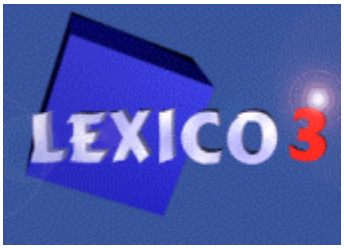
Outils de statistiques textuels



Manuel d'Utilisation

Equipe CLA2T

Cette nouvelle version de Lexico3 a été réalisée par Cédric Lamalle et André Salem

Manuel provisoire, établi par Aude Maisondieu et Andréa Kuncova

ILPGA Université Paris 3, Sorbonne Nouvelle
19 rue des Bernardins, 75005 Paris - France
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
lexico@msh-paris.fr

Table des Matières

I-Installation

1-Avertissement

Cette version de Lexico3 est en test. Nous vous demandons de signaler les éventuelles anomalies à :
Lexico2 / ILPGA
19, rue des Bernardins
75005 Paris
France
Joignez à votre envoi : le corpus de texte sur lequel vous aurez constaté des dysfonctionnements ainsi que le fichier atrace.txt créé au moment de l'exploitation (dans le répertoire où se trouve le corpus analysé).

Configuration requise

Configuration minimale :
Windows 3.1
Processeur 386, 4Mo de mémoire vive
3 Mo libres sur le disque dur

Configuration conseillée

Windows 3.1 ou 3.11 ou Windows 95
486 DX2, 8Mo de mémoire vive
3 Mo libres sur le disque dur
Lexico2 fonctionne sous Windows 95, et Windows NT 3.51 et 4.0.
Nous conseillons vivement de regrouper programme et corpus sur le disque dur.

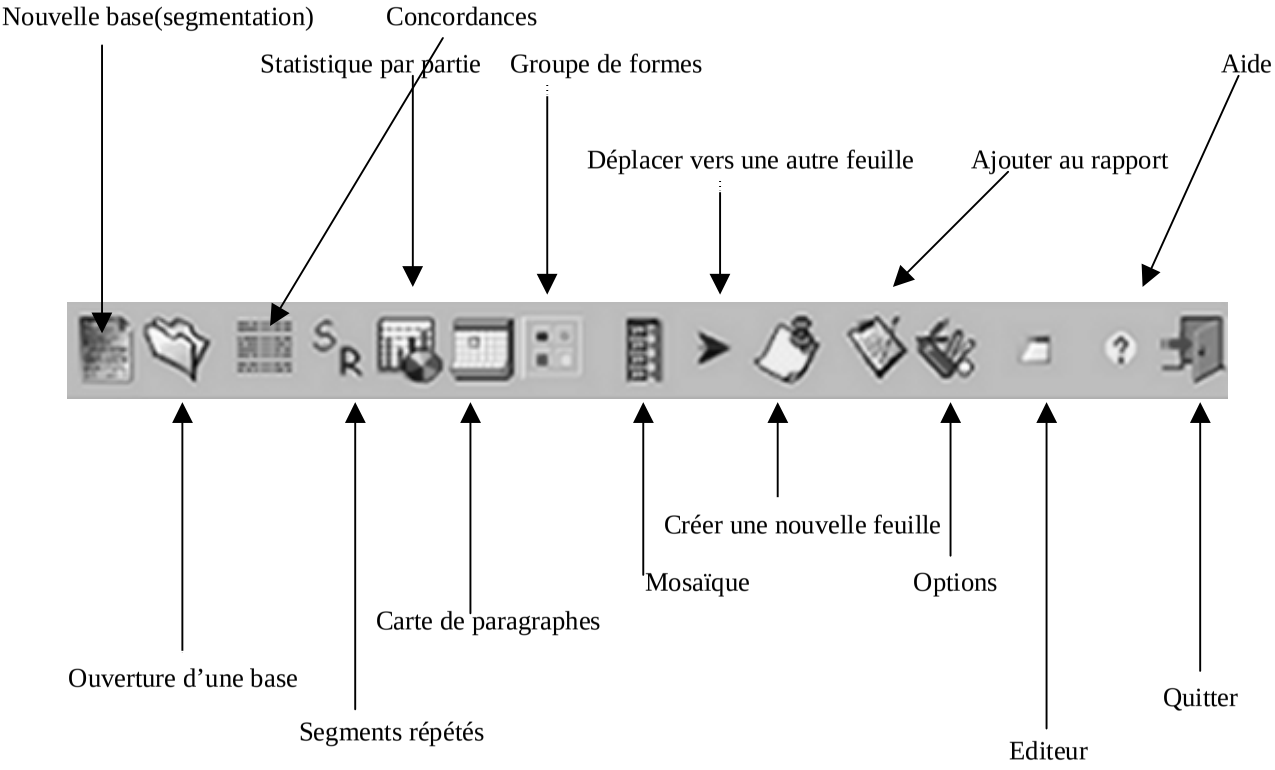
2-Installer le logiciel

Pour installer LEXICO3

a. Insérer le CD ROM dans votre lecteur.

b. Exécuter le fichier SETUP.EXE qui se trouve sur ce CD-ROM.
c. Laissez vous guider par le programme d'installation
d. Le message : ***Lexico3 a été installé*** signale la fin de l'installation.

II-Menu principal



III-Préparation du texte

1 Normes d'enregistrement

La norme basse

L'analyse statistique d'un texte se base sur l'unité textuelle qu'est la forme pour estimer ses occurrences, délimiter les séquences de mots où elle apparaît et comparer ses fréquences dans une partie donnée du corpus par rapport à une ou plusieurs autres. Pour rendre possible et fiable cette segmentation il est nécessaire de définir des normes de saisie qui assurent la cohérence du texte et de son codage.
Le texte doit être enregistré sous la forme d'un fichier de texte seul (*.txt). On écarte les fichiers de type document (*.doc) et autres formats créés par traitement de texte car ceux-ci intégrent au début de l'enregistrement un en-tête renfermant diverses informations sur la mise en forme notamment. Dans une étude lexicométrique qui s'intéresse principalement aux formes, aux phrases et à leur ventilation, ces données sur la mise en page sont superficielles et, qui plus est, elles peuvent être une source d'erreurs lors de la segmentation.

Problèmes de codage

Traitement des majuscules, apostrophes

On choisit de saisir le texte en minuscules pour permettre un classement plus juste dans les listes paradigmatiques. En effet, si l'on maintient les majuscules le module de segmentation distinguera la forme simple il et la forme Il en début de phrase, ce qui fausserait les fréquences d'apparition.
Toutefois, ces conversions ne sont pas sans risque puisqu'elles peuvent introduire des ambigüités dans les listes qui associeront alors certains noms propres et noms communs homographes (par exemple : une barre et Raymond Barre). On peut désambigüiser ces formes en accollant au nom propre un signe de distinction (par

exemple : ~barre). Toutefois, ce signe ne doit alors pas figurer dans la liste des délimiteurs.

Ponctuations, délimiteurs

Afin que le programme d'analyse LEXICO3 puisse reconnaître les formes du corpus que l'on segmente, on affecte un statut à chaque signe présent dans le texte.
On distinguera d'une part les délimiteurs :
-—_:;/.,?¿!¡*$"+=(){} ainsi que l'espace. (Cette liste est donnée à titre indicative et peut être modifiée).
D'autre part, sont appelés non-délimiteurs tous les caractères restants contenus dans la police : a, b, c,...
Toute suite de caractères non-délimiteurs bornée à ses deux extrémités par des délimiteurs est considérée comme une occurrence, une forme à relever et à classer.

Clés

On rencontre dans un texte un certain nombre de délimitations inhérentes telles que des délimitations chronologiques, des délimitations marquant un changement d'auteur ou encore indiquant la séquence des parties (les chapitres d'un livre).
Au cours d'une étude lexicométrique, on cherchera souvent à comparer les fréquences du vocabulaire dans une partie par rapport à une autre en modifiant les découpages du corpus. Pour que ces comparaisons s'effectuent efficacement, le texte doit comporter des balises informatives indiquant ces délimitations logiques sous une forme codée compréhensible par la machine.
Cette méta-information doit figurer dans le fichier sans cependant risquer d'influer sur les comptages statistiques. On introduit donc ces diverses marques sous formes de clés isolées du corpus par les chevrons < et >.
Ces clés peuvent être alpha-numériques. En voici un exemple :
<auteur = césar> <partie = 2>

Paragraphes

Le découpage du corpus peut aussi se faire selon l'ordre des paragraphes. Dans ce cas, on codera chaque paragraphe en ajoutant

devant celui-ci un caractère réservé à cet effet et déclaré dans la liste des délimiteurs, par exemple : §.

2 Tutoriel : Text 1

Dans cet extrait du corpus Text1.txt, plusieurs types de codage, mis en évidence pour notre exemple par une fonte plus grande :
- la clef Epg distingue 3 parties qui rendent compte de la pagination de l'édition originale du Père Duchesne
- le caractère paragraphe § distingue 4 paragraphes
- le caractère * permet d'identifier des lettres (à l'origine) en majuscules

Tableau 2.1 : Exemple de codage de corpus

<Sda=1793> <S01=220> <S03=0> **<Epg=1>** <Sat=0>
§ la grande colère du *****père *****duchesne , de voir que les mouchards de *la-*fayette et tous les fripons soudoyés par la liste civile, veulent rétablir les compagnies de grenadiers et de chasseurs, pour égorger les *sans-culottes et les chasser des assemblées de *section .ses bons avis aux *lurons des *faubourgs pour qu' ils arrachent les moustaches postiches à ces grenadiers de la vierge *marie , qui veulent rétablir la royauté.
<S03=1>
§ millions de tonnerre, nous ne mettrons donc jamais les fripons à la raison?ils **<Epg=2>**ont laissé tomber leurs masques et nous les voyons à nu. serons nous encore dupes des fripons? quand je voulais faire la conduite de *grenoble à tous les talons rouges quand je disais, du soir au matin, que tous les ci-devant ne cesseraient de nous trahir, n' avais je pas raison, foutre?
§ je me suis toujours plus défié des nobles convertis que des émigrés.c' est pour nous frapper de plus près que ces gredins sont restés au milieu de nous.ils ont fait les chiens couchants pour mieux nous tromper.jamais, foutre, ils n' ont cessé de s' entendre avec les ennemis du dehors. ce sont eux qui nous ont mis à chien et à chat, qui ont brouillé les cartes dans les trois assemblées nationales, et corrompu les représentants du peuple.si nous avions eu assez d' estoc pour les envoyer tous à *coblentz au commencement de la

révolution, nous n' aurions pas acheté notre liberté par des flots de sang;nous aurions depuis longtemps une constitution; la paix et le bonheur régneraient dans notre république.
§ dans le fond de mon coeur j' ai toujours détesté *philippe d' *orléans ; je le regardais comme un hypocrite qui tôt ou tard nous **<Epg=3>** tournerait casaque; mais comme cet infâme scélérat servait notre cause en prodigant son or pour donner des croc en jambes à *louis le traître, je pensais, comme tous les patriotes, qu' il fallait s' en servir comme d' une chemise que l' on quitte quand elle est sale. je ne le redoutais pas, car il est trop vil et trop méprisable pour croire que jamais les *sans-culottes pourraient se donner un pareil roi. je me doutais bien, foutre, qu' au premier faux pas que le *capon ferait, il se casserait le col.un viédase qui s' était caché au fond de cale, au combat d' *ouessant, ne pouvait jamais devenir un chef de parti.

IV- Outils de l'utilisateur

Segmentation

1-La segmentation automatique

Le module segmentation crée une base de données textuelles à partir d'un corpus fourni par l'utilisateur sous forme de fichier texte.

2-Mise en oeuvre

Depuis la console, cliquez sur l'icône du module : **Nouvelle base (*Segmentation*).**
Le programme vous propose de choisir un fichier texte dans un répertoire selon les procédures habituelles de Windows.

Figure 1: Dialogue de sélection de fichier texte

Sélectionnez le corpus à segmenter qui est obligatoirement un fichier texte (de type .txt).

Une boîte de dialogue apparaît alors qui vous permet de régler les paramètres de la segmentation.

Figure 2: Dialogue de sélection des paramètres de segmentation

Vous pouvez éventuellement modifier la liste de délimiteurs de forme ce qui aura une incidence sur le découpage des formes graphiques.
Lancez ensuite la segmentation en cliquant sur le bouton OK.

Vérification des clés

étape au cours de laquelle le programme vérifie la conformité du corpus d'entrée aux normes décrites plus haut. Ce module signale notamment les clés erronées :

clé non fermée	<S01=chirac
espace dans le type ou le contenu de clé	<S 01= chi rac>
balise de fermeture isolée	La vie est > belle.
absence de signe =	<S01balladur>
absence de contenu de clé	<S01=>
absence de type de clé	<=jospin>

Figure 3: Dialogue d'avertissement d'erreur de codage de clé

Des informations plus détaillées des erreurs sont fournies par le fichier de compte-rendu, **atrace.txt** qui indique de plus, le numéro de la ligne incriminée.

Tableau 3.4 : Compte-rendu de la segmentation

*****COMPTE-RENDU DE LA SEGMENTATION*****
Fichier -- C:\LEXICO3T\TEXTES\DUCH.TXT -- ouvert pour vérification
L 2 Clé incorrecte (espace dans contenu) : <Sda=17 93>
L 94 Clé incorrecte (pas de contenu) : <Epg=>
L 5709 Clé incorrecte : Mauvais emplacement de balise de fermeture
L 5845 Clé incorrecte (espace dans type) : <Ep g=3>
L13277 Clé incorrecte : <S02=330 <
L13496 Clé incorrecte (pas d'=) : <Epg8>

L…. indique la ligne fautive

Segmentation du texte

Lorsque les lignes fautives ont été corrigées, on relance le programme comme indiqué plus haut. S'il n'y a plus d'erreurs , une jauge permet de suivre la progression de la segmentation du texte.

Création, tri et enregistrement des dictionnaires

Ces étapes permettent de constituer la liste des formes présentes dans le corpus, et pour chacune de ces formes de calculer le nombre de ses occurrences ainsi que son rang alphabétique.

A l'issue de la segmentation, une boîte de dialogue informe de la fin normale de l'opération.

3-Fichiers de sortie

Plusieurs fichiers sont créés et enregistrés sur disque dur dans le même répertoire que le texte-source :
Le fichier *corpus.par* : contient les principaux décomptes effectués par le programme (formes, occurrences, etc...) ainsi que le rappel des caractères délimiteurs choisis lors de la segmentation.

Tableau 3.6 : Exemple fichier paramètres

Lexico2.1 PC DUCH
nbetiq=0
196125 196125 11023 142185 10859 6130 4953 5000000 14 8 143 0 0
*** Résultat de la segmentation du fichier: DUCH.TXT ***
Délimiteurs #-—:;/\\.,?¿!¡*$\"' +=(){}[]§
nombre des occurrences : 142185
nombre des formes : 10859
frequence maximale : 6130
nombre des hapax : 4953
nombre des clés(type) : 8
nombre des clés(ctnu) : 143
*** Fin de la segmentation du fichier: DUCH.TXT ***

Le fichier ***corpus.alp*** : contient un index des formes graphiques classés par ordre alphabétique. Il n' est créé que si vous avez coché la case *Index Alphabétique* dans la boîte de dialogue du module Segmentation.

Tableau 3.7 : Extrait de l'index alphabétique

frq forme
27 834 a

28 2632 à
29 1 abandonnant
30 4 abandonne
31 10 abandonné
32 1 abandonnées
33 3 abandonnent
(…) (…) (…)

Le fichier *corpus.dic* : contient le dictionnaire des formes classés par fréquence (un enregistrement pour chaque forme).

Tableau 3.8 : Extrait du dictionnaire numérisé

frq rang lex. forme
6130 2703 de
4749 6033 les
4298 5909 la
3773 4216 et
(…) (…) (…)
1 10809 voudrai
1 10817 vouer
259 10859 !
198 10860 "
49 10861 $
(…) (…) (…)
766 10873 Epg
96 10874 S01
(…) (…) (…)
97 10882 01
1 10883 02

Le fichier ***corpus.num*** : contient le texte numérisé, c'est à dire sous une forme codée où articles, occurrences, ponctuations, et autres éléments du corpus sont répertoriés de façon compacte. Ce fichier à usage strictement interne ne peut être consulté avec un éditeur de texte.

Le fichier ***atrace.txt*** : contient un rapport détaillé des opérations effectuées par le programme (mémoire allouée, paramètres pris en compte, fichiers lus et écrits, ...). En cas d'échec du traitement, ce fichier peut fournir des indications permettant de situer la cause du problème.

Tableau 3.9 : Extrait du fichier de trace (*atrace.txt*)

*****COMPTE-RENDU DE LA SEGMENTATION*****
Fichier -- C:\LEXICO2\TRAVAIL\DUCH.TXT -- ouvert pour vérification
MAXART (Maximum d' articles traités) = 50000
Allocation de 4687K
Fichier d' initialisation = C:\LEXICO2T\TEXTES\LEXICO2.INI
Fichier -- C:\LEXICO2\TRAVAIL\DUCH.TXT -- ouvert pour segmentation
Fichier xxxx.txn ouvert
Délimiteurs : .,:;!?/_-\"' ()[]{}
== nbf= 11016, nhap= 5079, nbcle= 8, nbctnu= 143
-- Tri lexicom. de 11016 formes
-- Fin Travail Dictionnaire
Fichier Dict : DUCH.dic :
-- Fin EcriDicNum 11178 articles
Fichier Texte : DUCH.num : 192083 items.
Fichier Param DUCH.par :
*****FIN DE LA SEGMENTATION*****

Ouverture d'une base

En cliquant sur cette icône vous avez la possibilité d'ouvrir des textes déjà segmentés, il s'agit des fichiers avec l'extension «.par».

Note:

Vous pouvez également ouvrir un texte déjà segmenté en le glissant sur l'icône de Lexico3.

Statistique par partie

Pour comparer les variations dans l'usage du vocabulaire entre les différentes parties du corpus, on utilise le module "Statistique par parties" qui opère une série de calculs statistiques. Le corpus aura été préalablement traité par le module "Segmentation".

1-Mise en œuvre du programme

Cliquer sur l'icône "Statistique par partie", une boîte de dialogue apparaît . Elle vous permet de régler les paramètres de la partition.

Figure 4: Choix de la partition

Sélectionner la forme (dictionnaire) que vous désirez faire apparaître dans le graphe, puis glisser-la sur le graphe.

Figure 5: Graphique

Principales caractéristiques de la partition

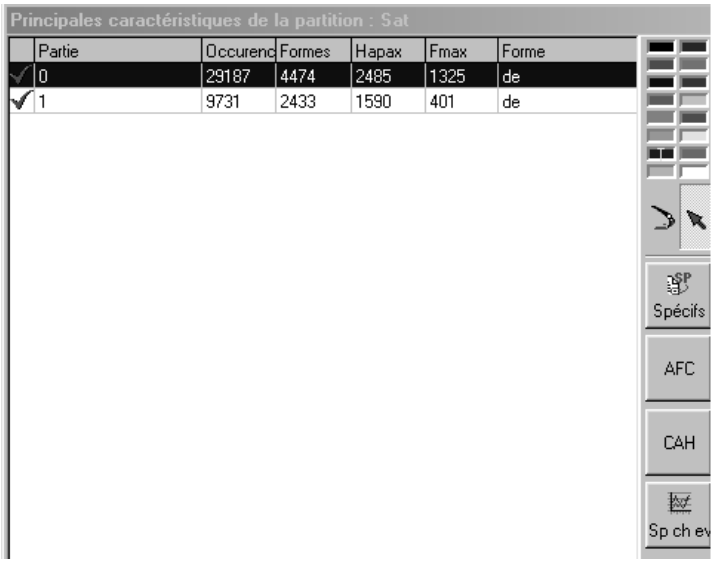
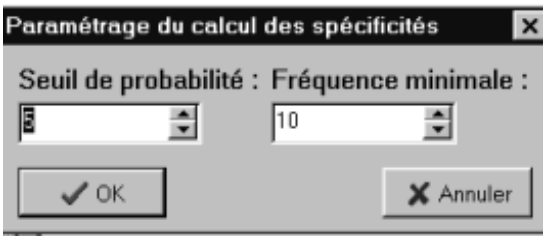
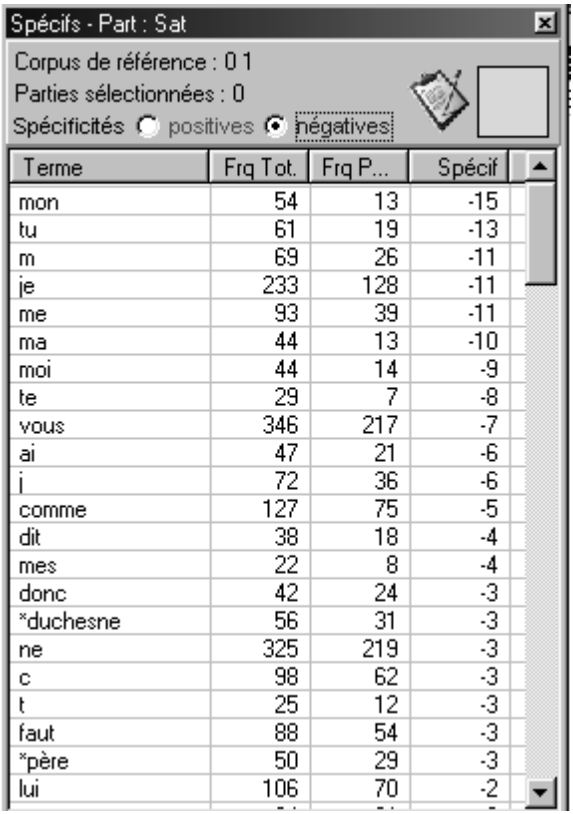


Figure 6: Caractéristiques de la partition

Spécificité

L'analyse des spécificités : Ce fichier (*.spf) indique la ventilation des formes dans les parties et adjoint aux fréquences remarquables un indice de spécificité. Ce dernier s'interprète comme suit : d'abord le signe + ou – qui indique un sur-emploi ou un sous-emploi.




Spécifs - Part : Sat

Corpus de référence : 0 1
Parties sélectionnées : 0
Spécificités ○ positives ◉ négatives

Terme	Frq Tot.	Frq P...	Spécif
mon	54	13	-15
tu	61	19	-13
m	69	26	-11
je	233	128	-11
me	93	39	-11
ma	44	13	-10
moi	44	14	-9
te	29	7	-8
vous	346	217	-7
ai	47	21	-6
j	72	36	-6
comme	127	75	-5
dit	38	18	-4
mes	22	8	-4
donc	42	24	-3
*duchesne	56	31	-3
ne	325	219	-3
c	98	62	-3
t	25	12	-3
faut	88	54	-3
*père	50	29	-3
lui	106	70	-2

Figure 7: Spécificités

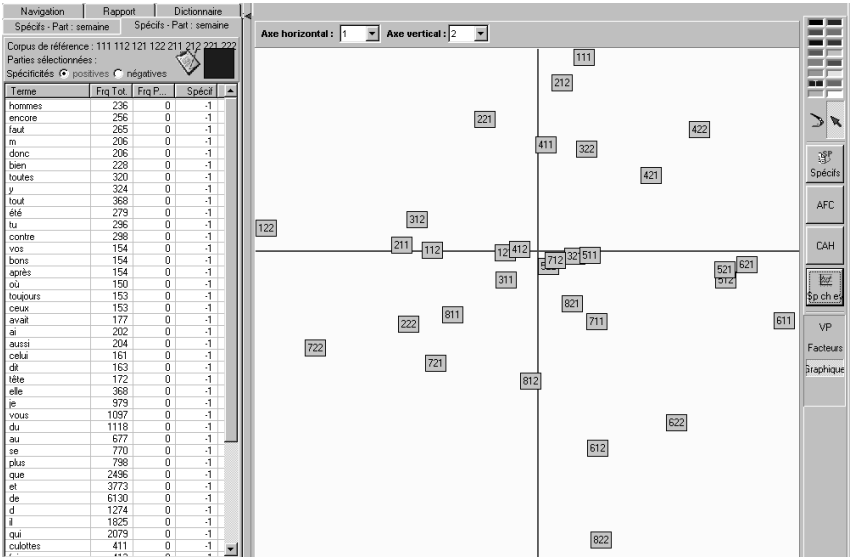
A la suite de cet index par formes vient un index des spécificités, duch.spf, par partie où l'on retrouve classées les spécificités positives et négatives. L'indice d'originalité * (astérisque) indique que la forme n'est présente que dans la partie courante.

Spécificités sur une partie donnée (par rapport à l'ensemble)

On sélectionne une (ou plusieurs) partie(s) (clic gauche de la souris, avec majuscule ou contrôle activé suivant le nombre de parties à sélectionné (comportement windows habituel)), et on lance les

spécificités. On peut ensuite réitérer les opérations de "glissement de mots" sur la carte des sections via la souris.

AFC



CAH

Sp ch ev

2-Le fichier de sortie

Le fichier est un fichier **.par** où sont enregistrés successivement toutes les requêtes effectuées lors de la dernière session d'utilisation.

Segments répétés

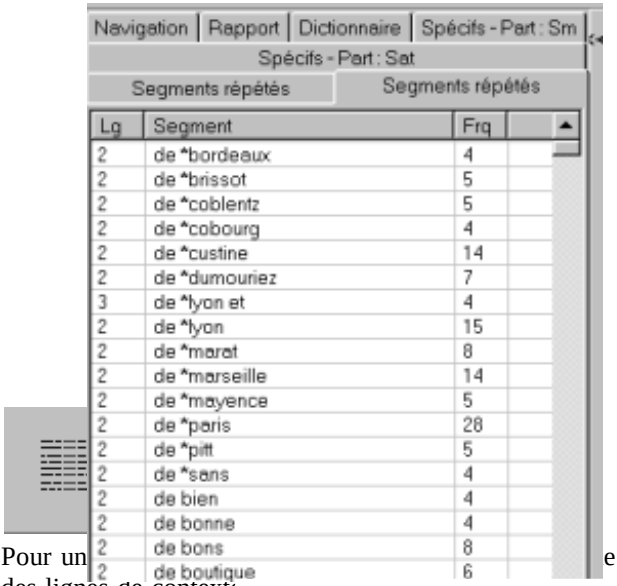
Les segments répétés sont des suites de formes dont la fréquence est supérieure à 2 dans le corpus.

1-Mise en œuvre du programme

Pour créer la liste des segments répétés cliquez sur l'icône, une boîte de dialogue apparaît, vous pouvez paramétrer le calcul des segments répétés (cf figure 5):
Vous avez la possibilité de choisir les délimiteurs de séquence et la partition.
Vous pouvez fixer une fréquence minimum en-dessous duquel les formes et les segments ne seront pas retenue. Ce seuil est par défaut fixé à 10.
Ensuite cliquez sur OK.

Figure 8: Paramétrage du calcul de segments répétés

La liste des segments répétés apparaît dans la partie gauche de la fenêtre, vous pouvez la consulter en cliquant sur l'onglet « Segments répétés ». (cf figure 6)



Pour un e l'ensemble des lignes de contexte permet un retour au te forme, qui permet de fréquemment autour du pole.

Figure 9: Liste des segments répétés

' occasion d ' être utile à un **homme** aussi intéressant . quel est votre
nsi donc , au lieu d ' être un **homme** sanguinaire , quand je faisais
ant de fois accusé d ' être un **homme** féroce , était le meilleur des
maladie de l ' adoration . un **homme** , tel qu ' il soit , n ' est plus
il soit , n ' est plus qu ' un **homme** à nos yeux ; tant qu ' il va droit
s coups ; reçois nos regrets , **homme** pur et courageux ; longtemps
nous

Tableau : Extrait d'une concordance autour de la forme-pôle *homme* dans le corpus *Duchesne*.

Les tris

Les différents contextes relatifs à une même forme peuvent être ordonnés de trois manières différentes. Le tri de ces contextes peut être effectué en fonction de :
- l'occurrence qui précède la forme-pôle (tri avant)
- l'occurrence qui suit la forme-pôle (tri après)
- l'ordre dans lequel les occurrences de la forme-pôle apparaissent dans le texte.

Le module Contextes permet de lancer plusieurs requêtes documentaires consécutives à partir d'une base de données textuelles créée par le module Segmentation.

1-Mise en oeuvre

Cliquer sur l'icône Concordance, une boîte de dialogue apparaît :

 Soit entrez la forme dont vous souhaitez étudier les contextes dans la zone d'édition 'forme pivot' (ex : homme).
Soit sélectionner la forme (dictionnaire) ou le segment répété(segments répétés) dont vous souhaitez étudier les contextes, puis glisser le dans la fenêtre de droite.
 Précisez l'ordre de tri des contextes grâce au menu déroulant « tri »(avant, après, aucun).
 Regroupement : vous permet d'effectuer des concordances par partition.

 Largeur : Vous pouvez choisir le nombre de caractères (espaces inclus) qui doivent apparaître avant et après chaque pôle. Pour le modifier après une première recherche, changer la largeur et cliquer sur rafraîchir.(Figure5)
 Lancez l'exécution de la requête en appuyant sur entrée.

Il est possible d'effectuer des requêtes sur plusieurs formes à la fois. Pour visualiser les différentes formes il suffit de cliquer sur les bouton "Précédent" et "Suivant".

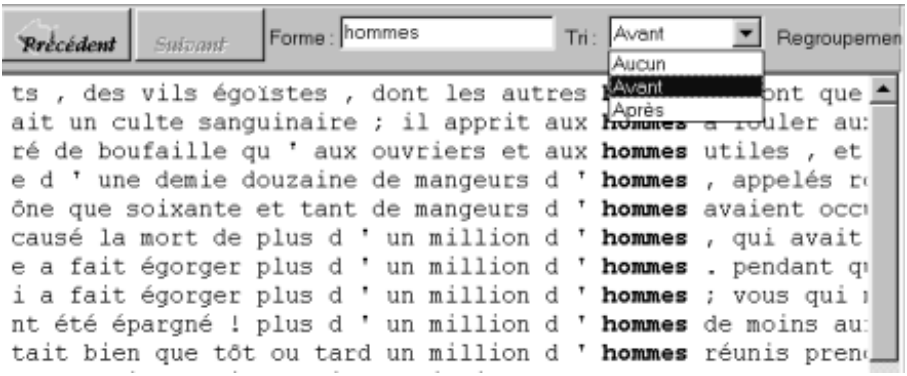


Figure 10: Boîte de dialogue "Concordances"

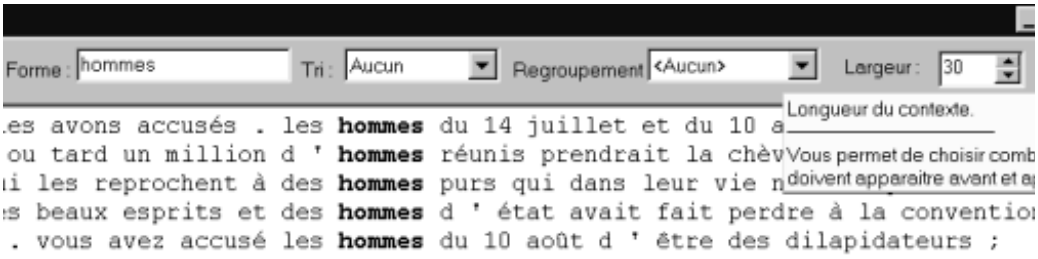


Figure 11: Largeur du contexte

2-Le fichier de sortie

Le fichier **conc1.html** est un fichier html ou sont enregistrés successivement toutes les requêtes effectuées lors de la dernière session d'utilisation de Concordances.

Carte des paragraphes

On peut créer une carte des sections, la segmentation se fait à partir des délimiteurs sélectionnés: paragraphes, point...

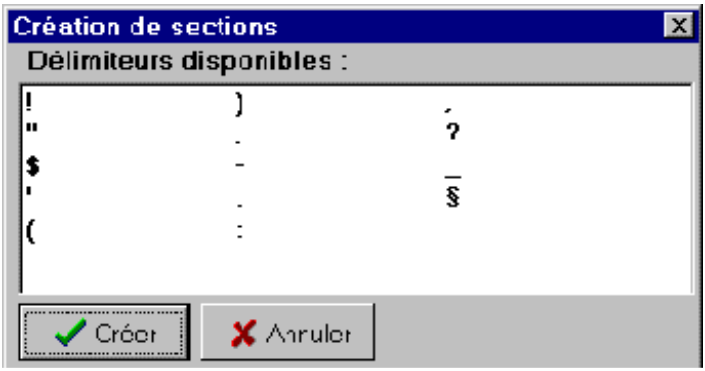
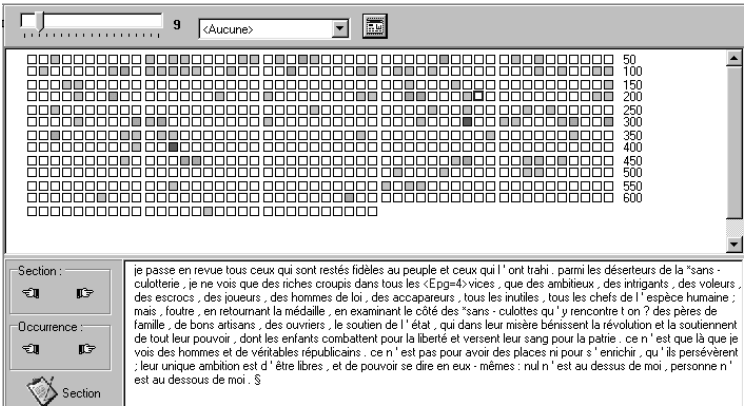


Figure 12: Délimiteurs disponibles

Carte des paragraphes des mots du dictionnaire

Sélectionner la forme (dictionnaire) que vous désirez faire apparaître sur la carte des sections, puis glisser-la sur la carte (clic gauche maintenu du segment vers le graphique).

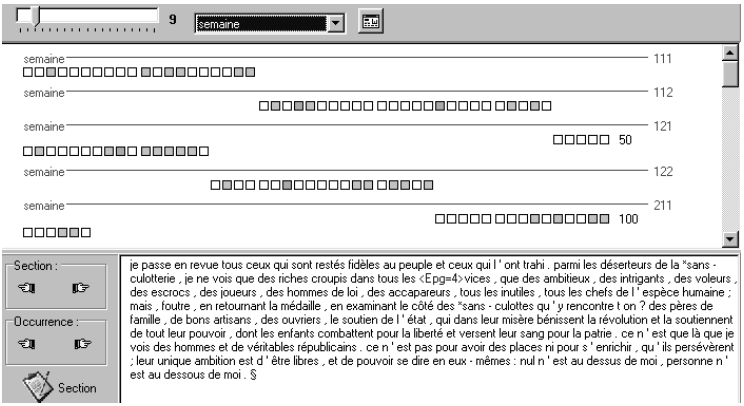


Carte de groupe de mots (segments répétés)

Après avoir activé le bouton "Segments répétés", Lexico produit un "dictionnaire" des segments dans la partie gauche. On peut manipuler les formes contenues dans le dictionnaire et par exemple les faire glisser sur la carte (clic gauche maintenu du segment vers le graphique).

Partitions, sections et retour au texte : le texte à géométrie variable

On peut visualiser le texte via la carte des sections (section sélectionnée au préalable), via le découpage des partitions initialement intégré dans le codage du corpus, et le texte lui même (fenêtre du bas)



Note

Vous pouvez vous déplacer à l'intérieur de texte soit par section (les carrés non-colorés) soit par occurrence (les carrés colorés).

Pour ajouter les cartes au rapport, cliquer sur "Section".

Groupe de formes

Il est possible d'effectuer des requêtes sur plusieurs formes à la fois, en basant les requêtes sur des préfixes, des suffixes, des expressions régulières (type egrep/grep), ou des suites des caractères graphiques.

1-Mise en œuvre

Entrez le nom du groupe de formes.
Entrez la forme que vous désirez rechercher.
Cliquez sur rechercher.
L"objet" résultant peut ensuite être manipulé comme une forme "normale", en cliquant sur la flèche rouge du groupe (clic gauche maintenu), on "glisse" le groupe sur la carte de la partition. **cf image**
Si vous effectuez une nouvelle recherche, vos résultats se concatènent aux précédents.

Mosaïque

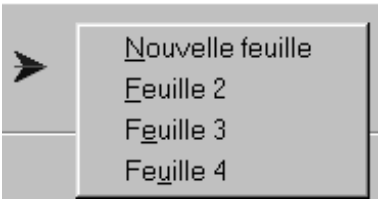
En cliquant sur cette icône vous réorganisez plusieurs applications (fenêtres) sur la même feuille.

Créer une nouvelle feuille

Pour éviter la superposition des différentes applications en cours, vous pouvez créer de nouvelles feuilles en cliquant sur l'icône.
Les feuilles de travail s'empilent sur la droite de la fenêtre principale.
Les onglets "Feuille n°i" permettent de passer de l'une à l'autre.

Déplacer vers une autre feuille

Pour déplacer une application vers une nouvelle feuille, sélectionnez la, cliquez sur l'icône et sélectionnez la feuille désirée.



Ajouter au rapport

Toutes les fenêtres produites au cours d'une exploration textuelle peuvent être intégrées dans un rapport au format HTML.
Pour sauvegarder l'application en cours cliquez sur "ajouter au rapport" et dans l'onglet rapport cliquez sur "enregistrer".
Le répertoire "Rapport" se crée automatiquement, il contient le fichier "Rapport.html" où vous trouverez, sous forme de liens hypertexte, toutes les données sauvegardées.

Figure 15: Rapport.html

Options

Ce bouton permet de fixer des seuils lors du traitement

de gros corpus, il permet aussi d'indiquer si le corpus traité a été préalablement étiqueté.

Aide

L'aide en ligne

Aide générale

Le fichier d'aide de Lexico3 peut être consulté à tout moment à partir de la console en cliquant sur l'icône *Aide*.

Aide contextuelle

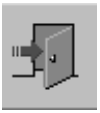
Lors de l'exécution des modules, l'utilisateur peut faire apparaître une aide contextuelle en rapport avec le traitement en cours en cliquant sur le bouton *Aide* dans la boîte de dialogue active.

Glossaire

La définition de quelques notions de base en statistique textuelle est reprise dans l'aide en ligne.

Editeur

Pour visualiser un texte ou vos résultats à partir de Lexico 3, cliquez sur l'icône "Editeur" et à partir de l'icône "Ouvrir" sélectionnez votre document.



Quitter

Avant de quitter Lexico3, vérifiez que vous avez sauvegardé vos données dans le rapport.
Pour quitter Lexico3 cliquez sur l'icône.

V-Astuces

Navigation

Sélectionner/Glisser

On sélectionne une (ou plusieurs) forme(s) (clic gauche de la souris, avec majuscule ou contrôle activé suivant le nombre de parties à sélectionné (comportement windows habituel)), et on glisse les mots. On peut ensuite réitérer les opérations de "glissement de formes" dans la fenêtre droite via la souris...

Plein écran

Pour visualise la fenêtre de droit en plein écran, cliquez sur la flèche rouge située entre les fenêtres gauche et droite.

Dictionnaire

Pour la lecture du dictionnaire vous pouvez choisir entre l’ordre lexicométrique ou l’ordre lexicographique.

Glossaire pour la statistique textuelle

NB : Les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :
ac Analyse factorielle des correspondances
acm Analyse des correspondances multiples
cla Classification
sp Méthode des Spécificités
sr Analyse des segments répétés
ling Linguistique
stat Statistique
sa Segmentation automatique

accroissement spécifique - (sp) spécificité* calculée pour une partie d'un corpus par rapport à une partie antérieure
analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.
analyse des correspondances (stat)- méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou c2).
caractère (sa) - signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.
caractères délimiteurs / non-délimiteurs (sa) - distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).
On distingue parmi les caractères délimiteurs:
- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de

ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.
- les caractères **délimiteurs de séquence** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.
- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.
classification (stat) - technique statistique permettant de regrouper des individus ou observations entre lesquels a été définie une distance.
classification hiérarchique (cla) - technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.
concordance (sa) - l'ensemble de lignes de contexte se rapportant à une même forme-pôle.
contribution absolue (ou contribution) - (ac) contribution apportée par un élément au facteur . Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.
contribution relative (ou cosinus carré) - (ac) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.
cooccurrence (sa) - (une c.) - présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.
corpus (ling) - ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.
(lexicométrie) ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.
délimiteurs de séquence - (sa) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général - le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

dendrogramme - (cla) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.
discours/langue - La langue est un ensemble virtuel qui ne peut être appréhendé que dans son actualisation orale ou écrite; "discours" est un terme commode qui recouvre les deux domaines de cette actualisation.
distance du chi-2 - distance entre profils* de fréquence utilisée en analyse des correspondances* et dans certains algorithmes* de classification*.
éditions de contextes (sa) - éditions de type concordanciel dans lesquelles les occurrences d'une forme sont accompagnées d'un fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce contexte est définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle.
éléments d'un segment (sr) - chacune des formes correspondant aux occurrences qui entrent dans sa composition. ex : A, B, C sont respectivement les premier, deuxième et troisième éléments du segment ABC.
éléments actifs- (ac ou acm) ensemble des éléments servant de base au calcul des axes factoriels, des valeurs propres relatives à ces axes et des coordonnées factorielles.
éléments supplémentaires (ou illustratifs)- (ac ou acm) ensemble des éléments ne participant pas aux calculs des axes factoriels, pour lesquels on calcule des coordonnées factorielles qui auraient été affectées à une forme ayant la même répartition dans le corpus mais participant à l'analyse avec un poids négligeable.
énoncé/énonciation - (ling) à l'intérieur du texte un ensemble de traces qui manifestent l'acte par lequel un auteur a produit ce texte.
facteur- (ac ou acm) variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.
forme- (sa) ou "**forme graphique**" archétype correspondant aux occurrences* identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale - (sp) pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie .
forme caractéristique - (d'une partie) synonyme de spécificité positive*.
forme commune - forme attestée dans chacune des parties du corpus.
forme originale- (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.
fréquence (sa) - (d'une unité textuelle) le nombre de ses occurrences dans le corpus.
fréquence d'un segment (sr) - (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.
fréquence maximale (sa) - fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de").
fréquence relative (sa) - la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).
gamme des fréquences (sa) **-** suite notée Vk, des effectifs correspondant aux formes de fréquence k, lorsque k varie de 1 à la fréquence maximale.
hapax - gr. hapax (legomenon), "chose dite une seule fois".
(sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).
identification - (stat, ling, sa) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.
index - (sa) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regouper les références* relatives à l'ensemble des occurrences d'une même forme.
index alphabétique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).
index hiérarchique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.
index par parties - ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.
lemmatisation - regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En

français, ce regroupement se pratique en général de la manière suivante :
_ les formes verbales à l'infinitif,
_ les substantifs au singulier,
_ les adjectifs au masculin singulier,
_ les formes élidées à la forme sans élision.
lexical - (ling) qui concerne le lexique* ou le vocabulaire*.
lexicométrie ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire* d'un corpus de textes.
lexique - (ling) ensemble virtuel des mots d'une langue.
longueur (sa) - (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. : partie, fragment, etc.). Synonyme de taille.
On note: T la longueur du corpus; t j celle de la partie (ou tranche) numéro j du corpus.
longueur d'un segment (sr) - le nombre des occurrences entrant dans la composition de ce segment.
occurrence (sa) - suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.
ordre lexicographique -
_ pour les formes graphiques :
l'ordre selon lequel les formes sont classées dans un dictionnaire.
NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple, rangées dans cet ordre, les formes : *mais, maïs, maison, maître* .
_ pour les polyformes:
ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante, les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.
ordre lexicométrique (sa) -
_ pour les formes graphiques :

ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes; les formes de même fréquence sont classées par ordre lexicographique.
_ pour les polyformes:
ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.
paradigme- (ling) ensemble des termes qui peuvent figurer en un point de la chaîne parlée.
paradigmatique- (sa) qui concerne le regroupement en série des unités textuelles, indépendamment de leur ordre de succession dans la chaîne écrite.
partie - (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.
partition - (d'un corpus de textes) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.
(d'un ensemble, d'un échantillon) division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.
partition longitudinale - (sa) partition d'un corpus en fonction d'une variable qui définit un ordre sur l'ensemble des parties
périodisation (sa) - regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.
phrase - (sa) fragment de texte compris entre deux séparateurs* de phrase.
polyforme (sr) - archétype des occurrences d'un segment; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.
ponctuation - Système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.
(*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.
pourcentages d'inertie - (ac ou acm) quantités proportionnelles aux valeurs propres* dont la somme est égale à 100. Notées ta.

profil - (stat et ac) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).
répartition (sa) - (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.
segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.
segment répété (sr) - (ou polyforme répétée) suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.
segmentaire - (sr) ensemble des termes* attestés dans le corpus.
segmentation - opération qui consiste à délimiter des unités minimales* dans un texte.
segmentation automatique - ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.
séparateurs de phrases - (sa) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).
séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur* de séquence.
seuil - (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).
sous-fréquence (sa) - (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.
sous-segments (sr) - pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.
spécificité chronologique - (sp) spécificité* portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale*.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.
spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.
stock distributionnel du vocabulaire - (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.
syntagmatique- (sa) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.
syntagme- (ling) groupe de mots en séquence formant une unité à l'intérieur de la phrase.
tableau de contingence (stat) - synonyme de tableau de fréquences ou de tableau croisé: tableau dont les lignes et les colonnes représentent respectivement les modalités de deux questions (ou deux variables nominales) , et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.
tableau lexical entier (TLE) - tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.
tableau des segments répétés (TSR) - tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).
tableau lexical- tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui

correspondent à des formes dont la fréquence est inférieure à un seuil donné).

taille- (sa) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme - (sr) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2,3, etc.

termes contraints / termes libres - Un terme S1 est contraint dans un autre terme S2 de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment S2. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

unités minimales (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent).

valeur modale - (stat) valeur pour laquelle une distribution atteint son maximum.

valeurs propres - (ac ou acm) quantités permettant de juger de l'importance des facteurs successifs de la décomposition factorielle. La valeur propre notée la. mesure la dispersion des éléments sur l' axe.a.

valeurs-tests - (ac ou acm) quantités permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur test dépasse 2 en valeur absolue, il y a 95 chances sur 100 que la position de l'élément correspondant ne puisse être due au hasard.

variables actives - variables utilisées pour dresser une typologie, soit par analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids des variables actives, qui doivent de ce fait constituer un ensemble homogène.

variables supplémentaires (ou illustratives) - variables utilisées a posteriori pour illustrer des plans factoriels ou des classes. Une variable supplémentaire peut-être considérée comme une variable active munie d'un poids nul.

variables de type T - variable dont la fréquence est à peu près proportionnelle à l'allongement du texte. (ex : la fréquence maximale)
variables de type V- variable dont l'accroissement a tendance à diminuer avec l'allongement du texte (ex : le nombre des formes, le nombre des hapax).
ventilation (sa) - (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences* de cette unité dans chacune des parties, prises dans l'ordre des parties.
vocabulaire (sa) - ensemble des formes* attestées dans un corpus de textes.
vocabulaire commun - (sa) l'ensemble des formes attestées dans chacune des parties du corpus.
vocabulaire de base - (sp) ensemble des formes du corpus ne présentant, pour un seuil fixé, aucune spécificité (négative ou positive) dans aucune des parties , (i.e. l'ensemble des formes qui sont "banales" pour chacune des parties du corpus).
vocabulaire original- (sa) (pour une partie du corpus) l'ensemble des formes* originales* pour cette partie.

voisinage d'une occurrence - (sa) pour une occurrence donnée du texte, tout segment (suite d'occurrences consécutives, non séparées par un délimiteur de séquence) contenant cette occurrence.

Références bibliographiques

Bécue M. (1988) - Characteristic repeated segments and chains in textual data analysis, COMPSTAT, 8th Symposium on Computational Statistics, Physica Verlag, Vienna.

Becue M., Peiro R. (1993) - Les quasi-segments pour une classification automatique des réponses ouvertes, in Actes des 2ndes Journées Internationales d'analyse des données textuelles, (Montpellier), ENST, Paris, p 310-325.

Benzecri J.-P.(1977) - Analyse discriminante et analyse factorielle, Les Cahiers de l'Analyse des Données, II, n °4, p 369-406.

Benzécri J.-P. & coll. (1973) - La taxinomie, Vol. I ; L'analyse des correspondances, Vol. II, Dunod, Paris.

Benzécri J.-P. (1982) - Histoire et préhistoire de l'analyse des données, Dunod, Paris.

Benzécri J.-P.& coll. (1981a) - Pratique de l'analyse des données, tome 3, Linguistique & Lexicologie, Dunod , Paris.

Benzécri J.-P. (1991a) - Typologies de textes grecs d'après les occurrences des formes des mots-outil, Les Cahiers de l'Analyse des Données, XVI, n°1, p 61-86.

Benzécri J.-P. (1992) - Correspondence Analysis Handbook, (Transl : T.K. Gopalan) Marcel Dekker, New York.

Bernet C. (1983) - Le vocabulaire des tragédies de Jean Racine, Analyse statistique, Slatkine-Champion, Genève 1983.

Bolasco S. (1992) - Sur différentes stratégie dans une analyse des formes textuelles : Une expérimentation à partir de données d'enquête, Jornades Internacionals d'Analisi de Dades Textuals, UPC, Barcelona, p 69-88.

Bonnafous S. (1991) - L'immigration prise aux mots. Les immigrés dans la presse au tournant des années quatre-vingt, Kimé, Paris.

Brunet E. (1981) - Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française, Slatkine-Champion, Genève-Paris.

Demonet M., Geffroy A., Gouaze J., Lafon P., Mouillaud M., Tournier M. (1975) - Des tracts en Mai 68. Mesures de

vocabulaire et de contenu, Armand Colin et Presses de la Fondation Nat. des Sc. Pol., Paris.
Dendien J. (1986) - La Base de données de l'Institut National de la Langue Française, Actes du colloque international CNRS, Nice, juin 1985, 2 vol., Slatkine-Champion Genève, Paris.
Geffroy A., Lafon P., Tournier M. (1974) - L'indexation minimale, Plaidoyer pour une non-lemmatisation, Colloque sur l'analyse des corpus linguistiques : "Problèmes et méthodes de l'indexation minimale", Strasbourg 21-23 mai 1973.
Gobin C., Deroubaix J. C. (1987) - Du progrès, de la réforme de l'Etat, de l'austérité. Déclarations gouvernementales en Belgique, Mots, n°15, p 137-170.
Guilbaud G.-Th. (1980) - Zipf et les fréquences, Mots N° 1, p 97-126.
Guilhaumou J. (1986) - L'historien du discours et la lexicométrie. Etude d'une série chronologique : Le père Duchesne de Hébert, juillet 1793- mars 1794, Histoire & Mesure , Vol. I, n° 3-4.
Guiraud P. (1954) - Les caractères statistiques du vocabulaire, P.U.F., Paris.
Guiraud P. (1960) - Problèmes et méthodes de la statistique linguistique, P.U.F., Paris.
Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction, in The prediction of personal adjustment (P. Horst, ed.), SSCR New York, p 251 - 264.
Habert B., Tournier M. (1987) - La tradition chrétienne du syndicalisme français aux prises avec le temps. Evolution comparée des résolutions confédérales (1945 - 1985), Mots, n°14.
Labbé D. (1990) - Le vocabulaire de François Mitterrand, Presses de la Fond. Nat. des Sciences Politiques, Paris.
Labbé D. (1983) - François Mitterrand - Essai sur le discours, La pensée sauvage, Grenoble.
Labbé D. (1990) - Normes de dépouillement et procédures d'analyse des textes politiques, CERAT, Grenoble.
Labbé D., Thoiron P., Serant D. (Ed.) (1988) - Etudes sur la richesse et la structure lexicales, Slatkine-Champion, Paris-Genève.

Lafon P. (1980) - Sur la variabilité de la fréquence des formes dans un corpus, Mots N°1 , p 127-165.
Lafon P. (1981) - Analyse lexicométrique et recherche des cooccurrences, Mots N°3 , p 95-148.
Lafon P. (1981) - Dépouillements et statistiques en lexicométrie, Slatkine-Champion, 1984, Paris.
Lafon P., Salem A. (1983) - L'Inventaire des segments répétés d'un texte, Mots N°6, p 161-177.
Lafon P., Salem A., Tournier M. (1985) - Lexicométrie et associations syntagmatiques (Analyse des segments répétés et des cooccurrences appliquée à un corpus de textes syndicaux). Colloque de l'ALLC, Metz -1983, Slatkine-Champion, Genève, Paris, p 59-72.
Lebart L. (1969) - L'Analyse statistique de la contiguïté, Publications de l'ISUP, XVIII- p 81 - 112.
Lebart L. (1982b) - L'Analyse statistique des réponses libres dans les enquêtes socio-économiques, Consommation, n°1, Dunod, p 39-62.
Lebart L., Salem A. (1988) - Analyse statistique des données textuelles, Dunod, Paris.
Lebart L., Salem A., Berry E. (1991) - Recent development in the statistical processing of textual data, Applied Stoch. Model and Data Analysis, 7, p 47-62.
Menard N. (1983) - Mesure de la richesse lexicale, théorie et vérifications expérimentales, Slatkine-Champion, Paris.
Muller C. (1964) - Essai de statistique lexicale : L'illusion comique de P. Corneille, Klincksieck, Paris.
Muller C. (1968) - Initiation à la statistique linguistique, Larousse, Paris.
Muller C. (1977) - Principes et méthodes de statistique lexicale, Hachette, Paris.
Muller C.(1967) - Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille, Paris, Larousse.
Pêcheux M. (1969) - Analyse automatique du discours, Dunod, Paris.
Peschanski D. (1988) - Et pourtant, ils tournent. Vocabulaire et stratégie du PCF (1934 - 1936), Klincksieck, Paris.
Petruszewycz M. (1973) - L'histoire de la loi d'Estoup-Zipf, Math. Sciences Hum., n°44.

Reinert M. (1990) - Alceste, Une méthodologie d'analyse des données textuelles et une Application : Aurélia de Gérard de Nerval, Bull. de Méthod. Sociol. n°26, p 24-54.
Romeu L. (1992) - Approche du discours éditorial de Ya et Arriba (1939 - 1945), Thèse Paris 3.
Salem A. (1984) - La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, Les Cahiers de l'Analyse des Données, Vol IX, n° 4, p 489-500.
Salem A. (1986) - Segments répétés et analyse statistique des données textuelles, Etude quantitative à propos du père Duchesne de Hébert, Histoire & Mesure, Vol. I- n° 2, Paris, Ed. du CNRS.
Salem A. (1987) - Pratique des segments répétés, Essai de satistique textuelle, Klincksieck, Paris.
Salem A. (1993) - Méthodes de la statistique textuelle, Thèse d'Etat, Université Sorbonne Nouvelle (Paris 3).
Sekhraoui M. (1981) - La saisie des textes et le traitement des mots : Problèmes posés, essai de solution, Mémoire, Ecole des hautes études en sciences sociales, Paris.
Tournier M. (1985a) - Sur quoi pouvons-nous compter ? Hommage à Hélène Nais, Verbum.
Tournier M. (1985b) - Texte propagandiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation, Mots N°11, p 155-187.
Tournier M. (1980) - D'ou viennent les fréquences de vocabulaire?, Mots N°1, p 189-212.
Van Rijckevorsel J. (1987) - The application of fuzzy coding and horseshoes in multiple correspondances analysis, DSWO Press, Leyde.
Warnesson I., Parisot P., Bedecarrax C., Huot C. (1993) - Traitements linguistiques et analyse des données pour une exploitation systématique des banques de données, Revue Française de bibliométrie, i 21.
Weil G.E., Salem A., Serfaty M. (1976) - Le livre d'Isaïe et l'analyse critique des sources textuelles, Revue (R.E.L.O) LASLA , N°2 , Liège.
Yule G.U. (1944) - The Statistical Study of Literary Vocabulary, Cambridge University Press, Reprinted in 1968 by Archon Books, Hamden, Connecticut.

Zipf G. K. (1935) - The Psychobiology of Language, an Introduction to Dynamic Philology, Boston, Houghton-Mifflin.