du ***Père Duchesne*** (1793)

3 Unités textuelles

Quelles sont les unités qui circulent dans un texte sociopolitique ? Quelles séquences doit-on constituer en unités insécables afin d'opérer des comptages dans les textes ? L'expérience du dépouillement informatisé des corpus de textes montre que ces interrogations constituent à chaque fois des questions centrales pour la recherche en cours et qu'elle ne peuvent être réglées une fois pour toutes et a priori.

Dans le corpus ***Duchn***, par exemple, on serait tenté de constituer en une seule unité le terme *sans-culottes*, pourvu d'une haute fréquence et qui renvoie à un référent assez clairement identifiable à l'époque. Sans doute, le tiret qui unit les deux formes graphiques n'est il pas de même nature que celui qui unit les formes dans *dit-il*. Une autre question se pose alors : Comment traiter le problème automatiquement sans être obligé de trancher au cas par cas ?

Notre expérience nous a conduit à privilégier dans un premier temps les dépouillements appuyés sur des caractères aisément automatisables (appartenance ou non de chacun des caractères à une liste préétablie délimiteurs/non-délimiteurs) et à repousser à une seconde phase l'observation d'unités plus complexes : séquences de formes, cooccurrences etc. Pour la séquence *sans-culottes* présentée plus haut, nos préfèrerons opérer dans un premier temps un dépouillement appuyé sur la segmentation en deux formes distinctes (tiret = délimiteur) laissant à d'autres procédures le soin de repérer ensuite la séquence des deux formes *sans culottes* aisément repérable du fait même de sa forte répétition dans le corpus.

Par ailleurs, au fil des recherches, est apparue la nécessité de généraliser fortement la définition du type d'unité textuelle prise en compte par les analyses textométriques. Le ***type généralisé*** ou ***Tgen*** est défini comme une sélection d'occurrences prise dans le texte. Cette définition permet de prendre en compte les types constitués à partir de critères de sélection difficiles à formaliser[6].

3.1 Le dépouillement en formes graphiques

La première phase de l'exploration textométrique est constituée par la segmentation du corpus textuel en unités qui serviront de base aux décomptes ultérieurs les ***occurrences*** (en anglais *tokens*). A l'issue de cette phase, une seconde phase d'identification constitue un dictionnaire des ***formes*** ou des ***types*** (en anglais *types*). Les *types* regroupent en une même unité chaque classe d'occurrences identiques d'après le critère d'identification retenu[7].

==== *Lexico3* ==== *Segmentation initiale*

- ✓ Lancer Lexico3
- ✓ Sélectionner l'icône ***Segmentation*** (1ère icône en haut à gauche)
- ✓ Choisir le fichier texte à segmenter (Duchn.txt)
- ✓ Accepter les délimiteurs de forme proposés « par défaut » (bouton ***OK***)

[6] Sur les types généralisés, cf. [Lamalle & Salem, 2002]

[7] Selon les études, on trouve des critères d 'identification dont la nature peut varier. Dans certains types de dépouillements, dits *dépouillement en forme graphiques*, on se base sur l'identité graphique des séquences considérées, d'autres formes de dépouillements font intervenir la nature grammaticale des occurrences isolées, voire des informations de type sémantique. On consultera sur ce sujet [Labbé xxx],

Différents outils textométriques que l'on décrira plus loin permettent d'apprécier la fréquence, la répartition, la spatialisation des occurrences relevant de chacun des types constitués à cette étape. Les résultats fournis par ces outils ne sont pas indépendants des types d'unités constitués, mais les mêmes outils s'appliquent à tous les types constitués de la sorte.

La qualité première d'une norme de dépouillement est d'être à la fois simple à énoncer et à automatiser. Le dépouillement du corpus ***Duchn*** en formes graphiques délimitées par les délimiteurs proposés par défaut conduit aux résultats suivants :

nombre des occurrences :	141 182
nombre des formes :	11 070
nombre des hapax :	5 056
forme la plus fréquente *de* :	6 130

3.2 Etude globale des types simples

Ces données sont accessibles en activant l'icône ***PCLC***, dès qu'une partition quelconque a été choisie. Sur le panneau qui apparaît alors on peut étudier l'accroissement du vocabulaire au fil du corpus en activant l'icône ***ACCV***.

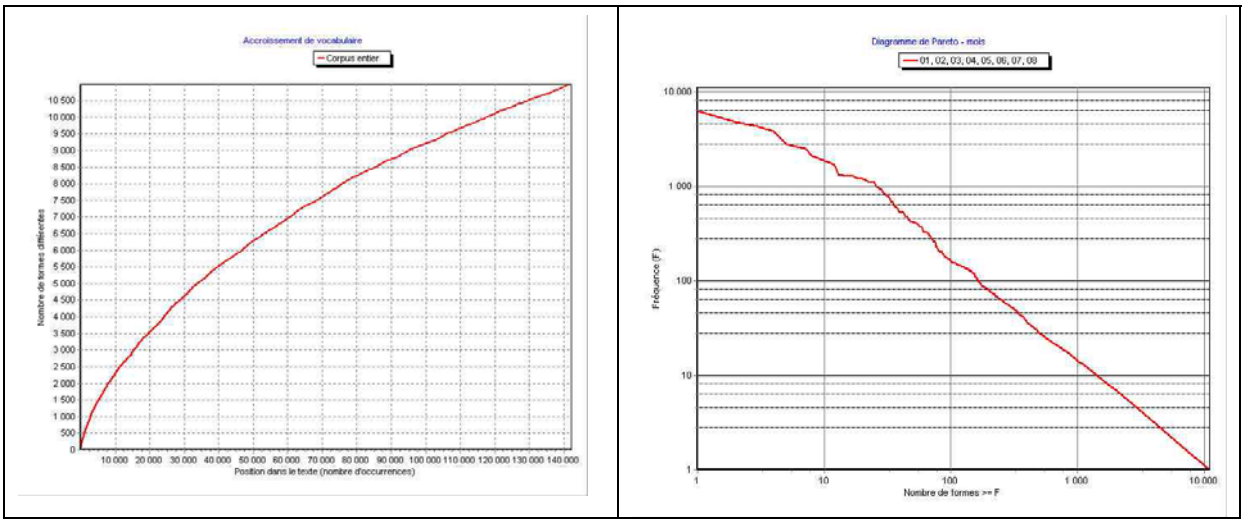


Figure 2 :

Accroissement du vocabulaire et structure de la gamme des fréquences

==== *Lexico3* ==== *Accroissement du vocabulaire*

- ✓ Sélectionner l'icône ***Statistiques par parties*** (5ème icône à partir de la gauche)
- ✓ Choisir un type de clé qui déterminera la partition active du corpus
- ✓ Sélectionner l'icône ***PCLC*** (5ème icône à partir de la gauche)
- ✓ Sélectionner, sur la droite du panneau (5ème bouton à partir du haut) le bouton ***AC (comme Accroissement du vocabulaire)***
- ✓ Le diagramme apparaît dans une fenêtre spécifique. On peut constituer le diagramme correspondant à chacune des parties, ou à un ensemble de parties en les sélectionnant l'une après l'autre et en les glissant sur la fenêtre du Diagramme d'accroissemenet.

Guide de lecture pour la figure 2

Le ***Diagramme d'accroissement du vocabulaire*** que l'on trouve sur la gauche permet d'observer l'apparition de nouvelles formes au fur et à mesure que l'on avance dans le corpus.
Comme c'est toujours le cas pour les corpus textuels, la courbe connaît une croissance rapide au début du corpus ; cette croissance ralentit à mesure que l'on avance dans le corpus. On remarque, par-delà cette caractéristique globale, des zones d'accroissement plus fort ainsi que des paliers durant lesquels l'apport de nouvelles formes est plus faible.

Le ***Diagramme de Pareto*** que l'on trouve sur la permet de visualiser la structure de la gamme des fréquences.

- L'axe vertical permet de représenter la fréquence *F* des formes du textes (laquelle varie de *1* à *Fmax*, fréquence maximale calculée pour le texte ***T***).
- Sur l'axe horizontal, on porte la quantité : *nombre de formes du texte dont la fréquence est supérieure à F*.
- Avant de tracer le Diagramme, on transforme chacune de ces quantités en son logarithme décimal.

Le Diagramme ainsi obtenu prend alors approximativement la forme droite que l'on appelle *Droite de Zipf* en l'honneur de Georges. Kingsley Zipf qui a montré que ce type de procédure réalisée à partir de larges catégories de textes permet de mettre en évidence une propriété statistique commune aux dépouillements en unités lexicales. Cette propriété est parfois présentée sous la forme excessivement simplifiée :

Rang x *fréquence* = *Constante*

3.3 Les types complexes

Les limites que l'on rencontre dès que l'on commence à explorer un corpus textuel à partir de formes isolées de leur contexte immédiat conduisent rapidement à la nécessité d'observer la répartition d'unités plus complexes.

Les segments répétés du Père Duchesne

La fonctionnalité ***Segments répétés*** permet d'établir la liste de toutes les séquences de formes répétées sans changement à différents endroits du corpus dont la fréquence totale dépasse un seuil minimal ***F*** préalablement fixé par l'utilisateur. Les segments ainsi sélectionnés peuvent ensuite être triés selon différents critères : longueur, fréquence, etc.

On retrouvera parmi les segments longs les expressions favorites du ***Père Duchesne***, comme :

```
employer le vert et le sec pour                    15
perdre le goût du pain                             12
ses bons avis aux braves sans-culottes             15
brouiller les cartes                               20
```

Parmi les segments plus courts et plus fréquents on retrouvera les unités composées évoquées plus haut comme :

```
*sans culottes                                    398
jean foutres                                      136
brigands couronnés                                 49
```

Une forme particulière de présentation des contextes du segment *tous les* qui compte 871 occurrences dans le corpus permettra de constater que cet opérateur textuel sert entre autres choses à introduire des entités présentées plutôt comme négatives et contre lesquelles le ***Père Duchesne*** propose de se mobiliser. On peut voir un extrait de cet inventaire au tableau 4.

Cependant l'ensemble constitué par la totalité de segments répétés qui se chevauchent de manière quasiment inextricable se révèle toujours d'une grande complexité et défie toute

description synthétique. En textométrie on utilise plutôt ce vaste ensemble pour en extraire des unités dont la répartition dans le corpus est particulièrement déséquilibrée. Du fait de leur longueur, ces séquences sont, dans l'ensemble, plutôt moins polysémiques que les formes simples isolées de leur contexte immédiat, ce qui facilite grandement l'interprétation des résultats.

Si l'on classe, par contre les lignes de cet inventaire d'après la fréquence de la forme qui suit, la séquence pôle, comme cela a été fait au tableau 4, on s'aperçoit que l'opérateur tous les introduit, la plupart du temps une notion appartenant à un registre négatif *(traîtres, brigands,* etc.*)* même si cette règle subit des exceptions notables[8].

Tableau 5 :

Début de l'inventaire distributionnel des segment répétés
pour la séquence *tous les* dans le corpus *Père Duchesne*.
(classement par ordre de fréquence décroissante de la forme qui suit)

```
871   --- tous les
       32 tous les hommes
       30 tous les traîtres
       29 tous les brigands
       26 tous les départements
       24 tous les ennemis
       21 tous les fripons
       20 tous les bons
       19 tous les scélérats
       15 tous les maux
       14 tous les patriotes
       13 tous les citoyens
       12 tous les bougres
       12 tous les muscadins
       12 tous les peuples
       12 tous les trônes
       11 tous les conspirateurs
       11 tous les coquins
       10 tous les jours
       10 tous les nobles
```

==== ***Lexico3*** ==== ***Segments répétés***

- ✓ Sélectionner l'icône ***Segments répétés*** (4ème icône à partir de la gauche)
- ✓ Sélectionner un seuil de fréquence minimal pour les segments
- ✓ Les segments apparaissent dans un onglet sur la partie gauche. Ils peuvent être triés selon différents critères (longueur, fréquence, ordre lexicographique) en cliquant sur le bandeau situé au-dessus de la colonne correspondante.
- ✓ Chaque sélection, simple ou multiple, réalisée dans la fenêtre des segments peut ensuite être analysée comme un tout, en transitant éventuellement par la fenêtre ***groupe de formes*** à l'aide des différents outils disponible (concordance, histogramme, carte des sections, etc.)

[8] Actuellement, les fonctionnalités de ***Lexico3*** ne permettent pas d'obtenir directement l'état présenté au tableau 5. Cet état a été obtenu en triant, à l'aide d'un tableur (Excel), les lignes du tableau 4.

Tableau 4 :

Début de l'inventaire distributionnel des segment répétés
après la séquence *tous les* dans le corpus *Père Duchesne*.
(classement par ordre lexicographique de la forme qui suit)

```
871 ---- ---- ---- ---- ----  tous les
      2 ---- ---- ---- ----  tous les *brissotins
      7 ---- ---- ---- ----  tous les *français
      3 ---- ---- ---- ----  tous les *jacobins
          17 ---- ---- ----  tous les *sans culottes
               2 ---- ----  tous les *sans culottes à
               3 ---- ----  tous les *sans culottes de
                    2 ----  tous les *sans culottes de *paris
               2 ---- ----  tous les *sans culottes se
      3 ---- ---- ---- ----  tous les aboyeurs
      7 ---- ---- ---- ----  tous les accapareurs
      2 ---- ---- ---- ----  tous les ambitieux
      7 ---- ---- ---- ----  tous les amis
               5 ---- ----  tous les amis de la
                    4 ----  tous les amis de la liberté
      7 ---- ---- ---- ----  tous les aristocrates
               2 ---- ----  tous les aristocrates et les
                    2 ----  tous les aristocrates tous les royalistes
      6 ---- ---- ---- ----  tous les autres
      6 ---- ---- ---- ----  tous les badauds
          4 ---- ---- ----  tous les badauds de
      6 ---- ---- ---- ----  tous les bandits
          2 ---- ---- ----  tous les bandits qui
      2 ---- ---- ---- ----  tous les beaux
      3 ---- ---- ---- ----  tous les biens
     20 ---- ---- ---- ----  tous les bons
               7 ---- ----  tous les bons *sans culottes
                    2 ----  tous les bons *sans culottes se
          7 ---- ---- ----  tous les bons citoyens
          4 ---- ---- ----  tous les bons républicains
     12 ---- ---- ---- ----  tous les bougres
                    2 ----  tous les bougres à poil qui ont
          8 ---- ---- ----  tous les bougres qui
               2 ---- ----  tous les bougres qui ont
      2 ---- ---- ---- ----  tous les boutiquiers
      2 ---- ---- ---- ----  tous les bras
      3 ---- ---- ---- ----  tous les braves
          2 ---- ---- ----  tous les braves bougres
     29 ---- ---- ---- ----  tous les brigands
          19 ---- ---- ----  tous les brigands couronnés
               2 ---- ----  tous les brigands couronnés ce
               3 ---- ----  tous les brigands couronnés et
                    2 ----  tous les brigands couronnés et les
               2 ---- ----  tous les brigands couronnés qui
          2 ---- ---- ----  tous les brigands et
          3 ---- ---- ----  tous les brigands qui
                    2 ----  tous les brouillards de la *tamise se
      2 ---- ---- ---- ----  tous les bureaux
      5 ---- ---- ---- ----  tous les châteaux
               3 ---- ----  tous les châteaux en *espagne
                    2 ----  tous les châteaux en *espagne que
      3 ---- ---- ---- ----  tous les chefs
          2 ---- ---- ----  tous les chefs de
      2 ---- ---- ---- ----  tous les chiens
          4 ---- ---- ----  tous les ci devant
     13 ---- ---- ---- ----  tous les citoyens
      2 ---- ---- ---- ----  tous les coeurs
          5 ---- ---- ----  tous les coins de
      7 ---- ---- ---- ----  tous les complots
          2 ---- ---- ----  tous les complots qu
                    3 ----  tous les complots que l on
     11 ---- ---- ---- ----  tous les conspirateurs
          3 ---- ---- ----  tous les contre révolutionnaires
     11 ---- ---- ---- ----  tous les coquins
          6 ---- ---- ----  tous les coquins qui
          18 ---- ---- ----  tous les coups de
              17 ---- ----  tous les coups de chien
                    2 ----  tous les coups de chien des ennemis
                    4 ----  tous les coups de chien qu
```

Cooccurrences pour un type donné

Si l'on se donne un découpage du corpus en sections (parties, paragraphes, phrases, groupes de phrases) et une forme-pôle (nous prendrons comme ci-dessus l'exemple de la forme : ***proie***) il est possible de constituer la liste des formes et des segments répétés qui trouvent, d'après un calcul statistique particulier[9], un nombre élevé d'occurrence dans les mêmes sections que la forme-pôle. Nous avons trouvé ici : *aux, gibet, oiseaux, perd*. Le retour aux contextes nous confirmera que ces formes entrent avec le pôle choisi dans des associations récurrentes insuffisamment stéréotypées, cependant, pour constituer des segments répétés, du type : *le gibet ne perd jamais sa proie*… etc.

Les calculs de cooccurrences fournissent, de manière symétrique, des listes d'unités textuelles qui trouvent au contraire, toujours d'après le même calcul statistique, très peu d'occurrences au voisinage d'une forme-pôle donnée. On pourrait appeler ces formes des formes *anti-cooccurrentes* ou des formes *évitées* ou *repoussées* par la forme-pôle. L'étude des listes de forme dont les occurrences sont repoussées par la présence dans un contexte proche d'une unité-pôle fixée peut parfois se révéler très instructive.

==== ***Lexico3*** ==== ***Cooccurrences***

- ✓ Demander une carte des sections (7ème icône à partir de la gauche)
- ✓ Choisir un délimiteur de section (paragraphe ou groupe de délimiteurs de phrase . !?)
- ✓ Faire glisser une forme sur la carte à partir du dictionnaire ou de toute autre liste
- ✓ Appuyer sur l'icône des cooccurrences, à l'extrême droite de la 2eme ligne d'icônes
- ✓ Choisir une fréquence minimale et un seuil de probabilité pour les cooccurrents
  NB : si la liste des segments répétés a été préalablement demandée, on obtiendra également les segments jugés cooccurrents spécifiques pour le pôle sélectionné.

Constituer des groupes de formes

On peut constituer des groupes de formes en associant plusieurs types élémentaires, par exemple : le singulier et le pluriel d'un même substantif, les différentes flexions d'un même verbe, les différentes formes d'un adjectif *(nouveau, nouvelle, nouveaux, nouvelles)*[10]. On peut également constituer des groupes à partir de toutes sortes de critères, grammaticaux, sémantiques, etc.

3.4 Les types généralisés (TGen)

Au-delà de ces constructions simples, l'outil *groupe de formes* permet également de constituer des unités qui correspondent au codage d'un thème particulier. Nous avons utilisé cette possibilité pour coder les occurrence d'un thème important chez le *Père Duchesne*, celui de la *mise à mort*. Pour repérer les occurrences de ce thème dans le corpus ***Duchn***, nous avons du

---

[9] Nous utilisons ici un simple calcul hypergéométrique pour comparer le nombre des occurrences du candidat cooccurrent dans les sections ou est attestée la forme-pôle avec sa fréquence dans l'ensemble du corpus. Pour des compléments sur les méthodes de calcul des cooccurrences, cf. par exemple [Lafon XX] et [Heiden XX].

[10] Cette possibilité offerte à l'utilisateur n'implique pas qu'il est toujours utile de rassembler dans tous les cas le pluriel et le singulier d'un même substantif lesquels peuvent avoir des répartitions très différentes dans le corpus. D'autre part le regroupement des types correspondant à l'adjectif *nouveau* mentionné plus haut absorbera également, dans l'état actuel de la fonctionnalité ***groupe de formes***, les occurrences qui correspondent aux formes substantivales *un nouveau, une nouvelle,* etc.

relire attentivement le texte d'un bout à l'autre en nous concentrant sur les seules expressions susceptibles de renvoyer à ce thème[11]

Au delà de la mention des substantifs *guillotine,échafaud, rasoir national*, etc., le recensement des formules susceptibles de constituer des occurrences du thème de la *mise à mort* permet de sélectionner les expression suivantes :

Tableau 2 :

Exemples d'expressions renvoyant au thème de la *mise à mort*
sélectionnées d'après une lecture cursive corpus ***Duchn***.

faire jouer X *à la main chaude*
avoir joué *à la main chaude*
(faire) *perdre le goût du pain* (numéro 272)
mettre la tête à la fenêtre (numéro 272)
jouer à la boule (numéro 280)
mettre la tête à la lunette (numéro 286)
(faire) *faire la bascule* (numéro 303)
faire la fatale culbute (numéro 304)
voyager dans la charrette de Samson (numéro 294)
*grimper (*ou *paraître) dans le vis-à-vis de maitre Samson* (numéro 296)
faire le voyage dans la voiture aux trente-six portières (numéro 321)
éternuer dans le sac (numéro 317)
cracher dans le sac (numéro 341)
avoir la tête dans le sac (numéro 304)
faire la grimace au pont rouge (numéro 319)

Il serait totalement déraisonnable d'espérer qu'une telle tâche puisse être confiée à une machine. Par contre, une fois repérées les séquences qui renvoient à ce thème, telle par exemple la séquence *la tête à la fenêtre* il est facile de repérer automatiquement toutes les occurrences du segment répété.

Tableau 3 :

Concordances du segment répété *la tête à la fenêtre* dans le corpus ***Duchn***.

```
 fallait , bon gré , mal gré , mettre la tête à la fenêtre , a tiré de sa manche à
ibunaux pour faire mettre promptement la tête à la fenêtre à la louve autrichienne
çoit pas d ' un pauvre bougre qui met la tête à la fenêtre . § cependant , foutre
t leurs véritables amoureux de mettre la tête à la fenêtre . § convention national
e vont dans cette semaine mettre tous la tête à la fenêtre , et six tribunaux comp
ue le dernier des *brissotins ait mis la tête à la fenêtre , foutre . § la grande
 comme son maître , va bientôt mettre la tête à la fenêtre . § il est donc vrai qu
de la convention , et il mettra aussi la tête à la fenêtre , le roi *coco . § les
punis . pas un conspirateur n ' a mis la tête à la fenêtre . le tribunal révolutio
fin à bon port . l ' ogre royal a mis la tête à la fenêtre , les *brissotins ne so
 pas échappé , et il aurait aussi mis la tête à la fenêtre . § lorsque sa foutue t
chicane , pour les empêcher de mettre la tête à la fenêtre ; mais j ' espère que t
ra pas plus à vous empêcher de mettre la tête à la fenêtre , qu ' elle n ' a pu s
 joie de voir bientôt ce butor mettre la tête à la fenêtre . ses bons avis aux bra
omme son confrère *capet , aurait mis la tête à la fenêtre , si l ' infâme *dumour
 allumer la guerre civile , aient mis la tête à la fenêtre . son grand discours au
ur qu ' elle fasse promptement mettre la tête à la lunette à l ' infâme *brissot ,
que tôt ou tard chacun d ' eux mettra la tête à la lunette comme leur confrère *ca
ps que nous aurions dû voir sa bougre de tête à la lunette . mieux vaut tard que j
```

[11] Notons qu'une bonne connaissance du corpus et de la période concernée peuvent se révéler indispensable pour repérer certaines de ces formules. Ainsi, le fait d'être informé par une source historique possiblement extérieure au corpus, que X a été exécuté dans une période précédente, permet de comprendre la formule *X a craché dans le sac* comme un équivalent de *X a été mis à mort*.

L'ensemble de ces mentions peut être rassemblé en un groupe de forme particulier dont on étudiera ensuite la variabilité au sein du corpus[12].

==== *Lexico3* ==== *Groupe de forme*

- ✓ Sélectionner l'icône ***Groupe de formes*** (8ème icône à partir de la gauche)
- ✓ Donner un nom au groupe (dans la boîte de dialogue supérieure)

Plusieurs possibilités s'offrent alors pour constituer le groupe

- ✓ Sélectionner un par un les constituants du groupe à partir du dictionnaire
- ✓ Utiliser les fonctionnalités génériques « est le début de ce que je recherche » etc.
- ✓ Sélectionner formes segments à l'aide d'une expression rationnelle[13].
- ✓ La flèche rouge située en haut à droite constitue un point d'accroche pour l'ensemble du groupe ainsi constitué. Elle peut être *traînée* vers tous les outils qui acceptent un TGen.

4 Etude la distribution d'un type

4.1 Les outils de base

L'outil concordances

L'outil ***concordances*** permet de rassembler toutes les occurrences relatives à un type donné en les munissant d'un petit fragment de contexte et de les trier selon différents critères, cf. tableau 1. En faisant varier la taille du contexte, l'ordre de présentation (ici les contextes sont triés en fonction de la forme qui suit le pôle sélectionné). A l'aide de cet outil, le chercheur peut opérer des rapprochements qu'une lecture cursive du texte ne lui aurait sans doute pas permis de saisir (ici, par exemple : *perdre sa proie* et *sa proie lui échappe*).

Tableau 1 :

Concordance de la forme *proie* dans le corpus ***Duchn***

pendant quelques instants ces oiseaux de	**proie**	avaient disparus , foutre , et depuis que
ès avoir rogné les ongles des oiseaux de	**proie**	de la finance ; après avoir détruit la mé
amée qui rugit quand on lui a arraché sa	**proie**	, elle poussait des cris affreux . " ains
fuite , mais le gibet ne perd jamais sa	**proie**	, et tôt ou tard les pigeons reviendront
e te dire que le gibet ne perd jamais sa	**proie**	? il y a plus de dix ans que tu aurais fa
er numéro que le gibet ne perd jamais sa	**proie**	. le jean - foutre est hors de la loi ,
s ' entre - déchiraient pour avoir leur	**proie**	, les *sans - culottes se fortifiaient ,
ut tout dévorer , tout engloutir ; si sa	**proie**	lui échappe , il devient enragé , et il
' examine ce tigre qui rugit de voir sa	**proie**	lui échapper . " me voilà au bout de mes
' aux tigres et aux ours de déchirer la	**proie**	qui tombe sous leurs griffes ; ils regard

[12] L'esquisse de procédure ainsi décrite ne garantit pas totalement que l'on a intégré aux comptages ***toutes*** les occurrences du textes susceptibles de relever du thème choisi. Un autre chercheur confronté au même texte disposant d'autres connaissances aurait peut-être inclus (ou exclu) d'autres occurrences susceptibles de modifier les comptages d'ensemble.

[13] Cf. sur ce point le manuel d'utilisation de ***Lexico3*** pg **xxxxxxxx**.

.

==== *Lexico3* ==== *Concordances*

- ✓ Sélectionner l'icône ***Concordances*** (3ème icône à partir de la gauche) et
- ✓ Entrer une forme dans la boite de dialogue *forme* ***(ex : proie)***
- ✓ Choisir l'ordre de présentation des contextes (Tri = après, avant, ordre du texte)
- ✓ Choisir [éventuellement] un regroupement par parties (si une partition a été sélectionnée)

L'outil statistiques par parties

L'outil ***statistiques par parties*** permet de juger de la répartition des occurrences relevant d'un même type dans les différentes parties d'une partition, cf. figure 2.

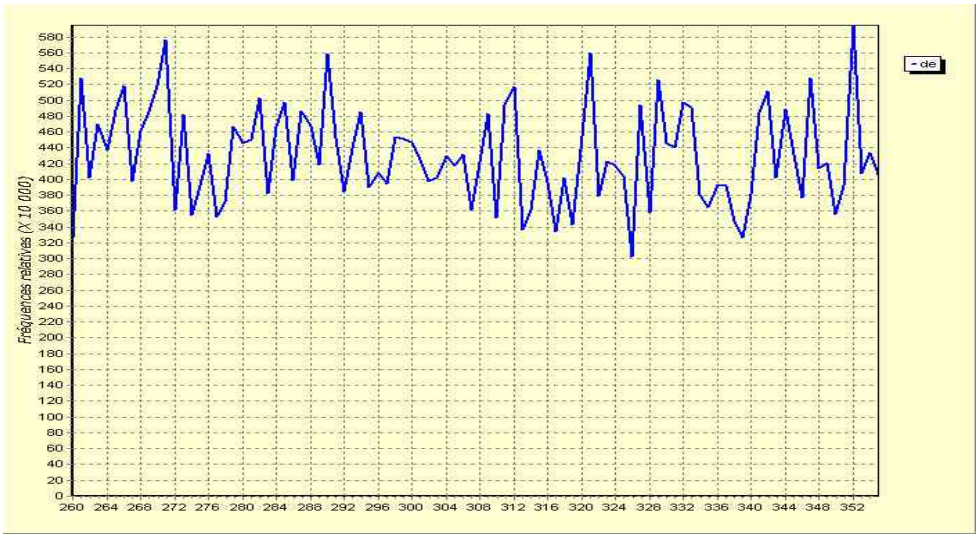


Figure 3 :

Ventilation des occurrences de la forme *de* en fréquence relative dans les 96 numéros du corpus ***Duchn***.

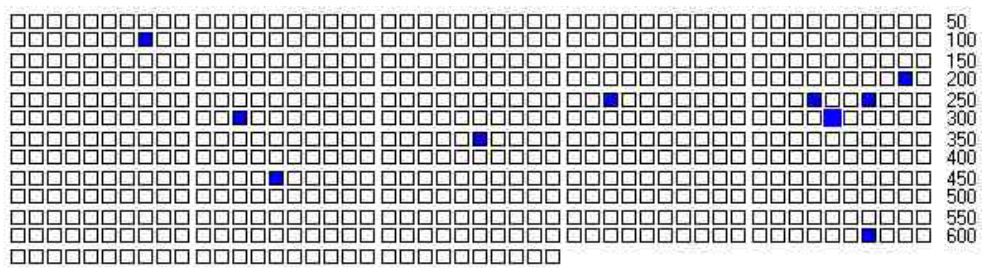
==== *Lexico3* ==== *Statistiques par parties*

- ✓ Sélectionner l'icône ***Statistiques par parties*** (5ème icône à partir de la gauche)
- ✓ Choisir le type de clé qui déterminera la partition active du corpus
- ✓ Faire glisser une forme à partir du dictionnaire ou de toute autre liste ***(ex : proie)***

L'outil carte des sections

L'outil *carte des sections* permet une visualisation globale de la répartition des occurrences qui relèvent d'un type donné dans l'ensemble du corpus. Chacun des carrés représente un élément particulier du texte découpé en sections. On a décidé, pour établir la carte présentée à la figure 4, de représenter chacun des paragraphes du texte, repérable, grâce à notre codage préalable, à ce qu'il s'ouvre sur un caractère §. La sélection à l'aide de la souris, d'un paragraphe particulier provoque son affichage dans une fenêtre située

sous la carte des sections. Comme on le verra plus loin (§ XX), il est possible, de matérialiser une partition sur ce type de carte.



[§ 295] te voilà enfin sur la fatale sellette , infâme *brissot ; quand je te prédisais que tu ferais une mauvaise fin , n ' étais je pas un bon prophète ! n ' avais je pas raison de te dire que le gibet ne perd jamais sa **proie** ? il y a plus de dix ans que tu aurais fait la grimace au pont *rouge , si l ' on t ' avait rendu justice , mais tes escroqueries d ' alors , toutes les bourses <Epg=4>que tu coupais dans l ' ancien régime , n ' étaient que des coups d ' essai des brigandages que tu as fait pendant la révolution.

Guide de lecture pour la figure 4

Chacun des carrés de la séquence du haut représente un des paragraphes du texte original. La numérotation de la colonne de droite permet de calculer le numéro de chaque paragraphe. La forme *proie* a été traînée sur la carte des sections provoquant le coloriage des sections ou elle est attestée. Le texte du paragraphe 295 sélectionné par l'utilisateur est affiché en bas de la figure. Les occurrences de la forme sélectionnée y sont mises en évidence.

Figure 4 :

Localisation des occurrences de la forme *proie*
sur une carte des sections du corpus ***Duchn***.

==== *Lexico3* ==== *Carte des sections*

- ✓ Sélectionner l'icône ***Carte des sections*** (5ème icône à partir de la gauche)
- ✓ Choisir un délimiteur de section qui servira à construire la carte
- ✓ Faire glisser une forme sur la carte à partir d'une liste ***(ex : proie)***
- ✓ Choisir [éventuellement] un regroupement par parties, si une partition a été sélectionnée

Intermède – utilisation de la partition en pages

La clé <*Epg=x*> ou x prend les valeurs 1, 2, 3, … , 8 permet de repérer les changements de page à l'intérieur de chaque numéro.[14] Comme c'est le cas pour chaque type de clé, il est possible d'utiliser la fonctionnalité ***Partition*** de *Lexico3* pour constituer, à partir de cette clé, un corpus en 8 parties. La partition réalisée à partir de la clé ***Epg*** rassemble donc en une même partie toutes les premières pages de chacun des 96 numéros, la seconde partie est composée de toutes les secondes pages et ainsi de suite jusqu'à la huitième partie qui rassemble les dernières pages de chaque numéro.

[14] Le contenu de la clé Epg : x – prend des valeurs de 1 à 8, car la publication, une grande feuille imprimée pliée en quatre par la suite est toujours composée de 8 pages.

Quel peut-être l'intérêt d'une telle partition au plan textométrique ?

Ce découpage du corpus, un peu curieux au premier abord, permet de mettre en évidence une particularité intéressante dans l'utilisation du vocabulaire. Comme on peut le voir sur la figure 5, la fréquence de la forme *foutre*, assez faible dans la première page, se maintient à un niveau stable dans les pages intérieures pour croître brutalement à l'intérieur de la dernière page. Ce déséquilibre traduit à coup sûr un procédé récurrent employé par l'auteur dans la conclusion de son périodique.

Une hypothèse explicative se présente immédiatement au vu de cette ventilation que des recherches ultérieures viendront conforter par la suite : la forme *foutre*, juron favori du ***Père Duchesne*** est utilisée assez modérément dans l'introduction de chaque livraison, sa fréquence relative reste stable dans les pages intermédiaires mais la conclusion du journal se fait sur un style plus « musclé » qui recourt largement à l'emploi de jurons et d'invectives. La visualisation des occurrences de *foutre* sur la carte des sections permet de localiser facilement des exemples de cette utilisation particulière.

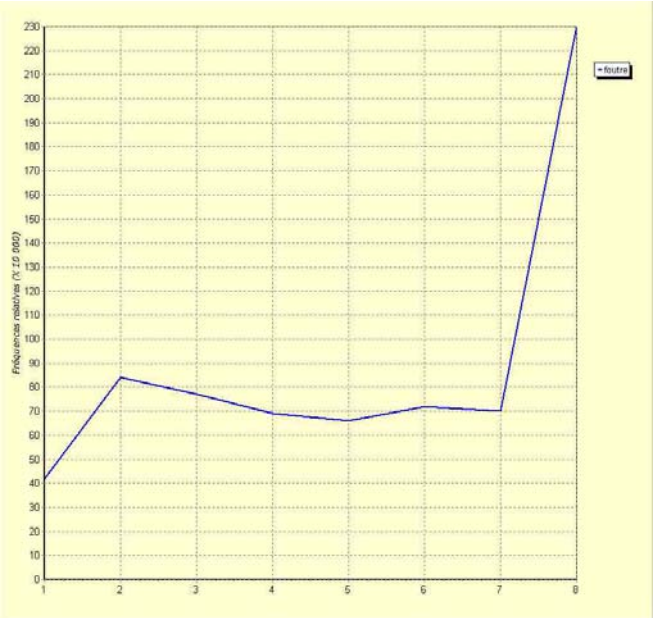


Figure 5

Ventilation des occurrences de la forme *foutre* dans les 8 pages du journal
(le numéro de page figure en abscisse sur le graphique)

On trouvera ci-dessous un exemple, parmi beaucoup d'autres possibles, d'une séquence prélevée dans la page qui clôt le *numéro 347* du corpus.

> **Numero 347** <Epg=8> imposture. ainsi donc, **foutre**, vive la raison vivent la vérité et l'humanité ! au **foutre** les prêtres, qui ne savent que mentir, tromper, voler et égorger, **foutre.**

L'analyse du vocabulaire spécifique de cette huitième partie nous permettra de dégager un ensemble de formes qui obéissent à ce même schéma d'utilisation : *vive, vos, soyez, peuple,* etc. En résumé, les résultats de cette expérience qui n'avait au départ d'autre finalité que celle de vérifier le fonctionnement correct du logiciel nous ont suggéré une possibilité d'exploration textométrique à laquelle nous n'avions pas pensé au départ. La mise en œuvre extrêmement simplifiée de la division du corpus en partie permet, on le voit, d'entreprendre à peu de frais, des expériences dont les résultats peuvent se révéler intéressants.

5 Méthodes textométriques

Plusieurs méthodes statistiques permettent d'éclairer la structure d'un corpus textuel à partir de comparaisons réalisées entre les fragments du corpus. La partition du corpus constitue une étape très importante dans l'analyse comparative des textes dans la mesure où les oppositions qu'il sera possible de mettre en évidence entre les parties soumises à comparaison dépendent étroitement du choix de la partition initiale.

Tableau 6 :

Tête du tableau lexical constitué par le décompte des 30 formes les plus fréquentes du corpus dans les 8 parties d'une partition en 8 mois

	Forme	P1	P2	P3	P4	P5	P6	P7	P8
0	de	886	875	853	753	746	757	669	591
1	les	641	687	569	549	534	687	555	526
2	la	641	550	593	579	479	502	484	447
3	et	449	524	480	530	463	467	461	399
4	le	348	376	398	374	355	321	327	265
5	à	382	384	350	361	319	308	266	262
6	que	349	390	351	298	317	287	287	217
7	qui	222	276	310	261	271	268	267	204
8	des	262	262	240	201	245	243	274	217
9	il	300	233	285	199	248	229	203	128
10	l	221	260	250	236	209	201	206	187
11	pour	214	249	252	220	194	181	183	172
12	en	171	199	153	192	167	169	147	111
13	qu	184	189	225	164	184	139	115	98
14	d	170	158	170	151	162	161	152	150
15	nous	180	250	167	132	155	100	182	104
16	un	156	190	155	157	186	157	135	131
17	est	162	137	171	142	181	190	134	115
18	tous	163	176	150	140	116	156	149	145
19	ils	145	181	147	137	183	104	160	130
20	ne	159	170	168	128	181	124	129	106
21	du	164	143	161	138	145	137	123	107
22	foutre	149	141	141	103	124	126	151	165
23	pas	140	189	187	125	160	103	105	91
24	vous	157	189	140	219	86	99	135	72
25	je	111	125	97	131	146	152	132	85
26	n	129	162	146	110	124	100	101	83
27	dans	129	133	128	123	115	96	111	89
28	on	87	153	110	77	151	137	82	59
29	a	118	119	131	106	95	118	81	66
30	plus	115	99	101	77	107	94	124	81

Le Tableau lexical

On commence par constituer un tableau qui compte autant de ***colonnes*** que la partition choisie compte de parties et autant de ***lignes*** que le vocabulaire du corpus compte de formes différentes. A l'intersection de la ligne *i* et de la colonne *j*, on notera le nombre d'occurrences que la forme *i* trouve dans la partie *j*, du corpus. Le tableau 6 présente les 30 premières lignes du *tableau lexical* réalisé à partir d'une partition du corpus ***Duchn*** en 8 parties dont chacune correspond à un mois de parution du journal[15].

[15] Un fichier coran.don est créé par ***Lexico3*** qui contient le tableau lexical, précédé de quelques paramètres nécessaires aux analyses multidimensionnelles.

Cette petite partie extraite du tableau lexical (8 parties x 11 070 formes) permet d'imaginer la difficulté qu'il y aurait à essayer d'analyser un tel tableau. Cependant, plusieurs méthodes statistiques permettent d'extraire de ces tableaux des faits particulièrement remarquables sur lesquels il est pratique de concentrer son attention dans une première approche. Pour ces méthodes et pour les machines qui les mettent en œuvre, la dimension des tableaux lexicaux ne constitue pas de difficulté particulière.

La division en 96 parties, numérotées de 260 à 355 selon la numérotation originale de la publication, paraît a priori la division la plus *naturelle* du corpus ***Duchn***. La clé <*numéro=x*> introduite lors du codage du corpus permet de réaliser cette partition en 96 numeros. Nous allons étudier cette partition en combinant deux méthodes d'analyse statistiques très complémentaires et couramment utilisées en textométrie : l'analyse factorielle des correspondances (AFC)et l'analyse des spécificités.

5.1 Etude de la partition du corpus Duchn en 96 numéros

On trouve sur la figure 6, une représentation de l'ensemble des 96 numéros fournie par l'analyse factorielles des correspondances à partir du tableau (96 numéros x 1420 formes de fréquence supérieure à 10).[16]

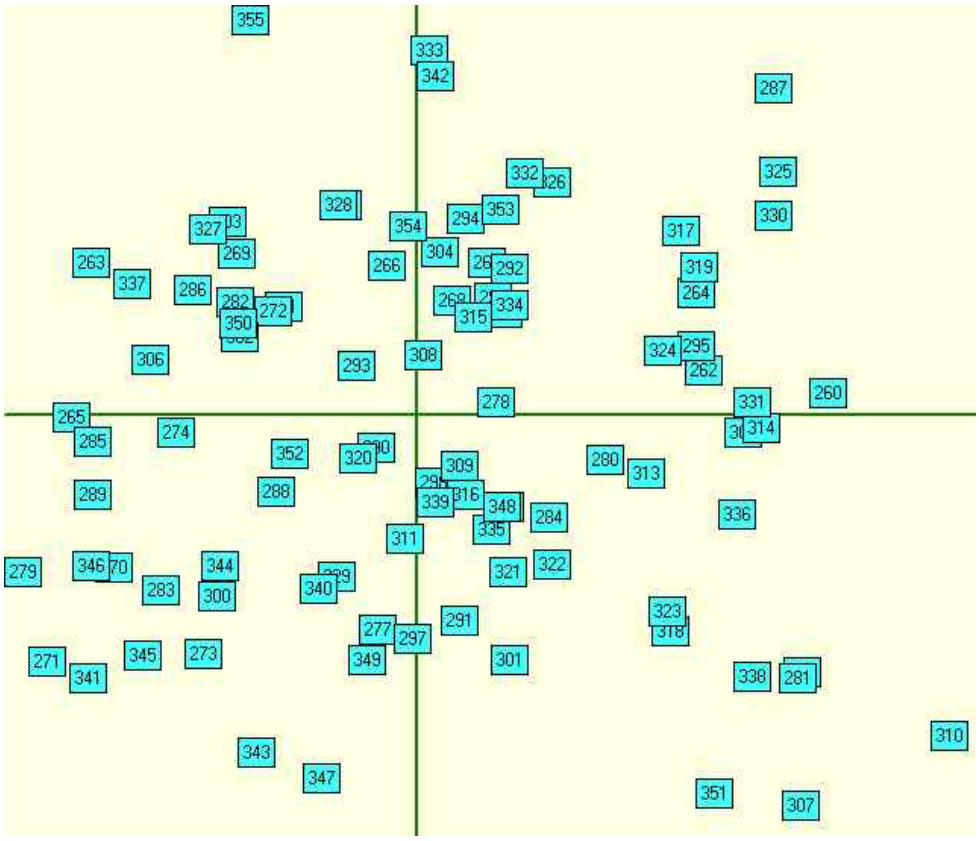


Figure 6
AFC sur le corpus ***Duchn***
96 numéros x 1420 formes de fréquence ≥10

La représentation proposée par l'AFC ne permet pas de repérer une quelconque évolution chronologique des parties. Pour tenter de comprendre les bases de l'opposition qui oppose les différents numéros opposés par le premier axe, nous pouvons consulter les longues listes de

[16] Les pourcentages d'inertie attachés aux deux premiers axes factoriels responsables de la représentation que l'on trouve au tableau 6, sont respectivement égaux à : τ_1=3%, τ_2=2%.

contributions aux facteurs fournis par les programmes d'AFC. Nous allons employer une méthode plus simple pour arriver à un résultat très proche.

==== Lexico3 ==== Analyse Factorielle des Correspondances (AFC)

- ✓ Vérifiez que vous avez opéré au moins une partition du corpus (cf. §xx)
- ✓ Sélectionner l'icône **PCLC** (5ème icône à partir de la gauche)
- ✓ Sélectionner une partition du corpus (ici : numero)
- ✓ Appuyez sur le bouton AFC ((à droite de l'écran)
- ✓ Choisissez un seuil de fréquence minimale (ou acceptez le seuil 10 proposé par défaut)
- ✓ Lancez l'analyse en appuyant sur le bouton **OK**

==== Repères méthodologiques ====

L'analyse factorielle des correspondances (AFC)

L'*analyse factorielle des correspondances* est une méthode statistique qui s'applique aux tableaux de contingence, tels par exemple les tableaux résultant du décompte de différents *types* de vocabulaire (lignes du tableau) dans les différentes *parties* (colonnes du tableau) d'un corpus de textes.

On commence par calculer une distance (dite *distance du chi-deux*) entre chacune des paires de textes qui constituent le corpus.

On décompose ensuite ces distances sur une succession hiérarchisée d'*axes factoriels*. La propriété remarquable de ce système d'axes factoriels est que les représentations limitées aux premiers axes de ce système sont celles qui déforment *le moins possible* les distances calculées entre chaque paire d'éléments. Des *pourcentages d'inertie*, dont la somme vaut 100, calculés pour chaque axe permettent d'apprécier la quantité d'information apportée par chacun des axes dans la décomposition.

Cette méthode d'obtenir des représentations synthétiques portant à la fois sur les distances calculées entre les textes et celles que l'on peut calculer entre les unités textuelles qui les composent. Les typologies obtenues sur chacun des deux ensembles mis en correspondance, sont intiment liées et peuvent être mise en relation grâce à des *représentations simultanées* sur les premiers axes factoriels.

L'intérêt principal de l'AFC réside dans sa capacité à extraire à partir de vastes tableaux de données difficilement appréhendables des structures simples qui rendent compte approximativement des grandes oppositions sous-jacentes dans un corpus de textes.

Pour en savoir plus :

Lebart, L., Salem, A. : Statistiques textuelles, Paris, Dunod, 1994.

5.2 Analyse des spécificités du corpus

L'analyse des spécificités permet de porter un diagnostic exprimé en probabilité sur l'effectif de chacune des cases d'un tableau lexical. [17]

==== Repères méthodologiques ====

La méthode des spécificités

A partir de l'effectif constaté à l'intersection de la ligne *i* et de la colonne *j* (le nombre d'occurrences de la forme *i* dans la partie *j*), étant donnés la fréquence totale de la forme F_i la longueur de la partie t_j et l'effectif total T, la méthode permet de tirer des conclusions sur l'effectif observé. Dans certains cas, la conclusion est que l'effectif observé correspond *à peu près* à ce que le modèle permettait de prévoir. On dira alors que la répartition de la forme est *banale* pour cette partie. Dans d'autres cas, le modèle amènera à conclure que l'effectif observé s'éloigne notablement des prévisions que l'on pouvait faire sous les hypothèses admises par le modèle.

On appelle *spécificités positives* les effectifs qui dépassent largement ce que le modèle laissait prévoir et *spécificités négatives* les effectifs qui se révèlent nettement inférieurs à ce que ce même modèle permettait d'espérer. On attache à ces diagnostic un *indice de spécificité*[18] qui permet de mesure les écarts constatés par rapport à ce que le modèle laissait prévoir. Plus ce diagnostic est élevé plus l'écart est jugé significatif par le modèle.

On peut étendre le calcul décrit ci-dessus pour les unités simples aux segments répétés d'un texte si l'on remarque que les occurrences d'un segment AB (ou A et B sont des formes simples) peuvent être vues comme un sous-ensemble des occurrences de la forme A pour lesquelles B succède immédiatement à A dans le texte. Le calcul simultané des spécificités sur les ensembles de formes et de segments répétés d'un même texte permet souvent de mettre en évidence des associations spécifiques composées de plusieurs formes dont les répartition particulières n'entraînent pas de diagnostic particulier.

Pour en savoir plus :

Pour un exposé et des exemples d'application de l'analyse des spécificités à l'étude des corpus de textes, on consultera par exemple :

Lafon, P. : *Dépouillements et analyses statistique en lexicométrie*, Paris, Klincksieck, 1984

Lebart, L., Salem, A. : *Statistiques textuelles*, Paris, Dunod, 1994.

---

[17] L'analyse des spécificités repose sur l'utilisation du modèle hypergéométrique pour l'analyse des tableaux de nombres à deux dimensions. Pour plus de détails sur le modèle des spécificités et ses applications à l'étude des corpus textuels, on consultera : [Lafon 1984] ou [Lebart et Salem 1994].

[18] Pour une spécificité *positive* et un effectif observé égal a *k*, un indice de probabilité ***x*** signifie que le modèle attache au phénomène constaté : effectif égal ou supérieur à *k*, une probabilité de l'ordre de 10^{-x}. Pour une spécificité *négative* cette probabilité s'attache à un effectif inférieur ou égal à *k*.

Pour comprendre l'opposition constatée sur le premier axe de l'AFC, on a calculé les spécificités, par rapport à l'ensemble du corpus, de deux groupes de numéros opposés par le premier facteur. Chacun des deux groupe est composé des 20 numéros les plus éloignés du centre sur la droite et sur la gauche du graphique. Les spécificités majeures pour chacun de ces groupe ont été rassemblées au tableau 6. L'analyse de ces listes nous fournira une piste pour expliquer la différence qui existe entre les deux groupes de textes.

Tableau 6 :

Formes et segments spécifiques positifs majeurs
pour les numéros opposés par l'AFC sur les 96 numéros

Spécificités positives de la partie gauche				***Spécificités positive de la partie droite***			
Forme	**Frq.Tot.**	**Partie**	**Coeff.**	**Forme**	**Frq. Tot.**	**Partie**	**Coeff.**
nous	1270	449	29	*je*	979	436	***
vous	1097	395	27	*me*	329	184	43
avez	171	94	21	*tu*	296	142	25
fermiers	28	24	13	*ma*	132	81	24
constitution	72	44	13	*m*	206	102	20
accapareurs	80	45	12	*moi*	144	80	20
est vous	24	21	12	*mon*	193	95	18
nos	348	132	12	*j*	281	123	18
vous avez	75	43	12	*ai*	202	91	14
c est vous	24	21	12	*me dit*	29	24	13
vous qui	42	28	10	*que je*	119	58	12
les	4748	1210	10	*dit*	163	72	11
subsistances	47	28	9	*j ai*	123	59	11
la constitution	40	26	9	*que j*	52	30	9
c est vous qui avez	10	10	8	**phélipotin*	13	12	8

Guide de lecture pour le tableau 6

Dans chacun des volets du tableau, on trouve les spécificités relatives à l'un des groupes de textes séparés par l'AFC.

- La première colonne du tableau indique le terme pour lequel le diagnostic de spécificité a été calculé ;
- la seconde *Frq. Tot.* donne la fréquence du terme dans l'ensemble du corpus ;
- la troisième *Partie* la fréquence de ce même terme dans la partie considérée ;
- la troisième *Coefft.* donne lle coefficient de spécificité calculé pour le terme.

Sur la partie droite du tableau 6 on trouve des formes comme *je, tu, me moi, mon* caractéristiques du dialogue, à gauche les contextes des formes comme *vous* renvoient moins au dialogue qu'a des monologues. On note également la présence de nombreux substantifs. Une analyse plus poussée de ces listes accompagnée de retours fréquents au contexte nous amènerons à la conclusion que l'écriture du *Père Duchesne* fait appel à deux types d'écritures distincts dans des proportions qui varient tout au long des huit mois sur lesquels s'étale le corpus et à l'intérieur de chaque numéro. Certains numéros relèvent plus particulièrement

d'un genre que nous appelons "parade"[19], caractérisé par la présence de nombreux effets scéniques empruntés au théâtre de foire, les autres sont de facture rhétorique plus classique. On trouve ci-dessous deux brefs extraits qui illustrent cette opposition :

Tableau 7 :
Deux extraits du corpus ***Duchn*** illustrant la différence
entre les genres *parade* et *classique*

Père Duchesne n°260 (exemple du *g*enre « facture classique »)
§ *marat n'est plus, foutre. peuple, gémis, pleure ton meilleur ami;il meurt martyr de la liberté. c'est le *calvados qui a vomi le monstre sous les coups duquel il vient de périr. une jeune fille, ou plutôt une furie armée par les prêtres, et pénitente, dit on, du cafard *fauchet ,part de *caen pour exécuter cet horrible attentat.

Père Duchesne n°262 *(*exemple du *g*enre « parade »)
§ voilà donc tes projets, infâme coquin; avais je tort, quand je foutais mes fourneaux sens dessus dessous, quand je brisais ma pipe toutes les fois que l'on m'annonçait qu'un noble avait été nommé à quelque place importante.
tu ne savais pas en défilant ton chapelet, archi-traître, que tu parlais au *père *duchesne? à moi mes gens, à moi mes aides de camp /…/

C'est cette alternance dans le style d'écriture qui explique pour l'essentiel l'opposition constatée sur le premier axe de l'AFC. Cette opposition intéressera sans doute à la fois les spécialistes de stylistique et les historiens qui étudient de près la rhétorique du *Père Duchesne*, cependant nos préoccupations plus centrées sur l'évolution du vocabulaire dans cette période nous ont entraînés à nous intéresser à des partitions regroupant plusieurs numéros consécutifs. De tels regroupements permettent de neutraliser les différences stylistiques opposant les livraisons que nous venons d'entrevoir et d'orienter les analyses vers l'observation des changements qui surviennent au cours du temps dans l'utilisation du vocabulaire.

==== ***Lexico3*** ==== ***Liste des spécificités pour une partie (ou un groupe de parties)***

- ✓ Sélectionner l'icône ***PCLC*** (5 ème icône à partir de la gauche)
- ✓ Sélectionner une partie ou un groupe de parties
- ✓ Appuyer sur le bouton ***Spécifs*** (à droite de la fenêtre)
- ✓ Les résultats apparaissent dans une fenêtre sur la gauche
- ✓ On obtient également les segments répétés spécifiques si la liste des segments répétés a été construite avant l'appel des spécificités (cf. §2.2).
- ✓ .On peut également appeler cette fonctionnalité en sélectionnant une ou plusieurs parties sur les plans factoriels produits par l'Afc ou des zones de texte de la carte des sections.

[19] A la suite de J. Guilhaumou [Guilhaumou 19xx].

6 Conclusion

L'exploration du corpus ***Duchn***, à l'aide des méthodes textométriques met en évidence une importante évolution du vocabulaire au cours des huit mois sur lesquels s'étend le corpus.

Les analyses quantitatives sur la partition en 96 livraisons, mettent en évidence des différences stylistiques liées à une alternance de genre entretenue par l'auteur du corpus. De ce fait, elles ne permettent pas d'apprécier l'évolution lexicale du corpus.

Un regroupement des livraisons en périodes de 30 jours consécutifs permet par sa part de cerner l'évolution lexicale de manière nettement plus satisfaisante. Les méthodes quantitatives permettent alors tout à la fois : de mettre en évidence un vocabulaire offensif qui trouvera un emploi particulièrement remarquable dans la période M6. Le retour au contexte permet de préciser ces analyses.

7 Références

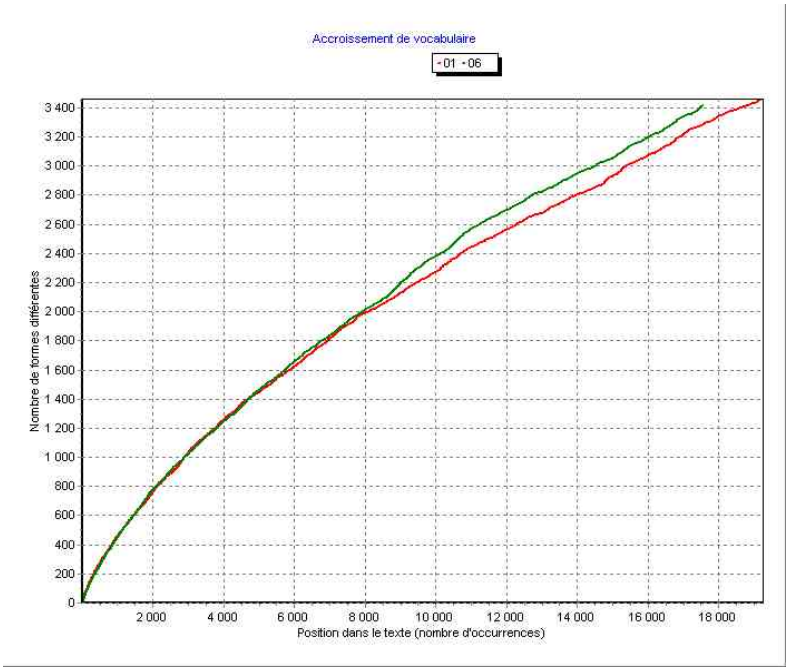
Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002

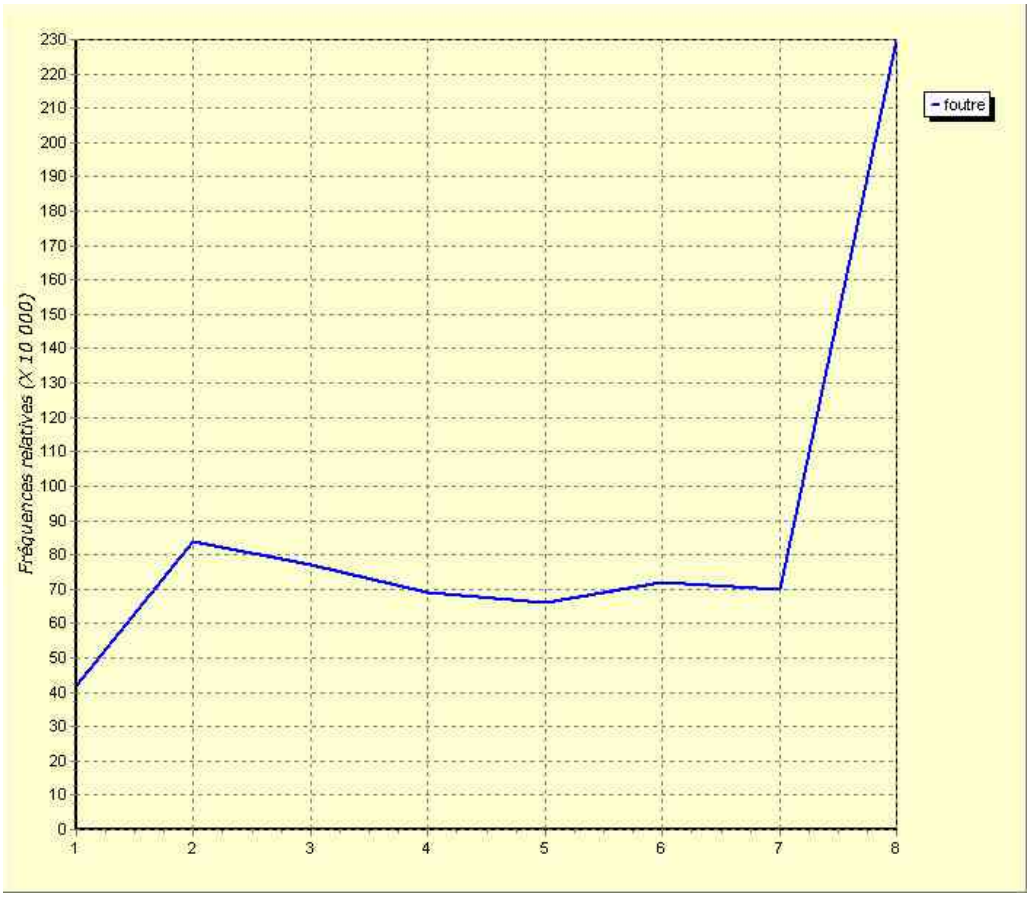
http://www.cavi.univ-paris3.fr/lexicometrica, 1997

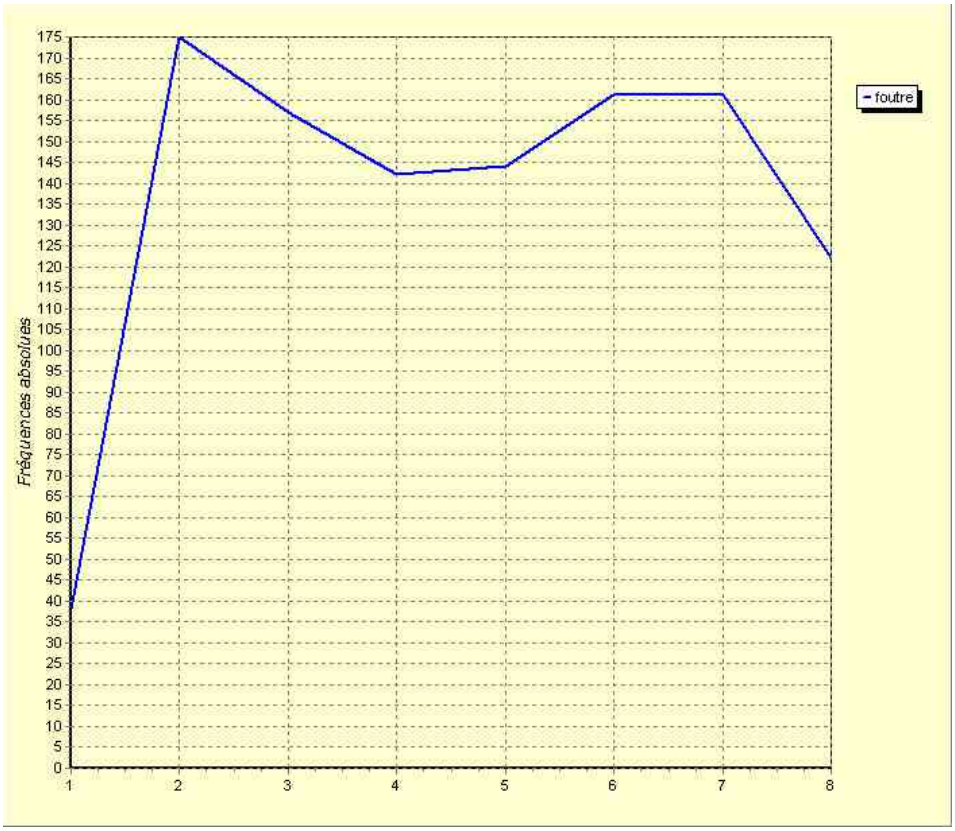
8 Principales fonctionnalités *Lexico3* utilisées

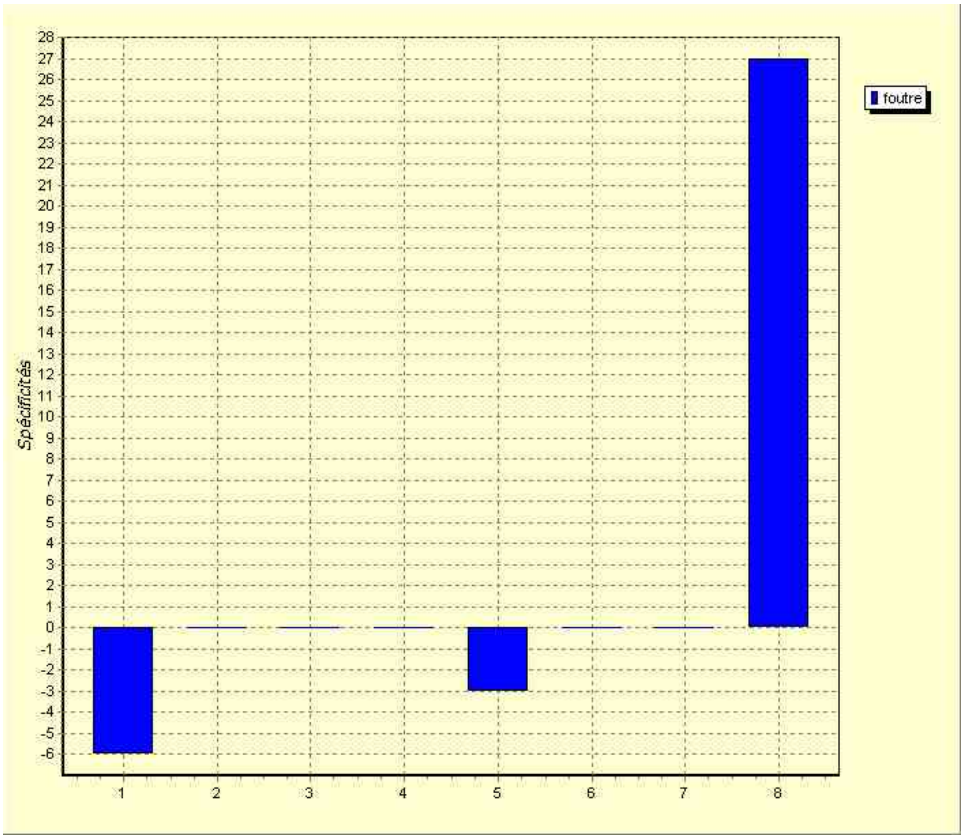
N°	*Fonctionnalité*	Résultat
2	Partition (clé a, pour année)	
5	Principales car lexicom (PCLC)	*Tableau 2*
5.6	Accroissement du vocabulaire (corpus)	*Figure 1*
5.6	Accroissement du vocabulaire (P92, P93)	*Figure 2*
4	Segments Répétés (seuil minimal =2)	
8	Sélection d'un Type (occurrence de SR long>10)	
7	Carte des sections (paragraphes, présence SR de long>10)	*Figure 3*

Annexe


Accroissement de vocabulaire
- 01 - 06
Nombre de formes différentes
3 400
3 200
3 000
2 800
2 600
2 400
2 200
2 000
1 800
1 600
1 400
1 200
1 000
800
600
400
200
0
2 000
4 000
6 000
8 000
10 000
12 000
14 000
16 000
18 000
Position dans le texte (nombre d'occurrences)


Fréquences relatives (X 10 000)
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
foutre





91 ---- ---- ---- ---- ---- les patriotes

```
89 ---- ---- ---- ---- ----  les hommes
88 ---- ---- ---- ---- ----  les plus
86 ---- ---- ---- ---- ----  les traîtres
75 ---- ---- ---- ---- ----  les aristocrates
66 ---- ---- ---- ---- ----  les autres
64 ---- ---- ---- ---- ----  les fripons
63 ---- ---- ---- ---- ----  les brigands
60 ---- ---- ---- ---- ----  les jean
58 ---- ---- ---- ---- ----  les ennemis
54 ---- ---- ---- ---- ----  les départements
46 ---- ---- ---- ---- ----  les bons
42 ---- ---- ---- ---- ----  les accapareurs
40 ---- ---- ---- ---- ----  les scélérats
37 ---- ---- ---- ---- ----  les uns
37 ---- ---- ---- ---- ----  les *français
37 ---- ---- ---- ---- ----  les *brissotins
35 ---- ---- ---- ---- ----  les rois
33 ---- ---- ---- ---- ----  les bougres
32 ---- ---- ---- ---- ----  les muscadins
31 ---- ---- ---- ---- ----  les riches
31 ---- ---- ---- ---- ----  les meilleurs
31 ---- ---- ---- ---- ----  les intrigants
30 ---- ---- ---- ---- ----  les prêtres
29 ---- ---- ---- ---- ----  les royalistes
```

Par page

22 foutre 38 175 157 142 144 161 161 122