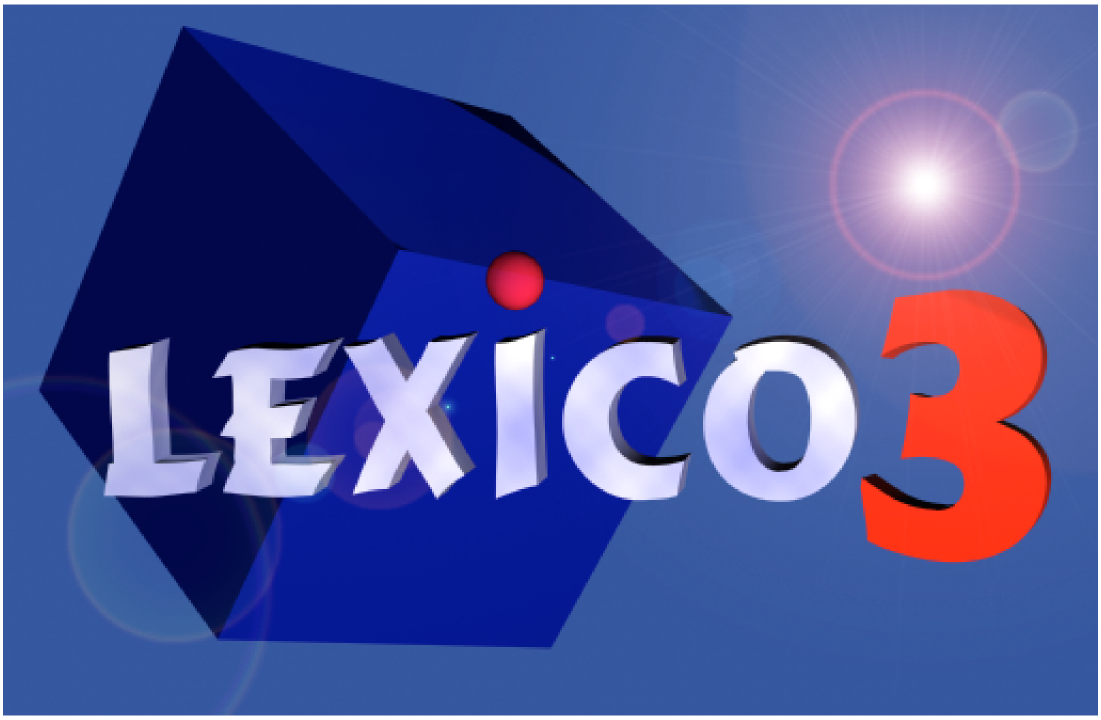
SYLED - CLA2T

Université de la Sorbonne nouvelle - Paris 3



Version 3.41 février 03

Textometric toolbox

Cédric Lamalle

William Martinez

Serge Fleury

André Salem

User's manual

Béatrice Fracchiolla

Andrea Kuncova

Bettina Lande

Aude Maisondieu

Maria Poirot Zimina

18/05/04

SYLED - CLA2T

Université de la Sorbonne nouvelle - Paris 3

Summary

Foreword **5**
Main improvements 5
Object-oriented version *5*
Establishing form groups *5*
Localization of lexicometric particularities *5*
To find out more 5
Upcoming developments 6
Installation **7**
0.1 Warning 7
Minimum hardware requirements *7*
0.2 Installing the software 7
1 Text corpora **8**
Quick tips 8
Introductory corpus authors.txt *8*
Your own test corpus *9*
1.2 Storage norms 9
Delimiting / non-delimiting characters *9*
Lower and upper case letters, apostrophes *10*
Sections of text *10*
Keys/Tags *11*
1.3 Choosing textual units 11
1.4 Example: the *Duchesne* corpus 12
2 Tools for textual exploration **14**
2.1 Segmenting a corpus 14
Operational set-up 14
Checking the keys 15
Segmentation of the text 16
Output files 16
2.2 Opening an existing database 19
2.3 Concordances 19
Select a form (or a type) 19
drag/drop 20
Display of the concordance 20
Sorting 20
2.4 Adding the results to the report 21
The report 21
Add to the report 21
2.5 Search for repeated segments 21
2.6 Form groups 23
Set-up 24
Regular expressions 24
2.6 *Word-store* 26
3 Tools for statistical analysis **27**
3.1 Partitioning 27
Distribution of a form (or Tgen) 27
Statistics by text part (PCLC) 28

3.2 Characteristic elements 29

Results of computation of characteristic elements 31

3.3 Chronological characteristic elements 31

Characteristic increments 31

3.4 Correspondence analysis (CA) 32

4 Tools for lexicometric browsing 34

4.1 Map of sections 34

Mapping of the sections for a Tgen 34

Statistical tools of the map of sections 34

Browsing with the help of the map of sections 35

4.2 Towards better use of the windows 36

Create a worksheet 36

Move to another worksheet 36

Mosaic 36

4.3 The report 36

4.4. Options – Help - Tips 37

Options 37

Browsing tab 38

Full screen 38

Help 38

Exit 38

5 Glossary of terms used in textual statistics 39

Bibliography 45

Web sites 48

Foreword

Lexico3 is the 2001 edition of the Lexico software, first published in 1990. Functions present from the first version (segmentation, concordances, measurements and counts based on graphical forms, computation of characteristic elements and correspondence analyses of forms and repeated segments) were maintained and for the most part significantly improved.

The Lexico series is unique in that it allows the user to maintain control over the entire lexicometric process, from initial segmentation to the publication of final results. The units that are then counted automatically originate entirely from the list of delimiters provided by the user, with no need for outside dictionary resources.

Beyond identification of graphical forms, the software allows for the study of the distribution of more complex units composed of form sequences: *repeated segments, pairs of forms in relation of co-occurrence*, etc. which are generally less ambiguous in terms of content than the graphical forms that make them up.

Main improvements

Object-oriented version

The main improvement found in this version concerns object-oriented program architecture. The different interactive modules are now able to exchange more complex data items (forms, repeated segments and co-occurrences _ upcoming).

Thus, it is now possible to send to the ***concordance*** module, or to any of the other modules, units established in the module of ***repeated segments***, lists of forms and segments established in the ***characteristic elements*** modules, etc. Hence, veritable lexicometric browsing becomes possible.

Establishing form groups

The study of most abrupt changes that occur in the distribution of a graphical form in different parts of a text corpus inevitably raises questions as to the identification of other related graphical units (different manifestations of the same lemma, forms related at the semantic level). New tools (based on regular expressions look-up facilities) have been included to simplify the search for such form groups.

Localization of lexicometric particularities

This new version allows for more precision in the characterization of different parts of a corpus according to the forms they contain in abundance by isolating sections of the text in which this sort of distribution is particularly evident. Mapping of these sections onto diagrams that represent the text allow the creation of a veritable textual topography.

To find out more

Concerning modifications, corrections, updates, the main source of information is the ***Lexico3*** website of the SYLED-CLA2T team at the Sorbonne-nouvelle University – Paris 3.

The website has previous versions of ***Lexico*** (***Lexico1***-MacIntosh, ***Lexico2*** PC) as well as various documents that can be downloaded, including this manual.

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

A general bibliography can be found in the appendix. References to the book
Lebart Ludovic, Salem André, ***Statistique textuelle***, Dunod, Paris 1994,
are noted (L&S, p. xxx).

Upcoming developments

Certain procedures currently used in lexicometric research could not be included in the present version. This is the case, for example, for Hierarchical Cluster Analysis (HCA) as well as for certain methods allowing the identification of networks of co-occurrences in a text. These procedures will be available in the next version of Lexico.

Installation

0.1 Warning

It is possible, in spite of all the care taken in the preparation of this version, that some errors remain. We ask you to point out any faults by writing us at the following address:

Lexico3 / ILPGA : 19, rue des Bernardins 75005 Paris, France

Please, include the text corpus where the problem was identified as well as the file *atrace.txt* automatically created in the directory where the corpus was located during the exploration. This file contains indispensable information for debugging.

Minimum hardware requirements

Windows 95
486 MHz processor, 4Mo RAM
3 Mo free on the hard disk
Lexico3 works under Windows 95 and later versions, and under Windows NT 3.51 and 4.0.
We heartily advise grouping program and corpus in a common directory on the hard disk

0.2 Installing the software

To install ***Lexico3***
Insert the CD-ROM
Double click on the file icon SETUP.EXE found on the CD-ROM
Follow the installation procedure
The message ***Lexico3 a été installé*** (***Lexico3*** has been installed) indicates that the installation is complete.

1 Text corpora

Lexicometric analysis compares counts resulting from the identification of occurrences of lexical units (forms, segments, generalised types, etc.) in the different parts of a text corpus.

This introduction presents some elementary examples (section 1.1), offering a rapid overview of the software. Problems involving automatic segmentation are presented in section 1.2. Section 1.3 treats the case of a real size corpus.

Quick tips

The following two sections are addressed to users who wish to rapidly go over the principal software functions.

Introductory corpus authors.txt

Using the introductory file *authors.txt* on the CD, we carry out a partition into three parts after which comparisons are made among the "texts" assembled in this corpus.

Tagging a corpus: ***the file*** **authors.txt**

```
<Author=Shakespeare>
From forth the fatal loins of these two foes
A pair of star-cross'd lovers take their life;
Whole misadventured piteous overthrows
Do with their death bury their parents' strife.
<Author=Blake>
O ROSE, thou art sick!
The invisible worm,
That flies in the night,
In the howling storm,
Has found out thy bed
<Author=Wilde>
The sea is flecked with bars of gray,
The dull dead wind is out of tune,
And like a withered leaf the moon
Is blown across the stormy bay.
```

The ***Author*** key allows for the division of the corpus into three parts, which will then be compared.

Proceed as follows:

- Run *Lexico3* by clicking on the icon of the software
- Select the file you wish to open in the **File** menu (in this case, ***authors.txt***)
- Accept the segmentation parameters (defined further on) by clicking on the **OK** button

Lexico3 then offers on the left side of the screen a list of forms identified in the corpus with their respective frequencies. You can now perform any of a series of lexicometric operations described further on in the manual using the buttons that call up the different software modules (cf. sections 2-4).

Your own test corpus

As in the previous example, insert several *tags* to delimit different parts of the corpus (for example: <part=1>, <part=2>, etc.).

Save your document in the directory ***Lexico3*** created during the installation of the software: use your own word processing software (Word, etc.) and choose the option ***text only*** (item ***Save as***...on the ***File*** menu).

Your test corpus is ready for analysis by ***Lexico3***. To start out, the simplest is to accept the *default segmentation* parameters proposed by the software (delimiting characters etc.)

1.2 Storage norms

New standards (XML, HTML etc.) are gradually being established for computerized storage of text corpora. However, corpora collected for lexicometric analysis are still made up of documents from different sources, often stored in different formats. To avoid variations among texts caused by different storage norms, it is useful to subject the texts to some minimal normalisation. Different software packages (including *MKCorpus*[1], offered on this CD-ROM), perform some of the necessary homogenisation work.

Lexicometric analysis studies the distribution of complex units within a text (*lemmas*, *repeated segments*, *co-occurrences*, *generalised types*). Nevertheless, segmentation into graphical forms is a prerequisite for carrying out a wide range of studies, allowing one to

- Obtain an initial estimate of the principal lexicometric characteristics of the corpus (number of occurrences, forms, hapax, maximum frequency);
- Create initial typologies on parts of the corpus;
- Identify errors that remain after first corrections.

To perform segmentation into graphical forms, norms need to be set. These norms are particularly simple in ***Lexico3***.

The text has to be saved as a file *text only* (*.txt)[2].

Delimiting / non-delimiting characters

In a corpus submitted to lexicometric analysis, a graphical form is a series of non-delimiting characters bounded by two delimiting characters. This means that the graphical forms, whose occurrences we will be counting, are entirely defined by the list of delimiting characters chosen by the user.
Identification occurs when the chains found between two delimiters are identical. If the text is not properly prepared, *Hen* will not be identical to *hen* and *openhearted* will be different from *open-hearted.*

[1] ***MKCorpus*** was developed by S. Fleury (Paris3-Ilpga-Syled).

[2] Word Document (*.doc) and other word processing formats are removed since they contain a header with information on formatting.

The technical part of automatic segmentation is considerably simplified by accepting a fairly straightforward principle stated below:

sign = status

This means that at the beginning of the procedure, each typographical sign can be assigned its status (delimiting or non-delimiting character).

Sometimes, these principles run into conflict with usual typographical norms. For example, the apostrophe in the proper name *O'Neil* should be considered a non-delimiting character but its status is different in the sequence *I'm*. (The same is true for points occurring within abbreviations: U.N.E.S.C.O., I.B.M., etc. and periods at the end of sentences).

Lexico3 provides a list of delimiting characters by default that can be modified by the user: `-—_:;/.,?!*$″+=(){}.` The space (blank) is added automatically to this list. Once the list of delimiting characters is established, the other characters: `a, b, c,...` become *non-delimiting characters*.

Any series of non-delimiting characters whose boundaries at both ends are delimiting characters is considered an occurrence. A form is then identified as a type corresponding to identical occurrences in a corpus of texts.

Lower and upper case letters, apostrophes

For special purposes, the user can combine the norms used in preparation of the text and the segmentation options to affect the form types produced by the segmentation procedure. For example, during preparation of the text, all the upper case letters can be replaced systematically by an asterisk followed by the same letter in lower case (ex. Me becomes *me). A segmentation containing the character * among the delimiting characters will not distinguish between the occurrences of the sequences *Me* and *me*; a segmentation which does not include the asterisk in the list of delimiting characters will produce separate counts for the two sequences.

Sections of text

Besides logical partitions, the text also contains marks for breathing (sentences, paragraphs, etc.). ***Lexico3*** offers the possibility of promoting one or several delimiting characters to the rank of *section delimiters*. Such pre-coding allows for the study of the distribution of occurrences of a lexicometric unit within the sections thus defined.

N.B.: The systematic insertion of section delimiters can be performed using the function *Replace* present in a word processing software.[3]

[3] The carriage return special characters will be replaced systematically by the following sequence: carriage return+blank+character §.

Keys/Tags

In lexicometric study, frequencies of forms in different sections of the corpus are compared. In order to make comparison possible, the text must include tags that indicate the logical structure delimiters of the corpus.

The sections defined by the user can be organised chronologically, as in the example from *Père Duchesne*, (cf. section 1.2, “Quick tips”), as well as thematically.

Codifying a key
A key (ex:<Author= Smith> is made up of 5 elements:

1	<	opening angle bracket
2	**Author**	**type** of the key
3	=	the “equal” sign
4	**Smith**	the **content** of the key
5	>	closing angle bracket

For example: <Year=1998>, <Author=Jean_de_la_Fontaine>

The insertion of keys is an important stage in the preparation of the text. The selected keys will allow the user to carry out comparisons of codified textual groupings (speakers, categories of speakers, authors, documents, etc.).

1.3 Choosing textual units

To proceed with statistical analyses of texts thus stored, it is necessary to define a norm, whose purpose is to isolate the various units within the chain of text upon which counts are carried out. How can occurrences of the same type be identified in the course of a text? Several norms are possible, depending on different fields of knowledge, practices and perspectives.

- Analyses based on ***graphical forms*** (automatic identification of identical occurrences of a series of non-delimiting characters) are simple to describe and to implement.
- ***Lemmatized*** analyses depend on external sources (dictionaries of lemmas, syntactic parsers).

Some software packages also offer analyses based on groupings of occurrences that contain a common ***root*** or a common ***n-gram*** using various identification procedures that are more or less automatic.

Beyond subdividing the text into graphical forms, ***Lexico3*** allows the identification of other types of textual units.

- ***Repeated segments***: series of consecutive forms found several times in the text.
- ***Co-occurrences***: simultaneous, but not necessarily contiguous, presence of occurrences of two forms in a given context (phrase, section, etc.).

- ***Generalised types*** or ***Tgen(s)***: textual units defined by the user with the help of tools which permit the automatic regrouping of occurrences in the text (ex: occurrences of forms that start with the sequence of characters *democra*: *democracy, democratic, democrat etc.*).

1.4 Example: the *Duchesne* corpus

Text1.txt is a file containing a fragment of the corpus *Père Duchesne*[4] (Duchn.txt). Both files are on the installation CD-ROM.

Here are the explanations of elements used to codify the text in the example files:

- The key Sda is a code for the year the text was published.
- The Numero key introduces an issue number, following the original edition of the text (96 issues numbered from 255 to 351 for the corpus DUCHn.txt, 6 issue numbers for the sub-corpus text1.txt).
- The Epg key moves to another page according to the pagination of the original edition of the text.
- The S03 key distinguishes among the portions of text that are titles and headings (S03=0) and so-called proper text (S03=1).
- The paragraph character § marks the beginning of each paragraph of the text.
- The character * identifies uppercase letters in the original document.

Table 1.1: Example of codified corpus

```
<An=1793> <Numero=220> <S03=0> <Epg=1>
```

[4] The *Père Duchesne* corpus, collected by Jacques Guilhaumou within the research centre *Lexicometrics and political texts* (ENS of Fontenay/St. Cloud), was used in a variety of methodological studies (cf. bibliography *infra*).

```
§ 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.
```

2 Tools for textual exploration

This section describes the functions of ***Lexico3*** that allow subdividing the texts into occurrences of the different textual units that can be constructed from the chain of text (*graphical forms, repeated segments, form groups, Tgens*).

2.1 Segmenting a corpus

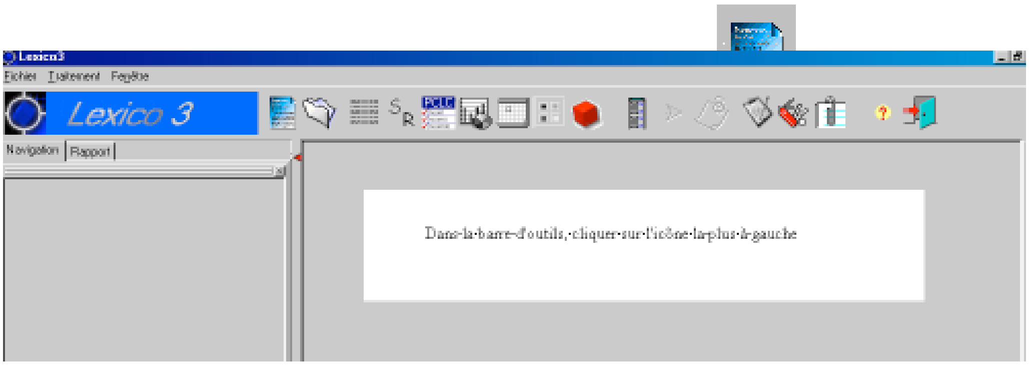
Segmentation creates a textual database from a corpus ***Mycorpus.txt*** furnished by the user. The database is made up of three files (***Mycorpus.dic, Mycorpus.par, Mycorpus.num***), the first two of which can be read using any word processing software.

Operational set-up

Run the software by double clicking on the icon:

Lexico3

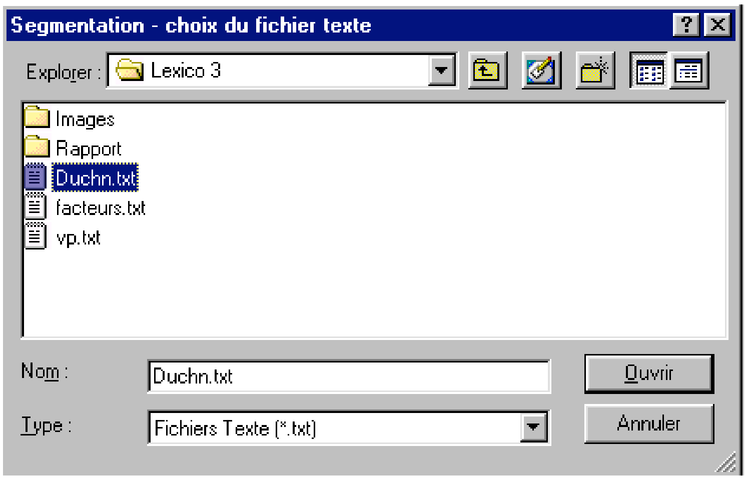
In the toolbar, click on the icon to the far left



Dans la barre d'outils, cliquer sur l'icône la plus à gauche

Click on the icon to open a text file

The program allows choosing a text file in a directory as any Windows software.

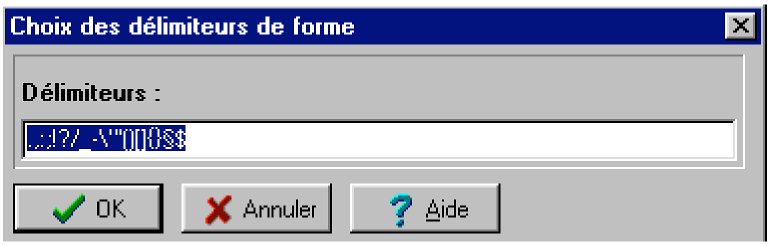


Nom : Duchn.txt

Type : Fichiers Texte (*.txt)

Figure 2.1: Selecting a text file

Select the file that contains the corpus for segmentation *Duchn.txt*. A dialog box appears in order to define segmentation parameters with the help of delimiting characters (cf.1-Preparation of text).



Délimiteurs : .,;:!?/_-\'"()[]{}§$

Figure 2.2: Segmentation parameters dialog box

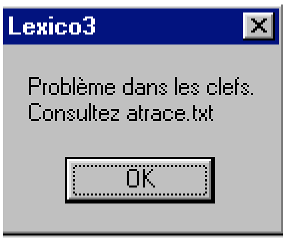
Reminder: It is possible to modify the list of delimiting characters.
Start the segmentation by clicking on the OK button.

Checking the keys

The program checks the conformity of the initial corpus with the norms described above. This module indicates the keys that are incorrectly codified:

Unclosed key	<S01=Alice
Space in the type or contents of the key	<S 01= Al ice>
Closing tag missing	she is < nice.
Absence of = sign	<S01Alice>
Key without contents	<S01=>

Undefined type of key	<=Alice>



Problème dans les clefs.
Consultez atrace.txt

Figure 2.3: Wrong key error message

For more detailed information on errors, see the report file **`atrace.txt`** (automatically created in the same directory as the text file), which indicates the line number at fault. Errors appear as follows:

Table 2.1: Segmentation Report
(Lxxx…. indicates the line at fault)

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

Segmentation of the text

When the faulty lines have been corrected, the program is launched again as above. If there are no more errors, a process bar allows you to follow the progress of the segmentation of the text.

At the end of segmentation, the left part of the screen displays the lexicometric list of the forms in the corpus with the frequency within the entire corpus indicated next to each form. Hapax means any form with a single occurrence within the corpus. To get an alphabetical listing, click on the column header (lexicographic order). A second click returns the list to its initial state (lexicometric order).

Output files

Several output files are created and stored on the hard disk in the same directory as the source text. If the corpus being segmented is called *genericname.txt*, the files are called respectively *genericname.par, genericname.dic, genericname.num.*

The file *genericname.par* contains the principal counts according to forms, occurrences, etc. as well as a reminder of the delimiting characters chosen for the segmentation.

Table 2.2: Example of the parameters file (.par)

```
Lexico3.1 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 ***
```

The file *mycorpus.dic* contains the dictionary of forms sorted by frequency (one entry for each form).
Next to the frequency of the form comes the lexicographic rank of the form (i.e. its number in the list of forms sorted in lexicographic order).
The file *mycorpus.num* contains the numeric coding of the text, that is, the occurrences, forms, punctuation marks, keys and other elements of the corpus in a coded, compact form. This file is for internal use only and can not be consulted using a text editor.
The file ***atrace.txt*** contains a detailed report of the operations carried out by the program (allocated memory, registered parameters, input and output files...). In case of process failure, this file can reveal the source of the problem.

Table 2.3: Excerpt of dictionary

```
frq    rang lex.     forme
6130     2703         de
4749     6033         les
4298     5909         la
3773     4216         et
(…)       (…)         (…)
   1    10967         voyager
   1    10987         zeté
-----------------------------   Fin de la zone des formes graphiques
      259
    10859  !
198 10860  "
 49 10861  $
-----------------------------   Fin de la zone des ponctuations
766 10873 Epg
 96 10874 S01
-----------------------------   Fin de la zone des types de clés
97 10882 01
 1 10883 02
-----------------------------   Fin de la zone des contenus de clés
```

Table 2.4: Excerpt of the trace file (*atrace.txt*)

```
LecParam

192000 192000 11169 142177 10988 6130 5056 5000000 14 8 159
Allocation de la mémoire :
Allocation de lexm réussie, 178720 octets
Allocation de tnum réussie, 768000 octets
Allocation de ftext réussie, 446800 octets
Allocation de list réussie, 24520 octets
Entrée dans OpenDicNum
Dictionnaire numérisé : Duchn.dic
Entrée dans OpenTextNumFichier Texte : DUCH.num : 192083 items.
Fichier Param DUCH.par :
```

2.2 Opening an existing database

We often conduct experiments on a corpus created during work sessions over a period of time. When reusing a database created at a previous session, we can be sure the segmentation parameters established at the first session are correct during a later session.

NB: It is possible to open a text already segmented by dragging it directly to the ***Lexico3*** icon.

2.3 Concordances

The **Concordance** tool allows for the visualisation in context of all the occurrences of a form or a *generalised type* (Tgen). Concordance lines enable to examine the immediate context around a given pivotal word.

Select a form (or a type)

Click on the icon ***Concordance***, a dialog box appears.
To obtain the concordance of a form, you have the choice of:

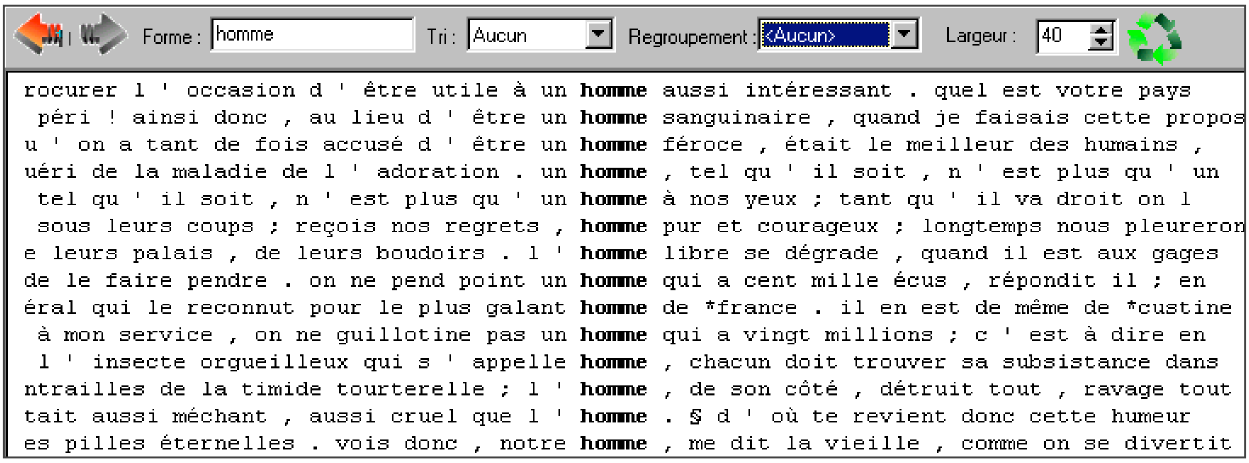
- Entering the form in the editing zone "pivotal word" (ex: homme), then hitting the return key.
- Dragging the form from the dictionary or the *Word-store* (*Garde-mots*) to the concordance window.
- Dragging a link from the window *group of forms* or a *repeated segment* (see section 2.5 'repeated segments'), of which you wish to study the contexts, and dropping it in the right window. The concordance of all the occurrences of the *Tgen* in context is displayed automatically.
- In a concordance window related to the given form, selecting any other form visible in the window and getting its concordance.

Launch the request by pressing the return key. The list of all the occurrences in context for the *type* under investigation is displayed on the screen.

drag/drop

select a form – click on the left mouse button.

keep the left button of the mouse pressed and drag the selected form to the desired place then drop (release the left mouse button).



Forme : homme Tri : Aucun Regroupement : <Aucun> Largeur : 40

```
rocurer l ' occasion d ' être utile à un homme aussi intéressant . quel est votre pays
 péri ! ainsi donc , au lieu d ' être un homme sanguinaire , quand je faisais cette propos
u ' on a tant de fois accusé d ' être un homme féroce , était le meilleur des humains ,
uéri de la maladie de l ' adoration . un homme , tel qu ' il soit , n ' est plus qu ' un
 tel qu ' il soit , n ' est plus qu ' un homme à nos yeux ; tant qu ' il va droit on l
 sous leurs coups ; reçois nos regrets , homme pur et courageux ; longtemps nous pleureron
e leurs palais , de leurs boudoirs . l ' homme libre se dégrade , quand il est aux gages
de le faire pendre . on ne pend point un homme qui a cent mille écus , répondit il ; en
éral qui le reconnut pour le plus galant homme de *france . il en est de même de *custine
 à mon service , on ne guillotine pas un homme qui a vingt millions ; c ' est à dire en
 l ' insecte orgueilleux qui s ' appelle homme , chacun doit trouver sa subsistance dans
ntrailles de la timide tourterelle ; l ' homme , de son côté , détruit tout , ravage tout
tait aussi méchant , aussi cruel que l ' homme . § d ' où te revient donc cette humeur
es pilles éternelles . vois donc , notre homme , me dit la vieille , comme on se divertit
```

Figure 2.4: Concordances
Part of the concordance around the pivotal word *homme*
in the *Duchesne* corpus.

Display of the concordance

The order for sorting by context can be chosen from the drop-down menu 'sort' (after, before, none).
With the drop-down box ***Regroupement*** (Grouping) the contexts can be regrouped according to a partition (for example, by speaker, month or year).
Largeur (Length): choice of the number of characters (including spaces) that should appear before and after each pivotal word. To modify it after one search, change the length and click on refresh (**Figure 2.4**).

Sorting

Various contexts related to a particular form can be arranged in three different ways. These contexts can be sorted in:

- alphabetical order of the occurrence that precedes the pivotal form (sort before)
- alphabetical order of the occurrence that follows the pivotal form (sort after)
- the order in which the occurrences of the pivotal form appear in the text.

The buttons ***Previous*** and ***Next*** (red arrows in the upper left corner of the window) allow the user to navigate among the concordances established for different forms, types, etc.

2.4 Adding the results to the report

all documents produced by ***Lexico3***, each concordance can be added to the final

The report

The results that interest the user for later study can be put together in a folder called ***Rapport*** (Report). This folder, easily handled with the help of a web browser (*Internet Explorer, Netscape*, etc.), contains a file *index.htm* permitting the user to browse through selected results. The report can be consulted at any moment on the condition that the user saved it (cf. section 4.3).

Add to the report

To add a document to the report, click on the icon *Ajouter au rapport* (Add to the report) described in this section. Generally, the icon in the toolbar is used. For certain documents (sections, lists, etc.) use a similar button located in the corresponding window.

2.5 Search for repeated segments

Repeated segments are series of consecutive forms whose frequency is greater than or equal to 2 in the corpus[5]. For example, in the *Duchesne* corpus are found the segments...

Segment	*lenght*	*frequency*
tirer les marrons du feu	5	6

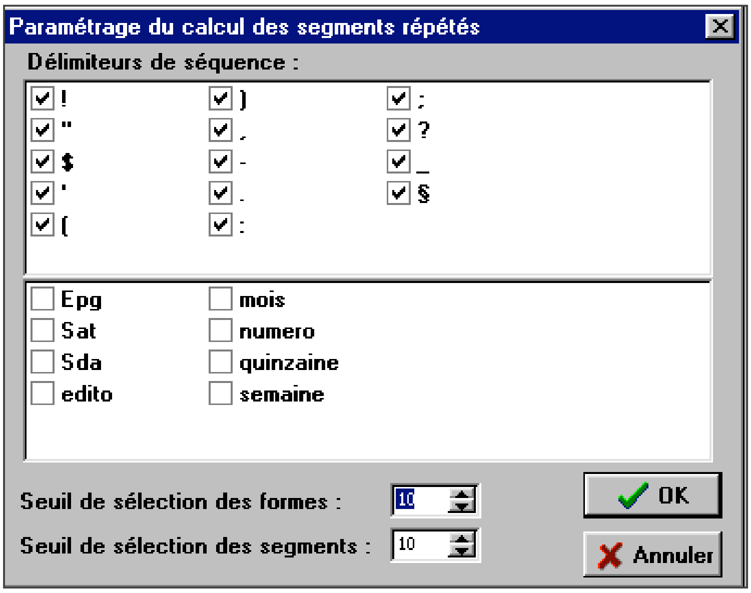
[5] (*L&S*, p.58)

To create a list of repeated segments click on the icon **SR**; a dialog box appears which allows you to choose parameters in the selection of repeated segments (**figure 2.5**):

The upper section of the window allows for the selection of the status of the delimiting characters in the text (the status by default is *sequence delimiter*. To change this status, uncheck the mark opposite the corresponding character). The segments listed will not overlap with this type of delimiter.

The lower section allows for the selection of the keys found in the corpus (here a segment can overlap with a key indicating a page-turn but not with a key indicating a change of section).

A minimum frequency is established below which forms and segments are not selected. The minimum frequency by default is 10.
The OK button starts the search for repeated segments.
The list of repeated segments found in the text appears on the left part of the window. To consult the list, click on the tab ***Segments répétés*** (Repeated segments).



Paramétrage du calcul des segments répétés

Délimiteurs de séquence :

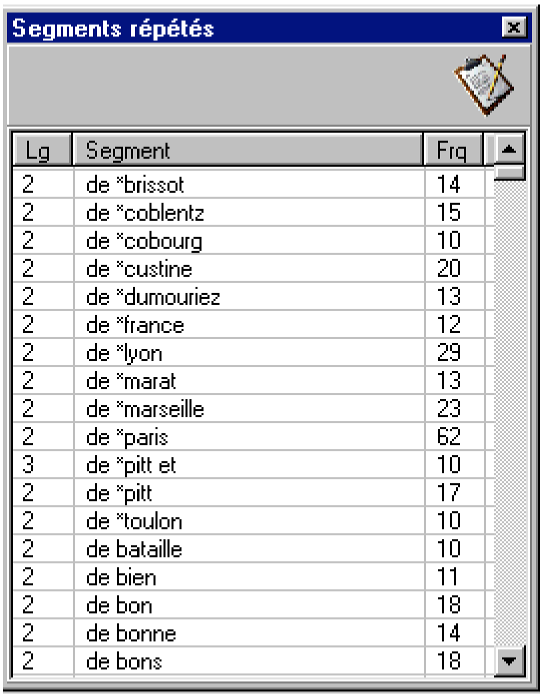
☑ !
☑ "
☑ $
☑ '
☑ (
☑)
☑ ,
☑ -
☑ .
☑ :
☑ ;
☑ ?
☑ _
☑ §

☐ Epg
☐ Sat
☐ Sda
☐ edito
☐ mois
☐ numero
☐ quinzaine
☐ semaine

Seuil de sélection des formes : 10

Seuil de sélection des segments : 10

Figure 2.5: Delimiters and threshold for forms



Lg	Segment	Frq
2	de *brissot	14
2	de *coblentz	15
2	de *cobourg	10
2	de *custine	20
2	de *dumouriez	13
2	de *france	12
2	de *lyon	29
2	de *marat	13
2	de *marseille	23
2	de *paris	62
3	de *pitt et	10
2	de *pitt	17
2	de *toulon	10
2	de bataille	10
2	de bien	11
2	de bon	18
2	de bonne	14
2	de bons	18

Figure 2.6: List of repeated segments

2.6 Form groups

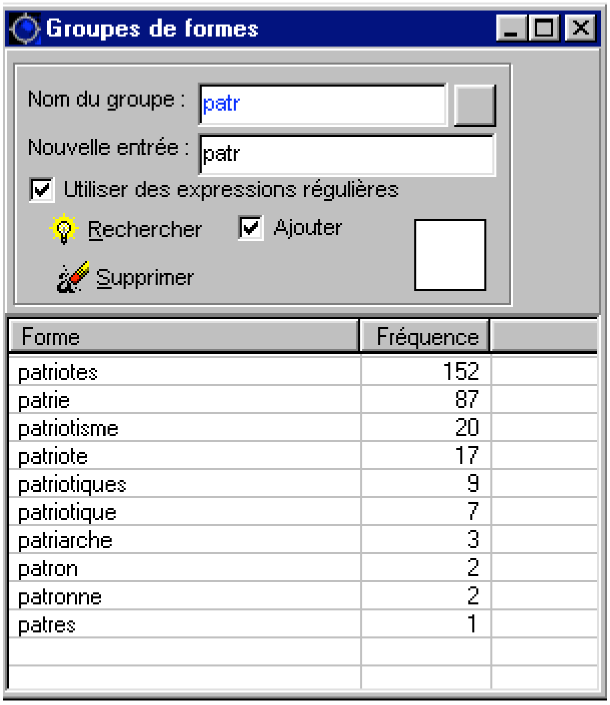
The tool ***Groupe de formes*** (Form group) allows the creation of *types* that collect occurrences of different graphical forms according to a common characteristic. For example, taking certain precautions, you can assemble the plural and singular of the same form, the tenses of the same verb, forms with a semantic connection, etc. Thus regrouped, the forms can then be processed like a unique entity *Tgen*.

At the same time a search is launched for multiple forms by introducing chains of characters corresponding to prefixes, suffixes and series of graphical characters.

Set-up

- Enter the name of the form group
- Enter the form to search for
- Click on *rechercher* (search)

The resulting “object” can then be processed like a “classical” form by clicking on the red arrow of the form group (and holding down the left mouse button), and dragging the group onto the map of the partition. **(Figure 2.7)**



Groupes de formes

Nom du groupe : patr

Nouvelle entrée : patr

☑ Utiliser des expressions régulières

☑ Ajouter

Forme	Fréquence
patriotes	152
patrie	87
patriotisme	20
patriote	17
patriotiques	9
patriotique	7
patriarche	3
patron	2
patronne	2
patres	1

Figure 2.7: Creation of the form groups

The button ***Supprimer*** (Delete) allows for the refinement of this list by eliminating, for example, after selecting them, the forms *patriarche*, *patron*, *patronne*, *patres*, etc.

Regular expressions

We have chosen the language of regular (or rational) expressions, frequently used in computing, to allow the user to establish groups[6].

To search forms (Tgen) using regular expressions, Lexico will introduce by default a search for words starting with a given chain.

For example: if you search for the pattern “pat”, the Tgen produced will be all the words that start with “pat” (patriot, pater…)

To specify the ending of words sought, use “\>”.

For example to search for all the words ending in “ism”, the pattern to use is “\<.*ism\>. This pattern can also be written “.*ism\>, if the search is by word.

[6] To learn more about regular expressions (xxxxx)

Operator	Function	Application
. (dot)	Replaces any character	“tr.e” can mean tree, trie…
*	0 or n occurrences of the preceding character	“com*e” searches for coe, comme, commme,…
+	1 or n occurrences of the preceding character	“com+e” searches for comme, commme,…
\<	Represents beginning of word	“\<capital” searches for capital, capitale, capitalism…
\>	Represents end of word	“.*ism\>” searches for syndicalism, capitalism…
[]	Represents a set of characters	“[aeiou]” represents one of the characters in the set of vowels. “[a-z]” represents one of the characters between a and z.
[^]	Represents the negation of the content of the set of characters	“[^aeiou]” represents the characters that do not belong to the set of vowels

2.6 Word-store

d-store allows for the memorization of forms, segments, *Tgen(s)* for later use.
To store a *Tgen* in the word-store, drag it to the icon of the red cube (cf. drag/drop *supra*).
To use a *Tgen* stored in the word-store, drag it from the red cube to the window (concordance, frequency, map of sections, etc.) where it should appear.

3 Tools for statistical analysis

This chapter includes various strategies used for statistical analysis, ranging from elementary descriptive methods (counts, histograms, etc.) to different types of multivariate textual data analysis (correspondence analysis, cluster analysis, textual time series).

3.1 Partitioning

The different keys introduced before automatic segmentation (cf. section 1 – Text corpora) allow us to carry out various partitions of the corpus.
To carry out a partition, select a key type; depending on the different values assigned to this key, the corpus will be divided into as many different parts.
Example: after segmenting the corpus ***Duchn.txt***, click on the icon ***Statistiques par partie*** (Statistics by part), a dialog box appears allowing for the selection of a partition key (**Figure 3.1**). Select, for example, the key *semaine* – week (double click or the button ***Créer*** - create).

A window opens allowing for a comparison of the frequency of textual units in the set of parts.

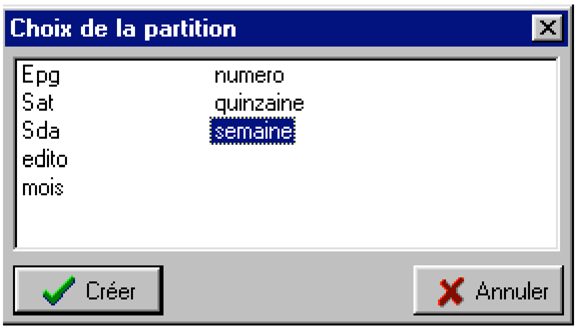


Figure 3.1: Choice of a partition

Distribution of a form (or Tgen)

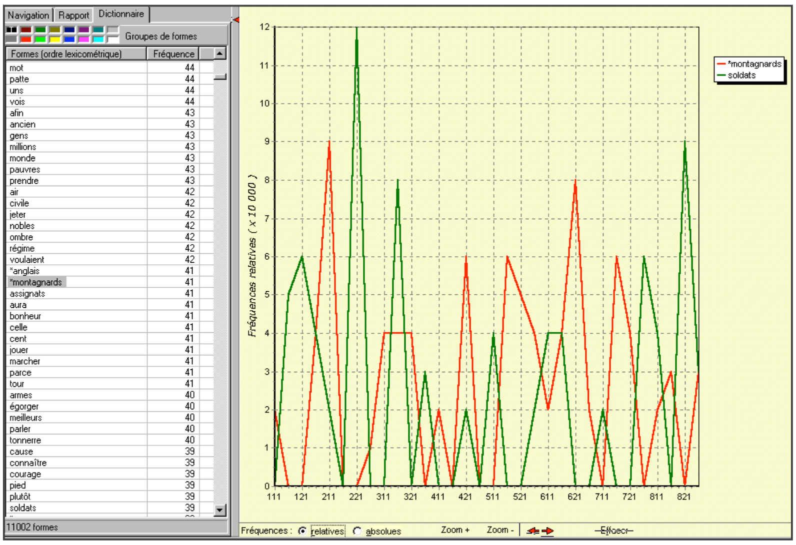
By dragging the forms and/or repeated segments (section 2.4) found in the windows at the left of the screen to this window, you get a distribution of the selected textual unit(s) in different parts of the corpus (Figure 3.2). You can also drag the form groups (section 2.5) from the corresponding window as well as the links saved in the word-store (section 2.6) to this window.

Choose a color to map the TGen by activating the paint box found at the top to the right of the dictionary (coordinating the *form group* window). If a color is not selected by the user, the software chooses different colors for each new distribution.

The mapping zone can be re-initialized at any time (button ***effacer*** - erase, for example after having recorded a graph in a report).

The distribution of several textual units found in parts of the corpus can be interpreted:

- as an absolute frequency (number of occurrences in the text part)
- as a relative frequency (number of occurrences in relation to the length of the text part)
- in terms of characteristic elements (as a result of a statistical calculation, section 3.2)



Formes (ordre lexicométrique)	Fréquence
mot	44
patte	44
uns	44
vois	44
afin	43
ancien	43
gens	43
millions	43
monde	43
pauvres	43
prendre	43
air	42
civile	42
jeter	42
nobles	42
ombre	42
régime	42
voulaient	42
*anglais	41
*montagnards	41
assignats	41
aura	41
bonheur	41
celle	41
cent	41
jouer	41
marcher	41
parce	41
tour	41
armes	40
égorger	40
meilleurs	40
parler	40
tonnerre	40
cause	39
connaître	39
courage	39
pied	39
plutôt	39
soldats	39

11002 formes

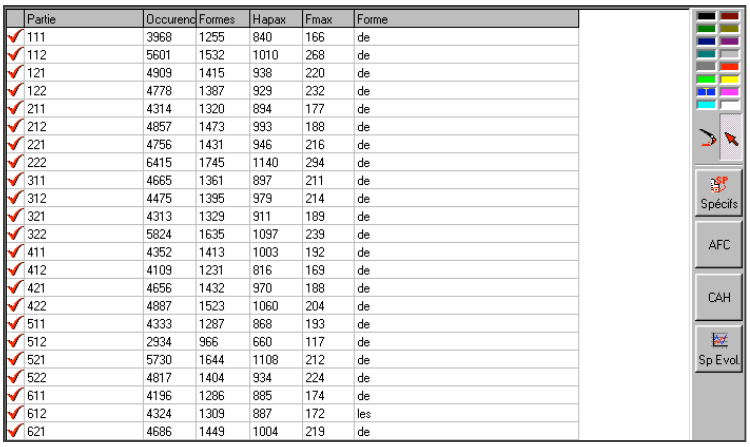
Figure 3.2: Distribution of forms in the text parts of a corpus

Statistics by text part (PCLC)

(principal lexicometric characteristics of the corpus and of the partition)

With the selection of the icon ***PCLC***, appear the principal characteristics by text part according to the partition chosen.

- A red check in the extreme right column indicates that the part is selected for the count of global frequencies in the corpus.
- The second column contains the names of the different text parts (here the number of the week).
- The ***occurrences*** column gives the total number of occurrences of the forms listed.
- The column ***formes*** gives the number of different graphical forms present in each part.
- The ***hapax*** column gives for each part the number of the forms that appear only once in the part.
- The ***Fmax*** (maximum frequency) column gives the number of occurrences of the most frequent form.



	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	111	3968	1255	840	166	de
✓	112	5601	1532	1010	268	de
✓	121	4909	1415	938	220	de
✓	122	4778	1387	929	232	de
✓	211	4314	1320	894	177	de
✓	212	4857	1473	993	188	de
✓	221	4756	1431	946	216	de
✓	222	6415	1745	1140	294	de
✓	311	4665	1361	897	211	de
✓	312	4475	1395	979	214	de
✓	321	4313	1329	911	189	de
✓	322	5824	1635	1097	239	de
✓	411	4352	1413	1003	192	de
✓	412	4109	1231	816	169	de
✓	421	4656	1432	970	188	de
✓	422	4887	1523	1060	204	de
✓	511	4333	1287	868	193	de
✓	512	2934	966	660	117	de
✓	521	5730	1644	1108	212	de
✓	522	4817	1404	934	224	de
✓	611	4196	1286	885	174	de
✓	612	4324	1309	887	172	les
✓	621	4686	1449	1004	219	de

Figure 3.3: Characteristics of the partition

The table allows for a rapid visual comparison of the parts with regard to their most important lexicometric characteristics.

3.2 Characteristic elements

The analysis of characteristic elements allows for an evaluation of the frequency of each of the textual units in each of the parts of the corpus[7].

The ***Spécifs*** button found on the right of the *PCLC* window (**Figure 3.3**) gives a table of the characteristic forms of a selected part (**Figure 3.5**) or a set of parts[8].

By default, the characteristic element index is calculated for all the units with a frequency of more than 10, with a *probability threshold* set at 5% (a window ***paramétrage du calcul des spécificités*** - *parameters for characteristic element computation* appears before calculation begins allowing the user to modify these parameters, if necessary).

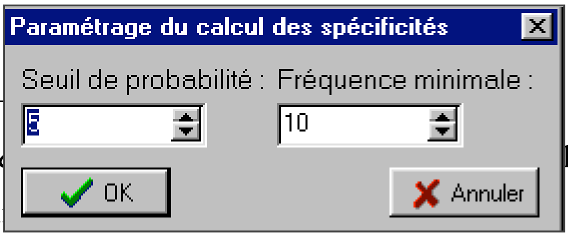
The *characteristic element diagnostics* contains two indications.

a) the sign (+ or -) indicating an over or under-use in the selected part(s) in comparison to the entire corpus.

b) an exponent that indicates the degree of significance of the difference (an exponent equal to *x* means that the probability of a distribution difference more than or equal to the difference found was of the order 10^{-x}).

Example: nous F=1270 f= 66 **+05**

indicates that the form *nous*, present 1270 times in the corpus and found 66 times in the texts of the week number 211, comes up more often than it might have been expected in a distribution "at random"[9].



Paramétrage du calcul des spécificités

Seuil de probabilité : 5

Fréquence minimale : 10

[7] See (Lafon, 1984) or (L lement computation.

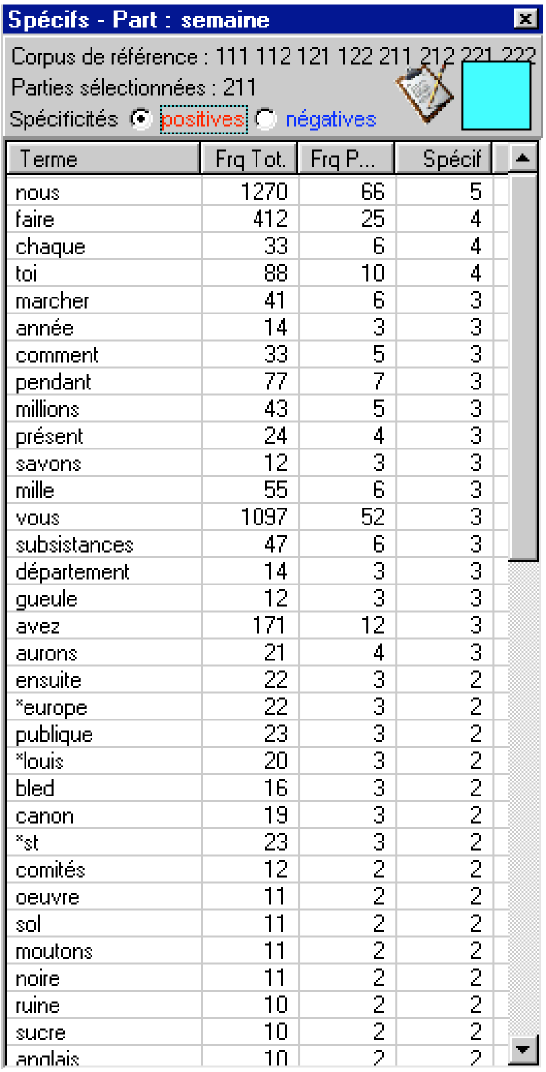
[8] To select a part, just cl a part to the set of parts already selected by pressing the *Control* key at the same time on.

[9] Under hypothesis of hypergeometric distribution with these parameters.

Figure 3.4: Parameters

NB: If the calculation of repeated segments was carried out ahead of time, the characteristic segments will also appear in the list of characteristic elements.

Results of computation of characteristic elements



Spécifs - Part : semaine

Corpus de référence : 111 112 121 122 211 212 221 222

Parties sélectionnées : 211

Spécificités ● positives ○ négatives

Terme	Frq Tot.	Frq P...	Spécif
nous	1270	66	5
faire	412	25	4
chaque	33	6	4
toi	88	10	4
marcher	41	6	3
année	14	3	3
comment	33	5	3
pendant	77	7	3
millions	43	5	3
présent	24	4	3
savons	12	3	3
mille	55	6	3
vous	1097	52	3
subsistances	47	6	3
département	14	3	3
gueule	12	3	3
avez	171	12	3
aurons	21	4	3
ensuite	22	3	2
*europe	22	3	2
publique	23	3	2
*louis	20	3	2
bled	16	3	2
canon	19	3	2
*st	23	3	2
comités	12	2	2
oeuvre	11	2	2
sol	11	2	2
moutons	11	2	2
noire	11	2	2
ruine	10	2	2
sucre	10	2	2
anglais	10	2	2

In the first column are the characteristic units in descending order of diagnostics indication. The next two columns show, respectively, the frequency of the form in the entire corpus and the frequency of the form in the selected part.

The *positive* and *negative* check buttons in the tab of characteristic elements allow you to inverse the order of presentation of the list, which, by default, starts with positive characteristic units.

3.3 Chronological characteristic elements

For textual time series (series of texts produced by the same textual source and spaced regularly over time, example ***Duchesne***), in addition to the analysis of the characteristic elements of each corpus part, chronological characteristic elements diagnostics reveals the vocabulary specific to longer periods of consecutive parts (cf. L&S p.197 and Salem 93).

Characteristic increments

For the selected corpus part, the ***SpEvol*** button enables to calculate characteristic elements (or ***characteristic increments***) of the corpus part relative to the preceding chronological periods (subsequent periods being temporally excluded from calculation). The final results of this computation are presented in a table of characteristic units similar to the one shown on **Figure 3.5**.

NB: Within the table, negative characteristic increment diagnostics reveals textual units that are likely to fall into disuse in the corpus period considered in relation to preceding periods.

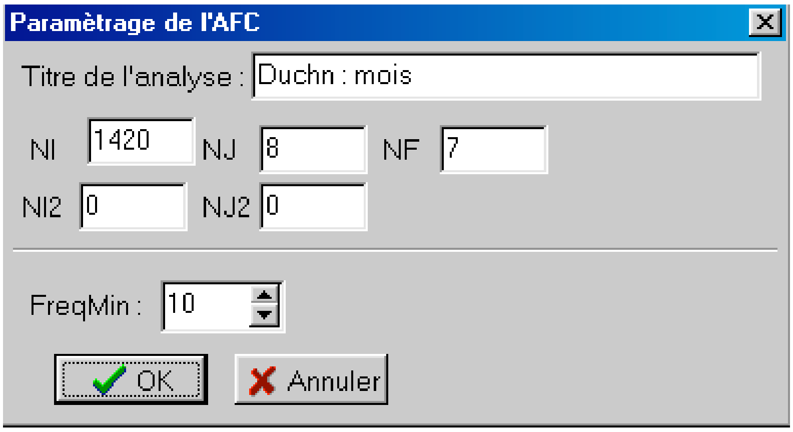
3.4 Correspondence analysis (CA)

The ***AFC*** button permits a correspondence analysis on all the parts of the corpus (excluding those where the red check has been removed)[10].

Use the parameters window (Figure3.6) to set, among other values:

- the number of textual units considered in the analysis
- the number of principal axes to be extracted

NB: By default, the analysis considers the units with a frequency of more than 10. The modification of the minimum frequency requires a new calculation of the number of units to be considered.



Titre de l'analyse : Duchn : mois

NI 1420 NJ 8 NF 7

NI2 0 NJ2 0

FreqMin : 10

Figure 3.6: CA parameters window

Click on the ***OK*** button to start the analysis. The parts of the corpus appear on the plane spanned by the first two dimensions of the correspondence analysis. Additional visualizations can be obtained selecting other principal axes (use drop-down boxes found above the graph).

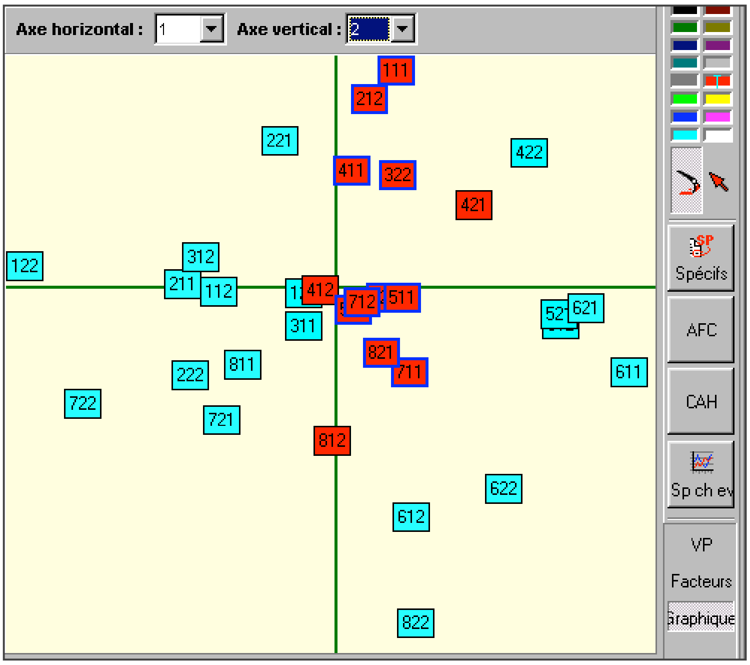
The different planar maps allow for an estimation of the proximities calculated between the selected parts regarding their vocabulary.

The analysis can be repeated after removing certain parts (right click – the parts removed from the corpus appear shaded in grey).

A part or a set of parts can be selected directly on the map (left click). The contours of the parts selected are highlighted. This allows you, for example, to calculate the characteristic elements of a set of parts.

[10] A complete discussion of this method can be found in (*L&S* p. 135).

Figure 3.7: CA (Correspondence Analyses) graph



Axe horizontal : 1 Axe vertical : 2

The paintbrush and the paint box found to the right of the graph allow you to associate a colour to a set of parts. With the arrow button you can return to selection mode.
With the last group of buttons you can browse among the results of the analysis:

- with ***VP*** (eigenvalue) consult the *histogram of eigenvalues*
- with ***Facteurs*** (principal axes) consult the principal axes table
- with ***Graphique*** (graph) return to the planar map.

4 Tools for lexicometric browsing

This section describes the functions that allow the user to move among the results produced by the different lexicometric methods and the original text.

4.1 Map of sections

The ***map of sections*** allows for the visualization of the corpus cut into sections by raising one (or several) characters (carriage return, period, etc.) to the rank of ***section delimiters***.

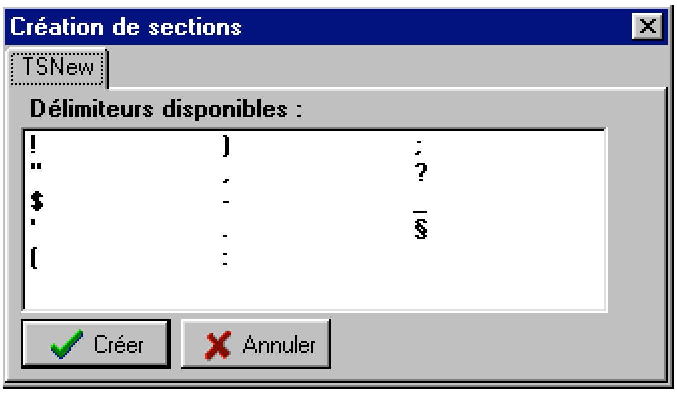


Figure 4.1: Choice of section delimiters

Mapping of the sections for a Tgen

Select the *Tgen* (from the dictionary, the *Word-store*, the list of repeated segments, etc. …) and drag it to the map (keeping left mouse button pressed).

- Select the section to be visualized in the lower window by clicking on the square representing it on the map of sections.
- Enlarge the squares representing each of the sections by moving the cursor (found at the top left of the window) to the right.
-
- Mark out an activated partition by selecting it in the drop-down box found immediately to the right of the pointer.
- Color the map of sections depending on the characteristic element diagnostics of the *Tgen* under study in each section. First check the box ***seuillage*** (threshold). The immediately preceding icon enables to set two probability thresholds, producing more or less dark section coloring. For a simultaneous representation of two *Tgen(s)*, this process can be reiterated (remember to change the color in the corresponding box). In this case, keep the *Control* key pressed while proceeding with the second *drag/drop* operation.

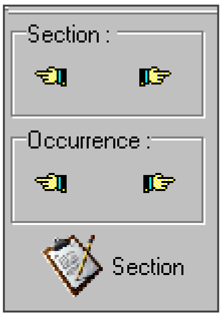
Statistical tools of the map of sections

The two icons found at the same level to the right of the window allow for the identification of the characteristic types of a set of sections (characteristic elements of the sections selected, cf. 3.2).

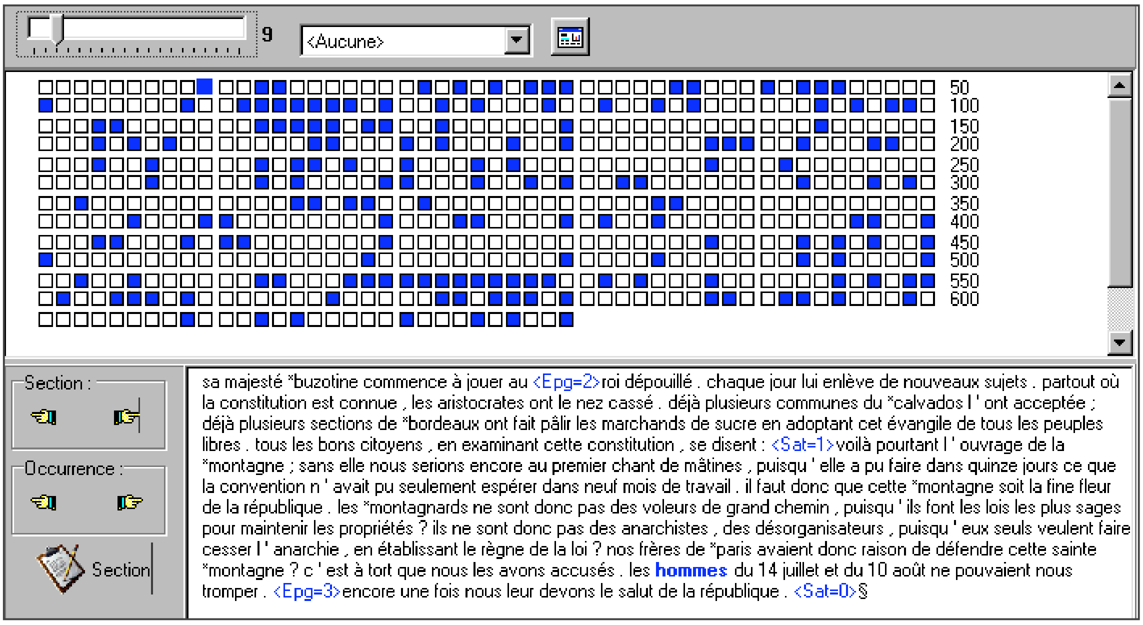
- The first button ***Cooccurrences*** (co-occurrences) automatically produces a selection of sections where the *Tgen* under study is found (this is the set of sections to be compared with the entire corpus).
- The second button ***Spécificités*** (characteristic elements) allows the user to produce an arbitrary selection of sections for subsequent study of its vocabulary (according to the rules set for *Windows*, the selection of the sections is made by simple mouse click while keeping the ***Control*** key pressed; the uppercase key allows for the selection of a group of consecutive sections).

As always, the lists of characteristic elements are displayed in the window at the left. The number of sections appears at the top of the window; the results can be saved using a button at the bottom of the window, ***ajouter au rapport Section*** (save to the Section report).

Browsing with the help of the map of sections



- By using the buttons (in the shape of hands) located to the left of the toolbox, you can go back or move forward or to the next or preceding section or to the next/preceding occurrence of the *Tgen* selected.
- With the icon ***ajouter au rapport Section*** (save to the Section report) you can record the section visualized in the lower window.



sa majesté *buzotine commence à jouer au <Epg=2>roi dépouillé . chaque jour lui enlève de nouveaux sujets . partout où la constitution est connue , les aristocrates ont le nez cassé . déjà plusieurs communes du *calvados l ' ont acceptée ; déjà plusieurs sections de *bordeaux ont fait pâlir les marchands de sucre en adoptant cet évangile de tous les peuples libres . tous les bons citoyens , en examinant cette constitution , se disent : <Sat=1>voilà pourtant l ' ouvrage de la *montagne ; sans elle nous serions encore au premier chant de mâtines , puisqu ' elle a pu faire dans quinze jours ce que la convention n ' avait pu seulement espérer dans neuf mois de travail . il faut donc que cette *montagne soit la fine fleur de la république . les *montagnards ne sont donc pas des voleurs de grand chemin , puisqu ' ils font les lois les plus sages pour maintenir les propriétés ? ils ne sont donc pas des anarchistes , des désorganisateurs , puisqu ' eux seuls veulent faire cesser l ' anarchie , en établissant le règne de la loi ? nos frères de *paris avaient donc raison de défendre cette sainte *montagne ? c ' est à tort que nous les avons accusés . les **hommes** du 14 juillet et du 10 août ne pouvaient nous tromper . <Epg=3>encore une fois nous leur devons le salut de la république . <Sat=0>§

Figure 4.2: Distribution of the form *hommes* within corpus paragraphs

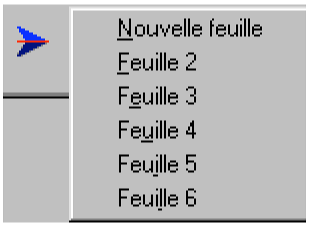
4.2 Towards better use of the windows

Create a worksheet

To avoid the splitting up of the main window, create new worksheets by clicking on this icon. The worksheets pile up at the right of the main window. With the tab “Feuille n°i” (sheet number) you can move from one to another. The *Tgen* links can be carried from one sheet to another using, for example, the *Word-store* function.

Move to another worksheet

To move a results window to a new sheet, select the window of your choice, click on this icon and select the desired sheet.



Mosaic

With this icon, several windows can be placed on the same sheet.

4.3 The report

The ***Rapport*** (report) folder contains the results selected by the user for later study. Easily handled with the help of a web browser (*Internet Explorer*, *Netscape*, etc.),

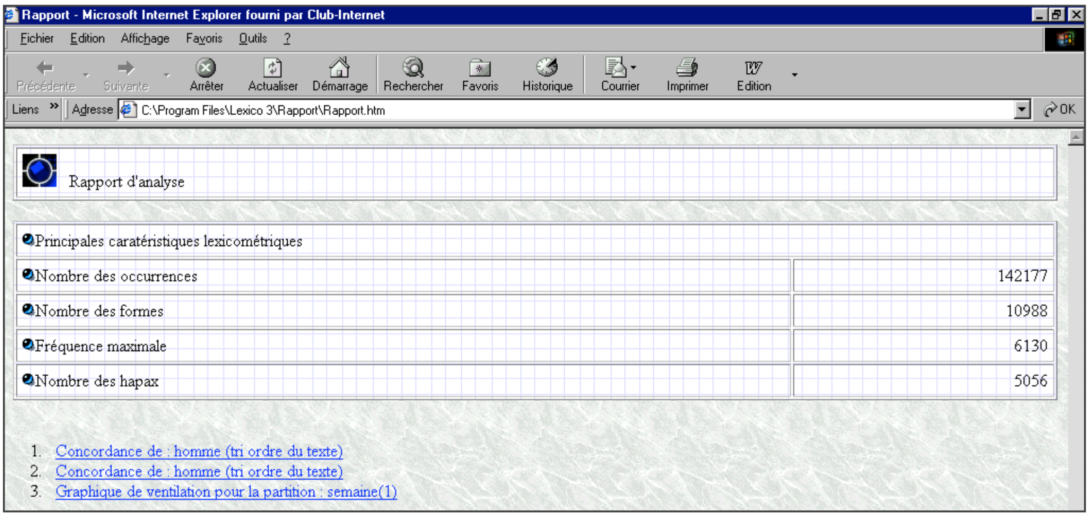
this folder contains a file *index.htm* that leads to the results.
The report can be consulted at any time on the condition that it has been saved by the user (button ***Enregistrer*** – *save* at the bottom of the ***Rapport***–*report* tab).

Editing the results

To visualize a text or the results obtained with ***Lexico3***, click on the icon "Editeur" (editor) and select the desired document under the icon "Ouvrir" (open).

To ensure that the documents are backed up during different sessions, save each time the folder ***Rapport*** in a different directory or under a different name.

The folder ***Rapport*** is found in the ***Lexico3*** directory created during the installation of the software.



C:\Program Files\Lexico 3\Rapport\Rapport.htm

Rapport d'analyse

Principales caratéristiques lexicométriques	
Nombre des occurrences	142177
Nombre des formes	10988
Fréquence maximale	6130
Nombre des hapax	5056

1. Concordance de : homme (tri ordre du texte)
2. Concordance de : homme (tri ordre du texte)
3. Graphique de ventilation pour la partition : semaine(1)

Figure 4.6: Report

4.4. Options – Help - Tips

Options

This button enables to modify the restrictions set for the software (approximately 100,000 distinct lexical forms) while processing large corpora (several millions of occurrences). You can also indicate whether the corpus was previously submitted to some sort of tagging (part-of-speech, sense etc.).

Several examples of corpora:

Corpus	***pages***	***occurrences***	***distinct forms***	***max. frequency***
Duchesne	350	142 177	10 988	6130 (*de*).
Coran (trad. Fr)				(*de*).
Duchesne				(*de*).

Browsing tab

With this tab, you can browse through the results produced by ***Lexico3*** as with Windows Explorer.

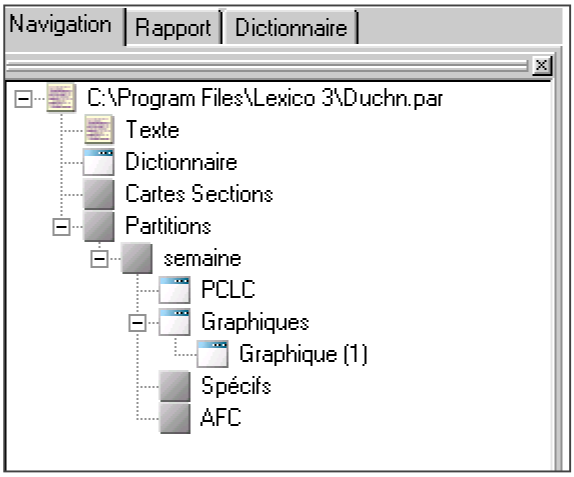


Figure 4.6: Browsing

Full screen

To display the right-hand window as a full screen, click on the red arrow mark found between the right and left windows.

Help

The help file of ***Lexico3*** (including the present manual) is available at any time on the screen by clicking on the help icon.

Exit

To quit ***Lexico3***, make sure all data have been saved in the report then click on the exit icon.

5 Glossary of terms used in textual statistics

The definitions of some basic notions of textual statistics are provided in the help on-line.

NB: The asterisks marks cross-references inside the same glossary. The abbreviations that follow in parenthesis indicate the field of study for which the definition is particularly appropriate.

Abbreviations:
as Automatic Segmentation
ca Correspondence Analysis
ced Characteristic Element Diagnostics
clu Cluster analysis
ling Linguistics
mca Multiple Correspondence Analysis
pa Principal Axes methods
rs Analysis of the Repeated Segments
stat Statistics

active elements. – set of elements used for calculating eigenvalues* and principal axes, as opposed to supplementary elements* that are positioned *a posteriori* on these principal axes.

active variables – variables used to produce a typology either by correspondence analysis or by cluster analyses. The typologies depend upon the choice and weight of active variables that are supposed to be uniform regarding these criteria.

alphabetical index – index* in which forms are arranged in lexicographic order* (as in dictionaries).

automatic segmentation – set of operations performed through computerized procedures, according to fixed rules, that result in dividing a text stored on a machine-readable device into distinct units, called minimal* units.

axe – (ca or mca) artificial variables constructed by the techniques of principal axes methods that approximately summarize initial active variables.

banal form – (ced) form that is, for a given corpus part, not “unusually frequent” or “unusually rare” (neither positive nor negative characteristic element of the part).

banal vocabulary – (ced) set of forms that are, for a given threshold, not “unusually frequent” (positive characteristic elements) or “unusually rare” (negative characteristic elements) in any of the corpus parts (i.e. set of forms that are "ordinary" for each of the corpus parts).

character (as) – typographic sign used for text encoding on a computer-readable device.

characteristic form – (of the part) syn.: positive characteristic element*.

characteristic increment – (ced) characteristic element* calculated for a corpus part relative to a preceding part.

chi-2 distance – distance between profiles* of frequencies, used in correspondence analysis* and in some cluster analysis* algorithms.

chronological characteristic element– (ced) characteristic element* calculated for a group of contiguous parts of a corpus with a longitudinal partition*.

chronological grouping (as) – grouping of natural corpus parts according to chronological order of writing, edition or publication of texts collected in the corpus.

cluster analyses (stat) – statistical technique for grouping observations or elements among which distances have been calculated.

common form – form found in each of the corpus parts.

common vocabulary – (as) set of forms found in each of the corpus parts.

concordance (as) – set of rows of contexts around a given pivotal-word.

confined terms/free terms – a term S1 is said to be confined to another term S2 of a superior length if all of its occurrences* stand for the sub-segments* of the segments corresponding to the occurrences of the segment S2. On the contrary, a term having several distinctive expansions that are not necessarily recurrent is a free term.

contingency table (stat) – (syn.: for table of frequencies or cross-tabulation) table whose rows and columns represent respectively the categories of two categorical (nominal) variables, and whose general term represents the number of individuals associated with each pair of categories.

co-occurrence (as) - (a c.) – simultaneous, but not necessarily contiguous, presence of occurrences of two given forms in a fragment of text (sequence, sentence, paragraph, neighbourhood of occurrence, corpus part, etc.).

corpus (ling) – limited set of texts upon which the study of a linguistic phenomenon is based.

In lexicometrics, a set of texts that are combined for comparison purposes, serving as a basis for a quantitative study.

correspondence analysis (stat) – principal axes method applied to contingency tables*. CA is mainly characterised by the use of special distance called chi-2 distance* (or c2).

delimiting/non-delimiting characters (as) – differentiation established among all characters within a text in order to make it possible for computerized procedures to segment the text into occurrences* (series of non-delimiting characters whose boundaries at both ends are occurrence delimiters).

There are several types of delimiting characters:

- **occurrence delimiters** (also called "**form delimiters**") which are generally: blanks, common punctuation marks, the signs of preliminary analysis possibly included in a text.

- **sequence delimiters**: subset of occurrence delimiters consisting of weak and strong punctuation marks.

- **sentence separators**: subset of sequence delimiters consisting, as a rule, of strong punctuation marks.

dendrogram – (clu) (syn.: hierarchical tree) graphic representation of a hierarchical cluster analysis*, showing the progressive inclusion of clusters.

discourse/language – language is an abstract object that can only be understood in its oral or its written implementation; "discourse" is a convenient term that encompasses both implementations.

distribution – (of occurrences* of an element in the parts of a corpus*) the series of *n* numbers (*n* = number of parts in the corpus) constituted by the sequence of sub-frequencies* of that element in each of the parts, shown in the same order as the parts.

distributional stock of vocabulary – (of a text fragment) the vocabulary* of the fragment together with frequency counts for each of the forms entering in its composition.

editions of contexts (as) – concordance type editions in which the occurrences of a given form are accompanied by a fragment of context corresponding to multiple rows of text around a given pivotal form. The context length is defined in number of occurrences before and after each occurrence of a given pivotal form.

eigenvalue – (ca or mca) quantities that make it possible to determine the relative importance of successive principal axes in a principal axes decomposition. The eigenvalue noted λ_α measures the variance of elements on axis α.

elements of a segment (rs) – each of the forms corresponding to the occurrences composing a segment. Ex: A, B, C are, respectively, the first, the second and the third element of a segment ABC.

entire lexical table (ELT) – contingency table* cross-tabulating forms and parts of a corpus. The general term *k(i,j)* of the ELT is equal to the number of times that form *i* is found in part *j* of the corpus. The rows of the ELT are sorted by lexicometric order* of the corresponding forms.

enunciation – (ling) a set marks, present within a text, revealing the way in which the author has produced this text.

form – (as) or "**graphical form**" type corresponding to identical occurrences* in a corpus of texts (occurrences composed of exactly the same non-delimiting characters).

frequency (as) – (of a textual unit) the number of its occurrences in a corpus.

frequency distribution (as) – sequence of forms of *k*-frequency noted *Vk*, where k varies from 1 to maximum frequency.

frequency of a segment (rs) – (or of a polyform) the number of occurrences of this segment within a corpus.

generalised types (Tgen*s*) – textual units defined by the user with the help of tools which permit the automatic regrouping of occurrences in the text (ex: occurrences of forms that start with the sequence of characters *democra*: *democracy, democratic, democrat etc.*).

hapax – gr. hapax (legomenon), "something stated once".

(as) form whose frequency is equal to one in the corpus (corpus hapax) or in one of its parts (part hapax).

hierarchical cluster analysis (clu) – special cluster analysis technique yielding, through progressive agglomeration, clusters that have the property of being either splitted or included in one another.

hierarchical index (as) – index* in which pivotal forms* are arranged in lexicometric order*.

identification – (stat, ling, as) recognition of identical elements through multiple usages in different contexts and situations.

index – (as) list consisting of a rearrangement of the forms of a text grouping together references* relative to the occurrences* of a given form.

index by parts – set of indexes (hierarchical or alphabetical) established separately for each corpus part.

lemmatization – grouping of occurrences of text into a canonical form (generally on the basis of a dictionary). In English, this grouping is generally done in the following way:

- verbal forms in the infinitive,
- nouns in the singular,
- elided forms to a form without elision.

length (as) – (of a corpus, of a part of that corpus, of text fragment, of a segment, etc.) number of occurrences contained in that corpus (or of that part, fragment, etc.); syn.: size.

Note: T – size of a corpus; t_j – size of the corpus part (or sub-division) of the number j.

length of a segment (rs) – the number of occurrences contained in that segment.

lexical – (ling) concerning the lexicon* or the vocabulary*.

lexical table – (LT) contingency table* obtained from the ELT by deleting certain rows (for example, those corresponding to the forms whose frequency is lower than a given threshold).

lexicographic order –

_ for graphical forms:

order in which words are arranged in a dictionary.

NB: The letters with diacritic marks have the same rank as similar non-diacritic characters, the diacritic sign being used for sorting only in case of complete homographs. For example, in dictionaries of the French language, one can find the following forms arranged in this order: *mais, maïs, maison, maître.*

_ for polyforms:

order resulting from sorting polyforms in lexicographic order of the first component; polyforms starting by the same graphical form are arranged according to lexicographic order of the second form, etc.

lexicometric order (as) –

_ for graphical forms:

order resulting form sorting the forms of a corpus in order of decreasing frequency; forms of the same frequency are arranged in lexicographic order*.

_ for polyforms:

order resulting form sorting in order of decreasing length of segments; segments of the same length are sorted according to their frequency; segments of the same length and of the same frequency are arranged by lexicographic order.

lexicometrics – series of methods that enable to operate formal reorganisations of textual sequence and to conduct statistical analysis based on the vocabulary* of a corpus of texts.

lexicon – (ling) entire set of words of a language.

longitudinal partition – ordered partitioning of a corpus*.

maximum frequency (as) – frequency of the most frequently occurring form in a corpus* (article “the” for most of the English texts).

minimal units (for a given type of segmentation) – units that are not to be subdivided further (ex: forms are not subdivided into characters during segmentation based on graphical forms).

modal value – (stat) value for which a distribution achieves its maximum.

negative characteristic element – (ced) for a given word i and part j , word i is said to be a negative characteristic element of part j (or an “anti-characteristic word” of that part) if its frequency is “unusually low” within that part; more precisely, if the sum of the probabilities calculated from the hypogeometric model for the values equal or inferior to the observed sub-frequency is smaller than the threshold fixed at the beginning.

neighbourhood of an occurrence – for a given occurrence of a text, each segment (any series of consecutive occurrences which is not separated by a sequence delimiter) containing this occurrence.

occurrence (as) – series of non-delimiting characters whose boundaries at both ends are occurrence delimiters (syn.: token).

ordinary form (ced) – form which is not unusually frequent or unusually rare in a given corpus part (not belonging to the list of positive or negative characteristic elements).

original form – (for a given corpus part) form whose occurrences are found exclusively in the present part.

original vocabulary – (as) (for a given corpus part) set of original* forms* found in the part.

paradigm – (ling) a set of terms, showing themselves at one point of speech series.

paradigmatic – (as) concerning grouping of textual units in series, regardless of their succession order in writing.

part – (of a corpus of texts) fragment of text corresponding to natural divisions of this corpus or to a grouping of such divisions.

partition – (of a corpus of texts) division of a corpus into *parts* composed of *consecutive text fragments* that do not have a common intersection and whose union is equal to the corpus.

(of a set, of a sample) division of a set of individuals or observations into disjunct *groups* whose union is equal to the complete set.

percentages of variance – (ca or mca) quantities proportional to eigenvalues*, whose sum is equal to 100.

phrase (ling) – **(syn.: syntagm)** sequence of words forming a unit within a sentence.

polyform (rs) – archetype of the occurrences of a segment; series of forms not separated by a sequence delimiter, which is not necessarily found in a corpus.

positive characteristic element – (ced) for a given word *i* and part *j* , word *i* is said to be a positive characteristic element of part *j* (or a characteristic word* of that part) if its sub-frequency is "unusually high" within that part. More precisely, if the sum of the probabilities calculated from the hypogeometric model for the values equal or superior to the observed sub-frequency is smaller than the threshold fixed at the beginning.

principal axes methods (stat) – (or: eigen-analyses) family of multivariate statistical methods whose aim is to extract principal axes that approximately summarize the information contained in the initial data table.

profile – (stat and pa) (of a row or of a column of a contingency table) vector composed of the counts contained in row (or column) of a contingency table, divided by the sum of the counts for that row (or column).

punctuation – system of marks indicating the divisions of a text, syntactic relations and/or conditions of enunciation.

(as) character (or series of characters) corresponding to a punctuation mark.

relative frequency (as) – frequency of a textual unit in a corpus (or in one of its parts) divided by the size of the corpus (or of the part).

repartition (as) – (of occurrences of a given form in the parts of a corpus) number of corpus parts in which this form is found.

repeated segment (rs) – series of consecutive forms whose frequency is greater than or equal to 2 in the corpus.

section – (rs) portion of text between two section delimiters (ex: paragraph, etc.).

segment – (rs) within a corpus, any series of consecutive occurrences which is not separated by a sequence delimiter* is a text segment.

segmentation – set of operations that result in dividing a text into minimal* units.

segmentorial – (rs) set of terms* within a corpus.

sentence - (as) fragment of text between two sentence separators*.

sentence separators – (as) subset of sequence* delimiters* consisting, as a rule, of strong punctuation marks (period, question mark, exclamation mark).

sequence – (as) series of occurrences of a text not separated by a sequence delimiter*.

sequence delimiters – (as) subset of form* delimiters* consisting of weak and strong punctuation marks (which are generally: period, question mark, exclamation mark, comma, semi-colon, colon, quotation marks, dashes and brackets).

size – (as) length* (of a corpus) measured in occurrences (of simple forms), see: length*.

squared correlation (ca) – (syn.: relative contribution) parameter that show the importance of the different axes in explaining the variance of an element. For a given element, the sum of squared correlation over all axes is equal to 1.

sub-frequency (as) – (of a textual unit in a corpus part, sub-division, etc.) number of occurrences of a unit in a given corpus part (sub-division, etc.).

sub-segment (rs) – for a given segment, all the segments of an inferior length contained in this segment represent its sub-segments, ex: AB and BC are the sub-segments of the segment ABC.

supplementary (or illustrative) elements – (ca or mca) set of elements that are positioned on principal axes *a posteriori*, but do not participate in calculations of these principal axes. A supplementary element can be considered as an active element with a null weight.

syntagmatic – (as) concerning the arrangement of textual units according to their succession order in writing.

table of repeated segments (TRS) – contingency table cross-tabulating repeated segments and parts of a corpus. The rows of the TRS are sorted by lexicometric order* of the repeated segments (i.e. decreasing length, decreasing frequency, lexicographic order).

term – (rs) generic name used for forms* and polyforms*. A form is a term of the length of 1. Polyforms have the length of 2, 3, etc.

test-values – (ca or mca) quantities that make it possible to determine the significance of the position of a supplementary* (or illustrative) element on a principal axis. Briefly, a test value can be considered as a standardized normal variable under hypothesis of independence between the element and the principal axis.

threshold – (stat) the quantity arbitrarily fixed at the beginning of an experience in order to select among a large number of results, those for which the values of a numeric index exceed the threshold (of frequency, probability, etc.).

type T variables – variables whose frequency is approximately proportional to the rate at which a text lengthens (ex: maximum frequency).

type V variables – variables whose growth has a tendency to diminish as the text lengthens (ex: number of forms, number of hapax).

vocabulary (as) – set of distinct forms* found in a corpus.

word – throughout the present manual, synonym of form* or type.

Bibliography

Baayen H. (2001) - “Word Frequency Distributions”, *Series: Text, Speech and Language Technology*, Volume 18, Kluwer Academic Publishers, Dordrecht Hardbound.

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

Bécue 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.

Benzécri J.-P. & coll. (1973) – “La taxinomie”, Vol. I ; *L'analyse des correspondances*, Vol. II, 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.& coll. (1981a) - “Pratique de l'analyse des données”, tome 3, *Linguistique & Lexicologie*, Dunod , Paris.

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

Biber D., Conrad S., Reppen R. (1998) - *Corpus Linguistics : Investigating language structure and use*, Cambridge University Press.

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.

Bouillon P. (1998), - *Traitement automatique du langage naturel*, Editions Duculot.

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.

Crochemore M., Hancart C., Lecroq T. (2001) - *Algorithme du texte*, Vuibert.

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.

Desgraupes B. (2001) *Introduction aux expressions régulières* , Vuibert.

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., Fabre C., Issac F. (1998) - *De l'écrit au numérique (constituer, normaliser et exploiter les corpus électroniques)*, InterEditions.

Habert B., Salem A., Nazarenko A. (1997) - *Les linguistiques de corpus*, Armand Colin, Paris.

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.

Jurafsky D., Martin J. H. (2000) - "Speech and Language Processing : An Introduction to Natural Language Processing", *Computational Linguistics, and Speech Recognition*, Prentice-Hall.

Labbé D. (1983) - *François Mitterrand - Essai sur le discours*, La pensée sauvage, Grenoble.

Labbé D. (1990) - *Le vocabulaire de François Mitterrand*, Presses de la Fond. Nat. des Sciences Politiques, Paris.

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.

Manning C., Schütze H. (1999) - *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge.

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.

Pierrel J.-M.(2000) - *Ingénierie des langues*, Traité IC2 -Série informatique et SI, Hermes

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. (1980) - “D'ou viennent les fréquences de vocabulaire? ”, *Mots* N°1, p 189-212.

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.

Van Rijckevorsel J. (1987) - *The application of fuzzy coding and horseshoes in multiple correspondances analysis*, DSWO Press, Leyde.

Véronis J.(2000) - “Annotation automatique de corpus : panorama et état de la technique”, *Ingénierie des langues*. J. M. Pierrel. Paris, Hermès.

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.

Web sites

Links

- FRANTEXT : http://zeus.inalf.cnrs.fr
- LEXICOMETRICA : http://www.cavi.univ-paris3.fr/lexicometrica/
- MARGES-LINGUISTIQUES : http://www.marges-linguistiques.com/
- ATALA : http://www.atala.org/

Software

- HYPERBASE : http://lolita.unice.fr/pub/hyperbase/
- TROPES : http://www.acetic.fr/
- SPHINX : http://www.lesphinx-developpement.fr/
- SPAD-T : http://www.cisia.com/
- ALCESTE : http://www.image.cict.fr/
- TALTAC : http://www.taltac.it/