SYLED - CLA2T
Université de la Sorbonne nouvelle - Paris 3

Textometric toolbox
Cédric Lamalle
William Martinez
Serge Fleury
André Salem

User’s manual
Béatrice Fracchiolla
Andrea Kuncova
Bettina Lande
Aude Maisondieu
Maria Poirot Zimina
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
Results of computation of characteristic elements

### 3.3 Chronological characteristic elements
Characteristic increments

### 3.4 Correspondence analysis (CA)

### 4 Tools for lexicometric browsing

#### 4.1 Map of sections
Mapping of the sections for a Tgen
Statistical tools of the map of sections
Browsing with the help of the map of sections

#### 4.2 Towards better use of the windows
Create a worksheet
Move to another worksheet
Mosaic

#### 4.3 The report

#### 4.4. Options – Help - Tips
Options
Browsing tab
Full screen
Help
Exit

### 5 Glossary of terms used in textual statistics

#### Bibliography

#### Web sites
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.

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

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 (*\texttt{.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 (*\texttt{.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:

\[ \text{sign} = \text{status} \]

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).

\textit{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: \texttt{a,b,c,...} become \textit{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.

\textbf{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 \textit{Me} and \textit{me}; a segmentation which does not include the asterisk in the list of delimiting characters will produce separate counts for the two sequences.

\textbf{Sections of text}

Besides logical partitions, the text also contains marks for breathing (sentences, paragraphs, etc.). \textit{Lexico3} offers the possibility of promoting one or several delimiting characters to the rank of \textit{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 \textit{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.

§ 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 Lexico\textsuperscript{3} 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 icon](image)

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

![Lexico3 interface](image)

Click on the icon to open a text file

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

![Figure 2.1: Selecting a text file](image)

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

![Figure 2.2: Segmentation parameters dialog box](image)

**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:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclosed key</td>
<td>&lt;S01=Alice</td>
</tr>
<tr>
<td>Space in the type or contents of the key</td>
<td>&lt;S 01= Al ice&gt;</td>
</tr>
<tr>
<td>Closing tag missing</td>
<td>she is &lt; nice.</td>
</tr>
<tr>
<td>Absence of = sign</td>
<td>&lt;S01Alice&gt;</td>
</tr>
<tr>
<td>Key without contents</td>
<td>&lt;S01=&gt;</td>
</tr>
</tbody>
</table>
Figure 2.3: Wrong key error message

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

\begin{tabular}{l}
\textbf{Table 2.1:} Segmentation Report  \\
(Lxxx…. indicates the line at fault)  \\
\end{tabular}

\begin{quote}
egin{verbatim}
*****COMPTÉ-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 ":") : <EpgB>  
\end{verbatim}
\end{quote}

\textbf{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).

\textbf{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 \texttt{genericname.txt}, the files are called respectively \texttt{genericname.par}, \texttt{genericname.dic}, \texttt{genericname.num}.

The file \texttt{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)

<table>
<thead>
<tr>
<th>Lexico3.1 PC DUCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>nbetiq=0</td>
</tr>
<tr>
<td>196125 196125 11023 142185 10859 6130 4953 5000000 14 8 143 0 0</td>
</tr>
</tbody>
</table>

*** 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

<table>
<thead>
<tr>
<th>frq</th>
<th>rang lex.</th>
<th>forme</th>
</tr>
</thead>
<tbody>
<tr>
<td>6130</td>
<td>2703</td>
<td>de</td>
</tr>
<tr>
<td>4749</td>
<td>6033</td>
<td>les</td>
</tr>
<tr>
<td>4298</td>
<td>5909</td>
<td>la</td>
</tr>
<tr>
<td>3773</td>
<td>4216</td>
<td>et</td>
</tr>
<tr>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
</tr>
<tr>
<td>1</td>
<td>10967</td>
<td>voyager</td>
</tr>
<tr>
<td>1</td>
<td>10987</td>
<td>zété</td>
</tr>
</tbody>
</table>

Fin de la zone des formes graphiques

259
10859  !
198 10860 "
49 10861 $

Fin de la zone des ponctuations

766 10873 Epg
96 10874 501

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 Lexico\textsuperscript{3} 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.

\textbf{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).

---

**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 report.

**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. For example, in the *Duchesne* corpus are found the segments...

<table>
<thead>
<tr>
<th>Segment</th>
<th>length</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>tirer les marrons du feu</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

*(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).

![Figure 2.5: Delimiters and threshold for forms](image-url)
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 $T_{gen}$.

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)

![Figure 2.7: Creation of the form groups](image)

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.

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 [link](http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/)

For further information: [link](http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/)
<table>
<thead>
<tr>
<th>Operator</th>
<th>Function</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>. (dot)</td>
<td>Replaces any character</td>
<td>“tr.e” can mean tree, trie…</td>
</tr>
<tr>
<td>*</td>
<td>0 or n occurrences of the preceding character</td>
<td>“com*e” searches for coe, comme, commme,…</td>
</tr>
<tr>
<td>+</td>
<td>1 or n occurrences of the preceding character</td>
<td>“com+e” searches for comme, commme,…</td>
</tr>
<tr>
<td>&lt;</td>
<td>Represents beginning of word</td>
<td>“&lt;capital” searches for capital, capitale, capitalism…</td>
</tr>
<tr>
<td>&gt;</td>
<td>Represents end of word</td>
<td>“.*ism&gt;” searches for syndicalism, capitalism…</td>
</tr>
<tr>
<td>[]</td>
<td>Represents a set of characters</td>
<td>“[aeiou]” represents one of the characters in the set of vowels. “[a-z]” represents one of the characters between a and z.</td>
</tr>
<tr>
<td>[^]</td>
<td>Represents the negation of the content of the set of characters</td>
<td>“[^aeiou]” represents the characters that do not belong to the set of vowels</td>
</tr>
</tbody>
</table>
2.6 Word-store

The word-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](image)

**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)
**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.

<table>
<thead>
<tr>
<th>Text Part</th>
<th>Formes</th>
<th>Occurrences</th>
<th>Hapax</th>
<th>Fmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>4</td>
<td>50</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>Week 2</td>
<td>5</td>
<td>60</td>
<td>1</td>
<td>150</td>
</tr>
<tr>
<td>Week 3</td>
<td>6</td>
<td>70</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>Week 4</td>
<td>7</td>
<td>80</td>
<td>5</td>
<td>300</td>
</tr>
</tbody>
</table>

![Graph showing distribution of forms](image-url)
The table allows for a rapid visual comparison of the parts with regard to their most important
dexicometric 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. The Specif's 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.

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.”

---

Figure 3.3: Characteristics of the partition

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. The Specif's 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.

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.”

---

7 See (Lafon, 1984) or (L&S p.171) for the method of characteristic element computation.
8 To select a part, just click on the name of this part. You can add 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.
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

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). Use the parameters window (Figure 3.6) 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.

![Figure 3.6: CA parameters window](image)

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).
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.
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.

![Image of Report]

**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.).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>pages</th>
<th>occurrences</th>
<th>distinct forms</th>
<th>max. frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duchesne</td>
<td>350</td>
<td>142 177</td>
<td>10 988</td>
<td>6130 (de).</td>
</tr>
<tr>
<td>Coran (trad. Fr)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duchesne</td>
<td></td>
<td></td>
<td></td>
<td>(de).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(de).</td>
</tr>
</tbody>
</table>
**Browsing tab**

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

![Figure 4.6: Browsing](image)

**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 \( (n = \text{number of parts in the corpus}) \) 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 $\lambda_{\alpha}$ measures the variance of elements on axis $\alpha$.

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 $V_k$, 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 (Tgens) – 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.
**LEXICO3  User's manual**

**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:
or 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, maitre*.

_ 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).

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

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


Habert B., Fabre C., Issac F. (1998) - De l'écrit au numérique (constituer, normaliser et exploiter les corpus électroniques), InterEditions.
Labbé D. (1990) - Normes de dépouillement et procédures d'analyse des textes politiques, CERAT, Grenoble.


Web sites

Links
- 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/