# 1 Subject

Implementing the multiple correspondence analysis (MCA) with Tanagra.

The multiple correspondence analysis is a factor analysis approach[1]. It deals with a tabular dataset where a set of examples are described by a set of categorical variables. The aim is to map the dataset in a reduced dimension space (usually two) which allows us to highlight the associations between the examples and the variables.

In this tutorial, we show how to implement this approach and how to interpret the results with Tanagra.

# 2 Dataset

We use the RACES_CANINES_ACM.XLS dataset, from the famous TENENHAUS[2] book (Table 1, page 200). The interest of this dataset is that we can compare our results with those described in the book (pages 212 to 222). We simply show the sequence of operations and the reading of the results tables in this tutorial. About the detailed interpretation, it is best to refer to the book.

The data table is the following:

| Chien | Taille | Poids | Velocite | Intelligence | Affection | Agressivite | Fonction |
|---|---|---|---|---|---|---|---|
| Beauceron | Taille+ + | Poids+ | Veloc+ + | Intell+ | Affec+ | Agress+ | utilite |
| Basset | Taille- | Poids- | Veloc- | Intell- | Affec- | Agress+ | chasse |
| Berger All | Taille+ + | Poids+ | Veloc+ + | Intell+ + | Affec+ | Agress+ | utilite |
| Boxer | Taille+ | Poids+ | Veloc+ | Intell+ | Affec+ | Agress+ | compagnie |
| Bull-Dog | Taille- | Poids- | Veloc- | Intell+ | Affec+ | Agress- | compagnie |
| Bull-Mastif | Taille+ + | Poids+ + | Veloc- | Intell+ + | Affec- | Agress+ | utilite |
| Caniche | Taille- | Poids- | Veloc+ | Intell+ + | Affec+ | Agress- | compagnie |
| Chihuahua | Taille- | Poids- | Veloc- | Intell- | Affec+ | Agress- | compagnie |
| Cocker | Taille+ | Poids- | Veloc- | Intell+ | Affec+ | Agress+ | compagnie |
| Colley | Taille+ + | Poids+ | Veloc+ + | Intell+ | Affec+ | Agress- | compagnie |
| Dalmatien | Taille+ | Poids+ | Veloc+ | Intell+ | Affec+ | Agress- | compagnie |
| Doberman | Taille+ + | Poids+ | Veloc+ + | Intell+ + | Affec- | Agress+ | utilite |
| Dogue All | Taille+ + | Poids+ + | Veloc+ + | Intell- | Affec- | Agress+ | utilite |
| Epag. Breton | Taille+ | Poids+ | Veloc+ | Intell+ + | Affec+ | Agress- | chasse |
| Epag. Français | Taille+ + | Poids+ | Veloc+ | Intell+ | Affec- | Agress- | chasse |
| Fox-Hound | Taille+ + | Poids+ | Veloc+ + | Intell- | Affec- | Agress+ | chasse |
| Fox-Terrier | Taille- | Poids- | Veloc+ | Intell+ | Affec+ | Agress+ | compagnie |
| Gd Bleu Gasc | Taille+ + | Poids+ | Veloc+ | Intell- | Affec- | Agress+ | chasse |
| Labrador | Taille+ | Poids+ | Veloc+ | Intell+ | Affec+ | Agress- | chasse |
| Levrier | Taille+ + | Poids+ | Veloc+ + | Intell- | Affec- | Agress- | chasse |
| Mastiff | Taille+ + | Poids+ + | Veloc- | Intell- | Affec- | Agress+ | utilite |
| Pekinois | Taille- | Poids- | Veloc- | Intell- | Affec+ | Agress- | compagnie |
| Pointer | Taille+ + | Poids+ | Veloc+ + | Intell+ + | Affec- | Agress- | chasse |
| St-Bernard | Taille+ + | Poids+ + | Veloc- | Intell+ | Affec- | Agress+ | utilite |
| Setter | Taille+ + | Poids+ | Veloc+ + | Intell+ | Affec- | Agress- | chasse |
| Teckel | Taille- | Poids- | Veloc- | Intell+ | Affec+ | Agress- | compagnie |
| Terre-Neuve | Taille+ + | Poids+ + | Veloc- | Intell+ | Affec- | Agress- | utilite |

The first column is the label of the examples. The active variables, used during the computation of the axes, are in green; the supplementary (illustrative) variables, used only for the interpretation of the results, are in blue.

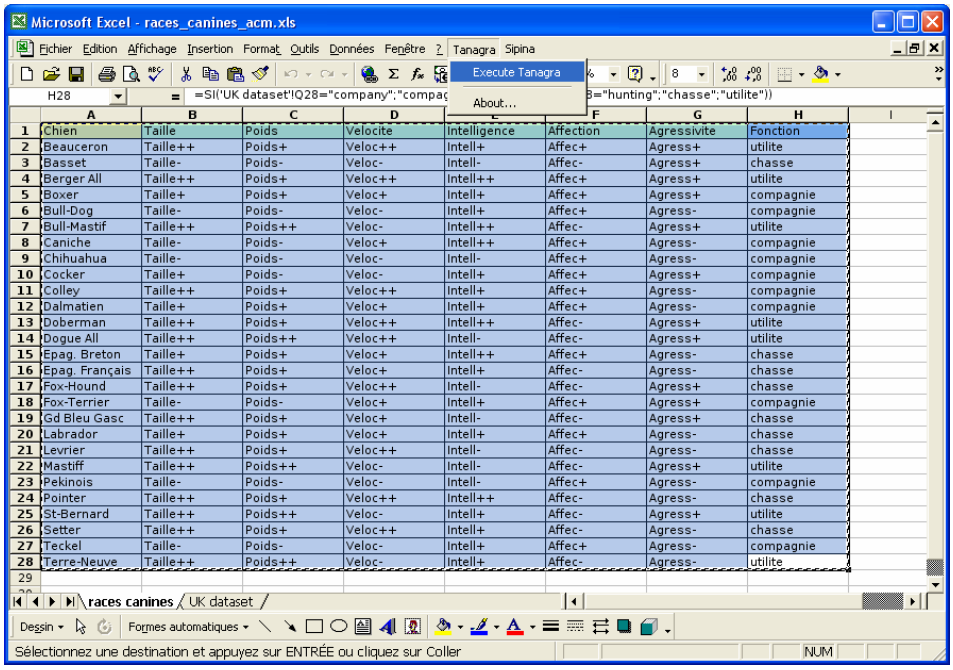
[1] http://faculty.chass.ncsu.edu/garson/PA765/correspondence.htm

[2] M. TENENHAUS, « Méthodes Statistiques en gestion », DUNOD, 1996 (in french).

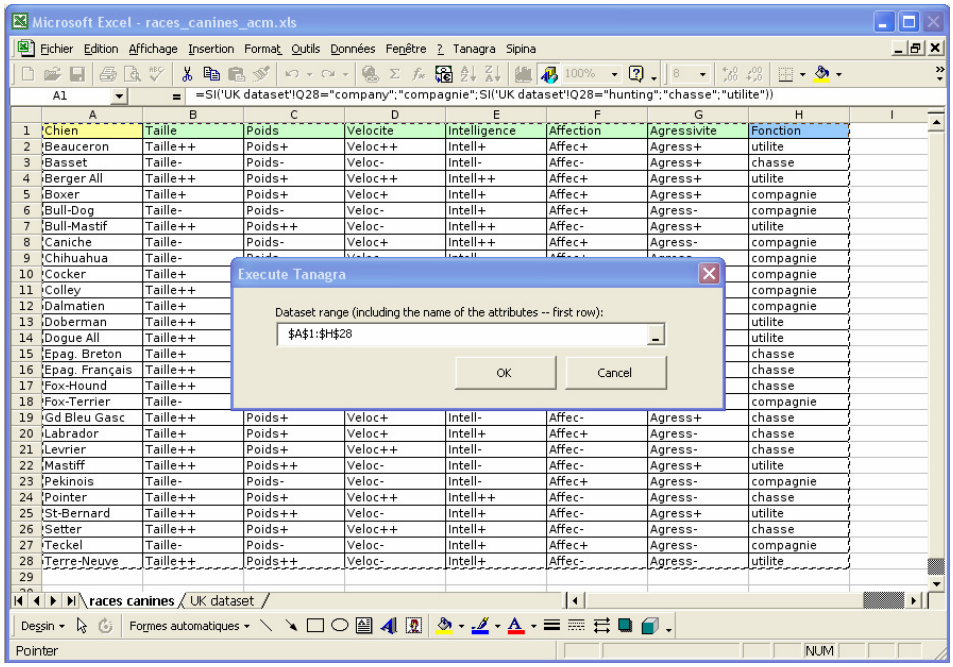
# 3 MCA with TANAGRA

## 3.1 Creating a diagram

We can launch Tanagra from Excel using an add-on[3]. We select the range of cells that contains the dataset. Then we click on the TANAGRA / EXECUTE TANAGRA menu.

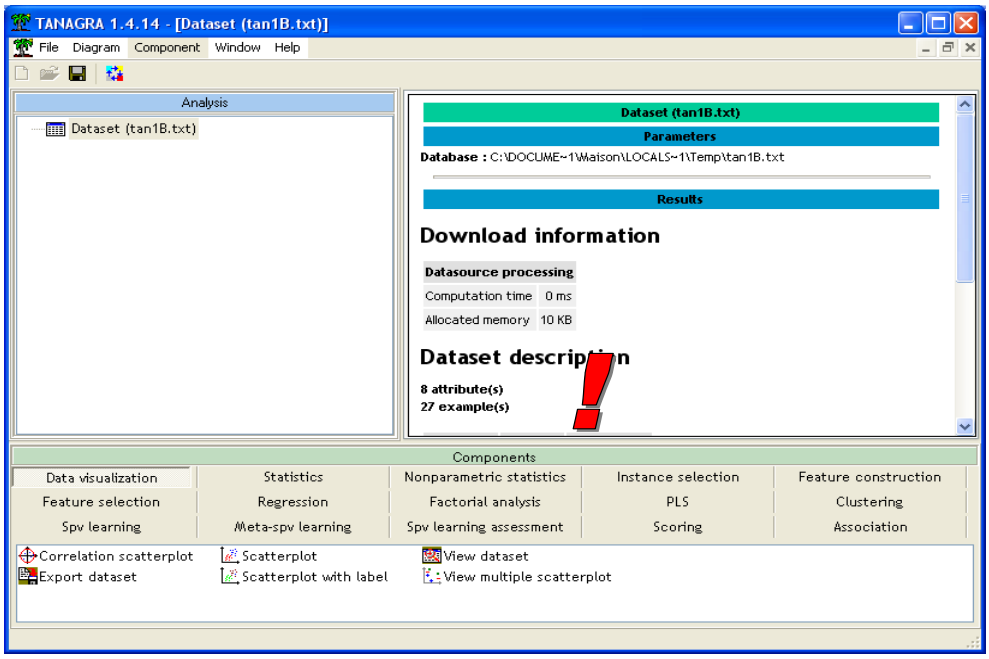



A dialog box appears. We click on OK if the selection is right.



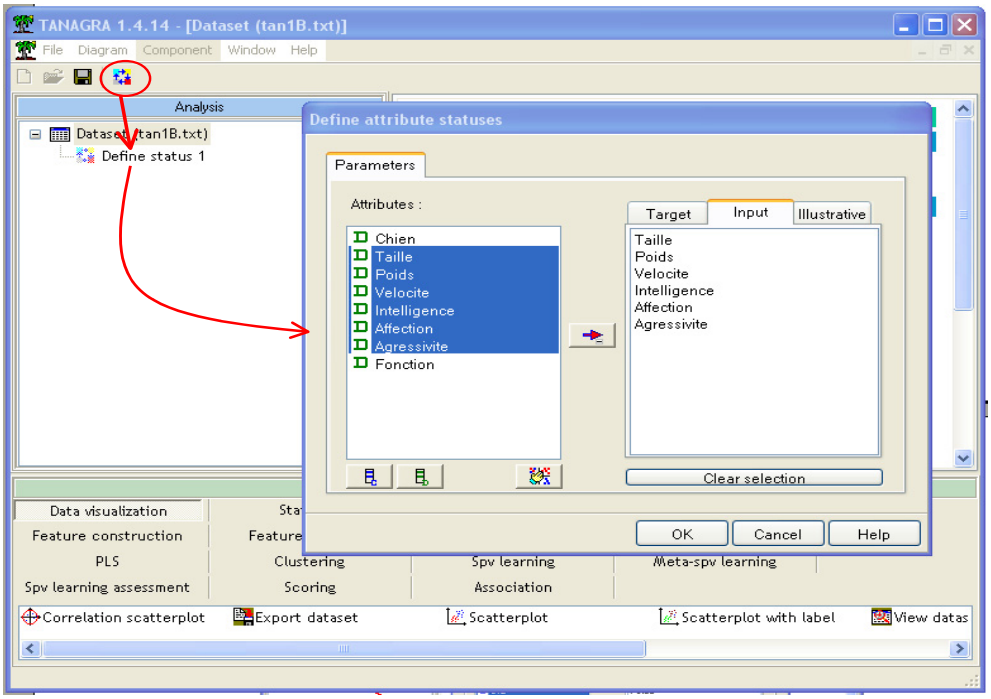

[3] http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html; we can also import the XLS data file even if Excel is not installed on our computer, see http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html.

TANAGRA is launched. We check that there are 27 examples and 8 variables[4].




## 3.2 MCA

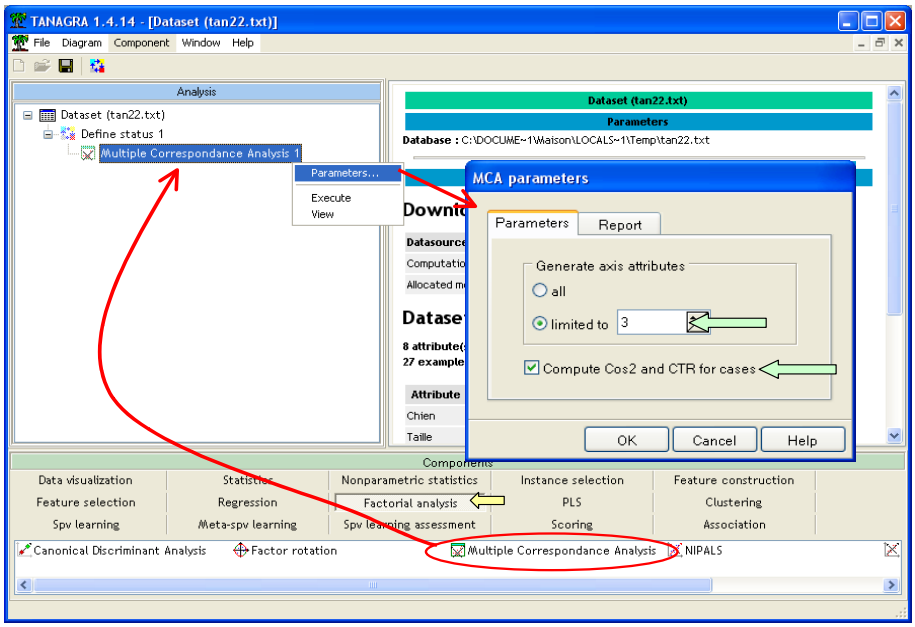
First, we must define the types of variables. We insert the DEFINE STATUS component into the diagram by clicking the shortcut in the toolbar. We set as INPUT the active variables. We see below how to use the illustrative variables.



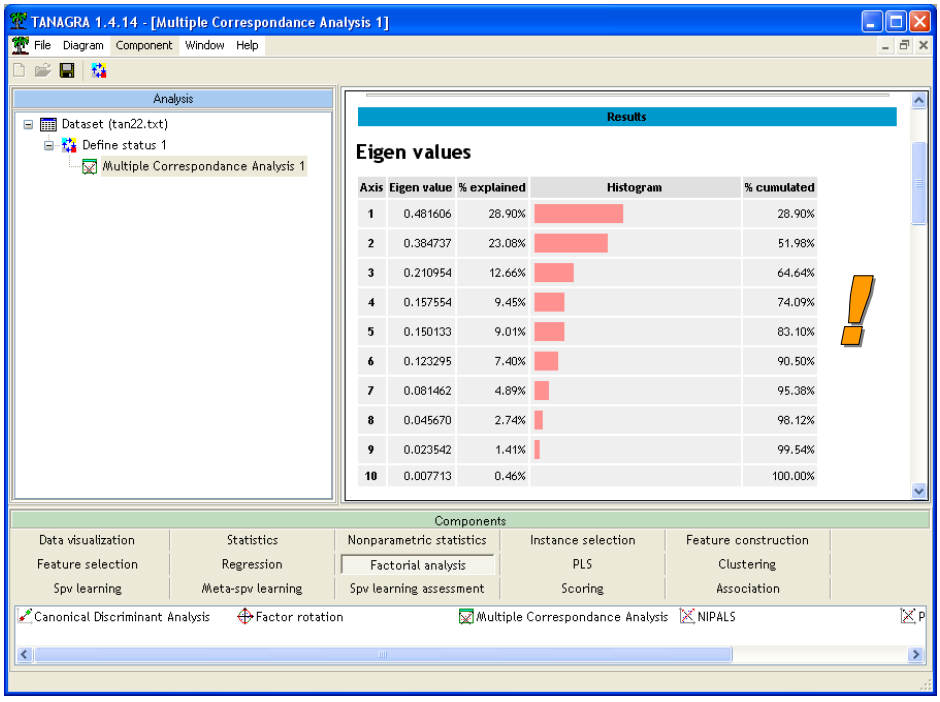

[4] Tanagra does not handle the label column. The first column is thus counted as a categorical variable. It is not a problem. Tanagra considers that each label is a value of the variable... but in this case, the number of different values is limited to 255.

Then we add the MULTIPLE CORRESPONDENCE ANALYSIS component (FACTORIAL ANALYSIS tab). We click on the PARAMETERS menu: we set the number of dimensions to supply (3 axes); we want to compute the COS2 (contribution of dimensions to points or squared correlations) and the CTR (contributions of points to dimensions).

We click on the VIEW menu in order to obtain the results.

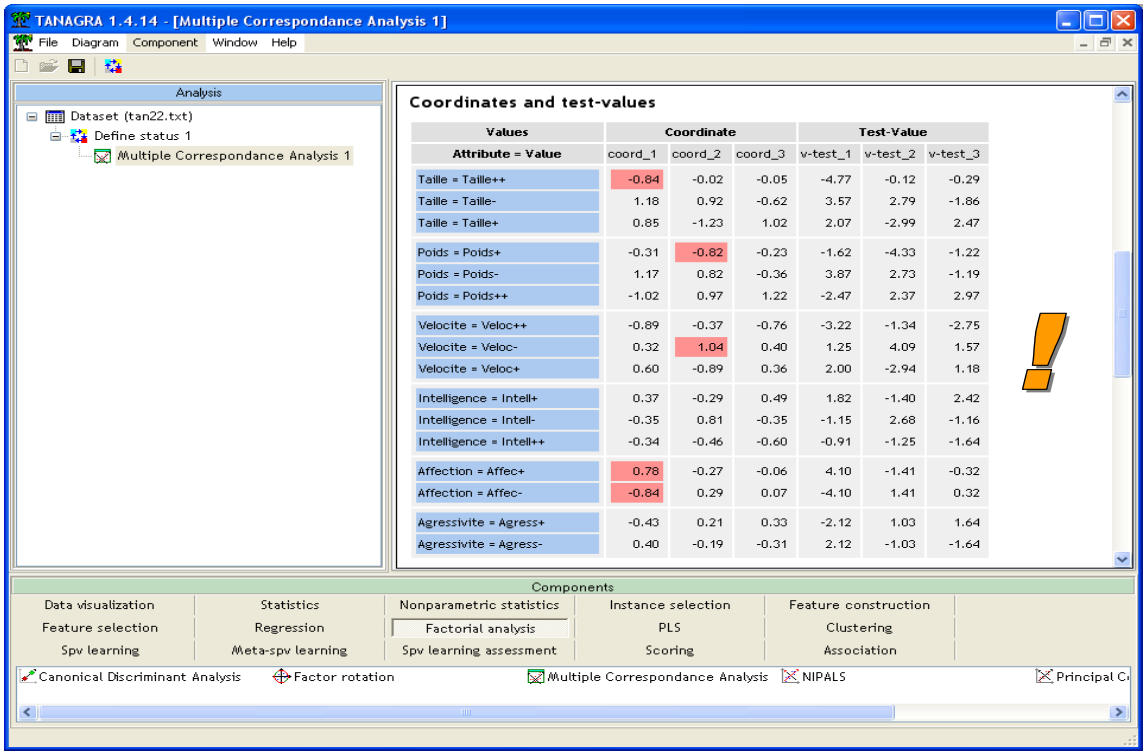



**Eigenvalues.** The first table describes the eigenvalues (Tableau 3, p.214; Tenenhaus, 1996). They reflect the importance of each dimension.



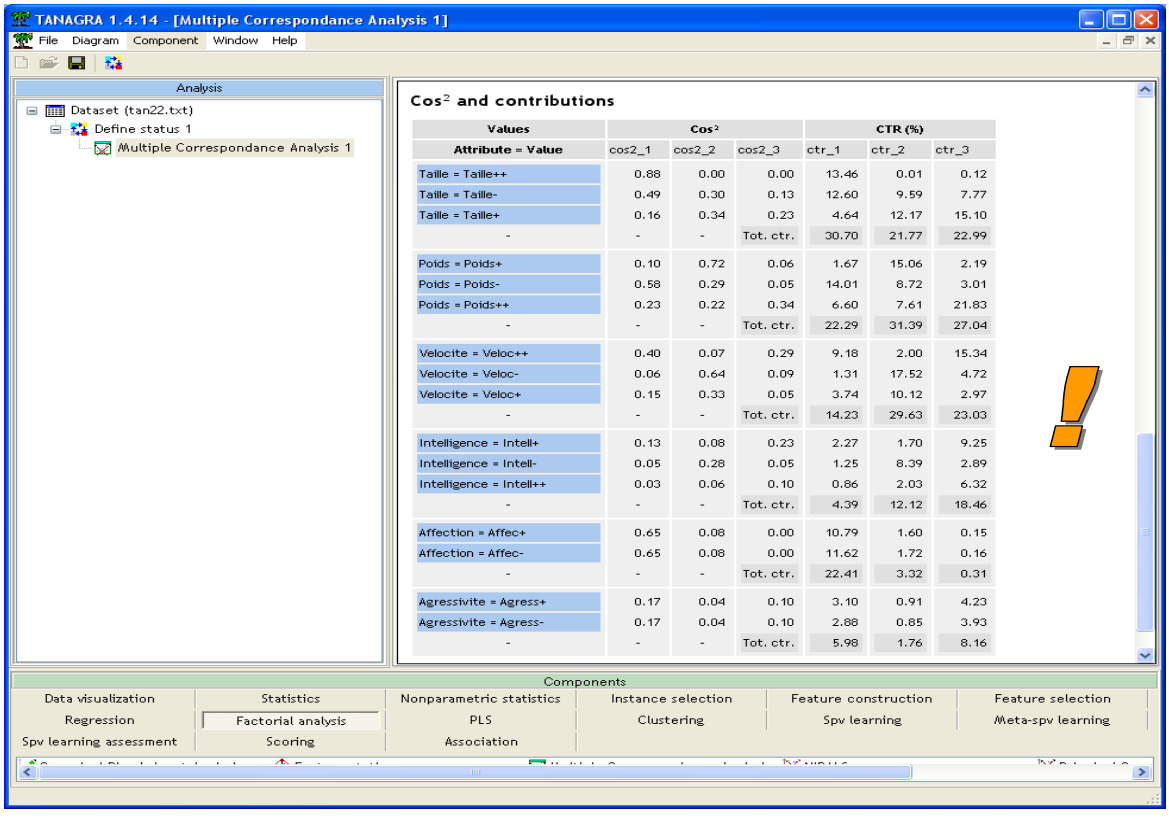

**Coordinates and test-values.** The second part of the results described the coordinates of the categories of the variables on each dimension (Tableau 5, p.216; Tenenhaus, 1996). The test value

is used for highlight strong associations. The cell is in red if test value is higher than a predefined threshold (the default cut value is 4).



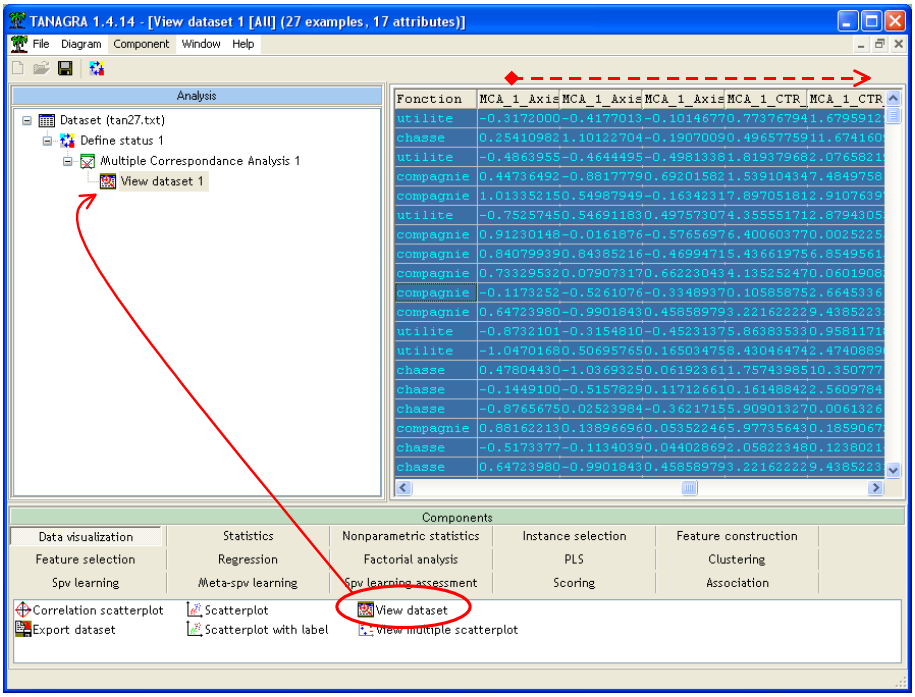

| Values | Coordinate | | | Test-Value | | |
|---|---|---|---|---|---|---|
| Attribute = Value | coord_1 | coord_2 | coord_3 | v-test_1 | v-test_2 | v-test_3 |
| Taille = Taille++ | -0.84 | -0.02 | -0.05 | -4.77 | -0.12 | -0.29 |
| Taille = Taille- | 1.18 | 0.92 | -0.62 | 3.57 | 2.79 | -1.86 |
| Taille = Taille+ | 0.85 | -1.23 | 1.02 | 2.07 | -2.99 | 2.47 |
| Poids = Poids+ | -0.31 | -0.82 | -0.23 | -1.62 | -4.33 | -1.22 |
| Poids = Poids- | 1.17 | 0.82 | -0.36 | 3.87 | 2.73 | -1.19 |
| Poids = Poids++ | -1.02 | 0.97 | 1.22 | -2.47 | 2.37 | 2.97 |
| Velocite = Veloc++ | -0.89 | -0.37 | -0.76 | -3.22 | -1.34 | -2.75 |
| Velocite = Veloc- | 0.32 | 1.04 | 0.40 | 1.25 | 4.09 | 1.57 |
| Velocite = Veloc+ | 0.60 | -0.89 | 0.36 | 2.00 | -2.94 | 1.18 |
| Intelligence = Intell+ | 0.37 | -0.29 | 0.49 | 1.82 | -1.40 | 2.42 |
| Intelligence = Intell- | -0.35 | 0.81 | -0.35 | -1.15 | 2.68 | -1.16 |
| Intelligence = Intell++ | -0.34 | -0.46 | -0.60 | -0.91 | -1.25 | -1.64 |
| Affection = Affec+ | 0.78 | -0.27 | -0.06 | 4.10 | -1.41 | -0.32 |
| Affection = Affec- | -0.84 | 0.29 | 0.07 | -4.10 | 1.41 | 0.32 |
| Agressivite = Agress+ | -0.43 | 0.21 | 0.33 | -2.12 | 1.03 | 1.64 |
| Agressivite = Agress- | 0.40 | -0.19 | -0.31 | 2.12 | -1.03 | -1.64 |

The COS2 and CTR are also supplied.

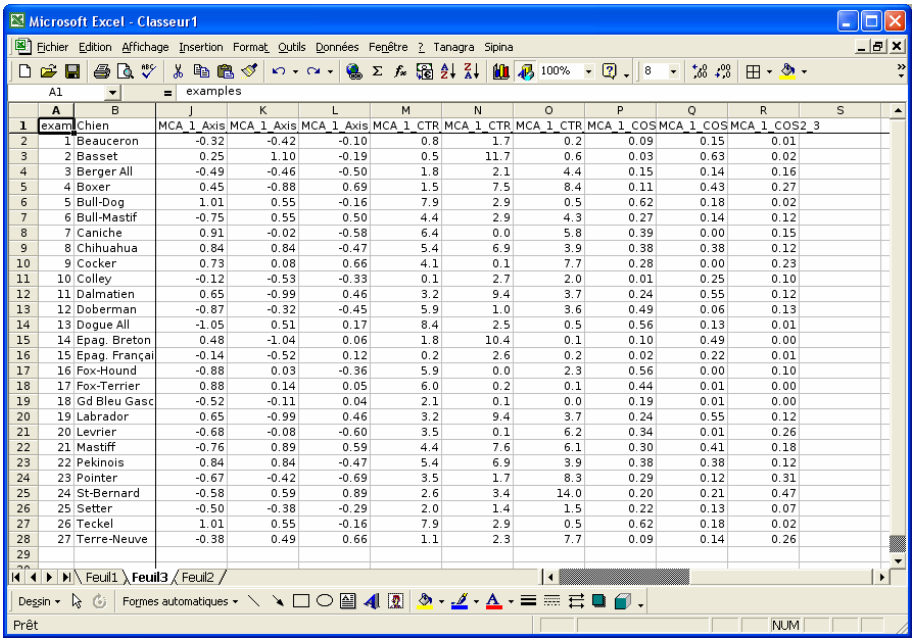



| Values | Cos² | | | CTR (%) | | |
|---|---|---|---|---|---|---|
| Attribute = Value | cos2_1 | cos2_2 | cos2_3 | ctr_1 | ctr_2 | ctr_3 |
| Taille = Taille++ | 0.88 | 0.00 | 0.00 | 13.46 | 0.01 | 0.12 |
| Taille = Taille- | 0.49 | 0.30 | 0.13 | 12.60 | 9.59 | 7.77 |
| Taille = Taille+ | 0.16 | 0.34 | 0.23 | 4.64 | 12.17 | 15.10 |
| - | - | - | Tot. ctr. | 30.70 | 21.77 | 22.99 |
| Poids = Poids+ | 0.10 | 0.72 | 0.06 | 1.67 | 15.06 | 2.19 |
| Poids = Poids- | 0.58 | 0.29 | 0.05 | 14.01 | 8.72 | 3.01 |
| Poids = Poids++ | 0.23 | 0.22 | 0.34 | 6.60 | 7.61 | 21.83 |
| - | - | - | Tot. ctr. | 22.29 | 31.39 | 27.04 |
| Velocite = Veloc++ | 0.40 | 0.07 | 0.29 | 9.18 | 2.00 | 15.34 |
| Velocite = Veloc- | 0.06 | 0.64 | 0.09 | 1.31 | 17.52 | 4.72 |
| Velocite = Veloc+ | 0.15 | 0.33 | 0.05 | 3.74 | 10.12 | 2.97 |
| - | - | - | Tot. ctr. | 14.23 | 29.63 | 23.03 |
| Intelligence = Intell+ | 0.13 | 0.08 | 0.23 | 2.27 | 1.70 | 9.25 |
| Intelligence = Intell- | 0.05 | 0.28 | 0.05 | 1.25 | 8.39 | 2.89 |
| Intelligence = Intell++ | 0.03 | 0.06 | 0.10 | 0.86 | 2.03 | 6.32 |
| - | - | - | Tot. ctr. | 4.39 | 12.12 | 18.46 |
| Affection = Affec+ | 0.65 | 0.08 | 0.00 | 10.79 | 1.60 | 0.15 |
| Affection = Affec- | 0.65 | 0.08 | 0.00 | 11.62 | 1.72 | 0.16 |
| - | - | - | Tot. ctr. | 22.41 | 3.32 | 0.31 |
| Agressivite = Agress+ | 0.17 | 0.04 | 0.10 | 3.10 | 0.91 | 4.23 |
| Agressivite = Agress- | 0.17 | 0.04 | 0.10 | 2.88 | 0.85 | 3.93 |
| - | - | - | Tot. ctr. | 5.98 | 1.76 | 8.16 |

**Mapping the examples in the new representation space. COS² and contributions.** TANAGRA computes the coordinates of the examples on each dimension. The new columns are automatically added to the current dataset. They are available in the subsequent part of the diagram. If we want to view the values, we can add the VIEW DATASET component (DATA VISUALIZATION tab).



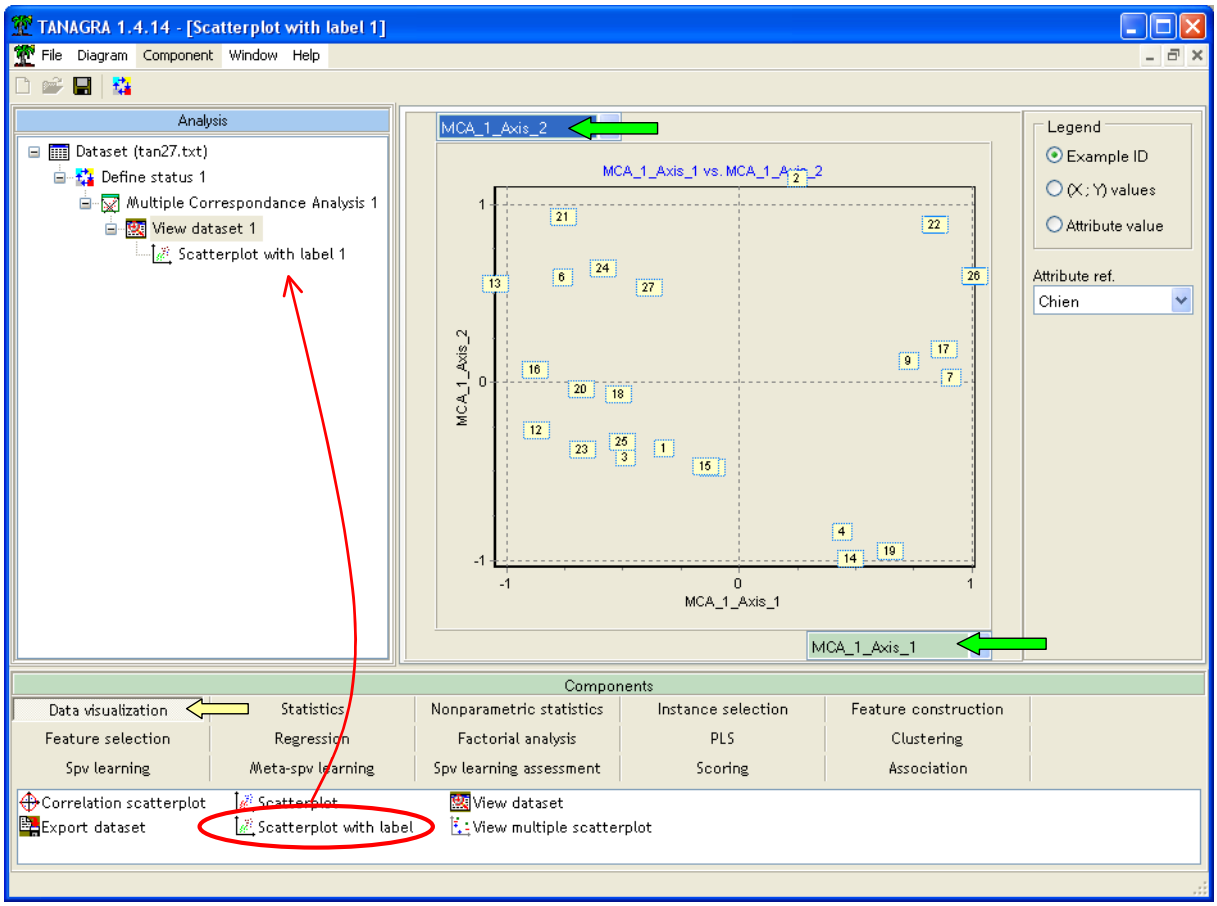

TANAGRA uses a scientific format. A simple way to obtain a more convenient presentation is to copy/paste the values into a spreadsheet (COMPONENT / COPY RESULTS menu) as the follows.




| | A | B | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | exam | Chien | MCA_1_Axis | MCA_1_Axis | MCA_1_Axis | MCA_1_CTR | MCA_1_CTR | MCA_1_CTR | MCA_1_COS | MCA_1_COS | MCA_1_COS2_3 | |
| 2 | 1 | Beauceron | -0.32 | -0.42 | -0.10 | 0.8 | 1.7 | 0.2 | 0.09 | 0.15 | 0.01 | |
| 3 | 2 | Basset | 0.25 | 1.10 | -0.19 | 0.5 | 11.7 | 0.6 | 0.03 | 0.63 | 0.02 | |
| 4 | 3 | Berger All | -0.49 | -0.46 | -0.50 | 1.8 | 2.1 | 4.4 | 0.15 | 0.14 | 0.16 | |
| 5 | 4 | Boxer | 0.45 | -0.88 | 0.69 | 1.5 | 7.5 | 8.4 | 0.11 | 0.43 | 0.27 | |
| 6 | 5 | Bull-Dog | 1.01 | 0.55 | -0.16 | 7.9 | 2.9 | 0.5 | 0.62 | 0.18 | 0.02 | |
| 7 | 6 | Bull-Mastif | -0.75 | 0.55 | 0.50 | 4.4 | 2.9 | 4.3 | 0.27 | 0.14 | 0.12 | |
| 8 | 7 | Caniche | 0.91 | -0.02 | -0.58 | 6.4 | 0.0 | 5.8 | 0.39 | 0.00 | 0.15 | |
| 9 | 8 | Chihuahua | 0.84 | 0.84 | -0.47 | 5.4 | 6.9 | 3.9 | 0.38 | 0.38 | 0.12 | |
| 10 | 9 | Cocker | 0.73 | 0.08 | 0.66 | 4.1 | 0.1 | 7.7 | 0.28 | 0.00 | 0.23 | |
| 11 | 10 | Colley | -0.12 | -0.53 | -0.33 | 0.1 | 2.7 | 2.0 | 0.01 | 0.25 | 0.10 | |
| 12 | 11 | Dalmatien | 0.65 | -0.99 | 0.46 | 3.2 | 9.4 | 3.7 | 0.24 | 0.55 | 0.12 | |
| 13 | 12 | Doberman | -0.87 | -0.32 | -0.45 | 5.9 | 1.0 | 3.6 | 0.49 | 0.06 | 0.13 | |
| 14 | 13 | Dogue All | -1.05 | 0.51 | 0.17 | 8.4 | 2.5 | 0.5 | 0.56 | 0.13 | 0.01 | |
| 15 | 14 | Epag. Breton | 0.48 | -1.04 | 0.06 | 1.8 | 10.4 | 0.1 | 0.10 | 0.49 | 0.00 | |
| 16 | 15 | Epag. Françai | -0.14 | -0.52 | 0.12 | 0.2 | 2.6 | 0.2 | 0.02 | 0.22 | 0.01 | |
| 17 | 16 | Fox-Hound | -0.88 | 0.03 | -0.36 | 5.9 | 0.0 | 2.3 | 0.56 | 0.00 | 0.10 | |
| 18 | 17 | Fox-Terrier | 0.88 | 0.14 | 0.05 | 6.0 | 0.2 | 0.1 | 0.44 | 0.01 | 0.00 | |
| 19 | 18 | Gd Bleu Gasc | -0.52 | -0.11 | 0.04 | 2.1 | 0.1 | 0.0 | 0.19 | 0.01 | 0.00 | |
| 20 | 19 | Labrador | 0.65 | -0.99 | 0.46 | 3.2 | 9.4 | 3.7 | 0.24 | 0.55 | 0.12 | |
| 21 | 20 | Levrier | -0.68 | -0.08 | -0.60 | 3.5 | 0.1 | 6.2 | 0.34 | 0.01 | 0.26 | |
| 22 | 21 | Mastiff | -0.76 | 0.89 | 0.59 | 4.4 | 7.6 | 6.1 | 0.30 | 0.41 | 0.18 | |
| 23 | 22 | Pekinois | 0.84 | 0.84 | -0.47 | 5.4 | 6.9 | 3.9 | 0.38 | 0.38 | 0.12 | |
| 24 | 23 | Pointer | -0.67 | -0.42 | -0.69 | 3.5 | 1.7 | 8.3 | 0.29 | 0.12 | 0.31 | |
| 25 | 24 | St-Bernard | -0.58 | 0.59 | 0.89 | 2.6 | 3.4 | 14.0 | 0.20 | 0.21 | 0.47 | |
| 26 | 25 | Setter | -0.50 | -0.38 | -0.29 | 2.0 | 1.4 | 1.5 | 0.22 | 0.13 | 0.07 | |
| 27 | 26 | Teckel | 1.01 | 0.55 | -0.16 | 7.9 | 2.9 | 0.5 | 0.62 | 0.18 | 0.02 | |
| 28 | 27 | Terre-Neuve | -0.38 | 0.49 | 0.66 | 1.1 | 2.3 | 7.7 | 0.09 | 0.14 | 0.26 | |
| 29 | | | | | | | | | | | | |

In addition, with the spreadsheet, we have multiple sorting options that allow to highlight the relevant information.

**Graphical representation of the examples.** The graphical representations are one of the main tools of the factorial methods. They allow us to visually evaluate the proximity between observations and interpret them. In our case, we project the observations in the first two dimensions. We can associate a label to each point. We insert the SCATTERPLOT WITH LABEL component (DATA VISUALIZATION tab) into the diagram. We set the first dimension for the horizontal axis and the second dimension for the vertical one. We note that we can easily modify the axes.

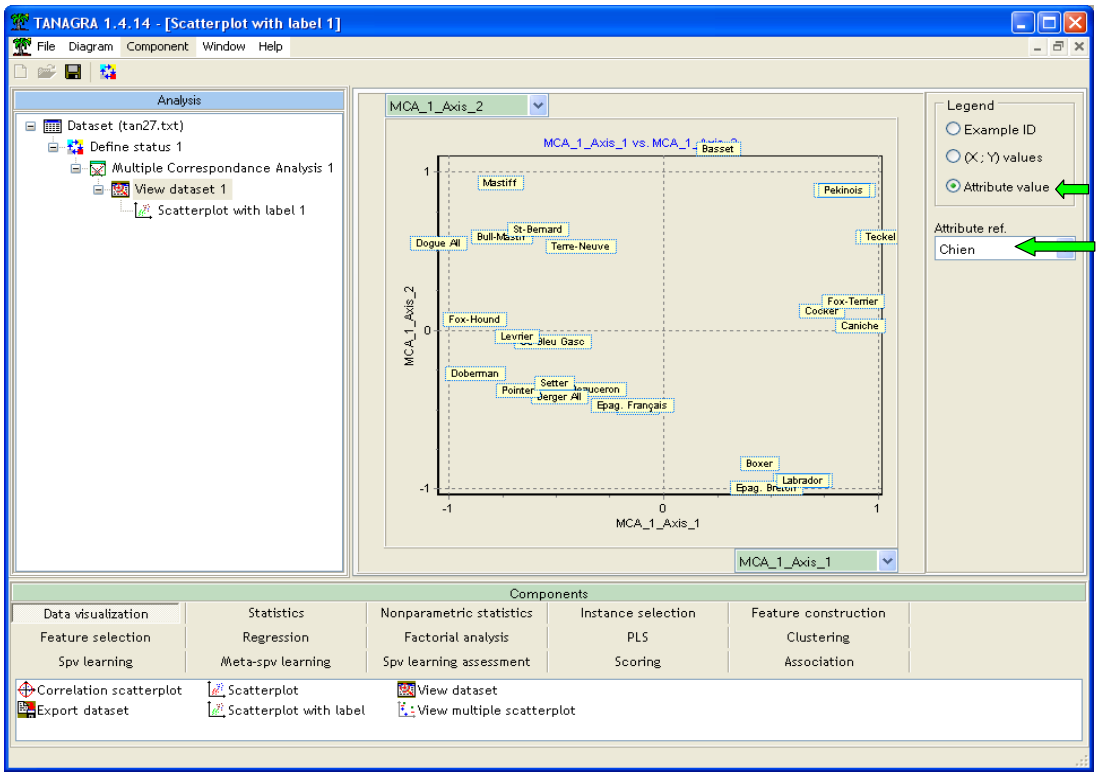



The points are automatically labeled by their number. We can modify this by clicking on LEGEND / ATTRIBUTE VALUE option. We select MODELE as reference attribute. We obtain the scatter plot on the two first dimensions (Figure 1, p.218; Tenenhaus, 1996).

Of course, this option is interesting if the number of points remains reasonable. Beyond of a certain number of observations, the graph would be unreadable.

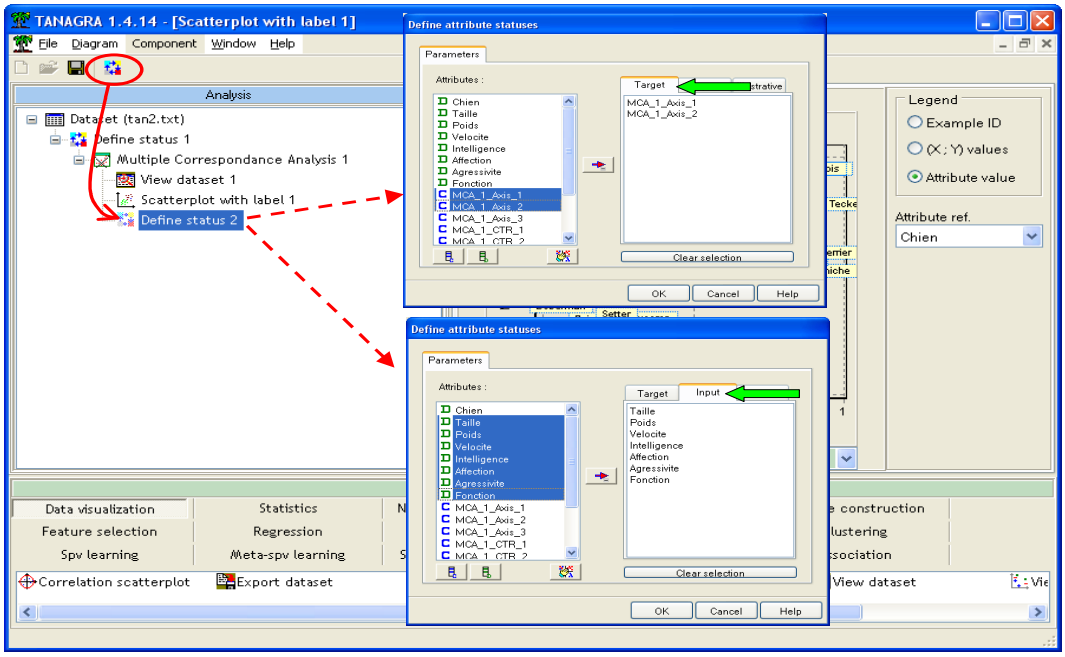
In our scatter plot, some points are superposed. In this case, copying the coordinates in a spreadsheet and ranking the examples according the dimensions is the best way to identify precisely each example.

We can modify the size of the labels with the shortcuts CTRL + W and CTRL + Q.

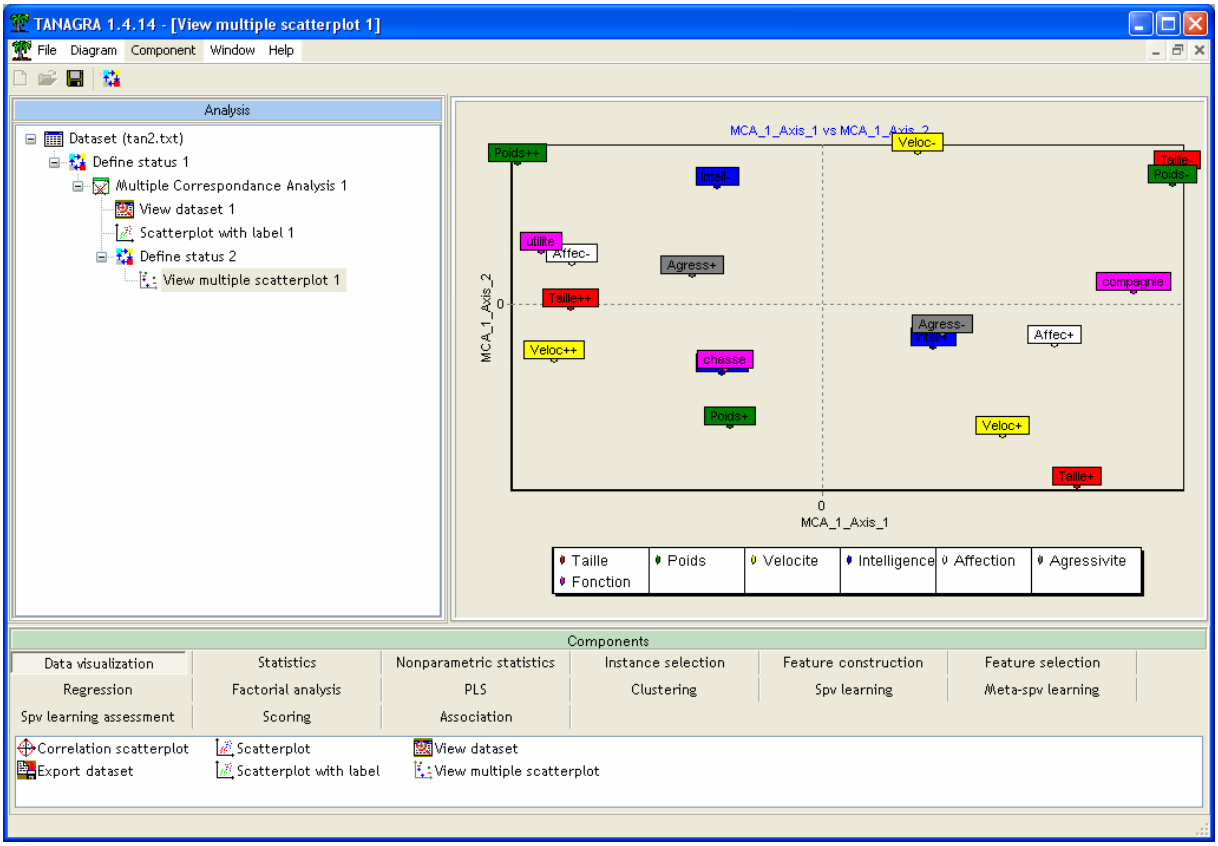



**Graphical representation of the variables.** We can also depict the variables in the first two (or other) dimensions. It is very useful when we want to interpret the dimensions and highlight associations between variables (or categories of variables). In this stage, we can use also the illustrative variables. Often, this functionality strengthens the interpretation of the dimensions.

We insert the DEFINE STATUS component. We set as TARGET the first two dimensions; as INPUT all the variables, including the illustrative ones.



We obtain the map of the variables (values of variables) in the first two dimensions of the representation space (Figure 2, p.220; Tenenhaus, 1996).




On the first axis, the dogs of big size, swift and not very affectionate, are opposed to the dogs of company, small size and low-weight (see the book for the detailed comments, pages 217-219).

**Illustrative examples.** We do not use this option in our tutorial, but we can also subdividing the dataset in a learning sample and an illustrative sample. The first is used for the computation of the new dimensions. The second corresponds to new instances that we want to locate into this new representation space, for instance when we want to apply the results on other sub-population.

# 4 Conclusion

The correspondence analysis, and in general the factor analysis, is useful to understand the underlying structure of a tabular dataset. In this tutorial, we show how to implement this approach with Tanagra.

The opportunity to copy/paste the results into a spreadsheet is certainly one of the most interesting functionalities of the software. Indeed, it gives us access to tools (sorting, formatting, etc.) in a well-known environment of the experts of the data processing. For example, the possibility of sorting the various tables according to the contributions and the COS2 proves really practical when one wishes to interpret the dimensions.