

DtmVic and textual data

Four more examples to practise DtmVic with textual data

Unlike Tutorial A, the following examples use existing command files (or: parameter files). Each example corresponds to a directory included in the directory “DtmVic_Examples_B_Texts” that has been downloaded with DtmVic.

Application examples B.1—B.4

[To select the examples, press the right button of the mouse](#)

Example B.1. [EX_B01. Text-Responses_1](#)

(Open questions in a sample survey: First exploration)

Page **B.2**

First processing of the responses to an open-ended question. Numerical coding of the responses. Correspondence Analysis (CA) of the sparse lexical table words x respondents, clustering of the responses, and a description of the obtained clusters through their characteristic words and responses. Kohonen map for words and for respondents.

Example B.2. [EX_B02. Text-Responses_2](#)

(Open questions in a sample survey: link with closed-end questions)

Page **B.8**

This example, involving 14 steps, contains the example B.1 but takes into account the information about closed questions. Numerical coding of the responses. Examples of modification of the frequency threshold for words. Example of concordances (syntactic context) for some words. CA of the lexical table words x respondents, clustering of the responses, and a description of the obtained clusters through their characteristic words, responses, and also through their characteristic categories (closed questions). Kohonen maps for words and for respondents.

Example B.3. [EX_B03. Text-Responses_3](#)

(Open questions and MCA in a sample survey)

Page **B.16**

Multiple Correspondence Analysis and Clustering of respondents using closed questions. Processing aggregated [and lemmatised] responses to open questions. Example B.3 illustrates another technique for grouping and processing responses to open question in a sample survey. In a first phase, a multiple correspondence analysis is performed on a set of selected categorical variables (i.e.: responses to closed-end questions). The principal axes visualisation is complemented with a clustering, followed by an automatic description of the clusters. These clusters are then used to aggregate the responses to an open question. The survey, the closed-end questions and the textual responses are the same as those of previous examples A.5 and B.1, B2.

Example B.4. [EX_B04. Text-Semantic](#)

(Visualization of the Semantic network of French verbs)

Page **B.25**

Visualisation of the semantic links existing between 829 French verbs. Each verb is described by a list of “synonyms”. This example is in fact similar to example B.1 (Responses to an open question). The “respondents” are here the 829 verbs. The open-ended question is “Which are your synonyms?”, and the textual response is constituted by a list of synonyms.

Example B.1: **EX_B01. Text-Responses_1** (*Textual Data Analysis: A single open question*)

Example B.1 aims at describing the responses to an open-ended question in a sample survey. The principal axes visualization is complemented by a clustering, with an automatic description of the clusters. This is a typical first outlook on the set of responses: to detect and describe the main groupings of responses. Such outlook is by no means an achieved processing.

Example B.2, below, will provide another point of view, making use of other pieces of information about the respondents.

To have a look at the data, search for the directory **DtmVic_Examples**.

In this directory, open the sub-directory **DtmVic_Examples_B_Texts** .

In that directory, open the directory of Example B.1, named **“EX_B01. Text-Responses_1”**.

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 2 files :

- a) the text file,
- b) the command file.

(in this particular context, there are neither data file nor dictionary file: the questionnaire comprises three open-ended questions, without closed-end questions)

a) Text file: TDA_TEX.txt

This file has already served as an example for Example A.5 of Tutorial 1. It contains the free responses of 1043 individuals to three open-ended questions.

Firstly, the following open-ended question was asked: *"What is the single most important thing in life for you?"*

It was followed by the probe: *"What other things are very important to you?"*.

A third question (not analysed here) has also been asked: *"What means to you the culture of your own country"*

These questions were included in a multinational survey conducted in seven countries (Japan, France, Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al.*, 1992). It is the United Kingdom survey which is presented here.

The format is very specific. Since the responses may have very different lengths, separators are used to distinguish between questions and between individuals (or: respondents). Individuals are separated by the chain of characters "--" (starting column 1) possibly followed by an identifier, and questions are separated by "++++" (column 1). The symbol "====" indicates the end of the file. Like all the data files involved in DtmVic as input files, that file is a raw text file (.txt). If the text file comes from a text processing phase, it must be saved beforehand as a ".txt file".

b) Command file: “TDA1_par.txt”

As shown in Tutorial A, another “command file” similar to the “command file” **“TDA1_par.txt”** can be also generated by clicking on the button **“Create the command file”** of the main menu (Basic Steps). A window **“Choosing among some basic analysis”** appears. Click then on the button: **“VISURESP – Visualization of Responses”** – located in the paragraph **“Textual data”**, and follow the instructions as indicated in Tutorial A.

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about parameters"**) and, with more details, below (Appendix B.1).

Running the example B.1 and reading the results

- 1) Click on the button: **“Open an existing command file”** (panel *Basic Steps* of the main menu)
- 2) Then, search for the sub- directory **DtmVic_Examples_B_Texts** in: **DtmVic_Examples**.
- 3) In that directory, open the directory of Example B.04: **“EX_B01. Text-Responses_1”**.
- 4) Open the command file: **TDA1_par.txt**

After identifying the textual data file, seven "steps" are performed: **ARTEX** (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text), **ASPAR** (correspondence analysis of the [sparse] contingency table “respondents - words”), **CLAIR** (Brief description of factorial axes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **MOTEX** (crosstabulating the partition produced by step PARTI with words: the obtained contingency table is called a lexical table), **MOCAR** (characteristic words, and characteristics responses for each class of the partition).

We will comment later on this command file (Appendix of the section) which commands the basic computation steps. Instead of editing this file, we will content ourselves here in going back to the main menu and execute the basic computation steps.

- 5) Return to the main menu (**“return to execute”**)

- 6) Click on the button: **“Execute”**

This step will run the basic computation steps present in the command file: archiving text, correspondence analysis of the lexical table, brief description of the axes, clustering procedure, thorough descriptions of clusters using characteristic words and responses.

- 7) Click the button: **“Basic numerical results”**

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation): **“imp_08.07.09_14.45.html”** means July 8th, 2009, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory.

This file is also saved under a simple text format, under the name **“imp.txt”**, and likewise with a name including the date and time of execution.

From the step **NUMER**, we learn for instance that we have 1043 responses, with a total number of words (occurrences or token) of 13918, involving 1368 distinct words (or: types). Using a frequency threshold of 8, the total number of kept words reduces to 11559, whereas the number of distinct kept word reduces (more drastically) to 208. The book “Exploring textual data” (L. Lebart, A. Salem, E. Berry. Kluwer, 1998) deals in details with this pre-processing and with all the processing that follow.

- 8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: **“VIC”**)

- 9) Click the button: **“Axesview”**

and ... follow the sub-menus. In fact, only two tabs are relevant for this example: **“Active variables”** [= words in the case of step **ASPAR**], **“Individuals (observations) [= respondents]”**. After clicking on **“View”** in both cases, one obtains the set of principal coordinates along each axis.

Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step **“CLAIR”**. Evidently, the use of

the Axeview menu is justified when the data set is large, which is the case here.

10) Click the button: **“PlaneView”**

and follow the sub-menus...

In this example, four items of the menu are relevant **“Active columns (variables or categories)”**, **“Active rows (individuals, observations)”**, **“Active columns + Active rows”**, **“Active individuals (density)”**. The graphical display of chosen pairs of axes are then produced.

The roles of the different buttons are straightforward, except perhaps the button: **“Rank”**, which is useful only in the case of very intricate displays, (which is the case here). Since the set “individual” has 1043 elements, it is possible to test, with this example, partial printings of the individuals in two subsets of 50% or four subsets of 25%...(subsets randomly drawn without replacement)

11) About the button: **“BootstrapView”**.

The implementation of the bootstrap in the step ASPAR is not yet completed... .

12). Click on **“ClusterView ”**

12.1 Choose the axes (1 and 2 to begin with), and **“Continue”**.

12.2 Click on **“View”**. The centroids of the 7 clusters (Step PARTI) appears on the first principal plane.

12.3 Activate the button **“Words”**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR. But we do have in front of us the pattern of clusters and their relative locations.

12.4 Activate the button **“Texts”**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

13) Click on **“Kohonen map”**

Select the type of coordinate.

13.1 Select: **“Active variables (columns)”**: these active variables are the words in this example.

13.2 Select a (5 x 5) map, and continue.

13.3 After clicking on two small check-boxes, press **“Draw”** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size (**“Font”**) and dilate the obtained Kohonen map (**“Dilat.”**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **“AxeView”**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on **“Kohonen map”** and choose the item **“Active observations”**

13.7 Select a (10 x 10) map, and redo the operations 13.3 to 13.5 for the observations.

In the context of this example, the other items of the menu are not relevant.

Appendix B1: (for advanced users)

The command file can be generated using the menu “Create_parameters”. Therefore, freshman practitioners could skip this appendix.

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "**Help about parameters**").

Let us remind that this set of commands comprises seven steps:

ARDAT (Archiving data), **SELEC** (selecting active and supplementary elements), **PRICO** (Principal components analysis) **DEFAC** (Description of factorial axes) , **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **DECLA** (Automatic description of the classes of the partition).

Now, we will exhibit the command file that contains comments (preceded by #). As seen previously, comments are also allowed in the (mandatory) line that immediately follows a statement "STEP xxxxx"

Command file “TDA1_par.txt”

```
# The Program DtmVic needs 2 files in this "open survey case"
# -----
# 1) The present file of commands, whatever its name.
# 2) The text file (NTEXZ).
#   Syntax: ">"= continuation, "#"= comments
#-----
LISTP = yes, LISTF = no # Global parameters(leave as it is)
#
NTEXZ = 'TDA_tex.txt' # name of text file (free name)
#
STEP ARTEX
==== Archive - Texts or responses to open ended questions
ITYP=2 NBQT=3 NCOL=80
#----- Comments about step ARTEX
# - ITYP: type of textual data file NTEXZ
#   ITYP = 2 ==> type of file = responses to open questions
# - NBQT: number of questions per respondent
#   NBQT = 3 ==> there are 3 open questions
# - NCOL: length (number of columns) of the records (80)
#-----
#
STEP SELOX
==== Selection of open questions (and of individuals)
NUMQ = LIST
1 2
#----- Comments about step SELOX
# - NUMQ: index of the selected question
#   if NUMQ = -1 or NUMQ = LIST : several questions
#   will be merged (the list of question numbers
#   must follow immediately next line)
#   here: questions 1 and 2 are selected
#-----
```

```

STEP NUMER
==== Numerical coding of words
NSEU = 8   LEDIT=TOT
weak ' " -
strong . ; : ( ) ! ? ,
end
#----- Comments about step NUMER
# - NSEU:   frequency threshold of the kept words
#           (here, only the frequencies > 8 will be kept)
# - LEDIT:  printing the words (0=no; 1=alphabetical order;
#           2=frequency order; 3= both 1 and 2).
# --- key-words:
# - weak    (weak separators) followed by those separators
#           [separators of words]
# - strong  (strong separators) followed by those separators
#           [separators of segments, for step SEGME]
# - end     ... indicates the end of key-words statements.
#-----

STEP ASPAR
==== Correspondence analysis of the table: Words X Responses
NAXE=8 LEDIT=0 NGRAF=5 NROWS=60 NPAGE=1 NBASE=12 NITER=20
#----- Comments about step ASPAR
# - NAXE:   number of requested principal coordinates
# - LEDIT:  printing the responses
#           (0 = no; 1 = coordinates of variables;
#           2 = 1 + coordinates of respondents)
# - NGRAF:  number of requested printer graphics
#           in the results file "imp.txt"
#           NGRAF = 5 means that we will get the printouts of
#           the planes spanned by the following pairs of axes:
#           (1, 2), (2, 3), (3, 4), (4, 5), (4, 6).
# - NPAGE:  number of pages of these graphics
# - NROWS:  number of lines of these graphics
#           The two following parameters concern an option
#           for diagonalizing very large matrices: (if NBASE > 0)
# - NBASE:  dimension of the approximation space
#           (NBASE = 0: main core diagonalization)
# - NITER:  number of iterations (if NBASE > 0)
#-----

STEP CLAIR
==== Brief description of NAXE principal axes
NAXE=6 LIGN=no NMAX=40
#----- Comments about step CLAIR
# - NAXE = ... number of axes to be described
# - LIGN = no means that lines (or rows, or individuals
#           or respondents are excluded)
# - NMAX = ... Maximum number of elements that will
#           be sorted to describe each axis
#-----

STEP RECIP
==== Clustering of respondents using reciprocal neighbours
NAXU=7 LDEND=DENSE NTERM=20 LDESC=no
#----- Comments about step RECIP
# This step carries out a hierarchical clustering
# using the reciprocal neighbours technique (recommended
# when dealing with less than 1000 individuals.
# - NAXU... number of axes kept from the
#           previous MCA .

```

```

# - LDEND...   printing dendrogram (0=no, 1=dense,
#             2=large).
# - NTERM...   number of kept terminal elements
#             NTERM = TOT means that all the elements are kept.
# - LDESC...   describing nodes of the tree (0=no, 1=yes).
#-----

STEP PARTI
==== Cut of the dendrogram to obtain 9 clusters
NITER=7 LEDIN=3
9      # number of classes of the partition
#----- Comments about step PARTI
# - NITER... number of "consolidation" iterations (0=no).
# - LEDIN... printing the correspondences classes-
# individuals (3 = printing of the correspondence
# classes->individuals and the correspondence
# individuals-->classes).
# The line immediately following the command must
# contain the sizes of the desired final partition
# (here: 9).
#-----

STEP MOTEX
==== cross-tabulating words and clusters
NVSEL=-1  LEDIT = 0
#----- Comments about step MOTEX
# NVSEL:   index of the categorical variable defining
#           the groupings of texts
#           the conventional value NVSEL = -1 means that
#           the categorical variable coincides with the
#           previously computed partition.
# LEDIT:   parameter for printing the table words*texts
#           (0=no, 1=yes).
#-----

STEP MOCAR
==== Characteristics words for each cluster
NOMOT=10  NOREP=6
#----- Comments about step MOCAR
# NOMOT:   number of requested characteristic words for
#           each text (i.e: for each cluster)
# NOREP:   number of characteristic responses for each text.
#-----

STOP
#-----

```

End of example B.1

Example B.2: EX_B02. Text-Responses_2 *(Open questions in a sample survey: link with closed-end questions)*

Example B.2 is a variant of Example B.1 aiming at describing the responses to an open ended question in a sample survey in relation with a set of categorical variables. The survey and the textual responses are the same as in example B.1, the closed-end questions are those of example A.5 (Tutorial A).

More explanation about this type of example and the corresponding methodology can be found in the book: "Exploring Textual data" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).

The sequence of steps is enriched by the following computations :

The numerical coding (step NUMER) is performed with a frequency threshold of 0 : all the words (types) are kept. This allow us to look at the complete distribution of words, and to carry out the new step CORDA, giving the concordance (contexts in the response) of two selected words : *life* and *money*.

Note that such "concordances" can be also generated by clicking on the button "Create the command file" of the main menu (Basic Steps). A window "Choosing among some basic analysis" appears. Click on the right hand side button: "Other analysis", and, then, click on the button: "CORDA", and follow the instructions.

To perform a grouping of responses upon a more significant statistical basis, the new step SETEX selects the words according to a larger threshold (8).

We can now take advantage of the presence of closed-end questions to describe the clusters, not only with characteristic words and responses (as done previously in Example B.1), but also with categories, selected after a step SELEX, and analysed through the step DECLA .

Another new step, POSIT, describes the location of these supplementary categories in the plane spanned by the first principal axes.

To have a look at the data, search for the directory **DtmVic_Examples**.

In this directory, open the sub-directory **DtmVic_Examples_B_Texts**.

In that directory, open the directory of Example B.2, named **"EX_B01. Text-Responses_2"**.

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application.

At the outset, such directory must contain 4 files :

- a) the data file,
- b) the dictionary file,
- c) the text file,
- d) the command file.

a) Data file: TDA_dat.txt (same as that of Example A.5)

This file contains responses to questions which were included in the multinational survey conducted in seven countries (Japan, France, Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al.*, 1992). It is the United Kingdom survey which is presented here.

It deals with the responses of 1043 individuals to 14 questions. Some questions concern objective characteristics of the respondent or his/her household (age, status, gender, facilities). Other questions relate to attitude or opinions. The data file " TDA_dat.txt" comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

b) Dictionary file: TDA_dic.txt (same as that of Example B.2)

The dictionary file "TDA_dic.txt" contains the identifiers of these 14 variables. In this version of DtmVic, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" should be used to facilitate this kind of format].

c) Text file: TDA_TEX.txt (same as that of examples A.5, and B.1)

This file is identical to that of the previous example B.1. Let us remind its characteristics. It contains the free responses of 1043 individuals to three open-ended questions.

Firstly, the following open-ended question was asked: "What is the single most important thing in life for you?". It was followed by the probe: "What other things are very important to you?".

A third question (not analysed here) has also been asked: "What means to you the culture of your own country". We refer to the previous example for comments about the data format.

d) Command file: TDA2_par.txt

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "Help about parameters") and, with more details, below (Appendix B.2).

Note that another "command file" similar to the "command file "TDA2_par.txt" can be also generated by clicking on the button "Create the command file" of the main menu (Basic Steps). A window "Choosing among some basic analysis" appears. Click then on the button: "VISURECA – Visualization of Responses..." – located in the paragraph "textual and numerical data", and follow the instructions.

Running the example B.2 and reading the results

- 1) Click on the button: "Open an existing command file" (panel *Basic Steps* of the main menu)
- 2) Then, search for the sub- directory **DtmVic_Examples_B_Texts** in **DtmVic_Examples**.
- 3) In that directory, open the directory of Example B.02, named "EX_B02.Text-Responses_2".
- 4) "Open the existing command file": "TDA2_par.txt"

After identifying the textual data file, 14 "steps" are performed: **ARDAT** (archiving data), **ARTEX** (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text: now, all the words are kept), **CORDA** (concordance for some selected words), **SETEX** (introducing a new threshold for the frequencies of words), **ASPAR** (correspondence analysis of the [sparse] contingency table "respondents - words"), **CLAIR** (Brief description of factorial axes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **MOTEX** (crosstabulating the partition produced by step **PARTI** with words: the obtained contingency table is called a lexical table), **MOCAR** (characteristic words, and characteristics responses for each class of the partition), **SELEC** (selecting active and supplementary elements), **DECLA** (systematic description of the classes of the partition produced by step **PARTI** using the other relevant categorical variables), **POSIT** (illustrating the principal spaces of responses with supplementary categorical variables).

We will comment later on this command file (Appendix of the section) which commands the basic computation steps. Instead of editing this file, we will content ourselves here in going back to the main menu and execute the basic computation steps.

- 5) Return to the main menu ("return to execute")

6) Click on the button: “Execute”

This step will run the basic computation steps present in the command file.

7) Click the button: “Basic numerical results”

The button opens a created (and saved) html file named “**imp.html**” which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name “**imp.html**” is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file “**imp.html**” is replaced for each new analysis performed in the same directory.

This file is also saved under a simple text format, under the name “**imp.txt**”, and likewise with a name including the date and time of execution.

From the step **NUMER**, with the new threshold of “0”, we check for instance that we still have 1043 responses, with a total number of words (occurrences or token) of 13 918, involving 1 368 distinct words (or: types).

Perusing the complete list of words shows some errors (inherent in real sized applications): for instance, the symbol “]” was absent from the list of separators, and creates some new “words”...

In this version of DtmVic, the results of the new steps **CORDA** and **POSIT** are confined to this “result file” (imp.txt).

8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: “VIC”)

9) Click the button “Axesview”

... and follow the sub-menus. In fact, only two tabs are relevant for this example: “**Active variables**” [= words in the case of step **ASPAR**], “**Individuals (observations)** [= respondents]”. After clicking on “**View**” in both cases, one obtains the set of principal coordinates along each axis.

Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step “**CLAIR**”.

10) Click the button: **PlaneView**, and follow the sub-menus...

In this example, six items of the menu are relevant “**Active columns (variables or categories)**” (principal coordinates of the active words), “**Supplementary categories**” (coordinates of the supplementary categories derived from the step “**POSIT**”), “**Active rows (individuals, observations)**” (coordinates of the respondents), “**Active columns + Active rows**”, “**Active individuals (density)**” and “**Active columns + Supplementary categories**”. The graphical display of chosen pairs of axes are then produced.

11) Click the button: “BootstrapView”...

The implementation of the bootstrap in the step **ASPAR** is not yet completed... .

12). Click on “ **ClusterView** ”

12.1 Choose the axes (1 and 2 to begin with), and “Continue”.

12.2 Click on “**View**”. The centroids of the 9 clusters (produced by the Step **PARTI**) appears on the first principal plane.

12.3 Activate the button “**Words**”, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step **MOCAR**. But we do have in front of us the pattern of clusters and their relative locations.

12.4 Activate the button **“Texts”**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

12.5 Activate the button **“Categorical”**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic categories of the selected category. This description is somewhat redundant with that provided in the results file (file “imp.txt” by the step DECLA. But we do have simultaneously in front of us the pattern of categories and their relative locations.

13) Click on **“Kohonen map”**

Select the type of coordinate.

13.1 Select: **“Active variables (columns)”**: these active variables are the words in this example.

13.2 Select a (5 x 5) map, and continue.

13.3 After clicking on two small check-boxes, press **“Draw”** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size (**“Font”**) and dilate the obtained Kohonen map (**“Dilat.”**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **“AxeView”**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on **“Kohonen map”** and choose the item **“Observations”**

13.7 Select a (10 x 10) map, and redo the operations 13.3 to 13.5 for the observations.

Appendix B2: (for advanced users)

A similar (but less sophisticated) command file can be generated using the menu **“Create_parameters”. Therefore, beginners could skip this appendix.**

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "Help about parameters").

Let us remind that this set of commands comprises 14 steps:

ARDAT (archiving data), **ARTEX** (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text: now, all the words are kept), **CORDA** (concordance for some selected words), **SETEX** (introducing a new threshold for the frequencies of words), **ASPAR** (correspondence analysis of the [sparse] contingency table “respondents - words”), **CLAIR** (Brief description of factorial axes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **MOTEX** (crosstabulating the partition produced by step **PARTI** with words: the obtained contingency table is called a lexical table), **MOCAR** (characteristic words, and characteristics responses for each class of the partition), **SELEC** (selecting active and supplementary elements), **DECLA** (systematic description of the classes of the partition produced by step **PARTI** using the other relevant categorical variables), **POSIT** (illustrating the principal spaces of responses with supplementary categorical variables).

Now, we will exhibit the command file that contains **comments** (preceded by #). As seen previously, comments could also be printed in the (mandatory) line that immediately follows a statement "STEP xxxxx" .

Command file: "TDA2_par.txt"

```
# -----Example EX_B02 : Textual Data Analysis 2 ----
# The Program DtmVic needs 4 files in this "open survey case"
# -----
# 1) The present file of commands, whatever its name.
# 2) The text file (NTEXZ).
# 3) The dictionary file (NDICZ).
# 4) The data file (NDONZ).
# Syntax: ">"= continuation, "#"= comments
# -----
LISTP = yes, LISTF = no # leave as it is...

NTEXZ = 'TDA_tex.txt'      # text file (same as in example TDA1)
NDICZ = 'TDA_dic.txt'     # dictionary file
NDONZ = 'TDA_dat.txt'     # data file

STEP ARDAT # Archiving data and dictionary
=====
  NQEXA =14 , NIDI = 1,  NIEXA =1043
#----- Comments about step ARDAT
# - NQEXA = ... number of questions (or variables)
# in both the dictionary and the data file
# - NIDI = 1...indicate the presence of an
# identifier
# - NIEXA = ... number of "individuals" (or rows)
# in the data file.
#-----

STEP ARTEX # Archiving responses to 3 open questions
=====
ityp = 2 nbqt = 3 nlig=5

# See Appendix B1 for the comments about this step
#-----

STEP SELOX # Selecting responses to questions 1 and 2
===== # See Appendix B1 for the comments about this step
NUMQ=LIST LDONA=1
1,2

STEP NUMER # extracting words : threshold= 0
===== # See Appendix B1 for the comments about this step
NSEU = 0, LEDIT = TOT NXMAX = 20000 coef = 10
weak -
strong . ? ; ( ) : , '
end

STEP CORDA # concordances
=====
LEDIT = 1
FORME life money
END
```

```

#----- Comments about step CORDA
#LEDIT:  printing identifiers of individuals
#          (0 = no printing, 1 = identifiers of
#          respondents are printed, default = 0)
# --- key-word of headings :
# FORME must be followed by the selected words
# END    end of the selection.
#-----

#---- selecting a new threshold for words -

NSPB ='NSPB'
# the file NSPB created by SETEX is given the name: 'NSPB'

STEP SETEX
=====
NSEU =8 NMOMI=0 NREMI=2 LEDIT =NEW

#----- Comments about step SETEX
# NSEU:  threshold of frequency for selecting words.
# NMOMI:  minimum number of letters of a kept word.
# NREMI:  minimum number of words of a kept response.
# LEDIT:  printing the dictionaries (0=no, 1=new, 2=tot).
#-----
NSPA = 'NSPB'
#---- the file 'NSPB' created by SETEX is substituted to
# the file NSPA that was created by NUMER.

#----- Comments about step ASPAR
# - NAXE:  number of requested principal coordinates
# - LEDIT:  printing the responses
#          (0 = no; 1 = coordinates of variables;
#          2 = 1 + coordinates of respondents)
# - NGRAF:  number of requested printer graphics
#          in the results file "imp.txt"
#          NGRAF = 5 means that we will get the printouts of
#          the planes spanned by the following pairs of axes:
#          (1, 2), (2, 3), (3, 4), (4, 5), (4, 6).
# - NPAGE:  number of pages of these graphics
# - NROWS:  number of lines of these graphics
# The two following parameters concern an option
# for diagonalizing very large matrices: (if NBASE > 0)
# - NBASE:  dimension of the approximation space
#          (NBASE = 0: main core diagonalization)
# - NITER:  number of iterations (if NBASE > 0)
#-----

STEP CLAIR
==== Brief description of NAXE principal axes
NAXE=6 LIGN=no NMAX=40

#----- Comments about step CLAIR
# - NAXE = ... number of axes to be described
# - LIGN = no means that lines (or rows, or individuals
#          or respondents are excluded)
# - NMAX = ... Maximum number of elements that will
#          be sorted to describe each axis
#-----

STEP RECIP
==== Clustering of respondents using reciprocal neighbours

```

```

NAXU=7 LDEND=DENSE NTERM=20 LDESC=no

# See Appendix B1 for the comments about this step
#-----

STEP PARTI
==== Cut of the dendrogram to obtain 9 clusters
NITER=7 LEDIN=3
9      # number of classes of the partition

# See Appendix B1 for the comments about this step
#-----

STEP MOTEX
==== cross-tabulating words and clusters
NVSEL=-1  LEDIT = 0

# See Appendix B1 for the comments about this step
#-----

STEP MOCAR
==== Characteristics words for each cluster
NOMOT=10  NOREP=6

# See Appendix B1 for the comments about this step
#-----

STEP SELEC          # Selection of nominal variables
===== Selects active, supplementary variables and observations
LSELI = TOT,  IMASS = UNIF, LZERO = REC, LEDIT = short
NOMI ILL 1  2  4--14
end

#----- Comments about step SELEC
# - LSELI = ... Parameter describing the selection
# of individuals, the value TOT means that all
# the rows are selected.
# - IMASS = ... weight (or: mass) of the individuals
#(rows); the value 0 (or: UNIF) means "uniform"
#(same weights)
# - LZERO = REC ... means that the value 0, which
# indicates a missing value, will be coded as an
# extra response item for the categorical variables.
# - NOMI ILL means illustrative nominal (or:
# categorical) variable. This key-word is followed
# by the list of the numbers of the variables.
# The key-word END indicates the end of the list.
#-----

STEP DECLA          # Description of partitions
===== Systematic description of clusters
CMODA = 5.0, PCMIN = 2.0, LSUPR = yes, CCONT = 5.0 >
LPNOM = NO, EDNOM = NO, EDCON = NO
9 # list of partitions (characterised by their numbers of clusters)

#----- Comments about step DECLA
# - CMODA... describing classes with categories
# (0=no; CMODA = 5.0 means a p-value <= 0.05 for
# the selection of characteristic categories).

```

```

# - PCMIN... minimum relative ( % ) weight for a
# category (categories whose relative weight is
# <=2% are discarded).
# - LSUPR... characteristic category if
#   %(cat./class) > %(cat./total) (0=no,1=yes).
# (LSUPR=yes means that only characteristic
# elements will be printed)
# - CCONT... describing classes with numerical
# variables. (0=no; CCONT = 5.0 means a
# p-value <= 0.05 for the selection of
# characteristic variables).
# - LPNOM... describing partition with questions
# (i.e: whole set of categories) (0=no, 1=yes).
# - EDNOM... printing the tables crosstabulating
# (classes * questions) (0=no).
# - EDCON... describing partition globally with
# numerical variables (0=no, 1=yes)
#-----

STEP POSIT
===== positions of cat. variables in principal space
naxe = 3 ngraf = 2

#----- Comments about step POSIT
# NAXE ... number of axes to be used
# NGRAF ... number of graphical displays (sketches of
#           printer displays appearing in file "imp.txt")
#
# The results of this step are printed in the file "imp.txt"
# and lead to the file "ngus_cat_sup.txt" DtmVic. The visualization is
# done via the item "Supplementary categorical variables" of
# the menu "PlaneView".
# The step posit is useful to interpret the groupings of
# responses, and to learn which categorical variable is
# linked to the responses. In fact, the "test values" produced
# by the step POSIT are the most commonly used outputs.
#-----

STOP
#-----

```

End of example B.2

Example B.3: EX_B03. Text-Responses_3 *(Open questions in a sample survey: link with closed-end questions)*

Example B.3 illustrates another technique for grouping and processing responses to open question in a sample survey. In a first phase, a multiple correspondence analysis is performed on a set of selected categorical variables (i.e: responses to closed-end questions). The principal axes visualisation is complemented with a clustering, followed by an automatic description of the clusters. These clusters are then used to aggregate the responses to an open question. The survey, the closed-end questions and the textual responses are the same as those of previous examples.

More explanation about this type of example and the corresponding methodology can be found in the book: "Exploring Textual data" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).

The sequence of steps is enriched by the following computations:

As in Example B.2, the numerical coding (step **NUMER**) is performed with a frequency threshold of 0 : all the words (types) are kept. We can then carry out the new step **CORTE**, allowing us to perform a "primary lemmatization" of the text. (see also the step **CORTEX** of the menu invoked by the button "Create a new command file", and the complementary command files whose names ends by "TEX").

We can now take advantage of the presence of both open ended and closed-end questions to describe the clusters, not only with characteristic words and responses (as done previously in Example B.4), but also with categories. Another new step, **POLEX**, describes the location of the words in the plane spanned by the first principal axes.

To have a look at the data, search for the directory **DtmVic_Examples**.

In this directory, open the sub-directory **DtmVic_Examples_B_Texts**.

In that directory, open the directory of Example B.3, named **"EX_B03. Text-Responses_3"**.

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 2 files :

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application.

At the outset, such directory must contain 4 files :

- a) the data file,
- b) the dictionary file,
- c) the text file,
- d) the command file.

a) Data file: TDA_dat.txt (same as that of Example B.2)

This file contains responses to questions which were included in the multinational survey [see also Example A.5] conducted in seven countries (Japan, France, Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al.*, 1992). It is the United Kingdom survey which is presented here.

It deals with the responses of 1043 individuals to 14 questions. Some questions concern objective characteristics of the respondent or his/her household (age, status, gender, facilities). Other questions relate to attitude or opinions.

The data file **"TDA_dat.txt"** comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

b) Dictionary file: TDA_dic.txt (same as that of Example B.2)

The dictionary file "**TDA_dic.txt**" contains the identifiers of these 14 variables. In this version of DtmVic, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" should be used to facilitate this kind of format].

c) Text file: TDA_TEX.txt (same as that of examples A.5, B.1 and B.2)

We refer to previous example for comments about the questionnaire and the data format.

d) Command file: TDA3_par.txt

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "**Help about parameters**") and, with more details, below.

Note that another "command file" similar to the "command file "**TDA3_par.txt**" can be also generated by clicking on the button "**Create the command file**" of the main menu (Basic Steps). A window "**Choosing among some basic analysis**" appears. Click then on the button: : "**MCA_Texts – Visualization of Responses...**" – located in the paragraph "**textual and numerical data**", and follow the instructions.

Running the example B.3 and reading the results

- 1) Click on the button: "**Open an existing command file**" (panel *Basic Steps* of the main menu)
- 2) Then, search for the sub- directory **DtmVic_Examples_B_Texts** in **DtmVic_Examples**.
- 3) In that directory, open the directory of Example B.03, named "**EX_B03. Text-Responses_3**".
- 4) Open the existing command file: **TDA3_par.txt**

After identifying the textual data file, 16 "steps" are performed: **ARDAT** (archiving data), **ARTEX** (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text: now, all the words are kept), **CORTE** (deleting some function words [or empty words], declaring as equivalent flections of a same lemma), **SETEX** (introducing a new threshold for the frequencies of words), **SELEC** (correspondence analysis of the [sparse] contingency table "respondents - words"), **MULTM** (Multiple correspondence analysis), **DEFAC** (Brief description of factorial axes), **POLEX** (projecting the words of the responses as supplementary elements in the principal planes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **DECLA** (systematic description of the classes of the partition produced by step **PARTI** using the other relevant categorical variables), **MOTEX** (crosstabulating the partition produced by step **PARTI** with words: the obtained contingency table is a "lexical table"), **MOCAR** (characteristic words, and characteristic responses for each class of the partition), **RECAR** (characteristic responses for each class of the partition using a different criterion of selection, allowing for lengthy responses).

We will comment later on this command file (Appendix B.3 of the section) which commands the basic computation steps. Instead of editing this file, we will content ourselves here in going back to the main menu and execute the basic computation steps.

5) Return to the main menu ("return to execute")

6) Click on the button: "Execute"

This step will run the basic computation steps present in the command file.

7) Click the button: “Basic numerical results”

The button opens a created (and saved) html file named “**imp.html**” which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name “**imp.html**” is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file “**imp.html**” is replaced for each new analysis performed in the same directory.

This file is also saved under a simple text format, under the name “**imp.txt**”, and likewise with a name including the date and time of execution.

From the step **NUMER**, with the new threshold of “0”, we check for instance that we still have 1043 responses, with a total number of words (occurrences or token) of 13 918, involving 1 368 distinct words (or: types). In this version of DtmVic, the results of the new step CORTE are confined to this “result file” (imp.txt).

8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: “VIC”)

9) Click the button “Axesview”

and ... follow the sub-menus. Here, four tabs are relevant for this example: “**Active variables**” [= categories in this MCA case], “**Supplementary categories**”, “**Individuals (observations) [= respondents]**”, and “**supplementary lexical units**” (provided by step **POLEX**). After clicking on “**View**” in both cases, one obtains the set of principal coordinates along each axis.

Clicking on a column header produce a ranking of all the rows according to the values of that column.

10) Click the button “PlaneView”

and follow the sub-menus...

In this example, seven items of the menu are relevant “**Active columns (variables or categories)**”, (Active categories of the Multiple Correspondence Analysis), “**Supplementary categories**”, (Supplementary categories of the same MCA), “**Active rows (individuals, observations)**”, “**Active columns + Active rows**”, “**Supplementary lexical units**” (projection of the words used by the respondent in their responses to the open question), provided by step **POLEX**, “**Active individuals (density)**” and “**Active columns + Supplementary categories**”. The graphical displays of the chosen pairs of axes are then produced.

11) Click the button “BootstrapView”...

This button opens the DtmVic-Bootstrap-Stability windows.

11 .1 Click “**LoadData**”. In this case (partial bootstrap), the two replicated coordinates file to be opened are named “ngus_var_boot.txt” and “ngus_sup_cat_boot.txt” (see the small panel reminding the names of the relevant files below the menu bar).

In fact, **ngus_var_boot.txt** contains both active and supplementary categories. The file **ngus_sup_cat_boot.txt** contains only supplementary categories, for which the bootstrap procedure is all the more meaningful.

11 .2 Click on “**Confidence Ellipses**”, submenu, and choose the pair of axes to be displayed (choose axes 1 and 2 to begin with).

11 .3 Click on “**Loading**” in the blue window that appears then, to obtain the dictionaries of variables. Tick the chosen white cases to select the elements the location of which should be assessed, and press the button “**Select**”. Select, for instance, the supplementary elements “male, female, less than 30 years old with high level of education, over 55 with high, and also with low, level of education.

11 .4 Click on “**Confidence Ellipses**” to obtain the graphical display of the active category points (in blue colour), and of the supplementary category points (in red).

In this display, we learn for example that in this principal space (built as a “space of opinions”, due to the selection of active questions), male and female do not occupy distinct locations (ellipses almost no

overlapping). As shown by the locations of other categories, age and education lead to distinct patterns of opinions.

11 .5 Close the display window, and, again in the blue window, press **“Convex hulls”**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary. Go back to the main menu.

12). Click on: **“ClusterView ”**

12.1 Choose the axes (1 and 2 to begin with), and **“Continue”**.

12.2 Click on **“View”**. The centroids of the 7 clusters (produced by Step PARTI) appears on the first principal plane.

12.3 Activate the button **“Categorical”**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic categories of the selected category. This description is somewhat redundant with that provided in the results file (file “imp.txt) by the step DECLA. But we do have simultaneously in front of us the pattern of categories and their relative locations.

12.4 Activate the button **“Words”**, and , pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR. But, again, we do have in front of us the pattern of clusters and their relative locations.

12.5 Activate the button **“Texts”**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

13) Click on **“Kohonen map”**

Select the type of coordinate.

13.1 Select: **“Active variables (columns)”**: these active variables are the words in this example.

13.2 Select a (4 x 4) map, and continue.

13.3 After clicking on two small check-boxes, press **“Draw”** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size (**“Font”**) and dilate the obtained Kohonen map (**“Dilat.”**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **“AxeView”**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis: large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on **“Kohonen map”** and choose the item **“Active observations”**

13.7 Select a (12 x 12) map, and redo the previous operations for the observations.

Appendix B.3 (for advanced users)

A similar (but not identical) command file can be generated using the menu “Create_parameters”. Therefore, beginners could skip this appendix

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "Help about parameters").

Let us remind that this set of commands comprises 16 steps:

ARDAT (archiving data), **ARTEX** (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text: now, all the words are kept), **CORTE** (deleting some function words [or empty words], declaring as equivalent flections of a same lemma), **SETEX** (introducing a new threshold for the frequencies of words), **SELEC** (selecting active and supplementary categorical variables for the forthcoming MCA), **MULTM** (Multiple correspondence analysis), **DEFAC** (Brief description of factorial axes), **POLEX** (projecting the words of the responses as supplementary elements in the principal planes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **DECLA** (systematic description of the classes of the partition produced by step **PARTI** using the other relevant categorical variables), **MOTEX** (cross-tabulating the partition produced by step **PARTI** with words: the obtained contingency table is called a lexical table), **MOCAR** (characteristic words, and characteristics responses for each class of the partition), **RECAR** (characteristics responses for each class of the partition using a different criterion of selection, allowing for lengthy responses).

Command file TDA3_par.txt

Now, we will exhibit the command file that contains **comments** (preceded by #).

```
# ----- TDA3_par.txt : Textual Data Analysis ----
# The Program DtmVic needs 4 files in this "open survey case"
# -----
# 1) The present file of commands, whatever its name.
# 2) The text file (NTEXZ).
# 3) The dictionary file (NDICZ).
# 4) The data file (NDONZ).
#     Syntax: ">"= continuation, "#"= comments
# -----
LISTP = yes, LISTF = no # leave as it is...

NTEXZ = 'TDA_tex.txt'      # text file (same as in example TDA1)
NDICZ = 'TDA_dic.txt'    # dictionary file
NDONZ = 'TDA_dat.txt'    # data file

STEP ARDAT # Archiving data and dictionary
=====
NQEXA =14 , NIDI = 1,  NIEXA =1043

# See Appendix B2 for the comments about this step
#-----

STEP ARTEX # Archiving responses to 3 open questions
=====
ityp = 2 nbqt = 3 nlig=5

# See Appendix B1 for the comments about this step
#-----

STEP SELOX # Selecting responses to questions 1 and 2
=====
NUMQ=LIST          LDONA=1
1,2

# See Appendix B1 for the comments about this step
#-----
```

```

STEP NUMER   # extracting words : threshold= 0
=====
NSEU = 4, LEDIT = TOT NXMAX = 20000 coef = 10
weak -
strong . ? ; ( ) : , '
end

# See Appendix B1 for the comments about this step
#-----

#----- example of pre-processing texts -----
NSPC = 'NSPC'
# the file NSPC created by CORTE is given the name: 'NSPC'

step CORTE
===== deletion and equivalence between words
LEDIT = 2
delet      a an and at but by etc for from if in into of on or >
           out over pp than the to up
equiv      two 2
equiv      be am m are re is been being was
equiv      child children
equiv      content contented
equiv      can could
equiv      would d
equiv      do doing don
equiv      enjoy enjoying
equiv      family families
equiv      get got getting
equiv      go going
equiv      have having ve
equiv      help helping
equiv      holiday holidays
equiv      job jobs
equiv      keep keeping
equiv      live living
equiv      look looking
equiv      see seeing
equiv      son sons
equiv      sport sports
equiv      thing things
equiv      work working
equiv      worry worries
end

#----- Comments about step CORTE
# step CORTE (correction of texts) helps us to perform
# what we may term a manual lemmatisation.
# In fact, the frequency threshold NSEU should be "0"
# I the preceding step NUMER..
# The deletions concerns mainly function words (or tool
# words, or auxiliary words, or grammatical words...).
# Many equivalences are found simply by looking at the
# alphabetical list of words provided by step NUMER.
# ledit:  printing of words  (0=no, 1=nspc, 2=tot).
# lclas:  printing sorted words (0=no, 1=yes).
#

```

```

# CORTE uses 3 key-words whose meanings are straightforward:
# delet, equiv, end
#-----
# IMPORTANT NOTE
# The previous series of deletions and equivalences can be
# generated via the step CORTEX:
# Click on the button "Create the command file" of the main
# menu (Basic Steps) and follow the proposed instructions
# (button: CORTEX, in the paragraph "Textual data").
#-----

```

```

NSPA = 'NSPC'
#----- the file 'NSPC' created by CORTE is substituted to
# the file NSPA that was created by NUMER.

```

```

#---- selecting a new threshold for words -
NSPB = 'NSPB'
# the file NSPB created by SETEX is given the name: 'NSPB'

```

STEP SETEX

```

=====
NSEU =15 NMOMI=0 NREMI=2 LEDIT =NEW

```

```

#----- Comments about step SETEX
# NSEU:   threshold of frequency for selecting words.
# NMOMI:  minimum number of letters of a kept word.
# NREMI:  minimum number of words of a kept response.
# LEDIT:  printing the dictionaries (0=no, 1=new, 2=tot).
#-----

```

```

NSPA = 'NSPB'
#----- the file 'NSPB' created by SETEX is substituted to
# the file NSPA that was created by NUMER and modified by CORTE.

```

STEP SELEC

```

===== Selects active, supplementary variables and observations
LSELI = TOT, IMASS = UNIF, LZERO = REC, LEDIT = short
NOMI ILL 1 2 11 14
NOMI ACT 4--10
end

```

```

# See Appendix B2 for the comments about this step
#-----

```

STEP MULTM

```

===== Multiple correspondence analysis
NAXE = 7, PCMIN = 2. , LBURT = TOT, LEDCO = yes NSIMU=10

```

```

#----- Comments about step MULTM
# - NAXE = ... number of computed principal axes
# - PCMIN ... threshold for "cleaning" the active
# categories (in percent). This means that the low-
# frequency active categories (less than 2% in this
# case) are eliminated, and the corresponding
# individuals are dispatched at random among the
# other categories of the same variable (to remedy
# a well known weakness of the chi-square distance).
#
# - LBURT... printing the Burt contingency table
#           (0=NO, 1=MASS, 2=TOT, 3=PROF).

```

```

# - LEDCO...   printing the correlations variable-
#             axes (0=no, 1=yes).
# - NSIMU...number of bootstrap replication (<=30)
#             (0 = no bootstrap)
#-----

STEP DEFAC      # Description of factorial axes
===== Multiple correspondence analysis
SEUIL = 40., LCRIM = VTEST, VTMIN = 2.0
VEC = 1--2 / MOD
end

#----- Comments about step DEFAC
# SEUIL = ... Maximum number of elements that will
#           be sorted to describe each axis
# LCRIM = ... Criterion for sorting the elements
#           (here VTEST means "test-values" (signed number
#           of standard deviations)
# VEC = ... list of axes to be described
# CONT = continuous variables , MOD = categories
# The key-word END indicates the end of the list.
#-----

STEP POLEX
==== projecting supplementary words
ngraf = 2

#----- Comments about step POLEX
# POLEX aims at positioning words on principal space
# (here: principal space provided by MCA of closed questions)
# ngraf =   number of requested graphics (on file imp.txt)
#-----

STEP RECIP
==== Clustering of respondents using reciprocal neighbours
NAXU=7 LDEND=DENSE NTERM=20 LDESC=no

# See Appendix B1 for the comments about this step
#-----

STEP PARTI
==== Cut of the dendrogram to obtain 7 clusters
NITER=10   LEDIN=3
7          # number of classes of the partition

# See Appendix B1 for the comments about this step
#-----

STEP DECLA
===== Systematic description of clusters
CMODA = 5.0, PCMIN = 2.0, LSUPR = no, CCONT = 5.0 >
LPNOM = no, EDNOM = no, EDCON = no
7 # list of numbers of classes of requested partitions

# See Appendix B2 for the comments about this step
#-----

STEP MOTEX
===== Cross-tabulating words and partition

```

```
NVSEL = -1, LEDIT = 1

#----- Comments about step MOTEX
# See Appendix B1 for the comments about this step
#-----

STEP MOCAR
==== Characteristic words for each cluster (criterion 1)
NOMOT=10      NOREP=6

# See Appendix B1 for the comments about this step
#-----

STEP RECAR
===== characteristic responses (criterion 2)
NOREP = 4

#----- Comments about step RECAR
# NOREP:  number of characteristic responses for each text.
#-----
STOP
#-----
```

End of example B.3

Example B.4: EX_B04. Text-Semantic *(Visualization of the Semantic network of French verbs)*

Example B.4 provides a visualisation of the semantic links existing between 829 French verbs. Each verb is described by a list of synonyms. This example is in fact very similar to Example B.1 (Responses to an open question). The “respondents” are here the 829 verbs. The fictitious open-ended question is “Which are your synonyms?”, and the textual “response” is constituted by a list of synonyms. The example is also similar to the “Japan Map” example, pertaining to Example C.4 (Descriptions of graphs) from Tutorial C.

The principal axes visualization is complemented by a clustering, with an automatic description of the clusters. This is a typical first outlook on the set of responses: to detect and describe the main groupings of responses. Such outlook is by no means an achieved processing...

For more information, please refer to the book: “La sémiométrie” (2003) [in French] by L. Lebart, M. Piron, J.F. Steiner; Publisher: Dunod, Paris. (can be downloaded from www.dtm-vic.com).

To have a look at the data, search for the directory **DtmVic_Examples**.

In this directory, open the sub-directory **DtmVic_Examples_B_Texts**.

In that directory, open the directory of Example B.04, named **“EX_B04. Text-Semantic”**.

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 2 files :

- a) the text file, **synotex.txt**
- b) the command file: **“syno_par.txt”**

(in this particular context, there are neither data file nor dictionary file: the fictitious questionnaire comprises one open-ended question, without closed-end questions).

a) Text file: synotex.txt

The format is typical of responses to open questions (see examples A.5, B.1, B.2). Since the “responses” (here: lists of synonym verbs) may have different lengths, separators are used to distinguish between these lists. Lists (in fact : responses) are separated by the chain of characters “----“ (starting column 1) possibly followed by an identifier. Like all the data files involved in DtmVic as input files, that file is a raw text file (.txt). If the text file comes from a text processing phase, it must be saved beforehand as a “.txt file”.

b) Command file: syno_par.txt

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "[Help about parameters](#)") and, with more details, below.

Note that another “command file” similar to the “command file “syno_par.txt” can be also generated by clicking on the button “Create the command file” of the main menu (Basic Steps). A window “Choosing among some basic analysis” appears. Click then on the button: : VISURESP – Visualization of Responses – located in the paragraph “Textual data”, and follow the instructions.

Running the example B.4 and reading the results

1) Click on the button: “Open an existing command file” (panel *Basic Steps* of the main menu)

2) Then, search for the sub- directory **DtmVic_Examples_B_Texts** in **DtmVic_Examples**.

3) In that directory, open the directory of Example B.04, named **“EX_B04. Text-Semantic”**.

4) Open then the command file: **syno_par.txt**

After identifying the textual data file, seven "steps" are performed:

ARTEX (Archiving texts), **SELOX** (selecting the open question), **NUMER** (numerical coding of the text), **ASPAR** (correspondence analysis of the [sparse] contingency table “respondents - words”), **CLAIR** (Brief description of factorial axes), **RECIP** (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method), **PARTI** (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained), **MOTEX** (cross-tabulating the partition produced by step **PARTI** with words: the obtained contingency table is called a lexical table), **MOCAR** (characteristic words, and characteristics responses for each class of the partition).

We will comment later on this command file (Appendix B.4 of this section) which commands the basic computation steps. Instead of editing this file, we will content ourselves here in going back to the main menu and execute the basic computation steps.

Return to the main menu (“return to execute”)

5) Click on the button: **“Execute”**

This step will run the basic computation steps present in the command file.

6) Click the button: **“Basic numerical results”**

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory.

This file is also saved under a simple text format, under the name **“imp.txt”**, and likewise with a name including the date and time of execution.

From the step **NUMER**, we learn for instance that we have 829 “responses”, with a total number of words (occurrences or token) of 17 446, involving 3 839 distinct words (or: types). Using a frequency threshold of 12, the total number of kept words reduces to 5 013, whereas the number of distinct kept word reduces (more drastically) to 280.

7) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: “VIC”)

8) Click the button: **“Axesview”**

and ... follow the sub-menus. In fact, only two tabs are relevant for this example: **“Active variables”** [= words in the case of step **ASPAR**], **“Individuals (observations) [= respondents]”**. After clicking on **“View”** in both cases, one obtains the set of principal coordinates along each axis.

Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step **“CLAIR”**. Evidently, the use of the **Axeview** menu is justified when the data set is large, which is the case here.

9) Click the button: **PlaneView ...** and follow the sub-menus.

In this example, four items of the menu are relevant **“Active columns (variables or categories)”**, **“Rows**

(individuals, observations) , **“Active columns + Rows”**, **“Individuals (density)”**. The graphical display of chosen pairs of axes are then produced.

The roles of the different buttons are straightforward, except perhaps the button: **“Rank”**, which is useful only in the case of very intricate displays, (which is the case here). Since the set “individual” has 829 elements, it is possible to test, with this example, partial printings of the individuals in two subsets of 50% or four subsets of 25%...(subsets randomly drawn without replacement)

10) About the button: “BootstrapView”

The implementation of the bootstrap in the step ASPAR is not yet completed.

11). Click on “ ClusterView ”

11.1 Choose the axes (1 and 2 to begin with), and **“Continue”**.

11.2 Click on **“View”**. The centroids of the 20 clusters (Step PARTI) appears on the first principal plane.

11.3 Activate the button **“Words”**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR. But we do have in front of us the pattern of clusters and their relative locations.

11.4 Activate the button **“Texts”**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

12) Click on “Kohonen map”

Select the type of coordinate.

12.1 Select: **“Active variables (columns)”**: these active variables are the words in this example.

12.2 Select a (8 x 8) map, and continue.

12.3 After clicking on two small check-boxes, press **“Draw”** on the menu of the large green windows entitled Kohonen map.

12.4 You can change the font size (**“Font”**) and dilate the obtained Kohonen map (**“Dilat.”**) to make it more legible. The words appearing in the same cell are often associated in the same verbs. This property holds, at a lesser degree, for contiguous cells.

12.5 Pressing **“AxeView”**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

12.6 Go back to the main menu, click on **“Kohonen map”** and choose the item **“Active observations”**

12.7 Select a (10 x 10) map, and redo the operations 12.3 to 12.5 for the observations.

In the context of this example, the other items of the main menu are not relevant.

13. Click on “Visualization”

13.1 A new window is displayed.

13.1 Click on **“Load coordinate”**

13.2 In the corresponding sub-menu, choose the file: **“ngus_ind.txt”**. The principal coordinates of the individuals (rows) are selected.

13.3 Click then on **“Select or Create Partition”**

13.4 In the corresponding sub-menu, choose **“no partition”**.

13.5 Click on **“MST”** (Minimum Spanning Tree). Choose then the number of axes that will serve to

compute the Minimum Spanning Tree: full space (for example).

13.6 Click on **“Load MST”**, to load the results for the forthcoming visualisation phase.

13.7 Click on **“N.N.”** (search for nearest neighbours – limited to 20 NN).

13.8 Click on **“Load N.N.”** (loading the nearest neighbours file)

13.9 Click on **“Visualisation”**.

13.10 Choose the axes 1 and 2 (default) in the small window “Description of classes” and click on **“Display”**.

13.11 In the new window entitled **“Contiguity Visualisation”** are displayed the individuals in the plane spanned by the selected axes. A random colour is attributed to each cluster (if any). The button **“Change colour”** allows you to try a new set of colour. When you estimate that the colours are sufficiently contrasted, you can press **“Lock colour”**.

About the window “Visualisation” (from the sub-menu Visualisation)

On the vertical tool bar, you can press each button to activate it (red colour), and press it again to cancel the activation (black colour)

- The button **“Density”**, for sake of legibility, replaces the identifiers of individuals by a single character reminding the cluster (the identifier and the cluster number can be obtained by clicking on the left button of the mouse in the vicinity of each point).
- The button **“C.Hull”** (Convex hull) draws the convex hull of each cluster.
- The button **“MST”** (Minimum Spanning Tree) draws the minimum spanning tree.
- The button **“Ellipse”** perform a Principal Components Analysis of each cluster within the two-dimensional sub-space of visualisation and draws the corresponding ellipses (containing roughly 95% of the points).
- The button **“N.N.”** (Nearest neighbours) joins each point to its nearest neighbours. Pressing afterwards the button **“N.N. up”** allows you to increment the number of neighbours up to 20 nearest neighbours.

Appendix B.4

The steps and the command file of example B.4 are the same as those of Example B.1 (if we except the name of the data file containing the input text).

The reader should then refer to Appendix B.1 to obtain the corresponding comments.

End of example B4

End of tutorial B