askia analyse
Mulitivariate Analysis & Lexicometry
User guide

Cpc – Analyse 5.3.3.X

The aim of this document is to provide step by step guidance on applying multivariate analysis and Lexicometry in askiaanalyse 5.3.3.

Content

1. Multivariate analysis

We use multivariate analysis when we want to retrieve more information than we can read in a cross table (Bivariate analyse).
Multivariate Data *Analyses* are powerful statistical *techniques* for analysing data with many variables simultaneously to identify patterns & relationships.
The multivariate analysis helps us to understand the dependence or the interdependence:

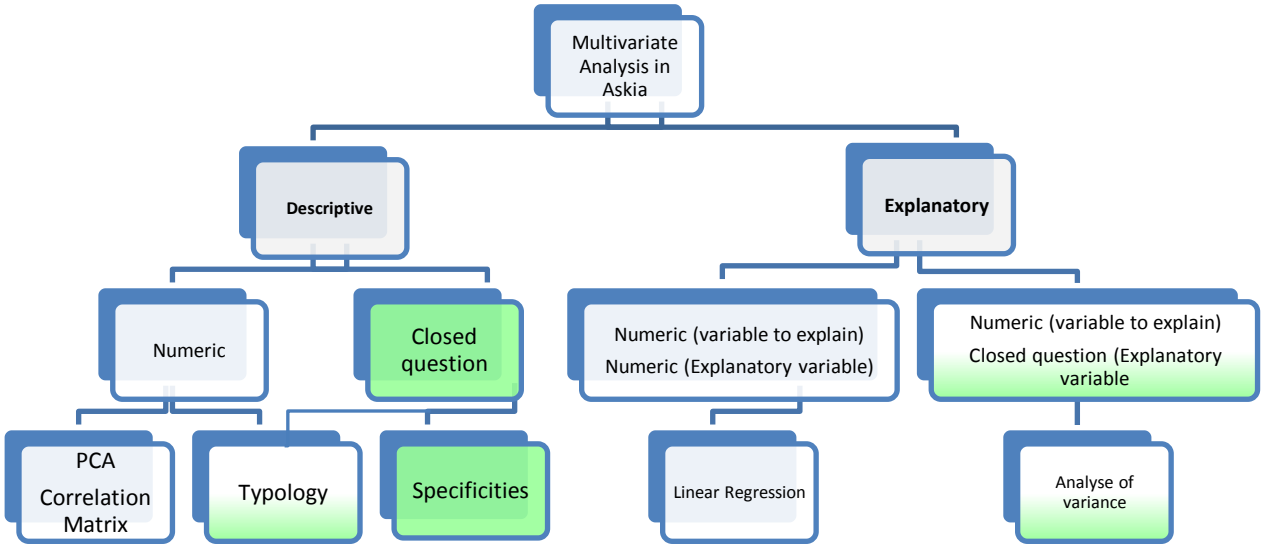
1. Why does the relationship between variables exist? What are the mechanisms and processes by which one variable is linked to another?
2. What is the nature of the relationship? Is it causal or non-causal?
3. How general is the relationship? Does it hold for people in general, or is it specific to certain subgroups?

We have 2 kinds of multivariate analysis:

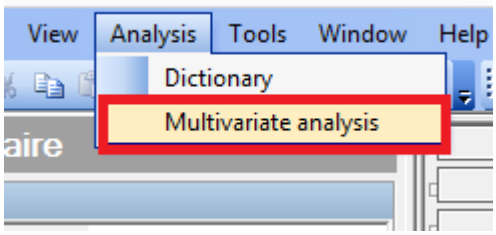
- Explanatory methods
- Descriptive methods

Using different form of data:

- Nominal/Categorical: Closed question (single or multicoded or scale responses)
- or metrics/Continuous: Numeric (numeric or scale responses)



To apply one of them, open your qes file in Analyse 5.3.3.X, select Analysis in the tool bar menu



1.1.Principal Component Analyse (PCA)

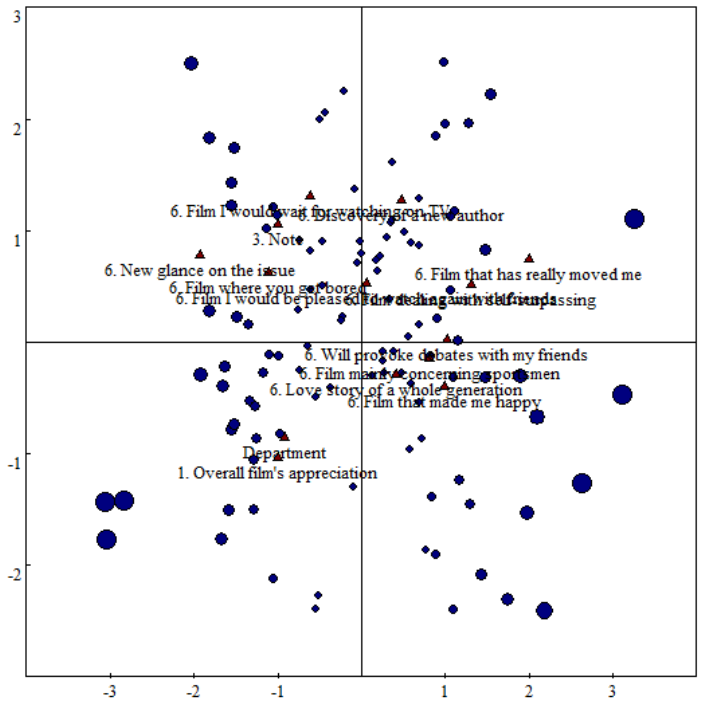
Definition

PCA is a descriptive analysis and it is used to analyse numeric questions (metrics).
We convert a set of observations **of possibly correlated variables into a set of values of linearly uncorrelated** variables **called principal components**.
This technique allows the researcher to reduce the number of correlated variables and decrease the redundant information. We will obtain a new representation in a new space where the first axis maximizes the amount of information that can be shown.

Data to process PCA uses a matrix *N* x *p* where:
N is the number of interviews
p is the number of questions.

At the intersection of the row *i* and the column *j*, we set the observation *i* to the question *j*.
To compare observations independently of scale and unit problems, we standardise and reduce the data matrix for each attribute.
We can then create an average "imaginary" point from which we can measure the differences between individuals using Euclidian distance. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

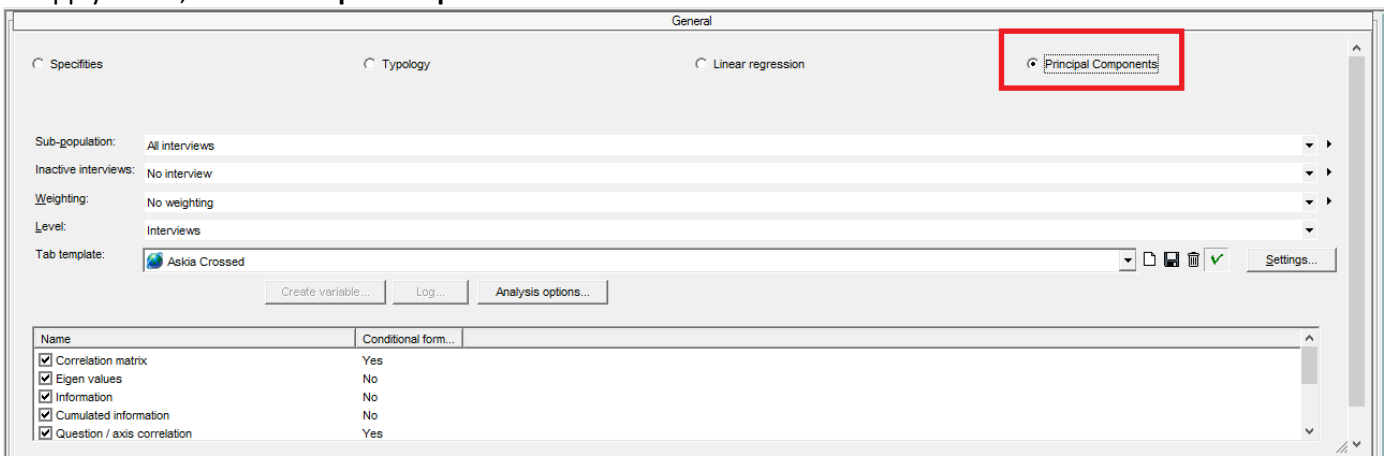
Doing the projection of each point on each axis, we get the coordinates of the points.
We then calculate correlations of questions to axis. The result allows a simultaneous representation of interviews and questions.



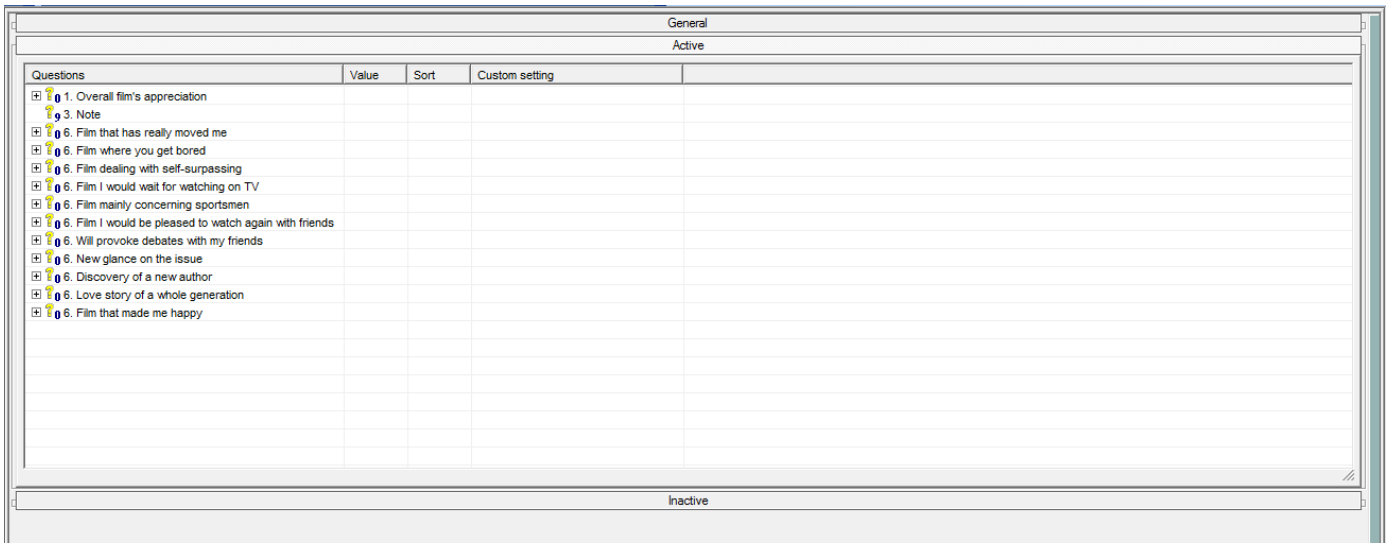
Note that PCA is sensitive to the relative scaling of the original variables.

How to do it:

To apply a **PCA**, select **Principal Components** in the General section:



Add Numeric questions (numeric or scale responses) in Active section



NB: if you want to display some other numeric variables but not include them in the calculations, add them in Inactive section

Then click on 

You will obtain as many pages as selected calculations:

The output

The correlation Matrix

A Correlation matrix describes correlation among *p* variables. It is a square symmetrical MxM matrix with the (ij)th element equal to the correlation coefficient r_{ij} between the (i)th and the (j)th variable. The diagonal elements (correlations of variables with themselves) are always equal to 1.00. The correlations value are always included between – 1.00 $< r_{values} <$ + 1.00.

Correlation matrix	3. Note	1. Overall film's appreciation	6. Film that has really moved me	6. Film where you get bored	6. Film dealing with self-surpassing	6. Film I would wait for watching on TV	6. Film mainly concerning sportsmen	6. Film I would be pleased to watch again with friends	6. Will provoke debates with my friends	6. New glance on the issue	6. Discovery of a new author	6. Love story of a whole generation	6. Film that made me happy
3. Note	1,00	0,08	0,02	0,18	-0,07	0,13	0,07	-0,04	-0,05	0,15	0,20	0,03	-0,03
1. Overall film's appreciation	0,08	1,00	-0,19	-0,08	-0,03	-0,03	-0,01	-0,10	-0,02	0,07	-0,14	0,05	0,08
6. Film that has really moved me	0,02	-0,19	1,00	-0,14	0,21	0,00	0,19	0,07	0,04	-0,11	0,13	0,06	0,11
6. Film where you get bored	0,18	-0,08	-0,14	1,00	-0,10	0,01	0,08	-0,06	0,17	0,20	-0,02	-0,07	-0,06
6. Film dealing with self-surpassing	-0,07	-0,03	0,21	-0,10	1,00	0,05	-0,12	-0,11	0,05	-0,08	0,13	-0,05	0,03
6. Film I would wait for watching on TV	0,13	-0,03	0,00	0,01	0,05	1,00	-0,21	0,09	-0,01	0,16	0,11	-0,02	0,00
6. Film mainly concerning sportsmen	0,07	-0,01	0,19	0,08	-0,12	-0,21	1,00	0,08	0,11	-0,09	-0,02	0,13	-0,06
6. Film I would be pleased to watch again with friends	-0,04	-0,10	0,07	-0,06	-0,11	0,09	0,08	1,00	-0,04	0,04	-0,09	0,02	-0,14
6. Will provoke debates with my friends	-0,05	-0,02	0,04	0,17	0,05	-0,01	0,11	-0,04	1,00	-0,16	0,07	0,09	0,13
6. New glance on the issue	0,15	0,07	-0,11	0,20	-0,08	0,16	-0,09	0,04	-0,16	1,00	0,02	-0,01	-0,11
6. Discovery of a new author	0,20	-0,14	0,13	-0,02	0,13	0,11	-0,02	-0,09	0,07	0,02	1,00	0,06	-0,03
6. Love story of a whole generation	0,03	0,05	0,06	-0,07	-0,05	-0,02	0,13	0,02	0,09	-0,01	0,06	1,00	0,07
6. Film that made me happy	-0,03	0,08	0,11	-0,06	0,03	0,00	-0,06	-0,14	0,13	-0,11	-0,03	0,07	1,00

The Eigen values and information

The Eigenvalues (λ) are a special set of scalars associated with a linear system of equations.
The information is the percentage supported by each axis, calculated as follow:

$$I = \frac{\lambda}{\sum_{i=1}^{p} \lambda}$$

N.B. The 1st axis has always the higher percentage
The cumulated information is the cumulated percentage of information

	Eigen values	Information	Cumulated information
Axis 1	1,63	12,5%	12,5%
Axis 2	1,45	11,2%	23,7%
Axis 3	1,39	10,7%	34,4%
Axis 4	1,30	10,0%	44,4%
Axis 5	1,15	8,8%	53,2%
Axis 6	1,04	8,0%	61,2%
Axis 7	0,91	7,0%	68,3%
Axis 8	0,87	6,7%	74,9%
Axis 9	0,84	6,5%	81,4%
Axis 10	0,68	5,2%	86,6%
Axis 11	0,66	5,1%	91,7%
Axis 12	0,57	4,4%	96,1%
Axis 13	0,51	3,9%	100,0%

The question/ axis correlation

Describes the correlation between questions and axis.

Question / axis correlation	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10	Axis 11	Axis 12	Axis 13
3. Note	-0,34	0,47	0,33	0,19	-0,28	-0,24	0,23	0,02	-0,32	-0,15	-0,34	-0,18	-0,22
1. Overall film's appreciation	-0,20	-0,34	-0,10	0,47	-0,47	-0,10	0,04	-0,48	-0,19	0,16	0,08	-0,01	0,28
6. Film that has really moved me	0,57	0,38	0,12	-0,27	-0,14	-0,07	0,41	-0,09	0,18	0,00	0,13	-0,35	0,27
6. Film where you get bored	-0,40	0,14	0,49	0,24	0,51	0,07	0,11	-0,05	0,20	-0,11	-0,18	0,14	0,37
6. Film dealing with self-surpassing	0,39	0,36	-0,40	0,09	0,16	-0,19	-0,03	-0,54	0,22	-0,08	-0,27	0,23	-0,13
6. Film I would wait for watching on TV	-0,28	0,51	-0,27	-0,04	-0,12	0,52	0,07	-0,11	-0,18	-0,36	0,33	0,14	0,03
6. Film mainly concerning sportsmen	0,24	-0,13	0,73	-0,15	-0,13	-0,22	0,14	-0,18	-0,03	-0,08	0,31	0,36	-0,17
6. Film I would be pleased to watch again with friends	-0,08	-0,05	0,14	-0,68	-0,15	0,44	0,07	-0,21	-0,18	0,30	-0,33	0,12	0,05
6. Will provoke debates with my friends	0,32	0,07	0,36	0,38	0,35	0,45	-0,20	-0,28	-0,16	0,17	0,09	-0,28	-0,20
6. New glance on the issue	-0,64	0,22	0,02	-0,02	-0,12	0,02	0,11	-0,13	0,55	0,33	0,18	-0,08	-0,21
6. Discovery of a new author	0,12	0,70	0,06	0,08	-0,07	-0,20	-0,36	0,20	-0,19	0,40	0,11	0,19	0,17
6. Love story of a whole generation	0,20	0,02	0,31	0,13	-0,58	0,26	-0,45	0,09	0,39	-0,22	-0,19	-0,01	0,05
6. Film that made me happy	0,34	-0,02	-0,10	0,51	-0,15	0,33	0,50	0,32	0,07	0,20	-0,11	0,24	-0,07

The interview coordinates

Gives the coordinates per interview on each axis

Interview coordinates	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10	Axis 11	Axis 12	Axis 13
Interview 1	-1,05	-0,31	-0,01	1,41	2,04	0,37	-1,76	-0,84	0,00	1,42	1,07	0,81	1,21
Interview 2	-0,15	0,69	0,98	-1,35	-0,74	-0,45	0,47	0,86	1,89	-1,86	0,50	0,01	0,79
Interview 3	0,95	0,55	-0,72	-0,13	0,13	0,92	-1,11	0,59	-0,59	-0,57	0,40	-1,97	-0,51
Interview 4	-1,84	2,09	-0,80	0,58	0,43	-0,75	1,95	0,62	-0,15	0,44	-0,39	-0,70	0,01
Interview 5	-0,88	-0,01	0,93	-1,12	0,65	-0,01	0,26	0,47	1,29	-0,31	-1,04	-0,39	1,05
Interview 6	0,05	1,01	-0,30	1,53	1,16	0,72	0,16	0,09	2,00	-0,13	-0,13	-1,39	-0,58
Interview 7	-0,87	-2,88	-1,35	-0,96	1,16	-0,74	0,71	-0,87	-0,28	0,76	-1,64	1,62	0,96
Interview 8	-0,03	-0,23	0,67	-1,53	-0,52	0,41	-0,33	-0,18	-0,13	-0,23	0,44	-2,16	0,73
Interview 9	-0,81	-0,19	0,78	-1,44	-0,39	-0,19	-0,46	-1,55	-1,18	1,20	0,65	-0,17	0,15
Interview 10	-1,51	-2,45	-2,27	-1,32	-0,63	0,24	-0,22	-0,40	-0,02	0,01	0,93	0,87	0,00
Interview 11	3,18	0,10	-1,08	-0,32	-0,51	1,13	0,23	0,33	-0,22	0,58	-0,16	0,19	0,63
Interview 12	0,51	1,02	0,44	-0,95	-0,80	0,31	-1,73	1,34	-0,99	-1,63	0,48	1,43	0,20
Interview 13	-2,01	0,92	0,18	-0,21	1,97	-1,42	0,04	0,97	0,48	1,18	0,28	0,42	0,50
Interview 14	0,16	-1,82	-0,99	-0,17	0,82	-2,08	-0,56	-0,57	0,99	0,14	1,05	-0,94	1,40
Interview 15	-1,11	-1,17	-1,76	0,69	0,45	0,24	1,30	-0,53	-0,53	0,53	-0,19	-0,65	-0,39

The interview representation quality

Gives the quality of representation of each interview on each axis
N.B. The sum per interview (across percentage) is equal to 100%

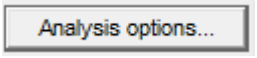
Interview representation quality	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10	Axis 11	Axis 12	Axis 13
Interview 1	6,7%	0,6%	0,0%	12,0%	25,1%	0,8%	18,7%	4,3%	0,0%	12,1%	6,9%	4,0%	8,8%
Interview 2	0,2%	3,6%	7,4%	14,1%	4,2%	1,6%	1,7%	5,8%	27,7%	26,9%	1,9%	0,0%	4,8%
Interview 3	9,9%	3,3%	5,6%	0,2%	0,2%	9,3%	13,4%	3,9%	3,7%	3,6%	1,7%	42,5%	2,8%
Interview 4	23,3%	30,1%	4,4%	2,3%	1,3%	3,9%	26,1%	2,6%	0,1%	1,3%	1,0%	3,4%	0,0%
Interview 5	10,1%	0,0%	11,3%	16,3%	5,5%	0,0%	0,9%	2,8%	21,7%	1,2%	14,0%	2,0%	14,3%
Interview 6	0,0%	8,7%	0,8%	20,0%	11,6%	4,5%	0,2%	0,1%	34,4%	0,2%	0,1%	16,6%	2,9%
Interview 7	3,5%	38,0%	8,3%	4,2%	6,2%	2,5%	2,3%	3,4%	0,3%	2,6%	12,4%	12,0%	4,2%
Interview 8	0,0%	0,6%	5,1%	26,2%	3,0%	1,9%	1,2%	0,4%	0,2%	0,6%	2,2%	52,7%	5,9%
Interview 9	6,9%	0,4%	6,4%	21,8%	1,6%	0,4%	2,2%	25,5%	14,7%	15,2%	4,4%	0,3%	0,2%
Interview 10	13,1%	34,3%	29,6%	10,0%	2,2%	0,3%	0,3%	0,9%	0,0%	0,0%	5,0%	4,4%	-
Interview 11	72,5%	0,1%	8,4%	0,7%	1,8%	9,2%	0,4%	0,8%	0,3%	2,4%	0,2%	0,3%	2,8%
Interview 12	1,9%	7,6%	1,4%	6,5%	4,6%	0,7%	21,6%	12,9%	7,0%	19,1%	1,7%	14,7%	0,3%
Interview 13	29,0%	6,0%	0,2%	0,3%	28,0%	14,5%	0,0%	6,8%	1,7%	9,9%	0,5%	1,3%	1,8%
Interview 14	0,2%	22,2%	6,5%	0,2%	4,5%	29,0%	2,1%	2,1%	6,6%	0,1%	7,3%	6,0%	13,1%
Interview 15	12,8%	14,3%	32,4%	5,0%	2,1%	0,6%	17,5%	2,9%	2,9%	3,0%	0,4%	4,4%	1,6%

The interview contribution

Gives the contribution of the interview for one axis.
N.B. The sum per axis (down percentage) is equal to 100%

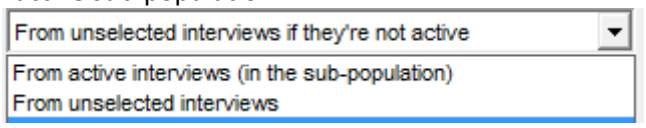
Interview contribution	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10	Axis 11	Axis 12	Axis 13
Interview 1	0,2%	0,0%	-	0,5%	1,2%	0,0%	1,2%	0,3%	-	1,0%	0,6%	0,4%	1,0%
Interview 2	0,0%	0,1%	0,2%	0,5%	0,2%	0,1%	0,1%	0,3%	1,5%	1,7%	0,1%	-	0,4%
Interview 3	0,2%	0,1%	0,1%	0,0%	0,0%	0,3%	0,5%	0,1%	0,1%	0,2%	0,1%	2,4%	0,2%
Interview 4	0,7%	1,0%	0,2%	0,1%	0,1%	0,2%	1,4%	0,2%	0,0%	0,1%	0,1%	0,3%	-
Interview 5	0,2%	-	0,2%	0,3%	0,1%	-	0,0%	0,1%	0,7%	0,0%	0,6%	0,1%	0,7%
Interview 6	0,0%	0,2%	0,0%	0,6%	0,4%	0,2%	0,0%	0,0%	1,6%	0,0%	0,0%	1,2%	0,2%
Interview 7	0,2%	2,0%	0,4%	0,2%	0,4%	0,2%	0,2%	0,3%	0,0%	0,3%	1,4%	1,6%	0,6%
Interview 8	0,0%	0,0%	0,1%	0,6%	0,1%	0,1%	0,0%	0,0%	0,0%	0,0%	0,1%	2,8%	0,4%
Interview 9	0,1%	0,0%	0,1%	0,5%	0,0%	0,0%	0,1%	1,0%	0,6%	0,7%	0,2%	0,0%	0,0%
Interview 10	0,5%	1,4%	1,3%	0,5%	0,1%	0,0%	0,0%	0,1%	0,0%	-	0,5%	0,5%	-
Interview 11	2,1%	0,0%	0,3%	0,0%	0,1%	0,4%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,3%
Interview 12	0,1%	0,2%	0,0%	0,2%	0,2%	0,0%	1,1%	0,7%	0,4%	1,3%	0,1%	1,2%	0,0%
Interview 13	0,8%	0,2%	0,0%	0,0%	1,2%	0,7%	0,0%	0,4%	0,1%	0,7%	0,0%	0,1%	0,2%

The interview management

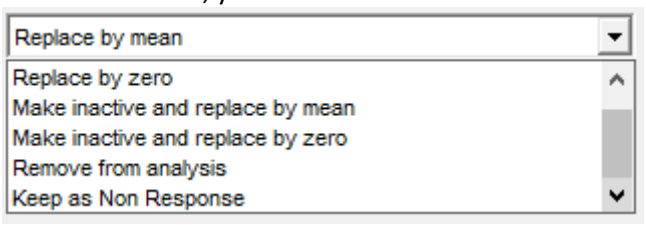
In 

you can choose how to manage the inactive interviews. (You might want to display them but not include the values given by these interviews.)

Inactive interviews:
By default, we apply the "unselected interviews if they're not active". But note that you can specify a sub-population in general section/Inactive sub-population.



Missing values
You have to specify what to do for the missing values.
Depending on the value included in the variable, you can define:

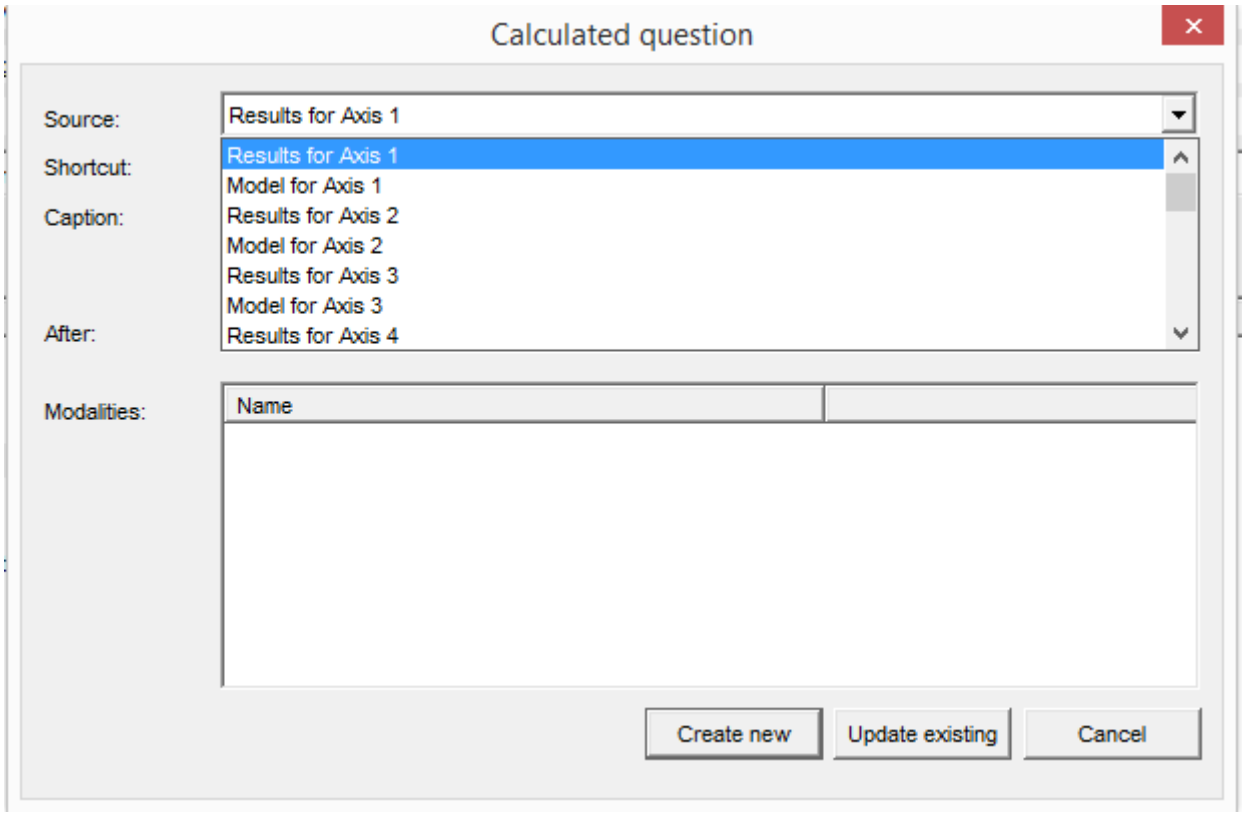


- All missing values will be replaced by the mean (by default): **a respondent's missing values will be replaced by the mean.** The interview will remain active.
- Replace by zero: **a respondent's missing values will be replaced by 0.** The interview will remain active.
- Make inactive and replace by the mean: **all values given by the respondent will be transformed into the mean; this mean will replace a respondent's missing values.** The interview will be inactive.
- Make inactive and replace by 0: **a respondent's missing values will be replaced by 0 .**The interview will be inactive.
- Remove from analysis: **The interview will be removed from active interviews if the respondent has given at least one missing value.**
- Keep as Non response: **only the missing value of the question will be converted into DK, the other values remain**

Create a variable

You can create a variable from the results for:

- Axis: the value of the contribution of each interview to the axis p
- Model: you will save the options of your PCA


Calculated question
Source:
Results for Axis 1
Shortcut:
Caption:
After:
Results for Axis 1
Model for Axis 1
Results for Axis 2
Model for Axis 2
Results for Axis 3
Model for Axis 3
Results for Axis 4
Modalities:
Name
Create new
Update existing
Cancel

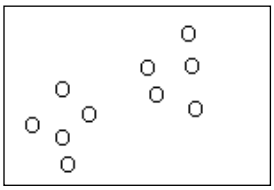
1.2.Typology

Definition

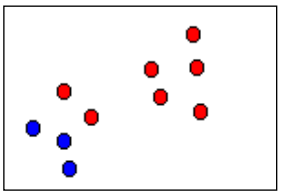
When we have a large volume of data, it can be useful to regroup similar interviews into groups (clusters) with similar objects, this is called **Cluster Analysis**. If the groups are very homogeneous, then the analysis of each group will sufficiently describe the population.

Just as for multivariate analysis, there are a great number of variants of the same method called the **k-means**

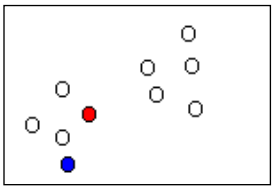
1) We start by arbitrarily choosing a number *k* of groups (we will return later on the methods of choosing this parameter) and we randomly select *n* prototype interviews from the population.
2) *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
3) We calculate the barycentre of each of the *n* groups.
4) We then repeat step 2) with the *n* average respondents instead of the prototype interviews. We continue this process until we reach a stable partition.



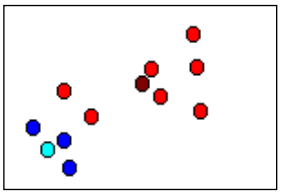
1) We look for a partition in 2 classes



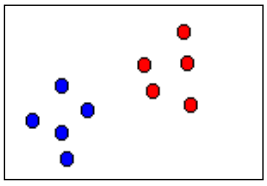
3) We associate the points to their closest centres



2) We randomly select 2 prototypes



4) We calculate the barycentres



5) We obtain a stable partition

The methods depend on the distance metrics used:

- Euclidian for numeric questions and
- Chi² for counts....

Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.
The number of clusters *k* is an input parameter: an inappropriate choice of *k* may yield poor results.
To measure the partition's quality, we calculate each scatter's inertia. We know, thanks to Huygens theory, that regardless of the partition, the sum of the Intra Group Variance and the Inter Group Variance is constant.

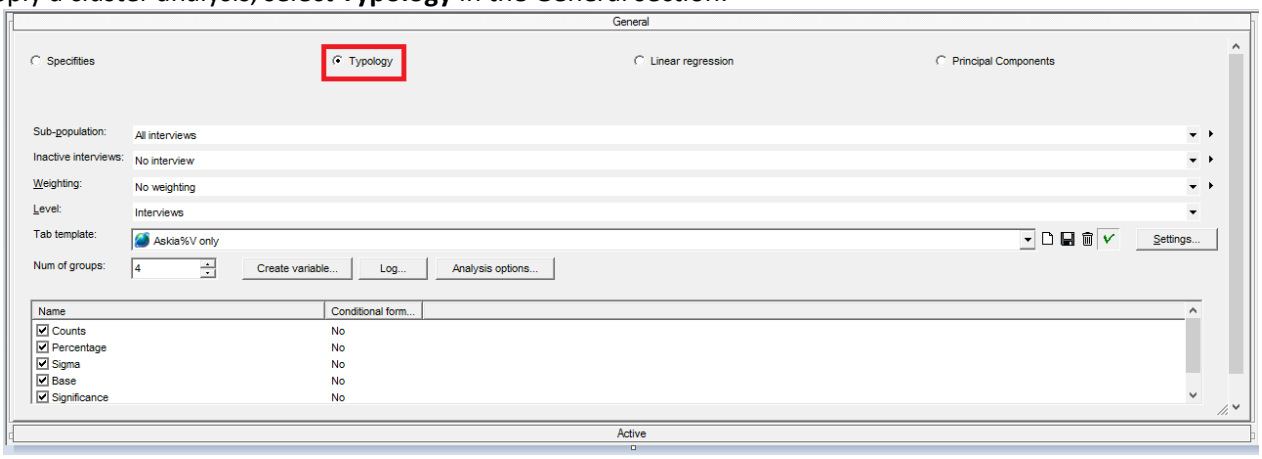
By calculating

$$info\ \% = \frac{Variance\ between\ clusters}{(Variance\ between\ clusters + variance\ within\ cluster)}$$

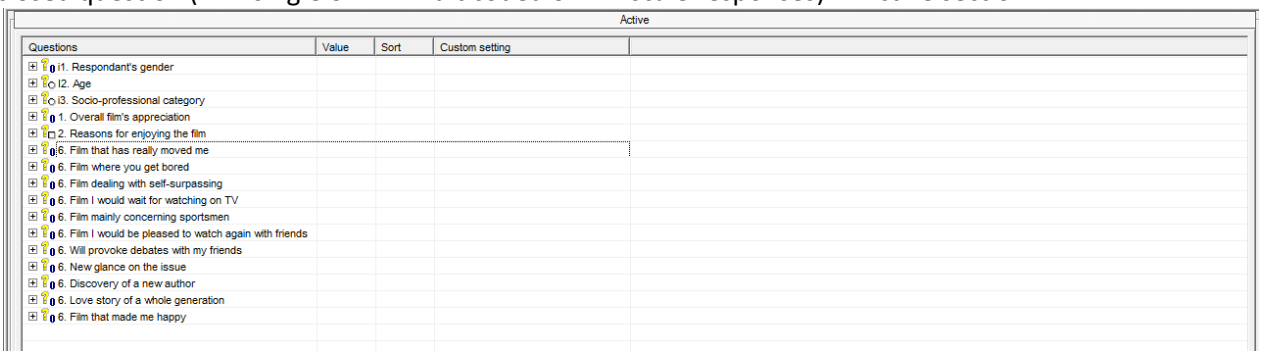
We will obtain the percentage of information remaining in spite of the regrouping. By applying a few typology tests with a different number of groups, we will be able to choose the most judicious partition.

How to do it:

To apply a cluster analysis, select **Typology** in the General section:

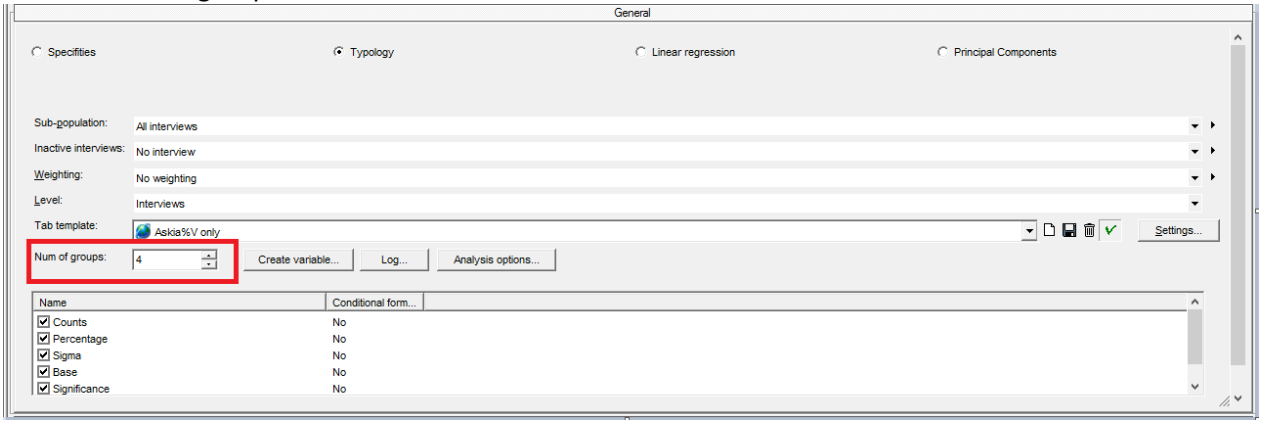


Add closed question (single or multicoded or scale responses) in Active section



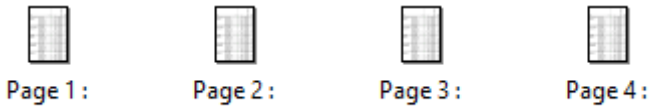
NB: if you want to display some other closed variables but not include them in the calculations, add them to the Inactive section

Select the number of groups:



Then click on 

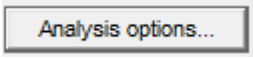
You will obtain as many pages as groups:



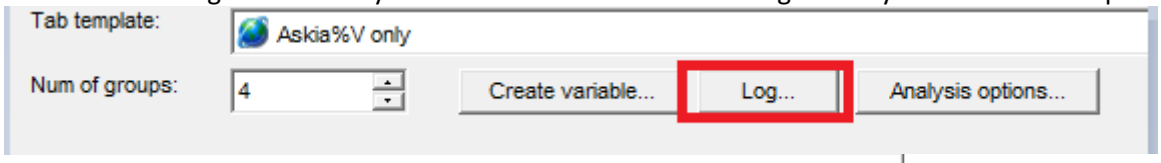
The output

By double clicking on page X, you will obtain the counts, percentage, Sigma, Base and significance within the group X.
In this example, the Group 1 (1) has 111 interviews, which represents 38% of the total base 292 (2).
The answer "rather no" of question Appreciation "Film I would be pleased to watch with friends" represents 66% among the total 97 in the whole population. This item is the first item which explains the Group 1 with a sigma* of 7.28 (4) at 100% (5)

Group 1 (1)		Counts	Percentage	Sigma	Base	Significance
Total		111	38,0%		292	(2)
6. Appreciation (Film I would be pleased to watch again with friends)	Rather No	64 (3)	66,0%	7,28 (4)	97	100,0% (5)
1. Appreciation	Not very much	49	65,3%	5,99	75	100,0%

*see Significancy test user guide for the calculation and output method in 

There is also a Log available to you in the General Section. This log allows you to review the quality of the Typology.



At the end of the log file, you will see the typology selected as the best run 

Run #9

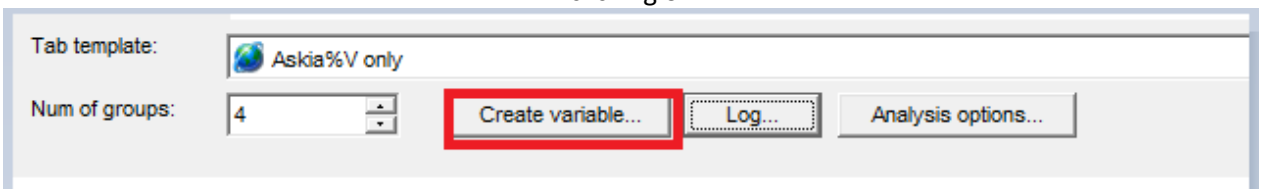
Iteration 6

Inertia 2228,84 (14% information retained)

Group	Size	Inertia %	Signature
1	111	39	01e7
2	49	16	00ed
3	74	25	0070
4	58	17	017a

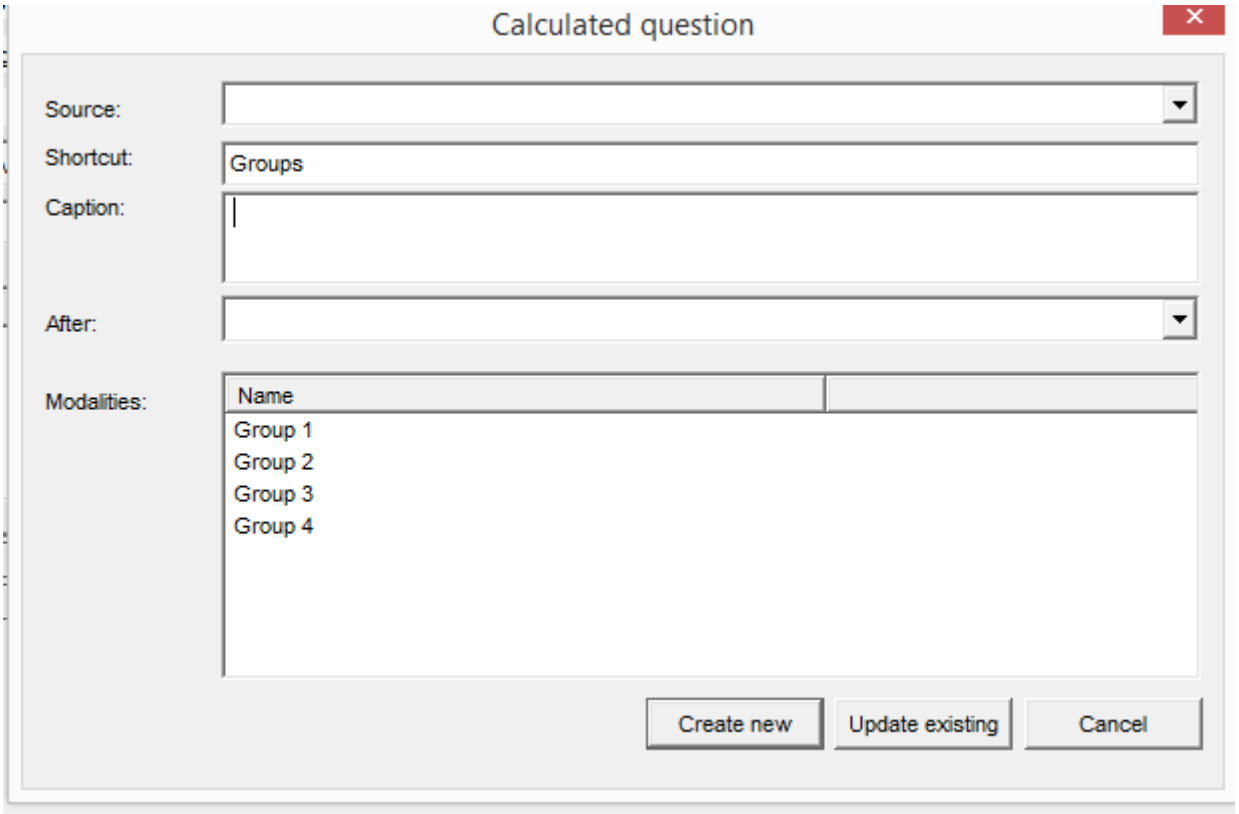
Create a variable

If you want to keep these groups and use them as a break in a tab definition, you can create a closed question by clicking on

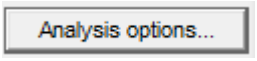


You can create a new one or update existing one.

Name it in 



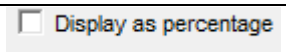
The interview management

In 

you can choose how to manage the inactive interviews. (You might want to display them but not include the values given by these interviews.)

Inactive interviews	Select the method to apply to inactive interview: • **From active interviews (in sub-population)**: the selected sub-population selects interviews from the active interviews and makes them inactive • **From unselected interviews**: the interviews that are in the active and inactive sub-population are marked as inactive • **From unselected interviews (if they are not active):** the interviews that are in the active and inactive sub-population are marked as active
Typology to run	The maximum number to run
Selection method	• **Typology with the maximum intergroup inertia:** • **Most frequent repartition:** pick the selection that appears the most often • **Most frequent then use lowest inertia** pick the selection that appears the most often, if there are equals, use inertia to pick the best
Converge when	• **Stop when the groups are stable** • **Stop when inertia variation is below** **Inertia variation %**: Insert the value
Max number of iterations	Insert the value (by default: 30)
Random seed	Is the seed selection for prototype interviews (by default: 10)

In the output, you can select the methods to calculate the sigma:

Sigma method	For the formulas (*see Significancy test user guide*): • **Using Chi²** • **Against all other columns with the unknown standard deviation** • **Against all other columns with the known standard deviation**
Sigma threshold	It could be the a value as: 1,64 – 1,96 – 2,58  or a percentage as 90% -95%- 99% 
Sort modalities	To order by count the items within the group
Show significantly low responses	The responses that do not explain the group will be added at the bottom of result page. Note that the sigma and the significancy percentage will be negative in the output.

1.3.Specificities

Definition

The descriptive analysis allows one **to prioritise the variables** that best describe a category of a population.

It is used on **closed variables** (single or multicoded or scale responses) and when we have *n* > 30 in each category

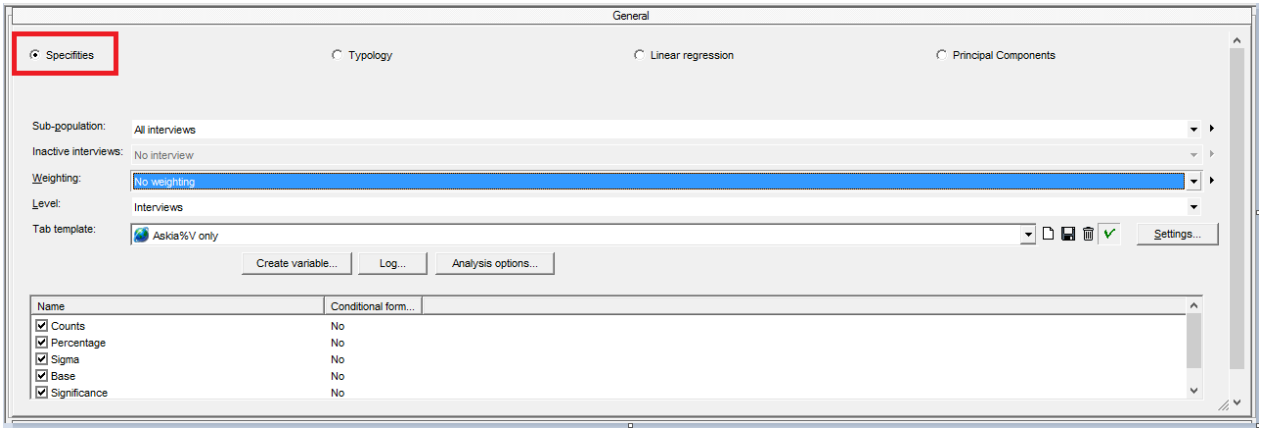
The higher the probability, the more the variable contributes (chi²) to this category The more the group depends on this category, the more this category will explain this group.
We obtain the % of each response in the population (group). This observation is normalised and displayed on the sigma (Z) and the probability
In terms of interpretation, the higher the sigma (> + 2), the more the modality explains the group (= high probability 100%).
When sigma is < -2, this means that this method does not explain the whole group (we also found a small number of people who have answered this code in the group)

$$S = \frac{(pi - p)}{\sqrt{\frac{p\,(1-p)}{N}}}$$

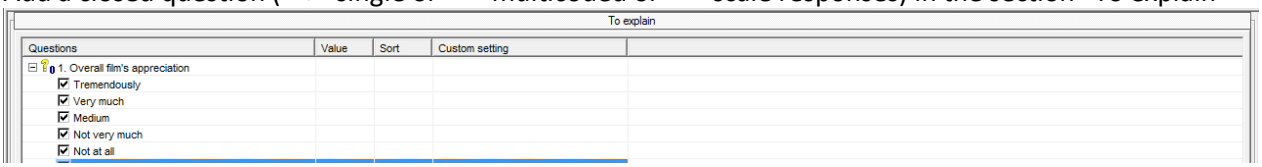
This statistic follows a standard normal law.

How to do it:

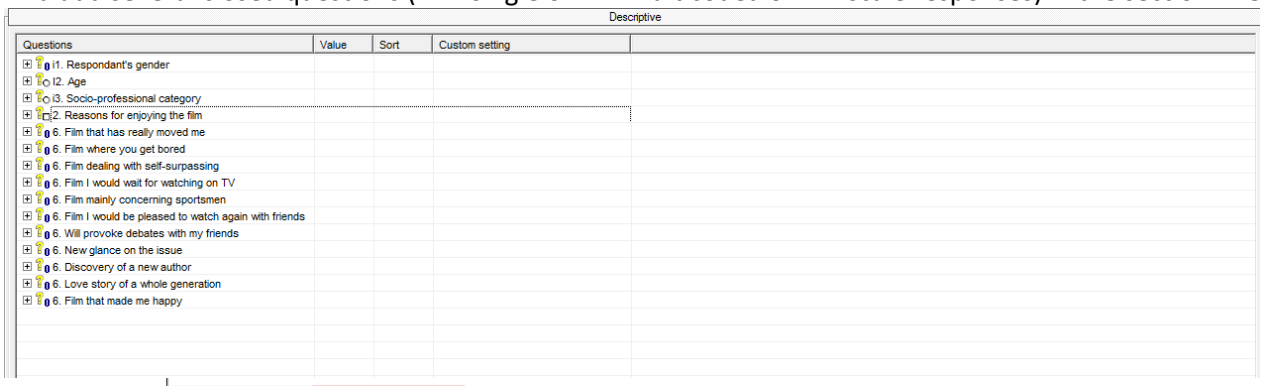
To prioritize the variables within the categories, select **Specificities** in the General section:



Add a closed question (single or multicoded or scale responses) in the section “To explain”

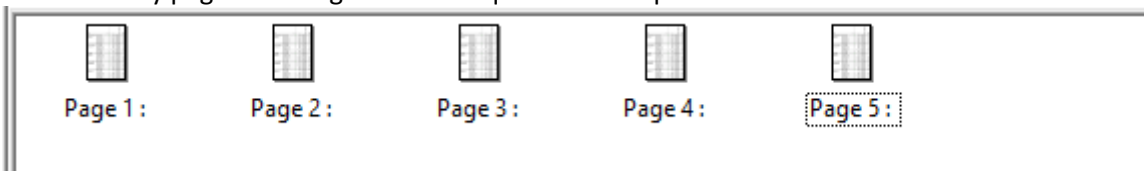


And add several closed questions (single or multicoded or scale responses) in the section “Descriptive“



Then click on 

You will obtain as many pages as categories in the question to explain:



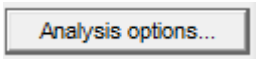
The output

By double clicking on page X, you will obtain the counts, percentage, Sigma, Base and significance within the group X.

In this example, we want to describe the category “Overall appreciation of the film: Not at all”

Not at all		Counts	Percentage	Sigma	Base	Significance
Total		28	11,0%		255	
6. Appreciation (Film I would wait for watching on TV)	Not at all	18	38,3%	5,20	47	100,0%
6. Appreciation (Love story of a whole generation)	Absolutely	13	24,5%	3,24	53	99,9%
6. Appreciation (Discovery of a new author)	Rather No	13	24,1%	2,78	54	99,5%
6. Appreciation (Will provoke debates with my friends)	Rather No	15	24,2%	2,69	62	99,3%
6. Appreciation (Film mainly concerning sportsmen)	Rather yes	13	22,8%	2,47	57	98,6%
6. Appreciation (Film that made me happy)	Rather No	19	21,1%	2,39	90	98,3%
i3. Profession	Worker, Employee	12	21,4%	2,39	56	98,3%
6. Appreciation (Film where you get bored)	Not at all	9	18,4%	2,30	49	97,8%
i3. Profession	Inactive, Jobless, Retired	6	26,1%	2,20	23	97,2%
6. Appreciation (Film that has really moved me)	Not at all	-	-	-2,03	37	-95,7%
i3. Profession	Managerial Staff, Executive Manager, Liberal Profession	1	1,7%	-2,18	59	-97,1%
6. Appreciation (Discovery of a new author)	Not at all	-	-	-2,70	62	-99,3%
6. Appreciation (Film I would wait for watching on TV)	Rather yes	1	1,1%	-3,15	93	-99,8%

Note that the first item which explains this category is “Film I would wait for watching on TV: Not at all”: 38, 3% of 255 (total population) are in this category. (We are sure at 100% because the sigma is high (5.20).

**see Significancy test user guide for the calculation and output method in* 

In the output, you can select the methods to calculate the sigma:

Sigma method	For the formulas (*see Significancy test user guide*): • **Using Chi²** • **Against all other columns with the unknown standard deviation** • **Against all other columns with the known standard deviation**
Sigma threshold	It could be a value such as: 1,64 – 1,96 – 2,58 ☐ Display as percentage or a percentage such as 90% -95%- 99% 95 ☑ Display as percentage
Sort modalities	To order by count, the items within the group
Show significancy low responses	The responses that do not explain the group will be added at the bottom of result page. Note that the sigma and the significancy percentage will be negative in the output.

1.4.Linear regression

Definition

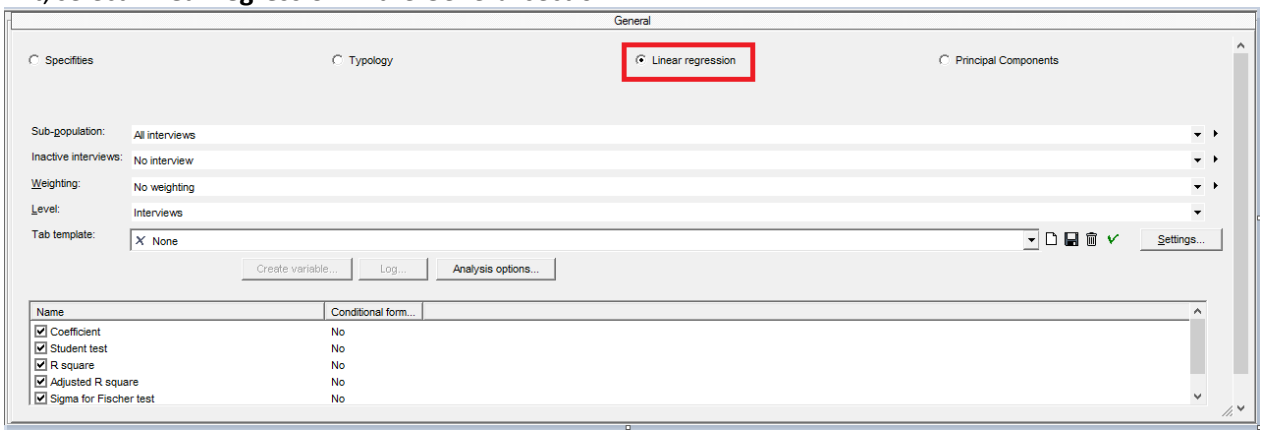
The aim of this analysis is to create a model which estimates the target question (To explain) by explanatory variables

It is to explain a **numeric** variable (numeric or scale responses)- e.g the question . ??3. Note?? by numeric explanatory variable: (e.g. item's appreciation)

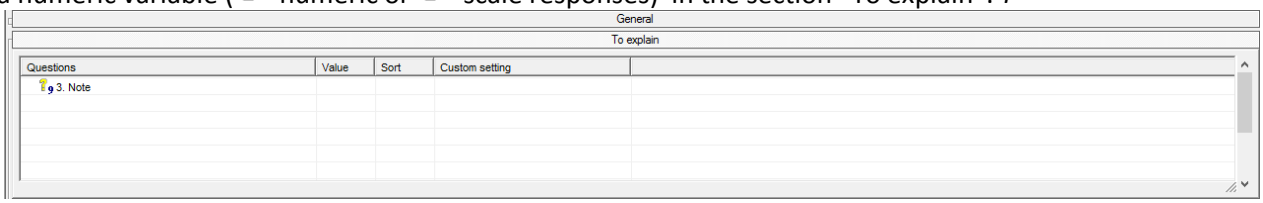
$$Y = a_i + cste$$

How to do it:

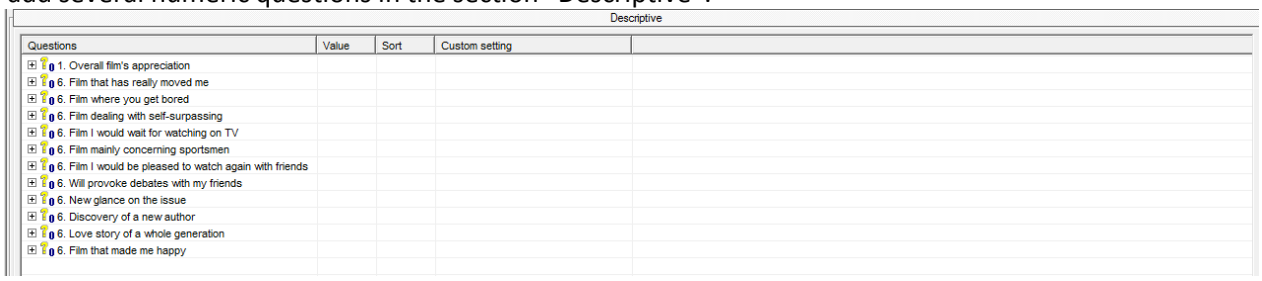
To run it, select **Linear regression** in the General section:



Add a numeric variable (numeric or scale responses) in the section "To explain": *Y*



Then add several numeric questions in the section "Descriptive":



Then click on 

You will obtain one single table:

The output

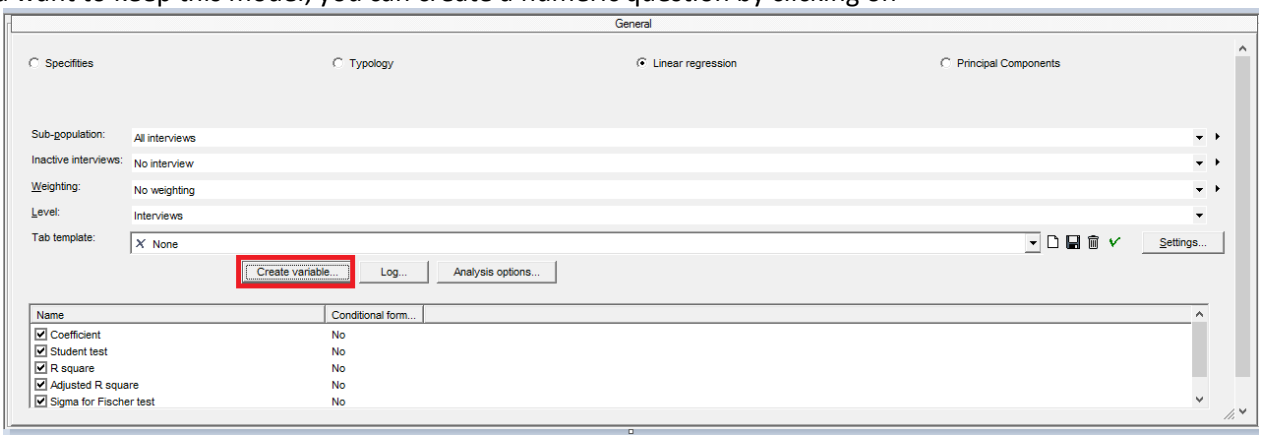
	Coefficient	Student test
Constant	2,99	100,0%
1. Overall film's appreciation	0,09	99,8%
6. Film that has really moved me	0,03	77,2%
6. Film where you get bored	0,12	100,0%
6. Film dealing with self-surpassing	-0,04	91,3%
6. Film I would wait for watching on	0,08	99,6%
6. Film mainly concerning sportsmen	0,06	95,8%
6. Film I would be pleased to watch	-0,02	53,0%
6. Will provoke debates with my	-0,06	97,7%
6. New glance on the issue	0,05	94,3%
6. Discovery of a new author	0,12	100,0%
6. Love story of a whole generation	0,01	23,9%
6. Film that made me happy	0,00	4,6%
R square	0,13	
Adjusted R square	0,09	
Sigma for Fischer test	3,30	

You will obtain:

- the coefficient a_i of the model $Y = a_i + cste$
- the constant of the model $Y = a_i + cste$
- The probability given by the Student test (when it's close to 100% , the item explains fully the variable)
- The R^2: (the coefficient of determination indicates how well data fit the statistical model. The coefficient lies between 0 and 1)
- The adjusted R^2 ($\overline{R^2}$): (measures the variability of *Y* when we increase the number of explanatory variables. The adjusted R^2 lies between 0 and 1))
- The Sigma for Fischer test

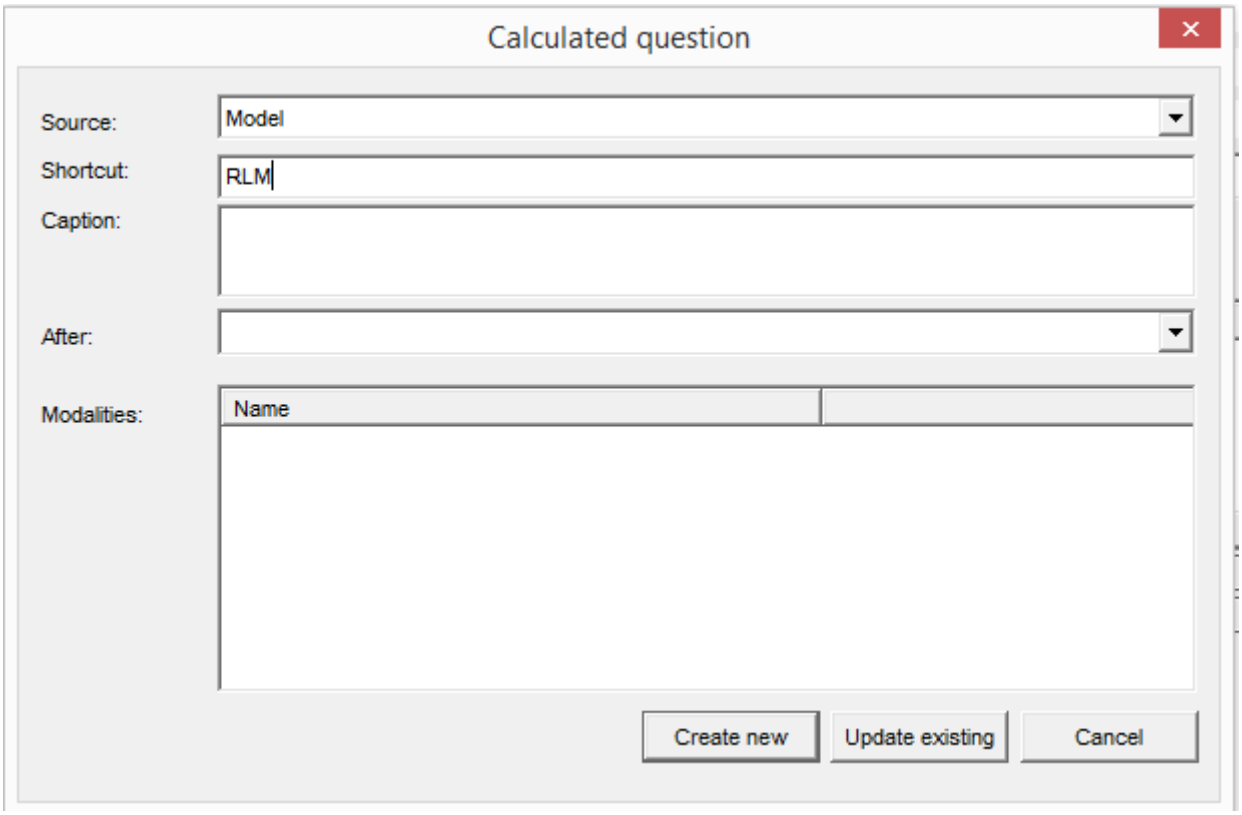
Create a variable

If you want to keep this model, you can create a numeric question by clicking on

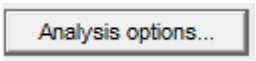


You can create a new one or update existing one.

Name it in 



The interview management

In 

you can choose how to manage the inactive interviews. (You might want to display them but not to include the values given by these interviews.)

Inactive interviews	Select the method to apply to inactive interview: • **From active interviews (in sub-population)**: the selected sub-population selects interviews from the active interviews and makes them inactive • **From unselected interviews**: the interviews that are in the active and inactive sub-population are marked as inactive • **From unselected interviews (if they are not active):** the interviews that are in the active and inactive sub-population are marked as active
Missing values	• **Replace by mean** • **Replace by 0** • **Make inactive and replace by mean** • **Make inactive and replace by 0** • **Remove from analysis** (at least on DK/NA at one item, the whole interview will be removed)

.

2. Lexicometry

2.1.Definition

Lexicometry is a method used in linguistics and statistics. It measures the frequency with which words occur in text. It's only focused on frequencies and sequences of words, and not on their meaning.

We will speak about:

- **Corpus**: A large and structured set of words.
- **Form**: Smallest unit (**the word**)
- Separator: Set of characters indicating the end of a word.
- **Repeated segment**: Set of consecutive forms found several times within the corpus
- **Frequency (of a form): Count** of the same form founded in the text
- **"Lemma"** is the **dictionary form** (the **Group)**,
  In English, for example, run, runs, ran and running are forms of the same lexeme, with run as the lemma.
- Hapax: Form of the frequency which is equal to 1
- Vocabulary: Set of distinct forms existing in the corpus
- **Specificities**: Correlation of the population with the form
- Word groups: Chart with different forms and form's size

To understand better, we will illustrate our purpose with the question ??5. Comments?? in the ex.qes .
We have 291 interviews and 266 comments (corpus)

We have:

T: size of the corpus (266)
V: number of forms (756 different words)
Hapax: number of forms stated once (503)
V_i: number of forms of i frequencies
F_{max}: the form with the highest frequency (émouvant = 52)

In order to simplify, we will apply a "lemmatisation" on this open-ended question by transforming all words into lemmas.

The adjectives, nouns, verbs, articles, singular, plural forms will be grouped:
e.g. Lemma Actors => acteur, acteurs, actrice, actrices
e.g Lemma Run => run, ran ,running , runs

When we have different languages, it's very difficult to do the lemmatisation automatically. It's the reason why it has not been implemented in askiaanalyse.
Therefore, you could do it through the Dictionary analyse: you visualize and add the forms into Group

4. Richness of a Vocabulary

Using the form number (V), hapax number and maximum frequency indicators enables us to measure the richness of a corpus.

In askiaanalyse, this analysis is named **Dictionary**.

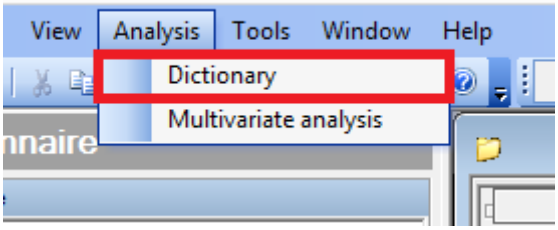
We will see also that we can create a variable by dictionary from the list of words and obtain the automatic codification of words.

Firstly, let's see the content of an open-ended question.

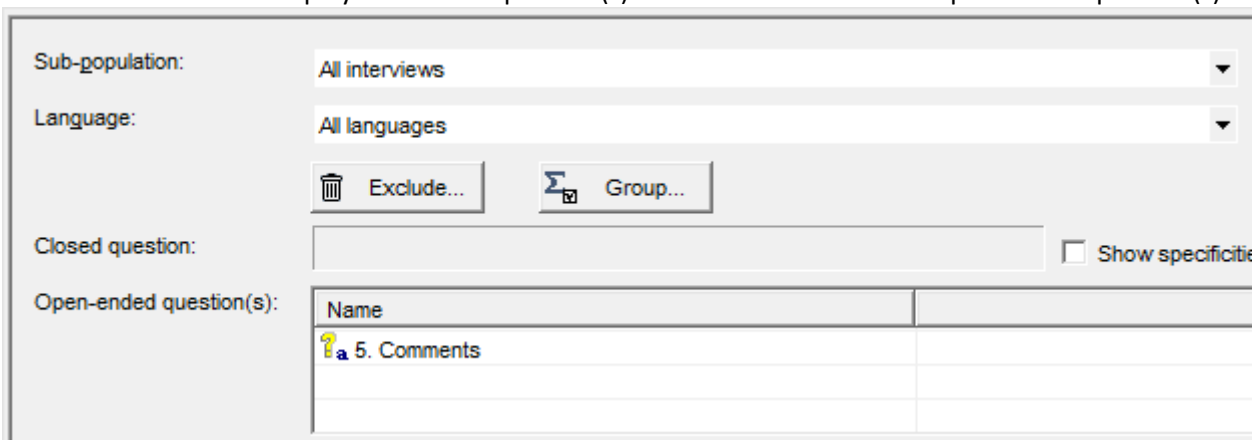
2.2.Dictionary

How to do it:

Select your open-ended question (s) in the tree view and then select Analysis in the tool bar menu



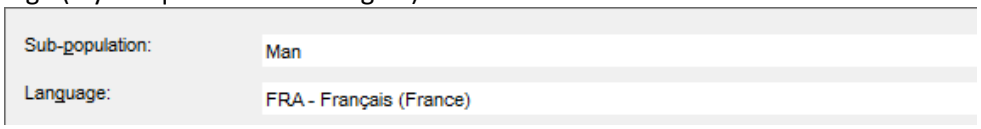
A new tab definition will be displayed with the question(s) selected in the window "Open-ended question(s)"



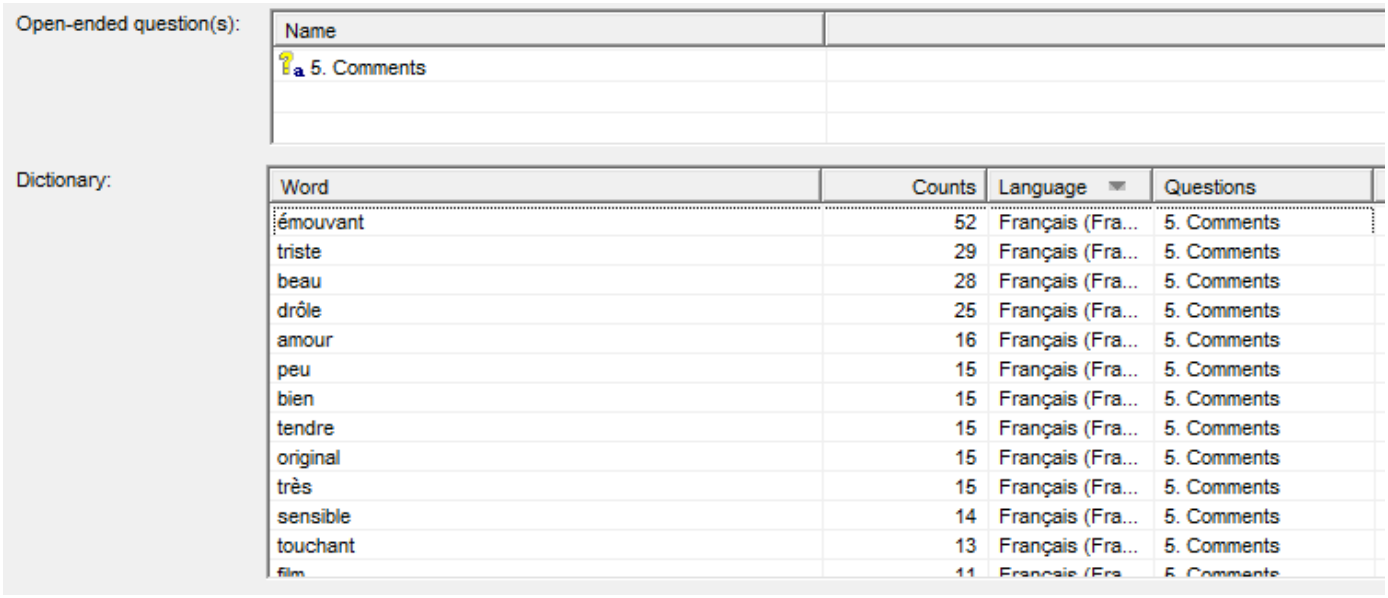
N.B. If you run the Dictionary analysis without any preliminary open-ended selection, you can drag and drop your question(s) directly into the "open-ended questions(s)" window.

You can apply some additional filters like:

- Sub – population
- Specific language (if your qes file is multilingual)



The system will list the words included in the verbatim order by the count:



Open-ended question(s):

Name
5. Comments

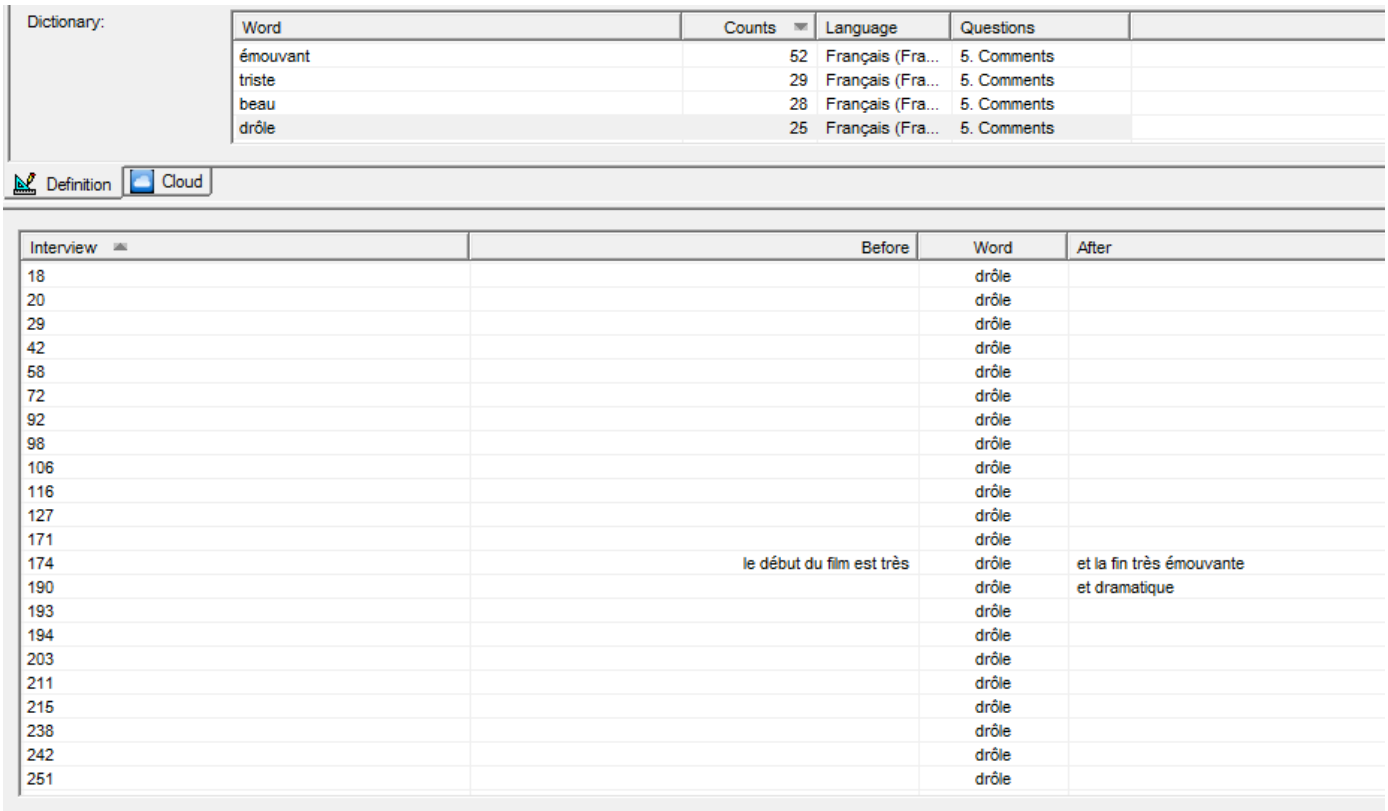
Dictionary:

Word	Counts	Language	Questions
émouvant	52	Français (Fra...	5. Comments
triste	29	Français (Fra...	5. Comments
beau	28	Français (Fra...	5. Comments
drôle	25	Français (Fra...	5. Comments
amour	16	Français (Fra...	5. Comments
peu	15	Français (Fra...	5. Comments
bien	15	Français (Fra...	5. Comments
tendre	15	Français (Fra...	5. Comments
original	15	Français (Fra...	5. Comments
très	15	Français (Fra...	5. Comments
sensible	14	Français (Fra...	5. Comments
touchant	13	Français (Fra...	5. Comments
film	11	Français (Fra...	5. Comments

Note that you can sort by Word (alphabetically), Counts, Language, Questions.

Concordance of a Form

It is always interesting to study the context in which a form is used. Instead of switching back and forth between the vocabulary and the corpus, we can visualise the location of the form in the corpus, the number of forms used “before” and “after” the form.



Dictionary:

Word	Counts	Language	Questions
émouvant	52	Français (Fra...	5. Comments
triste	29	Français (Fra...	5. Comments
beau	28	Français (Fra...	5. Comments
drôle	25	Français (Fra...	5. Comments

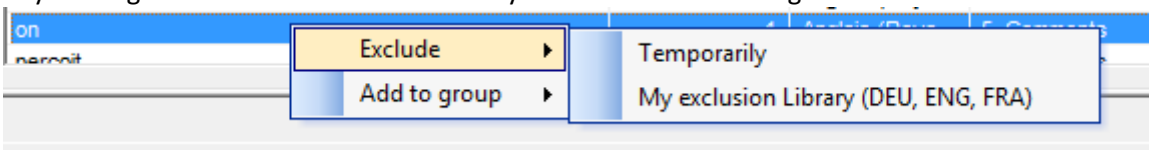
Definition | Cloud

Interview	Before	Word	After
18		drôle	
20		drôle	
29		drôle	
42		drôle	
58		drôle	
72		drôle	
92		drôle	
98		drôle	
106		drôle	
116		drôle	
127		drôle	
171		drôle	
174	le début du film est très	drôle	et la fin très émouvante
190		drôle	et dramatique
193		drôle	
194		drôle	
203		drôle	
211		drôle	
215		drôle	
238		drôle	
242		drôle	
251		drôle	

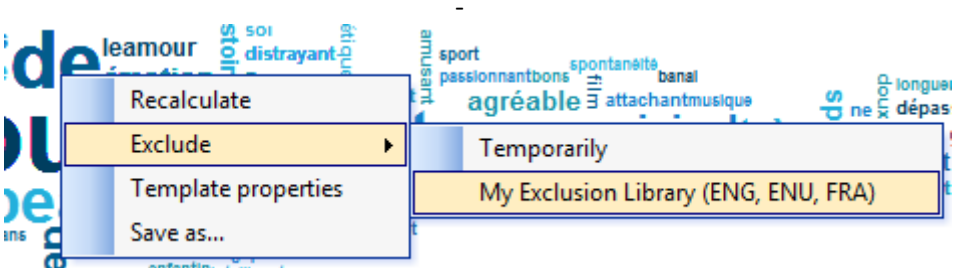
Exclude:

In order to simplify the view, you can exclude some words **temporarily** or using your **own exclusion library**:

- By adding directly in the text file (see below)
- Or by clicking on the word into the dictionary list and make a click right :

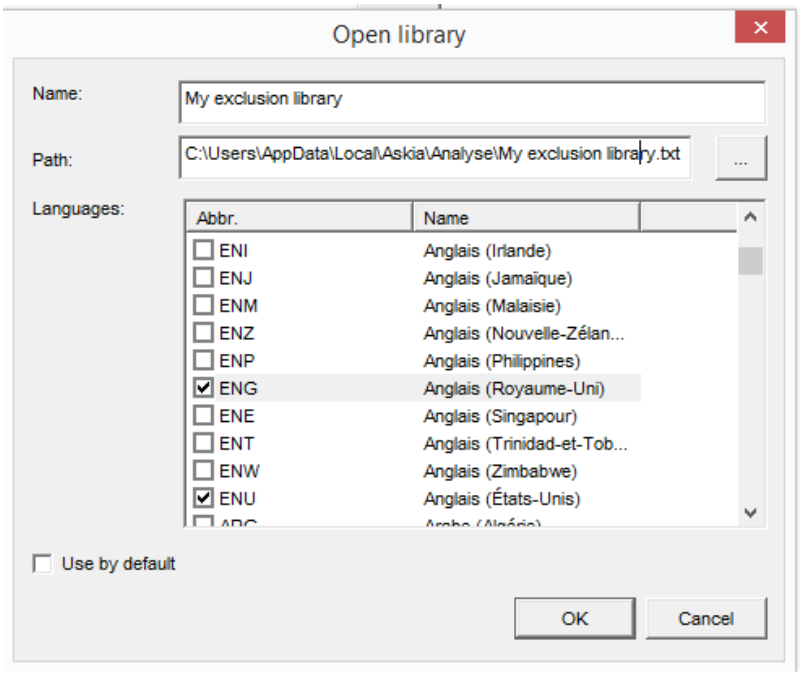


- By clicking on the word in the word cloud (see next chapter)



To create your own exclusion library, click on “Exclude” button

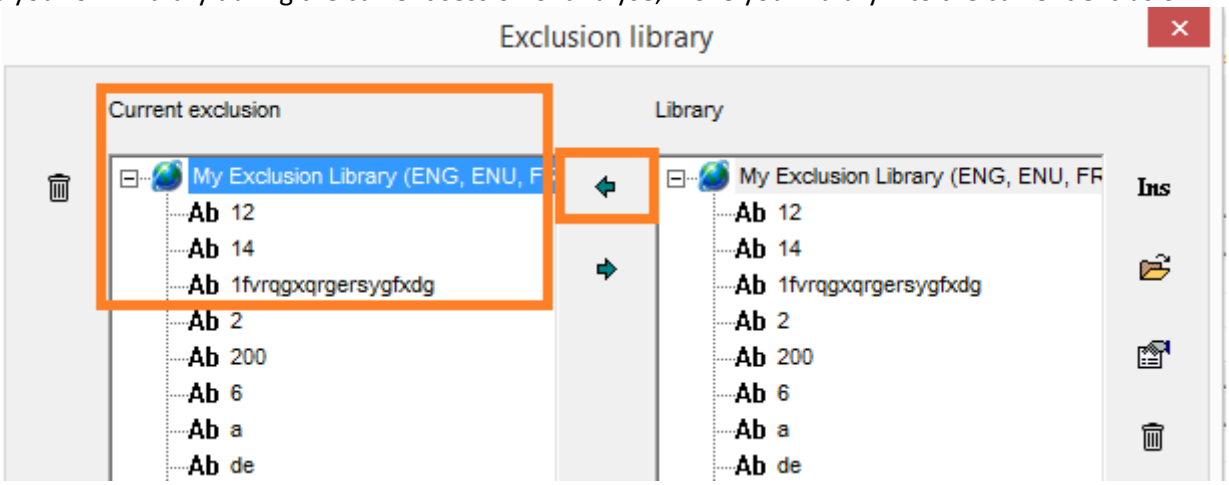
And click on the “Ins” button:



A text file will be displayed to enter the words. Add the excluded words as follows:

```
' Insert every excluded words on a new line or separated by spaces

12
14
1fvrqgxqrgersygfxdg
2
200
6
a
m
on
à
```

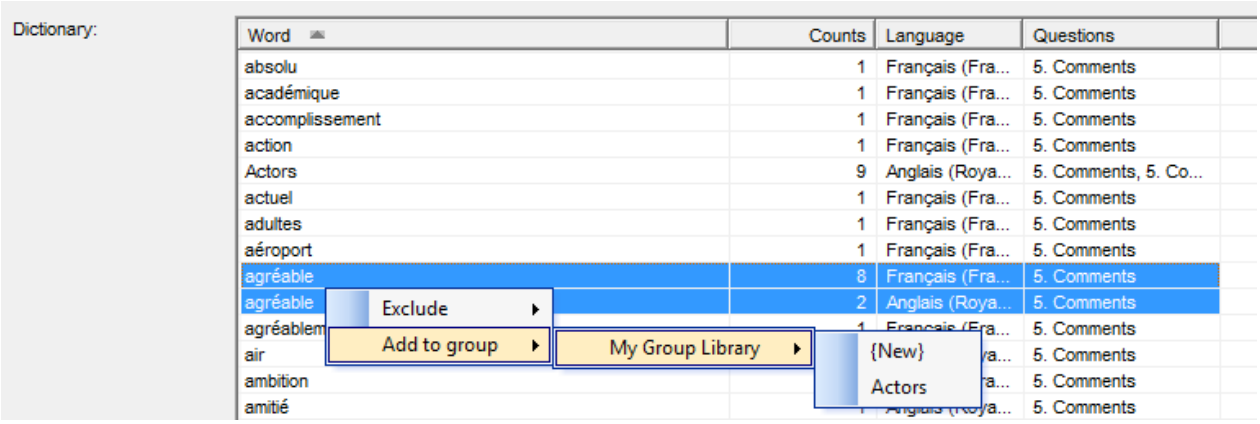
To use your own library during the current session of analyse, move your library into the current exclusion window



As soon your own library is posted in the current exclusion, the excluded words will disappear automatically from the dictionary list.

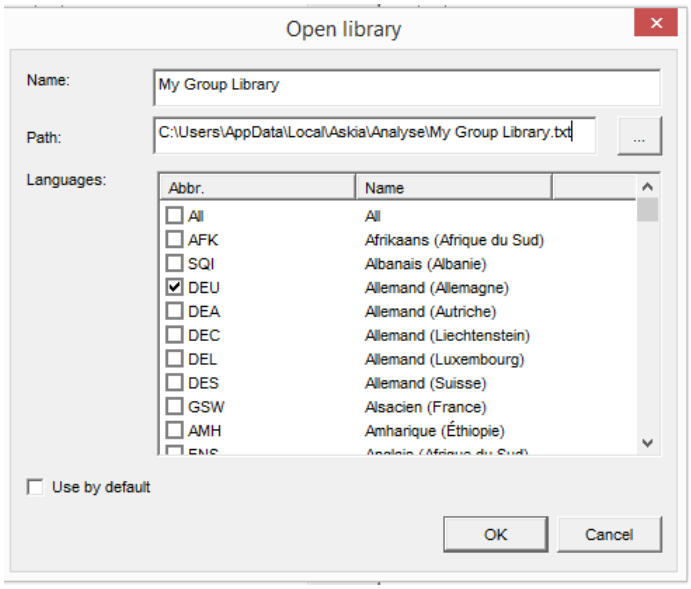
Group:

You can also group some words to simplify the view.
Select the words in the dictionary list and right click to “Add to group”



To create your own Group library, click on the “Group” button

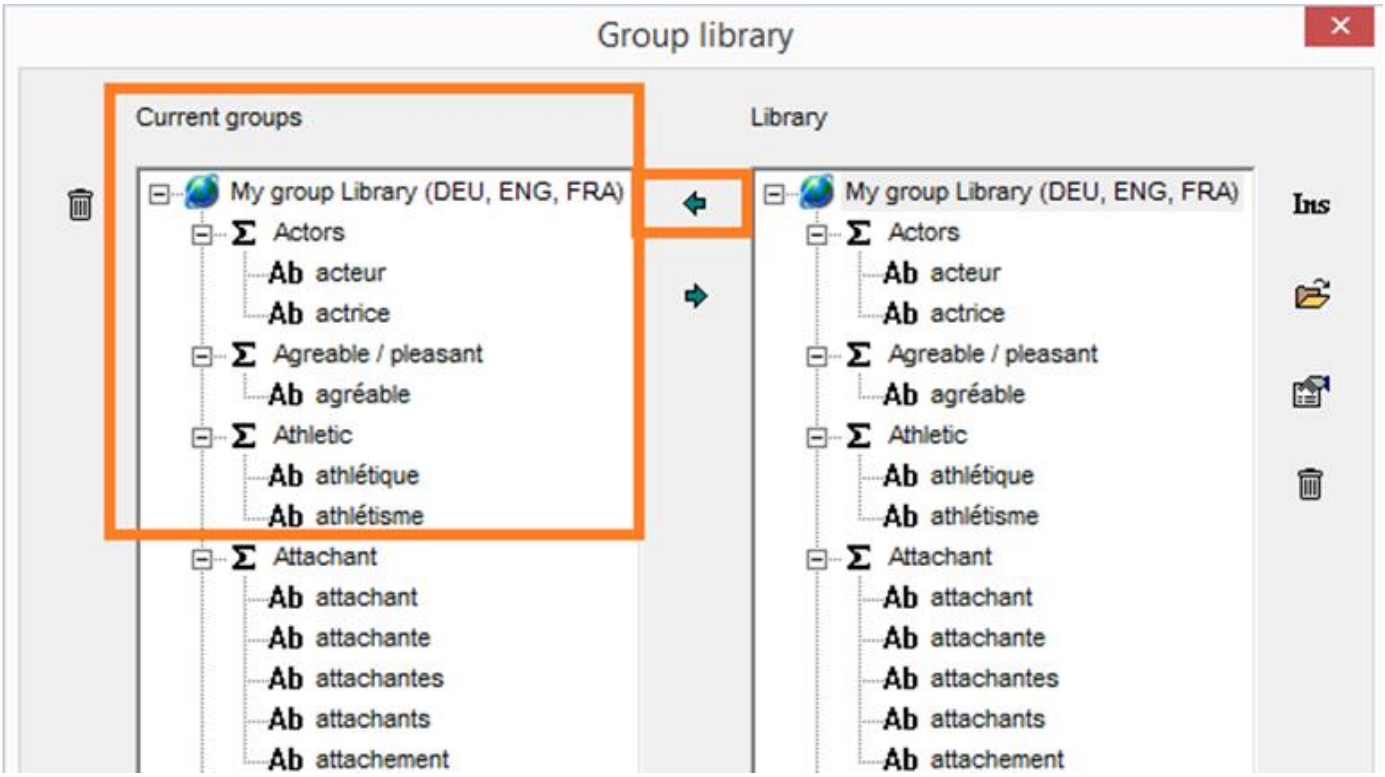
And click on the “Ins” button:



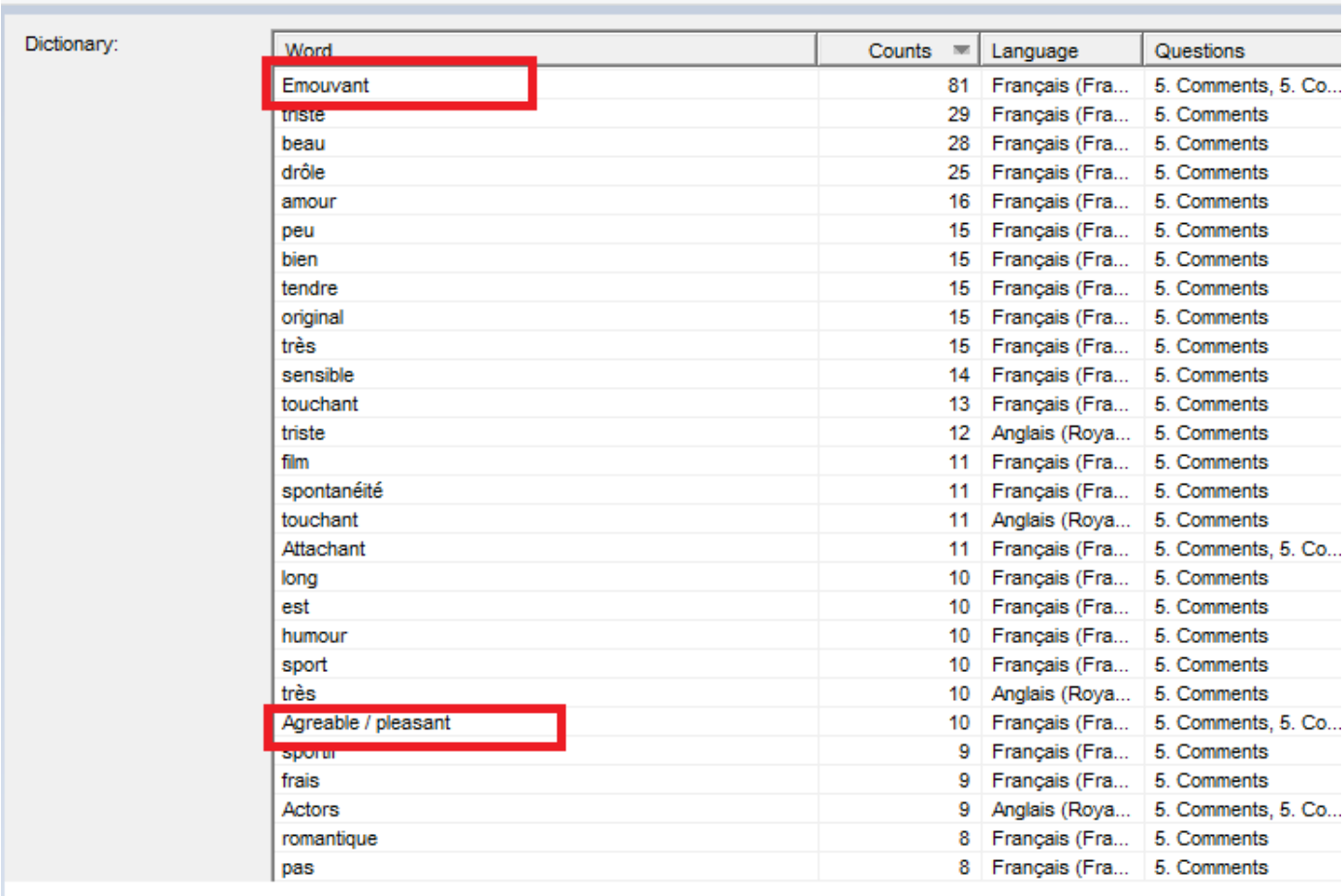
A text file will be displayed to enter the words. Add the grouped words as follows:

```
' Create every new group with a word followed by a semi-colon and each word on a line of its own eg:
'Positive:
'   Excellent 10
'   "Very good" 8
'   Good 5
'   OK
'Negative:
'   Awful 10
'   Bad

Actors:
        actrice
        acteur

Agreable / pleasant:
        agréable
        agréablement
```

To use your own library during the current session of analyse, move your library into the current groups window:



As soon as your own library is posted in the current group window, the grouped words will replace the single words automatically in the dictionary list:



Dictionary:

Word	Counts	Language	Questions
Emouvant	81	Français (Fra...	5. Comments, 5. Co...
triste	29	Français (Fra...	5. Comments
beau	28	Français (Fra...	5. Comments
drôle	25	Français (Fra...	5. Comments
amour	16	Français (Fra...	5. Comments
peu	15	Français (Fra...	5. Comments
bien	15	Français (Fra...	5. Comments
tendre	15	Français (Fra...	5. Comments
original	15	Français (Fra...	5. Comments
très	15	Français (Fra...	5. Comments
sensible	14	Français (Fra...	5. Comments
touchant	13	Français (Fra...	5. Comments
triste	12	Anglais (Roya...	5. Comments
film	11	Français (Fra...	5. Comments
spontanéité	11	Français (Fra...	5. Comments
touchant	11	Anglais (Roya...	5. Comments
Attachant	11	Français (Fra...	5. Comments, 5. Co...
long	10	Français (Fra...	5. Comments
est	10	Français (Fra...	5. Comments
humour	10	Français (Fra...	5. Comments
sport	10	Français (Fra...	5. Comments
très	10	Anglais (Roya...	5. Comments
Agreable / pleasant	10	Français (Fra...	5. Comments, 5. Co...
sportif	9	Français (Fra...	5. Comments
frais	9	Français (Fra...	5. Comments
Actors	9	Anglais (Roya...	5. Comments, 5. Co...
romantique	8	Français (Fra...	5. Comments
pas	8	Français (Fra...	5. Comments

Note that you can display:

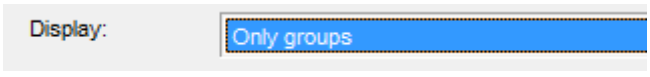
1) 

Display: Grouped words

The words and the group of words will be displayed at the same time

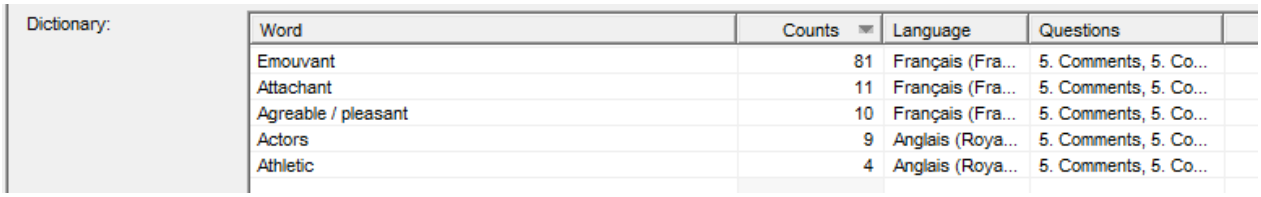
Dictionary:

Word	Counts	Language	Questions
Emouvant	81	Français (Fra...	5. Comments, 5. Co...
émouvant	52	Français (Fra...	5. Comments

2) Only groups 

Display: Only groups

Only groups of words will be displayed



Dictionary:

Word	Counts	Language	Questions
Emouvant	81	Français (Fra...	5. Comments, 5. Co...
Attachant	11	Français (Fra...	5. Comments, 5. Co...
Agreable / pleasant	10	Français (Fra...	5. Comments, 5. Co...
Actors	9	Anglais (Roya...	5. Comments, 5. Co...
Athletic	4	Anglais (Roya...	5. Comments, 5. Co...

Repeated segments:

There are some forms which make no sense when they are separated (e.g.: didn't). Other take on a different meaning when they are associated (e.g.: Social Security).

The repeated segments can appear for reasons of statistics or linguistics, (a number of forms are always stated together). When we ask the respondent to answer about a defined topic, we observe that the number of segments increases. The study of these segments is often more interesting than the study of the vocabulary.

It should be noted that the study of segments is a generalisation of the study of forms (a form is a segment with a length of 1)

You can display the repeated segments.

A repeated segment is a group of words generally stated together in the same sentence.
In our example (ex.qes, question I5. Comments), you will have "conte de fée" as a repeated segment. This repeated segment contains 3 words (Length=3)

Dictionary:

Word	Counts	Language	Questions	Length
conte de fée	4	Français (Fra...	5. Comments	3
un peu long	3	Français (Fra...	5. Comments	3
une belle histoire	3	Français (Fra...	5. Comments	3
un peu lent	2	Français (Fra...	5. Comments	3
un peu trop	2	Français (Fra...	5. Comments	3
c est un	2	Français (Fra...	5. Comments	3
histoire d amour	2	Français (Fra...	5. Comments	3
dépassement de soi	2	Anglais (Roya...	5. Comments	3

Specificities:

We use specificities when we want to see which word is most used by a specific population.
For example, we have asked the question:

"What is your opinion of this movie?
Why?"

We obtain the following responses:

What is your opinion on this movie	Why?
Positive opinion	I just adored it
Positive opinion	Very good
Positive opinion	Good. Amazing
Positive opinion	I just liked. Good
Positive opinion	Amazing
Negative opinion	Not good
Negative opinion	Too long
Negative opinion	Long
Negative opinion	Not liked

And the following dictionary

Forms	Positive Opinion	Negative Opinion	Total
Good	3	1	***4***
Just	2	0	***2***
Liked	*1*	*1*	***2***
I	*2*	*0*	***2***
Long	*0*	*2*	***2***
Not	*0*	*2*	***2***
Amazing	*2*	*0*	***2***
Adored	*1*	*0*	***1***
Very	*1*	*0*	***1***
Too	*0*	*1*	***1***
Total	***12***	***7***	***19***

The specifities are calculated as follows:

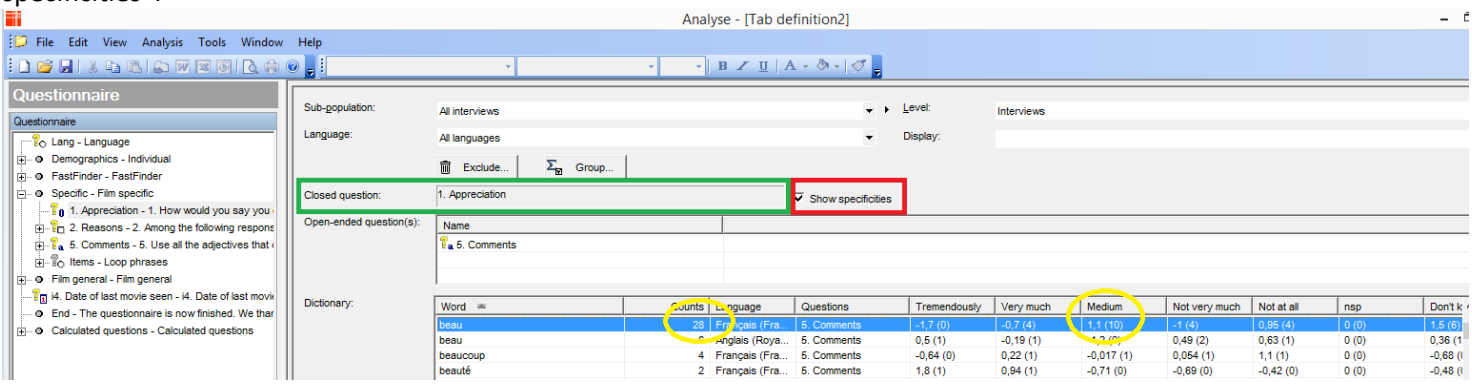
Forms	Positive Opinion	Negative Opinion
Good	*0.335 **	*-0.439*
Just	*0.693*	*-0.907*
Liked	*-0.247*	*0.324*
I	*0.693*	*-0.907*
Long	*-1.188*	*1.556*
Not	*-1.188*	*1.556*
Amazing	*0.693*	*-0.907*
Adored	*0.476*	*-0.624*
Very	*0.476*	*-0.624*
Too	*-0.816*	*1.069*

*The specificity of Good for positive opinions is:

$$S = \frac{\left(\frac{3}{12} - \frac{4}{19}\right)}{\sqrt{\frac{4}{19} * \left(\frac{1-\frac{4}{19}}{12}\right)}} = 0{,}335$$

When we order by decreasing specificities, we obtain the forms most relevant to a sub-population.

Drag and drop the closed question to explain (in our example ??1. Appreciation??) and then select “Show specificities”:



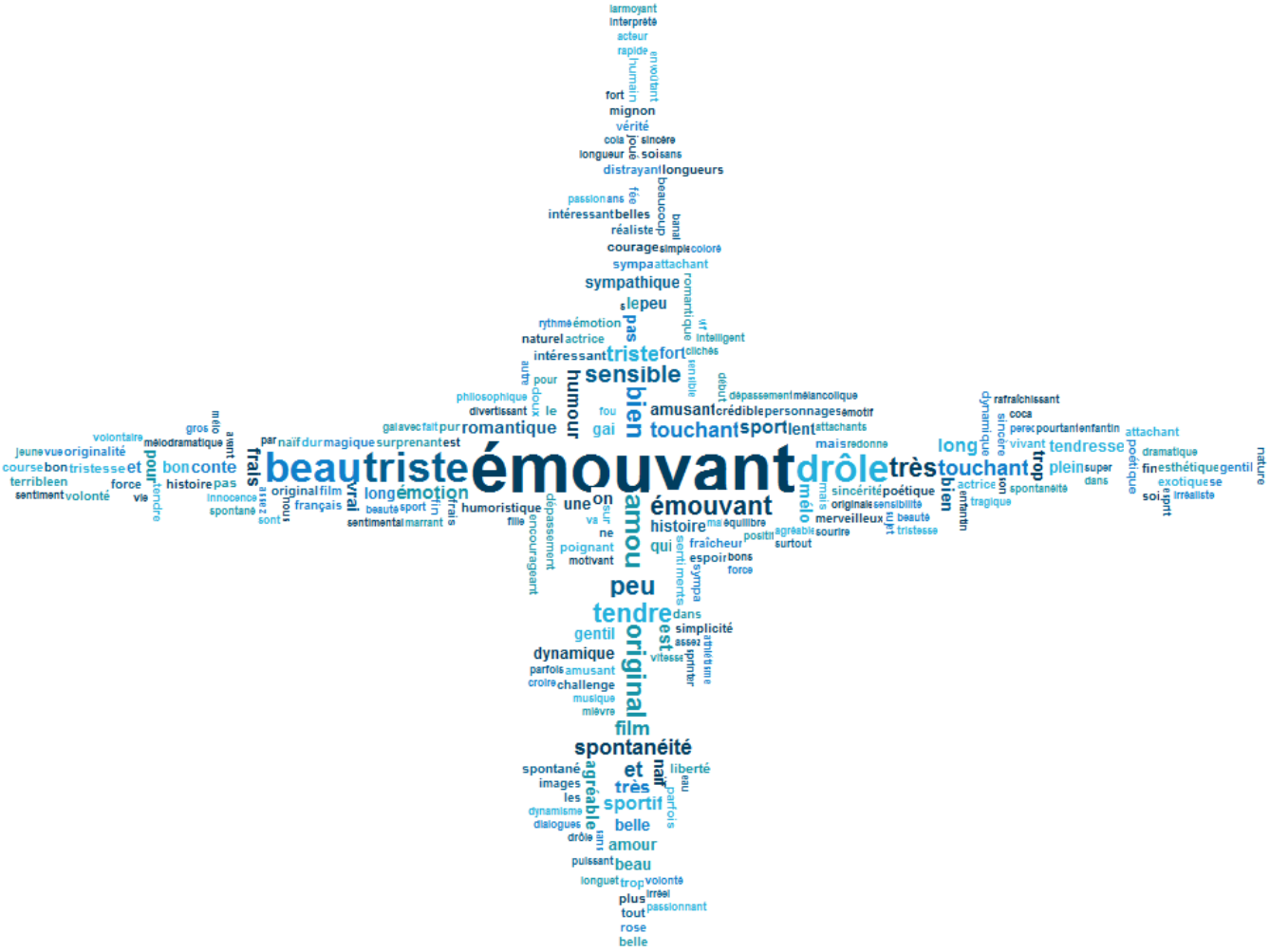
The word “beau” has been stated 28 times and 10 on these 28 have been stated by the “medium” population. So we can say that the word “beau” is more relevant for the “Medium” population. The coefficient “1,1” represents the specifities for this population.
You will also have the frequency of the word displayed in parenthesis

Word Cloud:

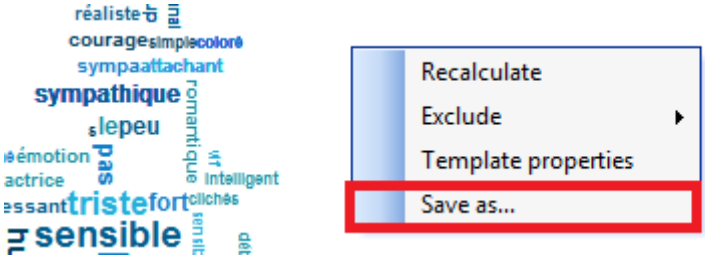
To visualize the word cloud, click on 

.

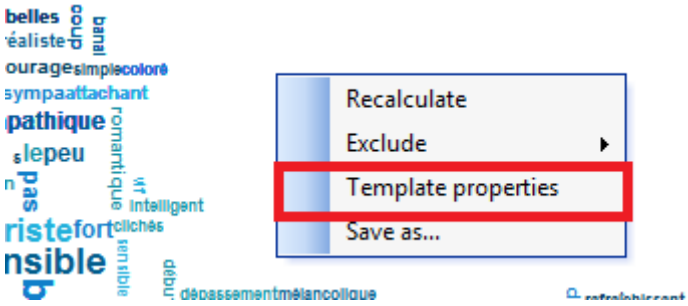
In the output, you will have by default a blue cross:



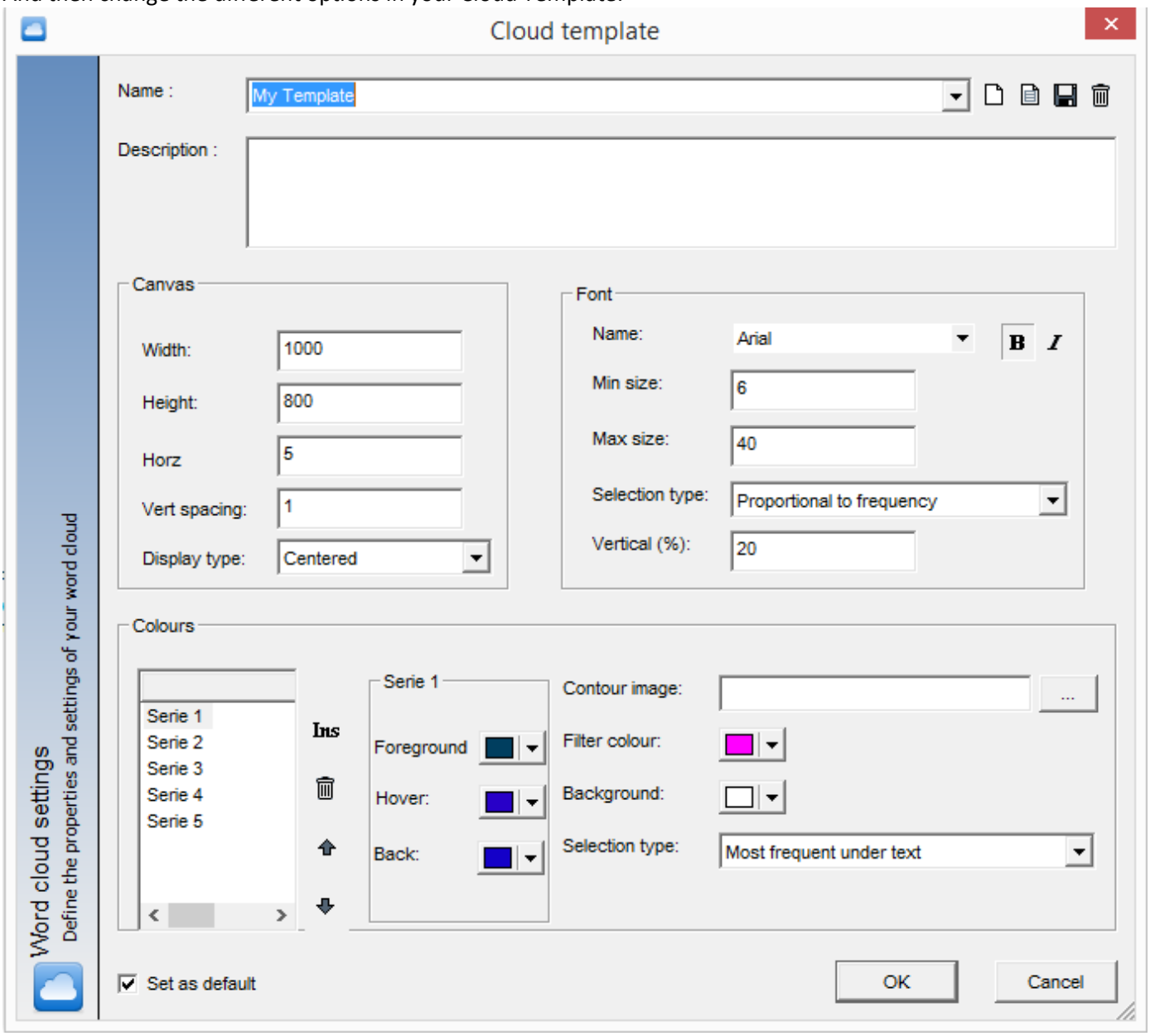
To **save the output as an image**, make a click right, “Save as”:



To **change the properties** (colour, shape, series etc.) right click: “Template properties”:

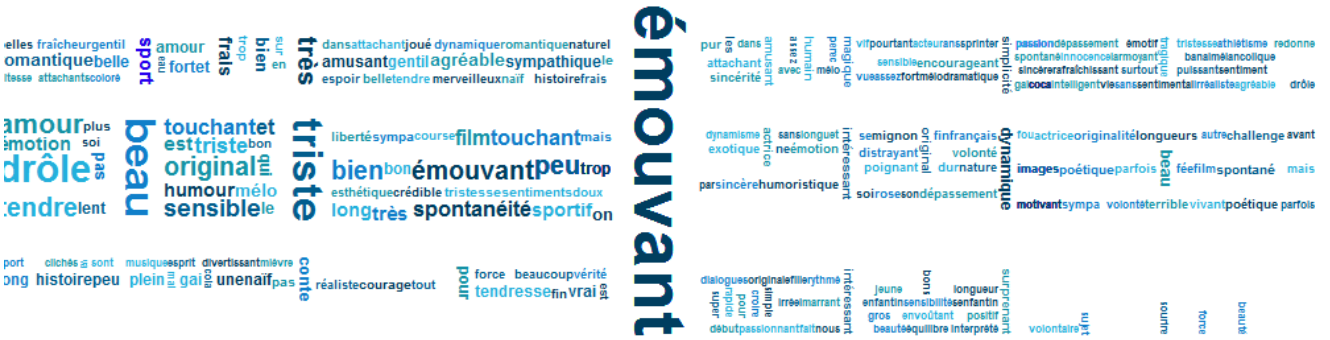


And then change the different options in your Cloud Template:

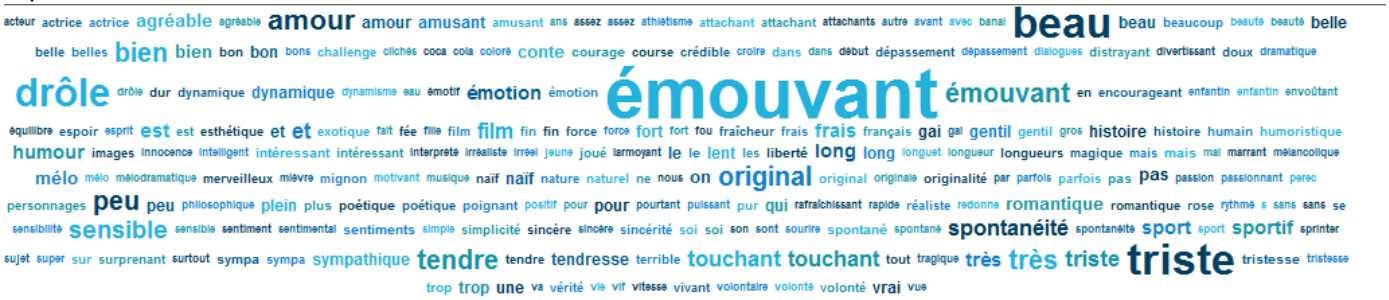


The **canvas** defines the proportions of your image (in pixels) and the **display type** of the word cloud.
By default, the display is centred.

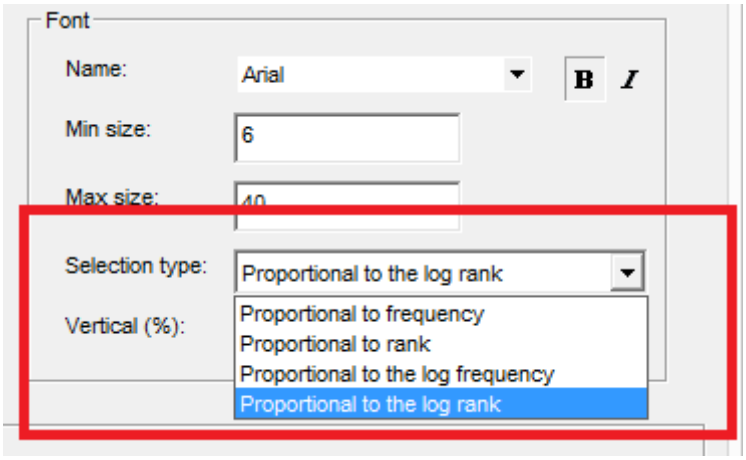
Excentered Word cloud:



Alphabetical Word Cloud:



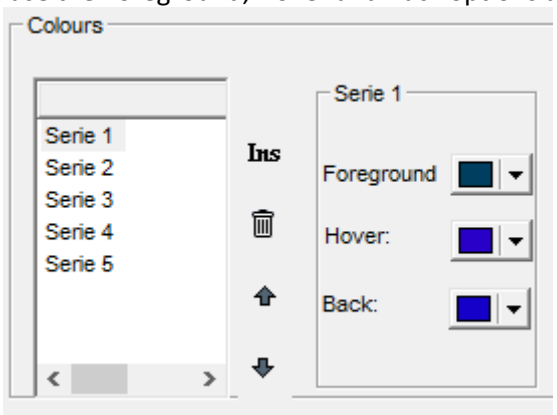
To change the wording size, you can modify the **font** and **the selection type**:



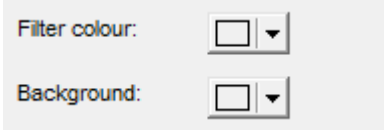
- Proportional to frequency: The more a word has been stated, the larger the wording.
- Proportional to rank: All words have been ranked from 1 to N. The 1st rank is allocated to the 1st stated.
- Proportional to the log frequency: We use the logarithmic frequency to decrease the visual impact of the frequency

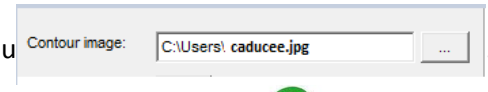
- Proportional to the log rank: We use the logarithmic rank to decrease the visual impact of the rank

To define the colours of series, use the Foreground, Hover and Back options as follow:



You can also define the background of your whole Word Cloud and the colours of the filter for each word.

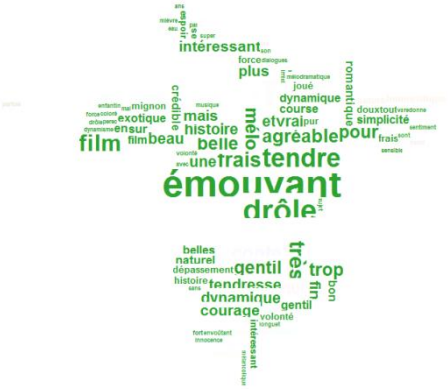


You can change the **shape** and using a logo. Select your resource by clicking on ... of the “contour image” menu 

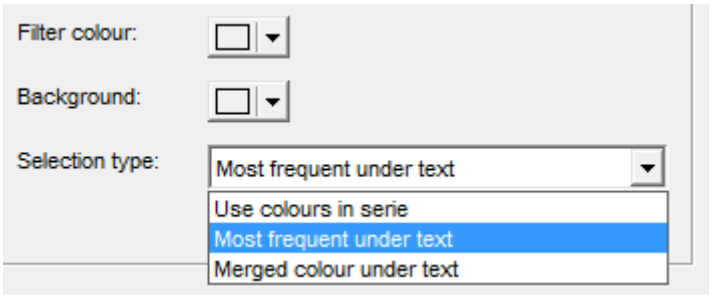
.

As an example, we have used 

So the output will be displayed as below using the logo’s colours and logo’s shape where possible:



Note if you select the option “Use colours in serie” in Selection type, the colours of the logo will be replaced by the colours of the series

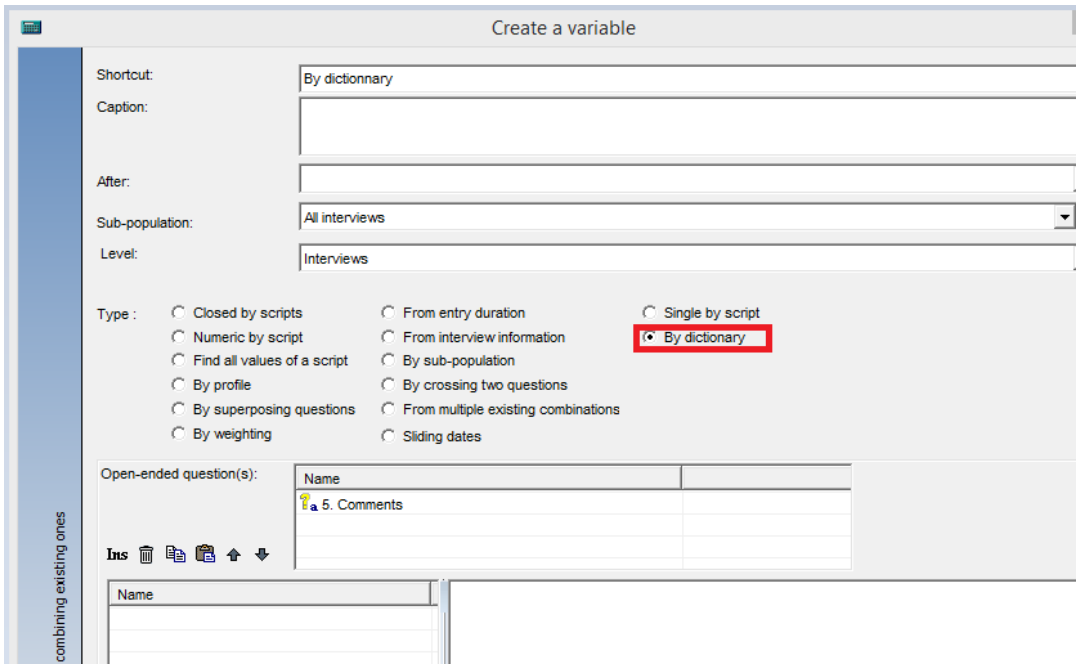


2.3. Create a variable by dictionary

If you want to have a multi-coded question including the list of words stated in the comment, you can create a closed question “By dictionary”.

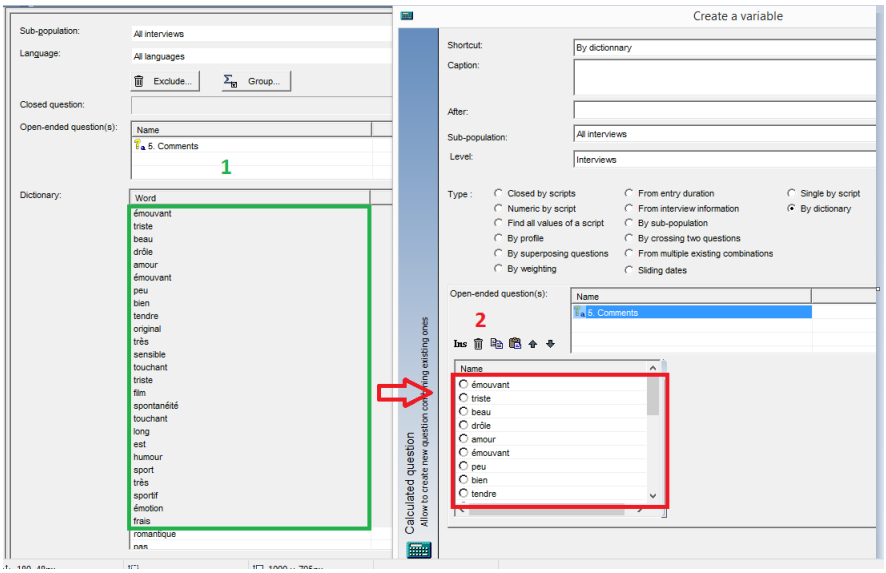
How to do it:

In the tool bar menu, go to “Create a variable” and select “By Dictionary”



Select your open-ended question(s) in the tree view and drag and drop your question(s) directly into the “open-ended questions(s)” window.

Run the Dictionary analysis, see Chapter “2.1. Dictionary”, and drag and drop the words (1) into the “Name” Window (2)



You can also create category by category by clicking on “Ins”.

For each modality created, add the corresponding word (be aware on the case sensitiveness)