IRAMUTEQ TUTORIAL

IRaMuTeQ

(*R INTERFACE for multidimensional analysis of texts and questionnaires)*

Brigido Vizeu Camargo e Ana Maria Justo *(Social Psychology Laboratory of Communication and Cognition - Federal University of Santa Catarina - Brazil)*

Translated by Teresa Forte *(European/International Joint PhD in Social Representation & Communication - SoReCom Joint-IDP - led by Sapienza University of Rome - Italy)*

IRAMUTEQ is a GNU GPL (v2) licensed software that provides users with statistical analysis on text corpus and tables composed by individuals/words. It is based on R software and on *python* language.

In order to install it for free, you must first download and install R software in www.r-project.org, (without it is not possible to run any analysis); then, download and install IRAMUTEQ software.

PICTURE 1- IRAMUTEQ INTERFACE

Software installation for Windows operating system - "IRAMUTEQ Kit"

Iramuteq Kit is available on: www.laccos.com.br. In the link "*Novidades*" press "*Clique AQUI*" and download the KIT, which includes the software, references and a tutorial.

1- **Install Open Office software package**

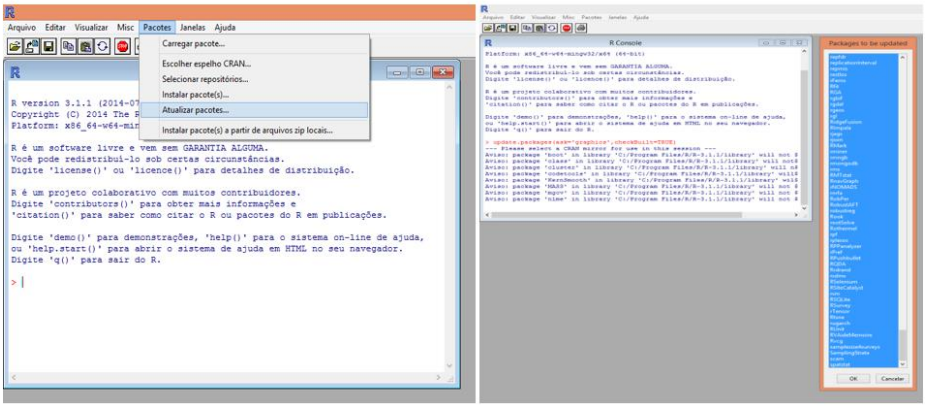
Free equivalent of Microsoft Office. Two software of this package are required: 1)OpenOffice.org Writer to create ".odt" text archives used to enter the corpus and read reports and results 2) OpenOffice.org Calc to create archives of ".ods" spreadsheets used to enter word association matrices and to read and export results as well.

Never open these archives or any other IRAMUTEQ output with Microsoft word applications (Word, Excel, WordPad or Notepad) they generate bugs with the Unicode used by IRAMUTEQ (UTF-8)

2- **Install "R" software, base system for IRAMUTEQ alongside with Python language**

3- **Update R packages**

Open R software. Choose "Packages" + "Update Packages (Picture 2). Choose: the nearest country/ state you are). Wait a few seconds (depending on the computer and internet it can take a little longer) and select OK when asked to update some blue items. Once updated, close the software.



Picture 2- Update of packages in R interface

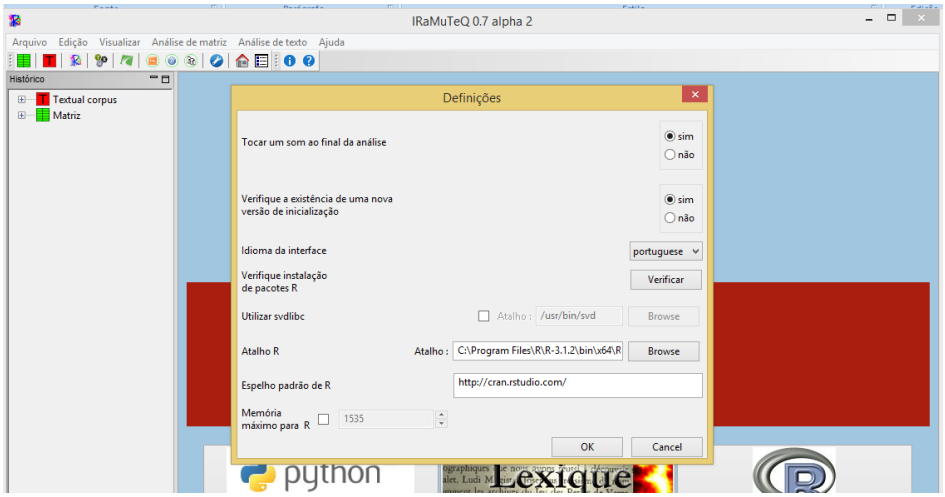
4- **Install IRAMUTEQ** software

5- **Update IRAMUTEQ libraries**

Click on IRAMUTEQ icon on your desktop. **You must be connected to Internet for this step.** Press ok on the message displayed about the installation status (Picture 3) and wait for R software archives update.

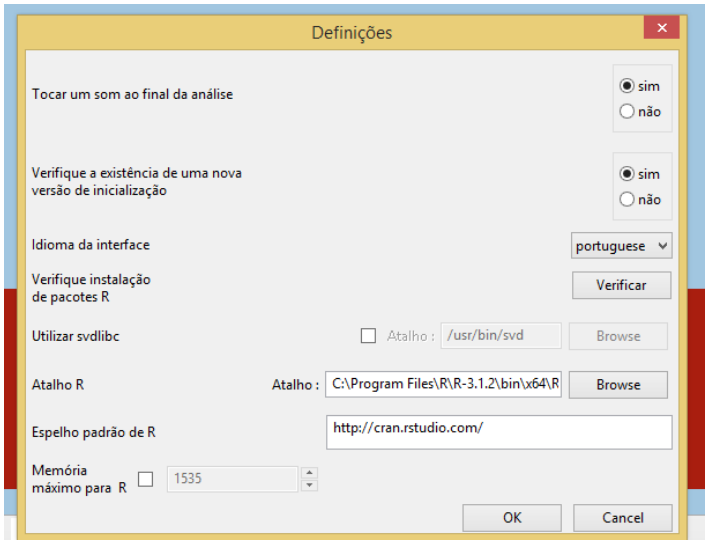
Picture 3- Libraries update

ATTENTION! If the Software fails to update automatically: open IRAMUTEQ; press “Edition” + “Preferences” + “R Path” and click on **browse** as shown in Picture 4. **Find R application** in: Program Archive/R/Bin. Save and close IRAMUTEQ



Picture 4- R path correction

Reopen IRAMUTEQ. Click “Edition” + “Preferences”, go to “Check installation of R packages”, click on “Check”, and wait for the installation process to be completed (Picture 5).



Picture 5- Verification of libraries installation on IRAMUTEQ

Introduction

This software provides the users with different text analyses, either simple ones, such as the basic lexicography related to lemmatization and word frequency; or more complex ones such as descending hierarchical classification, post- hoc correspondence factor analysis and similarity analysis. The vocabulary distribution is presented in a comprehensive and clear way with graphical representations derived from the lexicographic analysis.

These analyses can be performed using texts referring to a certain thematic (text corpus) grouped in one text archive; or data from spreadsheets (matrices with individuals in a row and words in a column), like the dataset derived from free evocation tests.

Part 1: Text corpus analysis

Text analysis enables to explore oral material transcribed, including texts, interviews, documents, etc, which can be individually or collectively produced. It's also useful for comparing different productions according to specific variables described by who produced the text. To understand Text analysis, some concepts need to be clarified first:

The concepts of Corpus, text and Text segments

Corpus

The corpus is created by the researcher and is the set of units to be analyzed. For instance, if a researcher wants to analyze beauty related news published on a magazine during 5 years, the set of these news is the corpus.

Text

These units are defined by the researcher, depending on the research. In the previous example, each beauty related news is a text. If the analysis is to be applied on a set of interviews, each interview is a text. If you intend to analyze the answers of “n” participants to an open question, each answer will be a text and there will be “n” text. In research on documents, letters, etc.; each document is a text.

A set of text units is a corpus of analysis. The ideal corpus for the Descending Hierarchical Analysis must be a groups of texts focused on one theme (monothematic), in order to avoid a replication of the initial structure.

In interviews, usually including larger texts, 20 to 30 texts are sufficient when the groups is homogenous (Ghiglione & Matalon, 1993). For a comparative design, it is recommended to have at least 20 texts per group.

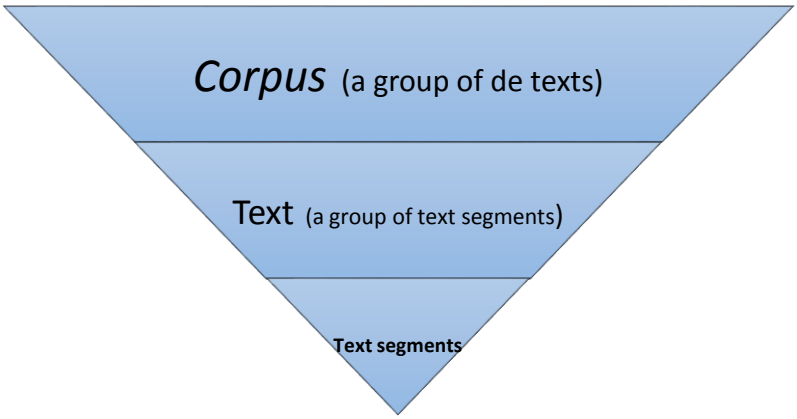
With open ended questions from a survey, it is recommended to aggregate answers to the same question, in order to assure they’re referring to the same theme. In case of questions related to different themes or aspects, it is necessary to perform a separate analysis for each question. As abovementioned, the analysis is sensible to how the stimulus producing the text material are structured, thus this process if not performed properly can lead to invalid conclusions. A higher number of answers is needed if they are 3 / 4 lines long each.

Command lines, also called “asterisk lines” divide the text. In interviews, for instance, given that each of them is a text, they start with a command line which includes the identification number of the interviewee and some important features (variables) for the research design (such as: sex, age, groups affiliation , socio-cultural level, etc.). This may vary according to the research. The number of levels of each variable also depend on the research design and number of interviews conducted.

Text segments

Generally, text segments (TS) have three lines, automatically sized according to the corpus extension. The text segments are the words contexts. They can be created by the researcher, or automatically by the software.

Although the texts are limited by the researcher, the corpus division in text segments (TS) is automatically done in a standard analysis. However, the researcher may adapt the segments division, for example: in case of a high frequency of short answers and an open question from a survey, it is recommended to define the text as a single TS.



Picture 6- Corpus, text and text segments

Creating a text corpus for analysis

1-Insert all texts (interviews, articles, texts, documents or answers to one question) in one text archive using OpenOffice.org (http://www.openoffice.org/) or LibreOffice (Http://pt-br. Libreoffice.org/). **Never open these archives or any other output of IRAMUTEQ with Microsoft word applications (Word, Excel, WordPad or Notepad)** they generate bugs with the Unicode used by IRAMUTEQ (UTF-8)

2- Divide the text with command lines (with asterisk). For example, for each interview to be recognized by the software as a text, it should start with a line like this: (NOTE: leave a clear line before the first command line.)

**** *ind_01 *ida_1 *par_2 *fil_2 *temp_2 *caus_1

Insert 4 asterisk (in a row, no space between them), a space, another asterisk and the name of the variable (no space between these), underscore and the code for the variable type (no space between these), a space followed by the asterisk

of the second variable and so on. This example was extracted from a research with sex workers about prevention of STD's and pregnancy. It indicates that the following text (interview answers) refers to individual nº 01 (2 digits because the sample has more than 10 and less than 100 individuals), age between 19 and 26 years old (coded as 1= 19 to 26 years old; 2= 27 to 47 years old); she doesn't have a fix sexual partner (boyfriend or husband) (coded as 1- have a partner and 2= don't have a partner); for how long has she been a sex worker, which in this case is between 13 and 36 months (1= 12 months, 2= from 13 to 36 months and 3= from 48 to 132 months) and the reason why she is a sex worker which in this case is "family issues" (1= family issues, 2= financial need, 3= family provider, 4= frustrated love relationship and 5= didn't answer). Immediately after this line press ENTER and without tabulation or clear line, enter the answer of this subject nº1.

3- You have 2 options for creating the text. The first, original or monothematic, where each line is followed by a joint text. The second one, thematic, contains two or more themes for line, subordinating lines to a main one. The analysis of a corpus with thematic divisions (different themes) provides information about relations between themes; and can be used as a preliminary exploratory analysis (in order to have a snapshot of the text material). Nonetheless, the monothematic analysis is still needed because it provides a more in depth understanding of the studied object.

Example of a monothematic corpus

*** *ind_01 *ida_1 *par_2 *fil_2 *temp_2 *caus_1
I use remedies to avoid pregnancy such as contraceptives or injections, which are easier, because for someone who drinks, is difficult to take remedy, sometimes the effect is nullified, so injections are a safer choice. I also use condoms because they are safer. I have to take care of myself. I think I must use contraceptives, can't only trust in condoms. I may have a child of someone I barely know, or have a STD and if this happens and I tell that person I'm bearing his child he will probably say that I'm a prostitute and the child can be anyone's. I'm aware of all STD, I learned about that in school. Nowadays one doesn't take care only if one doesn't want to. To me, all of them are risky so you have to prevent them all, is not easy to hang around with one, another, you must be careful, even drunk you must be aware. Sometimes, we go out with people we can't stand, we see a lot of nasty things. If you're not aware, things can happen and you will only noticed

afterwards. Sometimes, many people can't find a girl on the street or a girlfriend so they come here and we are forced to accept them. The prostitute is considered vulgar. I don't think of myself as vulgar. I have my reasons and don't accept that one comes here and call me vulgar, I don't accept, because I'm not, if I were I would be on the streets, doing everything. It is really different to work on night clubs, street or cabaret. I already had vaginal discharge, but not STD. Discharge is a natural thing that you can have from the condom or anything else, like soap, clothes, but never have STD.

*** *ind_02 *ida_2 *par_1 *fil_2 *temp_3 *caus_2
I always use condoms, because besides preventing pregnancy it prevents aids and other diseases, we need to use it. It's good for everything. I know many STD's such as gonorrhea, cancer, genital louse, there are so many. I had gonorrhea from a boyfriend, couldn't imagine this. As the time went by, I felt pain, went to a doctor, and had to do a surgery and found out I had gonorrhea. I'm not afraid to talk about this, everyone is at risk of getting these, all these diseases. To protect me from STD's, I do oral and regular sex with condom, and never do anal.
*** *ind_03 *ida_1 *par_2 *fil_1 *temp_1 *caus_2
I use contraceptives and condoms. Not the pill, but the injection, is easier to remember. That's why I take both, because if something happens (to be continued)

Example of a thematic corpus
**** *ind_01 *ida_1 *par_2 *fill_2 *temp_2 *caus_1
-*theme_prevention
I use remedies to avoid pregnancy such as contraceptives or injections, which are easier, because for someone who drinks, is difficult to take remedy, sometimes the effect is nullified, so injections are a safer choice. I also use condoms because they are safer. I have to take care of myself. I think I must use contraceptives, can't only trust in condoms. I may have a child of someone I barely know, or have a STD and if this happens and I tell that person I'm bearing his child he will probably say that I'm a prostitute and the child can be anyone's
-*theme_std
I'm aware of all the STD, I learned about that in school. Nowadays one doesn't take care only if one doesn't want to. To me, all of them are risky so you have to prevent them all, is not easy to hang around with one, another, you must be careful, even when you're drunk you must be aware. Sometimes, we go out with people we can't stand, we see a lot of nasty things. If you're not aware, things can happen and you will only noticed afterwards. Sometimes, many people can't find a girl on the street or a girlfriend so they

come here and we are forced to accept them. The prostitute is considered vulgar. I don't think of myself as vulgar. I have my reasons and don't accept that one comes here and call me vulgar, I don't accept, because I'm not , if I were I would be on the streets, doing everything. It is really different to work on night clubs, street or cabaret. I already had vaginal discharge, but not STD. Discharge is a natural thing that you can have from the condom or anything else, like soap, clothes, but never have STD.

*** *ind_02 *ida_2 *par_1 *fil_2 *temp_3 *caus_2

-*theme_prevention

I always use condoms, because besides preventing pregnancy it prevents aids and other diseases, we need to use it. It's good for everything

-*theme_std

I know many STD's such as gonorrhea, cancer, genital louse, there are so many. I had gonorrhea from a boyfriend, couldn't imagine this. As the time went by, I felt pain, went to a doctor, and, had to do a surgery and found out I had gonorrhea. I'm not afraid to talk about this, everyone is at risk of getting these, all these diseases. To protect me from STD's, I do oral and regular sex with condom, and never do anal.

*** *ind_03 *ida_1 *par_2 *fil_1 *temp_1 *caus_2

I use contraceptives and condoms. Not the pill, but the injection, is easier to remember. That's why I take both, because if something happens (to be continued)

Note: After creating the corpus, a careful reading is recommended, paying special attention to command lines. This must be verified by the researcher in order for the text to be processed.

1- **Correct and review the archive,** to avoid spelling mistakes or errors to be taken into account as different words
2- **Pay attention to the punctuation.** In case of doubts concerning the proper way to insert the paragraphs, it is better to avoid them.
3- In case of interviews or surveys, the questions and oral material from the interviewer (interventions and notes) must be suppressed, thus excluded from the analysis. Keep the referents.
4- Don't justify the text, or use bold, italic or other similar resource.
5- The use of acronyms and abbreviations must be consistent, either you use them always or write it in full with underscores connecting the words. For example: WHO or World_Health_Organization.

6- The words with hyphen are considered two separate words (the hyphen becomes a space). In case you need to analyze them, unite them with an underscore. Ex: “alto-mar” becomes “alto_mar”; “terça-feira” becomes “terça_feira”.
7- All verbs with pronouns must be in proclisis, because the dictionary doesn’t have the verb-pronominal inflexions. EX: “ tornei-me” becomes “me tornei”
8- If possible, avoid using diminutives, due to the dictionary features.
9- When writing numbers, insert digits, not words. EX: use “2013” and not “two thousand thirteen”; “70” and not “seventy”.
10- Don’t use in any part of the archive the following signs: quotes (“), apostrophe (‘), hyphen (-); dollar sign ($), percentage (%), suspension points (…) The asterisk (*) can only be used in the command lines.
11- The archive with the corpus from software OpenOffice.org or LibreOffice must be saved in a new folder on desktop, used only for analysis, with a short name and as coded text (archive_name.txt). In OpenOffice.org this option displays a window where you choose “keep the present format”. A second window appears where the options “group of characters” and “ Paragraph break” should be respectively “Unicode (UTF-8)” and “LF”.
12- Don’t reutilize the txt archive (coded text) when conducting a new analysis on the same corpus. Create a new one with odt archive (proper format to archive it).

Types of analysis on IRAMUTEQ text corpus

IRAMUTEQ provides the users with different text analyses, from simple ones such as basic lexicography (word frequency) to multivariate analysis (descending hierarchical classification).

I) **Classic lexicographical analysis**- identifies and formats text units, turns texts into TS, identifies word frequencies, medium frequency and hapax (words with frequency=1), searches for vocabulary and reduces the words to their primary lexical units (reduced forms) creates the reduced forms dictionary, identifies active and supplementary forms.

II) **Specificities and correspondence factor analysis**- associate texts with variables and allows you to analyze texts according to characterization variables. Allows the user to run correspondence factor analysis for variables with at least 3 levels.

III) **Method of Descending Hierarchical analysis (DHA) -** The TS are clustered according to their vocabularies and distributed according to the reduced forms frequencies.

Using matrices that cross reduced forms with TS (in repeated texts of X^2 type), the DHA method allows you to obtain a definitive classification. It is aimed at obtaining TS clusters with similar vocabulary within, but different from other segments. A dendrogram will be displayed showing clusters relations.

The *software* calculates descriptive results of each cluster conforming to its main vocabulary (lexic) and words with asterisk (variables). Furthermore, it provides the users with another way of presenting data, derived from a correspondence factor analysis. Based on the chosen clusters, the software calculates and provides the most typical TS of each cluster, giving context to them.

These word clusters and TS integrate several segments according to the vocabulary distribution. On the interpretative level, it depends on the theoretical scope of the research. Reinert (1990), when studying French literature, considered each cluster as a "world", a cognitive-perceptive framework with a certain temporal stability related to a complex environment. Research in linguistic considers these clusters as lexical fields (Cros, 1993) or semantic contexts. For research in social psychology, especially when interested in studying the common sense knowledge, which take into account the linguistic expressions, these clusters may indicate social representations, images about a certain object or aspects of a certain social representation (Veloz, Nascimento- Schulze & Camargo, 1999).

Generally, the number of clusters and the number of social representations involved are not the same as in the abovementioned study. What defines if they refer to different social representations or just to one social representation is its content and its relation with factors considered in the research design, through a differentiated selection of the participants according to their group affiliation, previous social practices, etc.

IV) Similarity analysis- This analysis, based on graph theory, is often used by social representations researchers. It allows to identify the words co-occurrences, providing information on the words connectivity thus helping to

identify the structure of a text corpus content. It also allows to identify the shared parts and specificities according to the descriptive variables identified in the analysis (Marchand and Ratinaud, 2012).

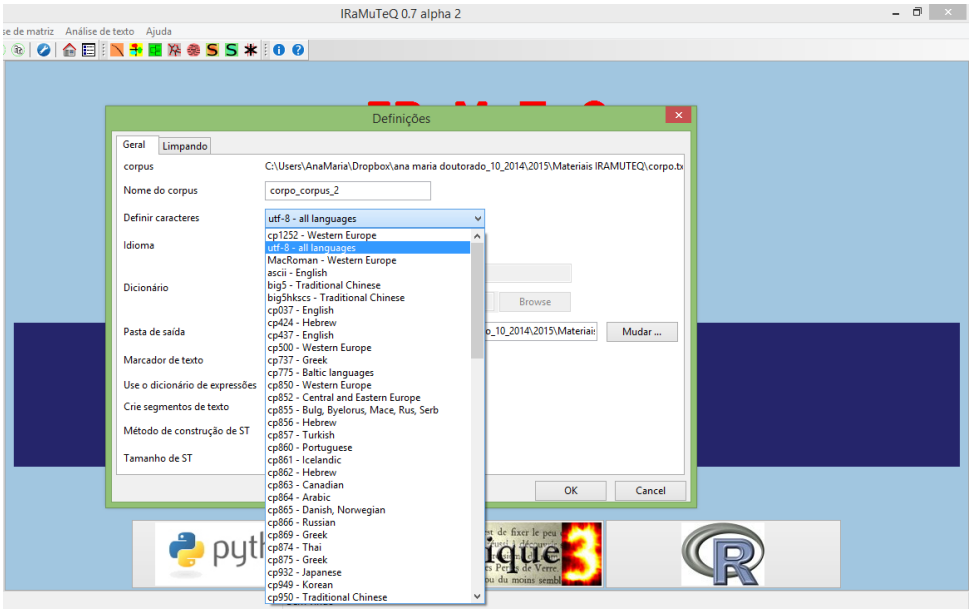
V) Word cloud- aggregates words and organizes them graphically according to its frequency. It's a simpler lexical analysis, however graphically interesting.

Running the analysis

Open the software and import the corpus. Click *File* and *Open a text corpus* on the upper toolbar (see Picture 7). Locate and select the *corpus* for analysis and click *Open.*

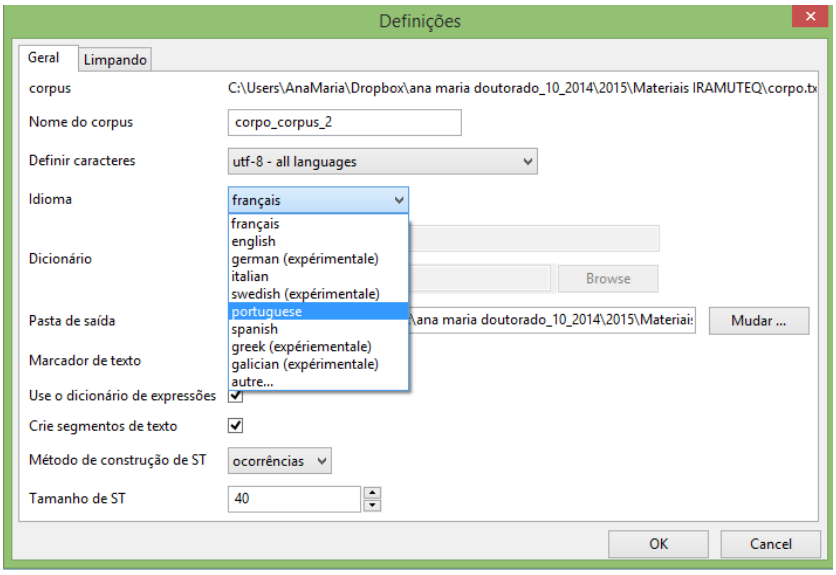
PICTURE 7- Corpus import

Once the software imports the corpus, a new window is displayed (Picture 8).



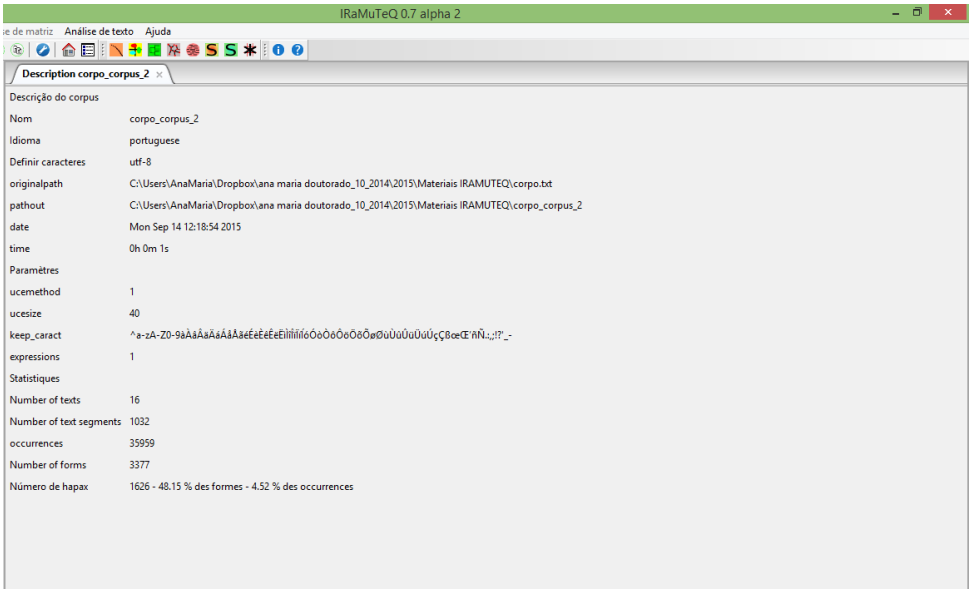
PICTURE 8- Analysis configuration- corpus coding

This window (Picture 8) presents the configuration for text analysis. Most of the configurations in *General* tab, may be kept according to the standard definitions, except the following 2: the text coding (*define characters*), in which you must choose the second option: "*uft-8-all languages*"; and the language (*Language*). According to Picture 9, select the language corresponding to the language of the text.



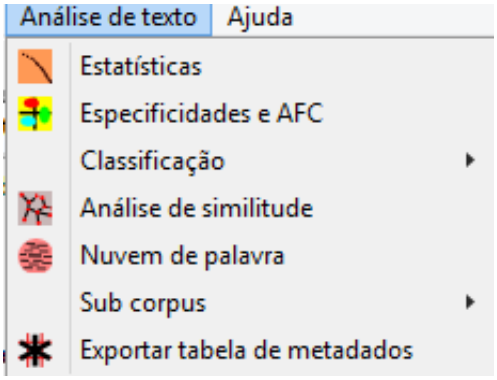
PICTURE 9- Analysis configuration- Language

Press OK and wait a few seconds for the data import. A brief description of the corpus will appear on the right large window (Picture 10), where you may verify the number of texts, text segments, identified forms, occurrences and *Hapax* frequency.



PICTURE 10- Preliminary results, corpus description

Once the corpus import is completed, you can run the analysis. By selecting "*Text Analysis*" in the upper toolbar the following options for analysis are displayed. (Picture 11).



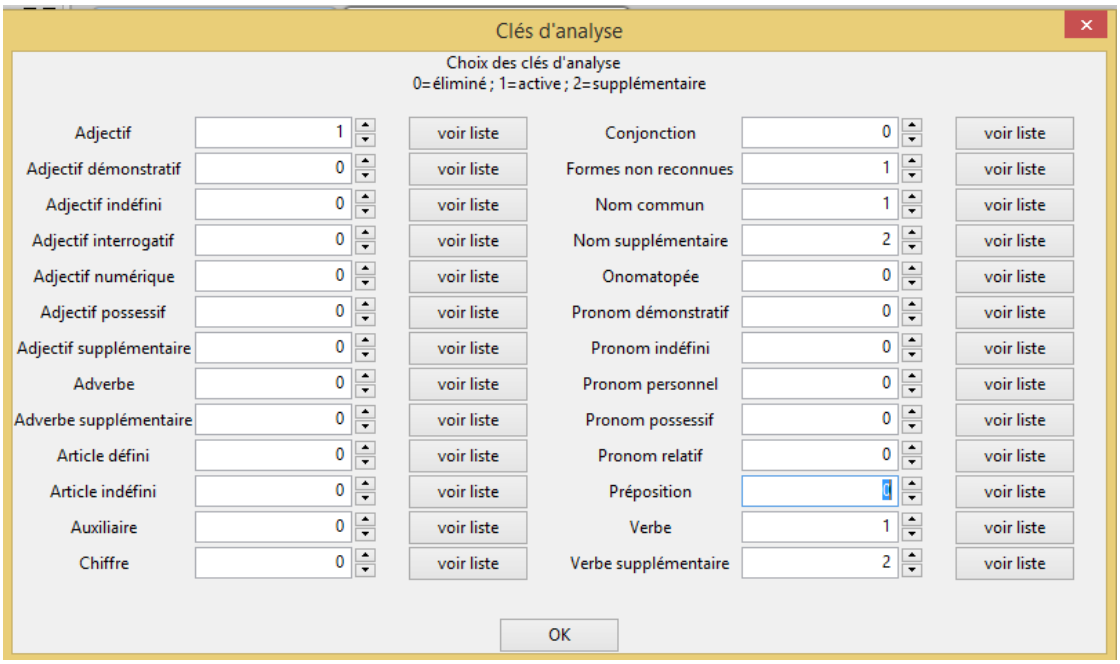
PICTURE 11- Text analysis options

Every time you pick an analysis, a new window appears asking if you want to keep the lemmatization. Keep YES selected so the software can use the reduced forms dictionary to run the analysis.

The window also allows you to edit the active and supplementary forms. If you wish to do so, press "Properties". It is recommended to select which grammatical clusters will be active in the analysis (0= eliminated words; 1= active words; 2= supplementary words). Once this change is performed it will be valid for all the further analysis on the same corpus. The researcher may change them again whenever he wants to. After choosing the grammatical cluster click OK twice.

Key Properties parameters:

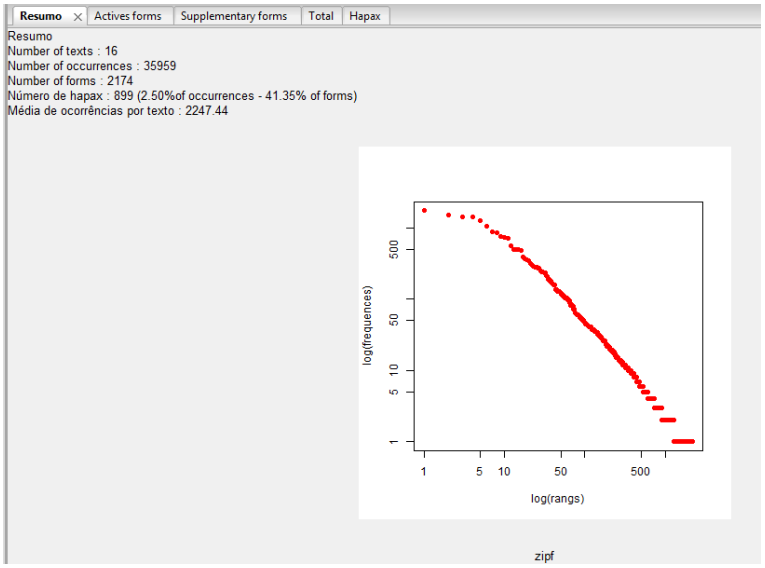
For research in Psychology, it is suggested to follow the example of Picture 12. These parameters are the most appropriate for research focused on the text contents. The idea is to work with some language elements as active: adjective, non -recognized forms, names, verbs; and names and auxiliary verbs as supplementary; eliminating the “tool words”. Moreover, select the words in similarity analysis and word cloud, and disregard the words with high frequency associated with questions.



PICTURE 12- Parameters of active, supplementary and eliminated words

Analysis: Text statistics

The first option, “Statistics”, provides the number of texts and texts segments, occurrences, medium frequencies, as well as the total frequency of each form, and its grammatical cluster, according to the dictionary of reduced forms. In the results interface it’s possible to visualize the Zipf diagram (Picture 13) which presents the word frequency in the corpus on a graphic with a X rang frequency distribution.



PICTURE 13- Zipf diagram

On the left column, you identify this analysis as: CORPUS NAME_stat_1. Click on this name with the mouse's right button to select other options, such as export the reduced forms dictionary, which will be saved in the same folder as the initial corpus, in a sub folder called: Corpus Name_stat_1.

The software classifies the words in grammatical forms, with the following coding, which will be used for every analysis henceforth:

Adj= adjective

Adj_num= numerical adjective

Adj_sup= adjective in supplementary form

Adv= adverb

Adv_sup= adverb in supplementary form

Art_def= definite article

Conj= conjunction

Nom= noun

Nom_sup= name in supplementary form

Nr= non recognized

Ono= onomathopea

Pro_ind= indefinite pronoun

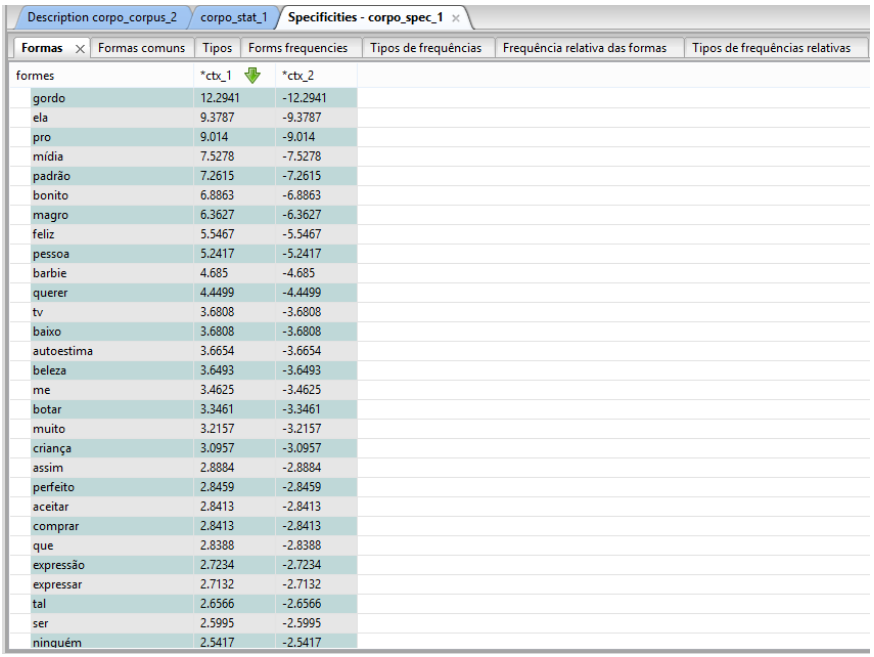
Pre= preposition

Ver= verb

Verbe_sup= verb in supplementary form

Analysis: specificities and CA

When selecting “Specificities and CA”, choose the categorical variable according to which you want to conduct the analysis. Select it and click OK. Wait for a few seconds for the results to appear in the main window (Picture 14).



formes	*ctx_1	*ctx_2
gordo	12.2941	-12.2941
ela	9.3787	-9.3787
pro	9.014	-9.014
mídia	7.5278	-7.5278
padrão	7.2615	-7.2615
bonito	6.8863	-6.8863
magro	6.3627	-6.3627
feliz	5.5467	-5.5467
pessoa	5.2417	-5.2417
barbie	4.685	-4.685
querer	4.4499	-4.4499
tv	3.6808	-3.6808
baixo	3.6808	-3.6808
autoestima	3.6654	-3.6654
beleza	3.6493	-3.6493
me	3.4625	-3.4625
botar	3.3461	-3.3461
muito	3.2157	-3.2157
criança	3.0957	-3.0957
assim	2.8884	-2.8884
perfeito	2.8459	-2.8459
aceitar	2.8413	-2.8413
comprar	2.8413	-2.8413
que	2.8388	-2.8388
expressão	2.7234	-2.7234
expressar	2.7132	-2.7132
tal	2.6566	-2.6566
ser	2.5995	-2.5995
ninguém	2.5417	-2.5417

PICTURE 14- Results, specificities and CFA

By pressing with the mouse’s right button on any word presented in the table (Picture 14) and in *Concordance,* a new window will appear where you can identify the text segments containing the word, hence getting back its context.

Analysis: Clustering (Descending hierarchical classification-DHC)

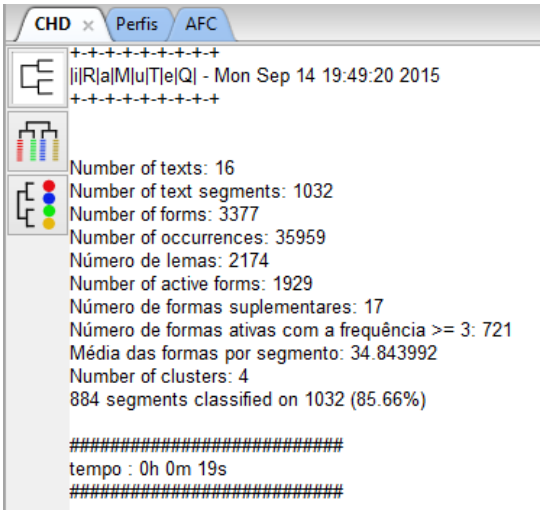
For Clustering (Descending hierarchical classification*, Reinert method*), you may choose between 3 options in the window displayed on IRAMUTEQ’s interface.

DOUBLE ON RST- not used due to a low exploration of the corpus

SIMPLE ON TEXT SEGMENTS- analogous to a TS analysis, defined by the software (Standard analysis), recommended for long answers

SIMPLE ON TEXTS- performs the analysis on texts, without dividing them in TS. Recommended for short answers[1]

Choose one of the options. You don't need to change any of the other parameters. Click OK and wait for a few seconds for the analysis to be completed. Some relevant data of DHA will be displayed (Picture 15), alongside with the dendrogram (Picture 16)

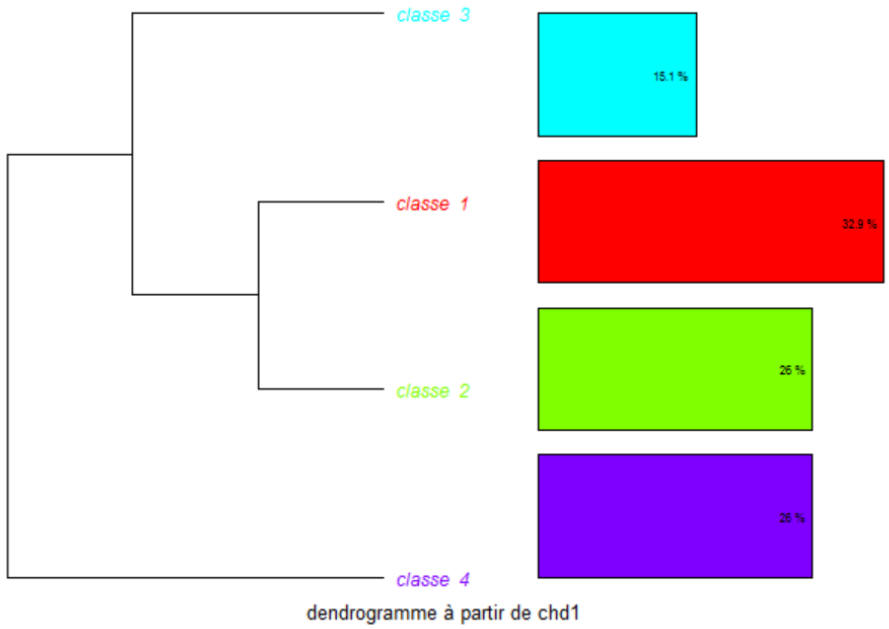


Picture 15- Main DHC aspects to consider

For the results description, the main features to be considered are:

- Number of texts= 16 (software recognizes the corpus division in 16 text units)
- Number of text segments= 1.032 (the software divides 1.032 text segments)
- Number of forms)3.377
- Number of occurrences =35.959
- Number of active forms: 1929
- Number of clusters = 4
- Text segments retention: **884 segments classified on 1.032 (85,66%)**

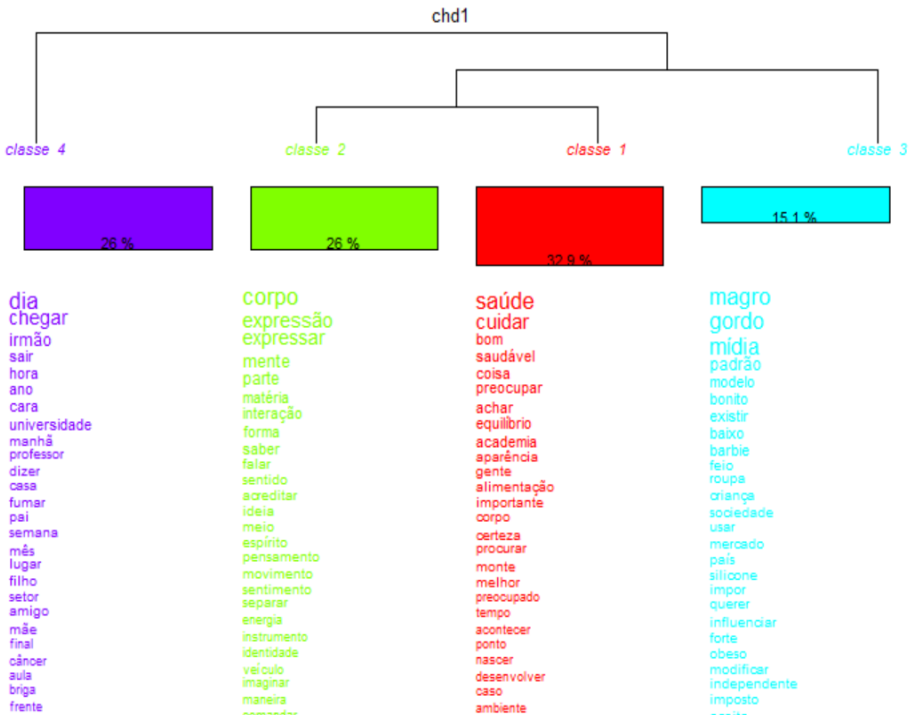
[1] In this case, you need the previous parameter. After the corpus import, besides indicating the coding and language, select "paragraphs" as the ***method to create TS***



Picture 16- DHC dendrogram

In the DHC results tab, you can access the dendrogram with the text divisions and final clusters. It is to be read from left to right. In the following example (Picture 16), the corpus “Body”, was divided (1º division, or iteration) in two sub corpus, separating cluster 4 from the rest. On a second moment, the larger sub corpus was divided, generating cluster 3 (2º division or iteration). On a third moment, another division generated clusters 1 and 2. The DHA stopped here, due to the 4 clusters stability, as text segments units with similar vocabulary.

The dendrogram can also be graphically presented as shown in Picture 17.

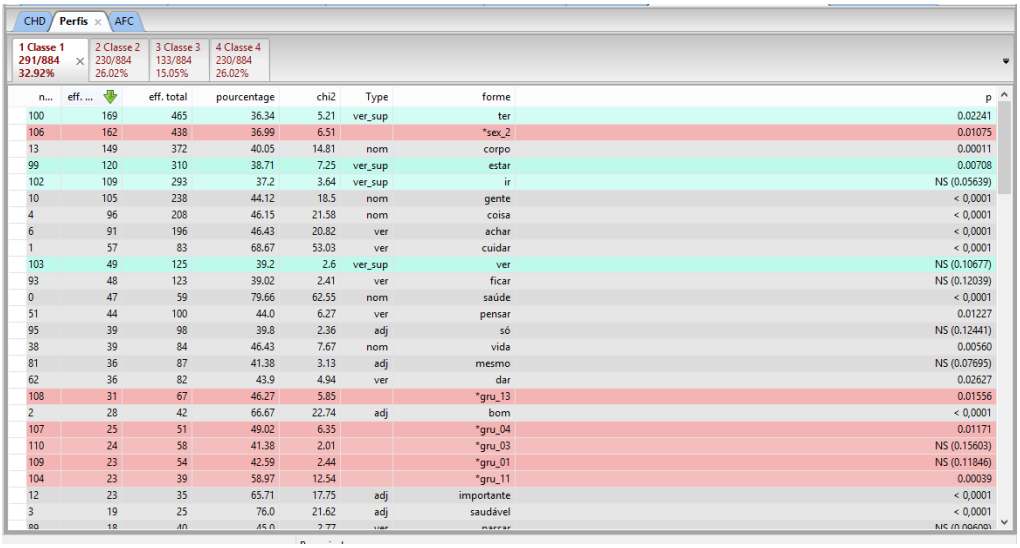


PICTURE 17- DHC dendrogram

DHC data exploration

This interface also provides users with the identification of lexical content of each cluster (click *Profiles*) and a factor representation of DHC (press *CA*)

The *Profile*s tab shows data of each cluster content: *n.* (number which organizes the words in the table); *eff. st* (number of text segments containing the word in the cluster); *eff. Total (* number of text segments containing at least once the cited word); *pourcentage (*percentage of word occurrence on the text segments of this cluster in relation to its occurrence in the corpus); *chi2* (X^2 of association between word and cluster); *Type (*grammatical cluster identifying the word in the forms dictionary*); Forme (*identifies the word) and *P (*identifies the significance level of the association between word and cluster). Picture 18 shows the "*Profiles*" bar.



n...	eff. ...	eff. total	pourcentage	chi2	Type	forme	p
100	169	465	36.34	5.21	ver_sup	ter	0.02241
106	162	438	36.99	6.51		*sex_2	0.01075
13	149	372	40.05	14.81	nom	corpo	0.00011
99	120	310	38.71	7.25	ver_sup	estar	0.00708
102	109	293	37.2	3.64	ver_sup	ir	NS (0.05639)
10	105	238	44.12	18.5	nom	gente	< 0,0001
4	96	208	46.15	21.58	nom	coisa	< 0,0001
6	91	196	46.43	20.82	ver	achar	< 0,0001
1	57	83	68.67	53.03	ver	cuidar	< 0,0001
103	49	125	39.2	2.6	ver_sup	ver	NS (0.10677)
93	48	123	39.02	2.41	ver	ficar	NS (0.12039)
0	47	59	79.66	62.55	nom	saúde	< 0,0001
51	44	100	44.0	6.27	ver	pensar	0.01227
95	39	98	39.8	2.36	adj	só	NS (0.12441)
38	39	84	46.43	7.67	nom	vida	0.00560
81	36	87	41.38	3.13	adj	mesmo	NS (0.07695)
62	36	82	43.9	4.94	ver	dar	0.02627
108	31	67	46.27	5.85		*gru_13	0.01556
2	28	42	66.67	22.74	adj	bom	< 0,0001
107	25	51	49.02	6.35		*gru_04	0.01171
110	24	58	41.38	2.01		*gru_03	NS (0.15603)
109	23	54	42.59	2.44		*gru_01	NS (0.11846)
104	23	39	58.97	12.54		*gru_11	0.00039
12	23	35	65.71	17.75	adj	importante	< 0,0001
3	19	25	76.0	21.62	adj	saudável	< 0,0001

PICTURE 18- Forms associated with cluster 1

For the descriptive analysis of each cluster, 2 criteria should be considered: 1) pay attention to the non- instrumental words with a higher frequency than the medium frequency of the entire corpus' set of words (in this example 35.959 occurrences divided by 3.377 distinct forms, resulting in 10.65) and 2) consider the words with X^2 of cluster association ≥ 3.84 (hence $p < 0,05$).

More results are presented in the left column, by clicking the mouse's right button on the analysis - Corpus name_alceste_1. The most important ones are:

- **Colorful corpus**- opens an interface of your Internet's browser allowing you to visualize the typical text segments of each cluster, identifying the clusters by colors, according to the ones in the dendrogram (see Picture 19).
- **Report**- adds a document in .txt, called RAPPORT in the folder with the corpus, in a sub folder called CORPUS NAME_alceste_1. This document, which can be visualized in any text editor, has the lexical description of each cluster composed by DHC, like a simplified report of the analysis.

****** *gru_01 *ctx_1 *ida_1 *sex_2**

The word I thought of was visit card because of the video images and the idea is that the body is the visit card is what you present

So you need to have a presentable hair be well dressed fit It is also related to identity question because is how you present yourself

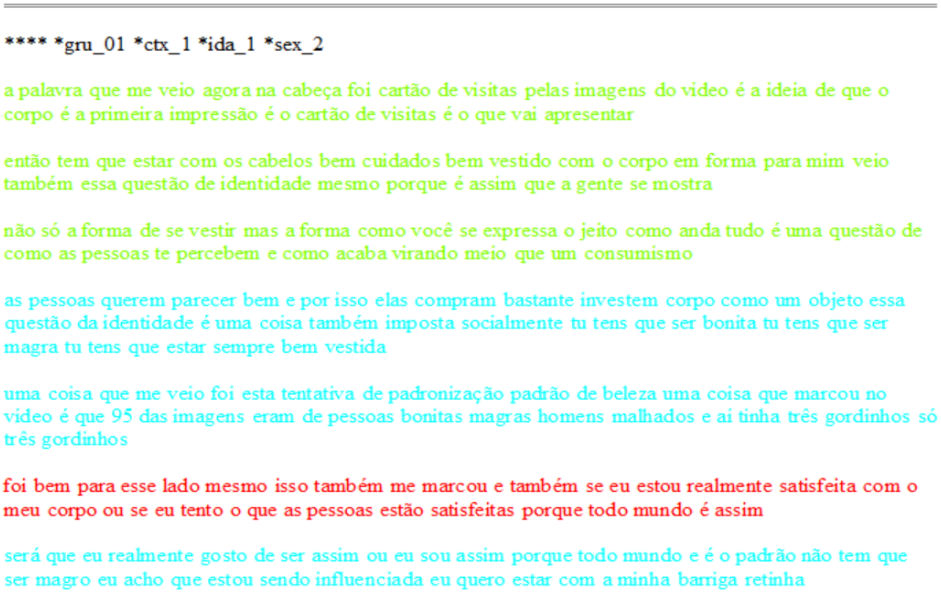
Not only how you dress up but the way you express yourself the way you walk is a question of how people perceive you and how it becomes consumerism

People want to look good thus they buy a lot invest on their bodies as an object this identity question is a thing also socially imposed, you have to be pretty, you have to be skin you also have to dress up.

The thing I thought was the attempt to create beauty patterns a thing patent in the video is that 95% of the shots were of beautiful people slim fit men and only three fatties, only three fatties

Exactly that also impressed me and also if I'm really pleased with my body or if I try everyone is satisfied because everyone is like that

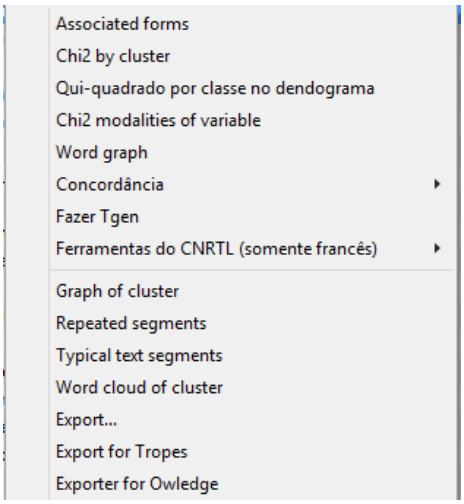
Do I really like to be like that or I'm like this because everyone is and is the pattern doesn't have to be slim I think I'm being influenced I want to be with my belly slim



PICTURE 19 – Colorful corpus

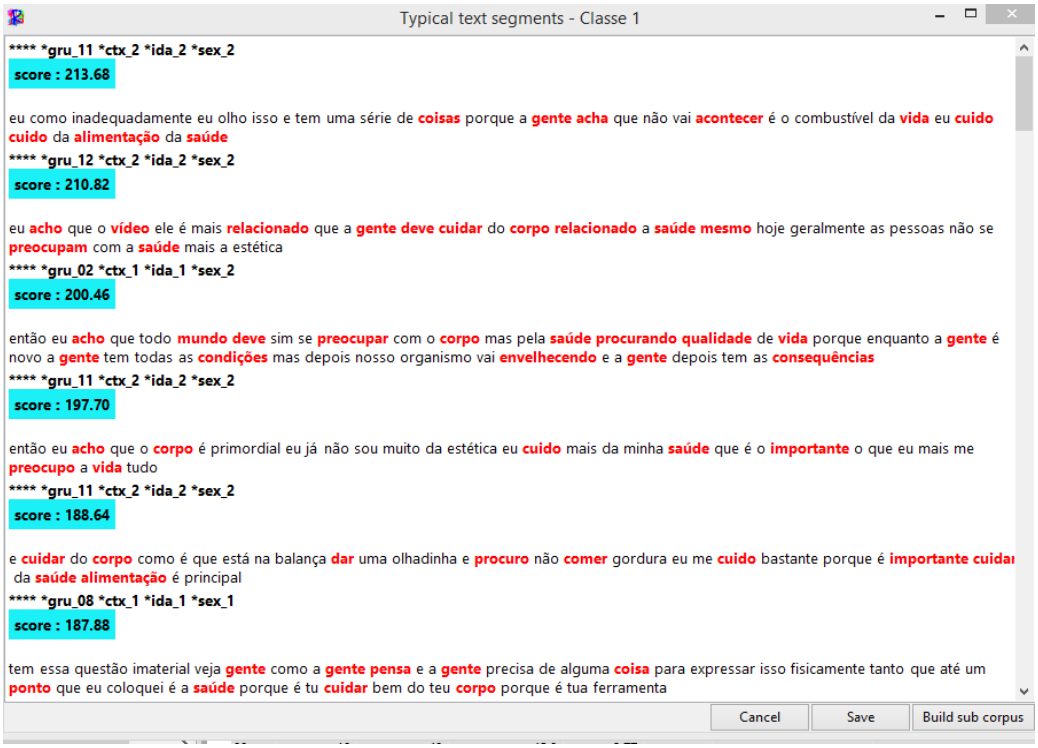
To be efficient, DHC requires a minimal retention of 75% of the text segments (some authors mention 70%). Any inferior retention is not acceptable corresponding to a partial classification.
On these cases (when text segments retention is inferior to 75%) it is suggested to drop out the method and apply other resources, such as specificities analysis.

Still in *Profiles*, the content of each cluster may be explored using other available resources (see Picture 20), shown when you click with the mouse's right button on any word pertaining to a cluster. On the top of the window you can consult more data referring to the selected word. The lower part provides information related to the respective cluster.



PICTURE 20- Resources to interpret each cluster

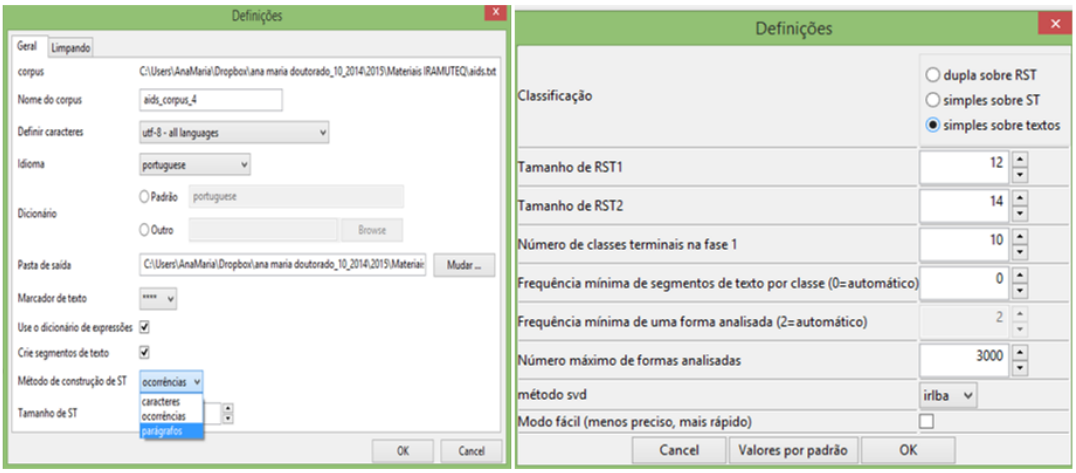
The resources presented in this window (Picture 20) provide you with the words related to form (from the reduced forms dictionary); the graphic visualization of frequency, association and co-occurrence of a specific word, as well as the text segments where the word appears in the cluster. It's also possible to visualize a graph of clusters, repeated segments, typical text segments (see Picture 21), as well as export the segments related to the cluster.



PICTURE 21- Typical text segments of cluster 1

Descending hierarchical classification on short answers (questionnaire)

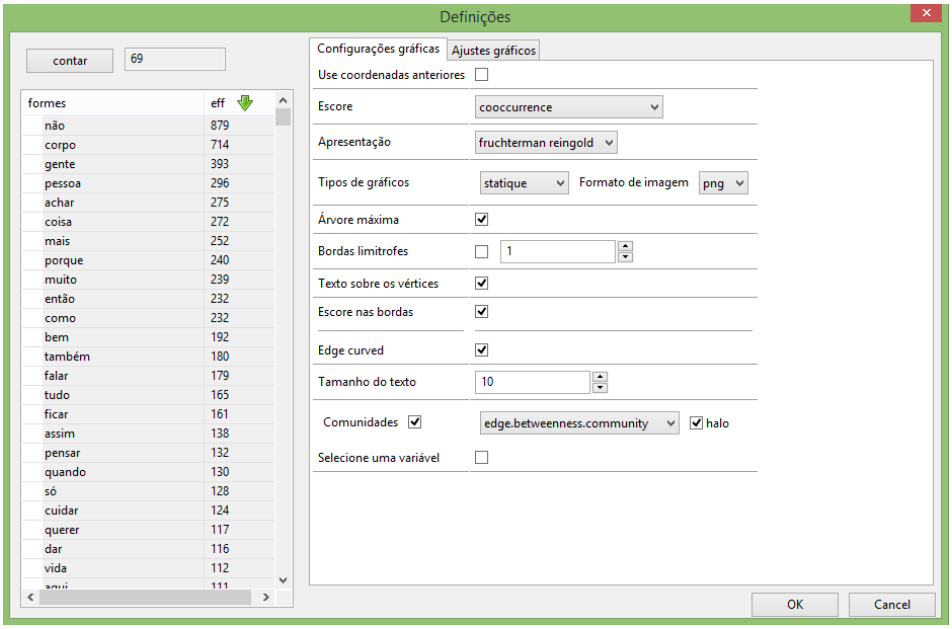
Whit many short answers to an open ended question, it is necessary to adapt the DHC (see picture 22). When importing this type of corpus, you should identify the coding and language and select “paragraphs” as method to create text segments (TS). Then, choose *Classification “*simple on texts” in order to avoid the segmentation of each answer. Thus, the text segment considered will be the text itself or the short answer to a questionnaire.



PICTURE 22- Method configuration for TS creation

Analysis: Similarity

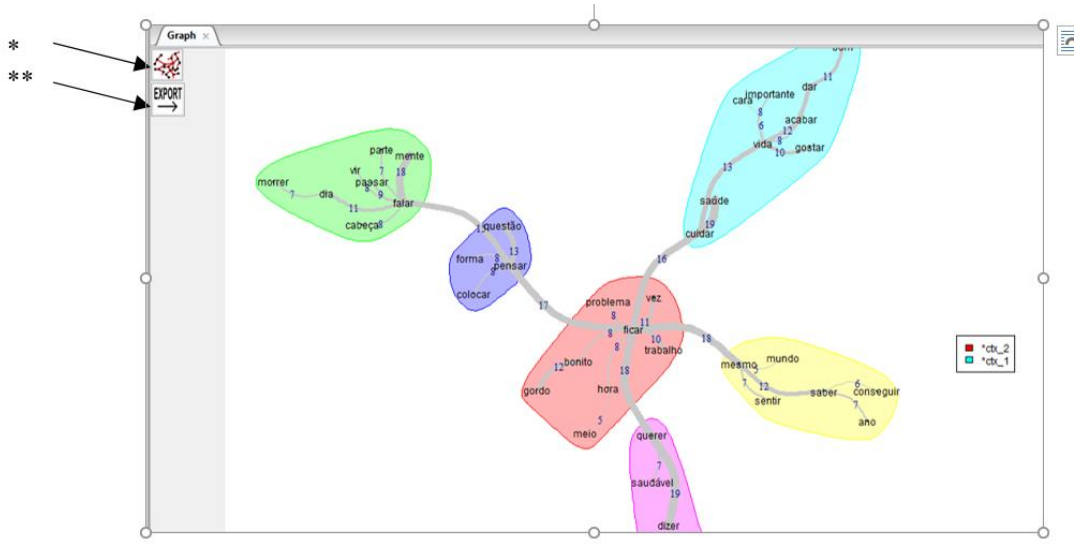
For the similarity analysis, a new window is displayed (Picture 23), where you can choose the criteria for the co -occurrences tree. In *Graph Settings,* you may edit the analysis, change the co- occurrences index, choose if it's a maximum tree or not, select a descriptive variable to be highlighted in the tree. Click on *Communities + Halo* where you may ask for the most related words to be presented in a colorful cloud. And on the *Score on Edges,* you can visualize the co –occurrences values.



PICTURE 23- Parameter edition window for similarity analysis

You can select words to integrate the analysis on the left column. Click on "*Select a variable*" to choose a categorical variable for the similarity analysis, identifying the differences between groups.

After choosing the criteria click OK and wait for the analysis to finish.



PICTURE 24- Similarity Analysis results

The tree is displayed on results interface which has 2 buttons on the upper left corner (Picture 24). The first one (*) with red lines and black dots allows you to change the analysis parameters, reopening the edition window. The second button (**), EXPORT, will export the image for the analysis folder, in a sub folder called **Corpus Name_simitxt_1.**

Analysis: Word cloud

A new window is displayed when choosing word cloud. Like the one of similarity analysis you may also choose some parameters, which don't need necessarily to be edited.

This is a simpler analysis which represents graphically the word frequency. After choosing the criteria, press ok on both windows and wait a few seconds.

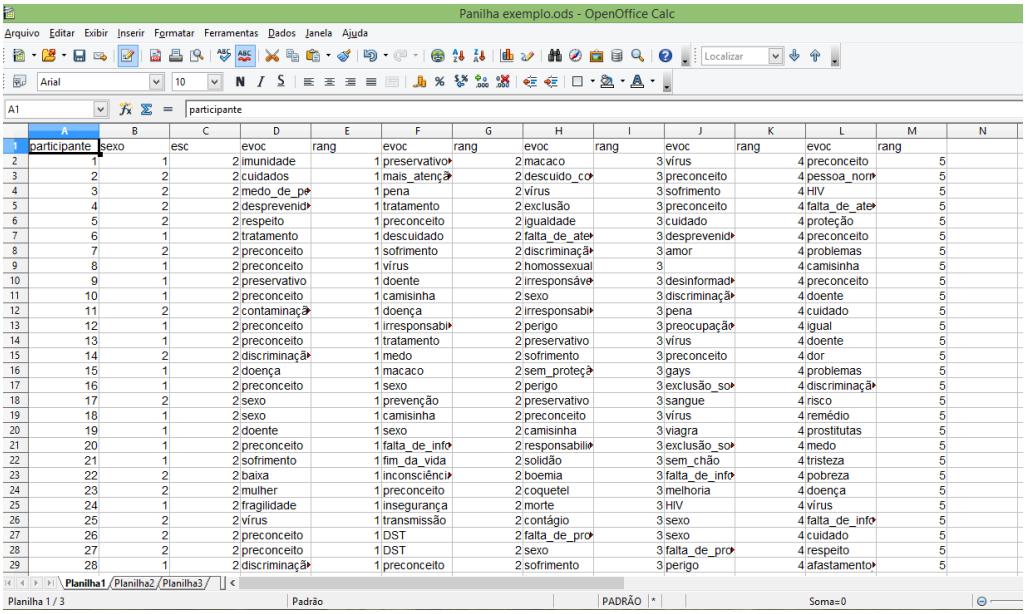
PICTURE 25- Word cloud results

You can visualize the word cloud in the results interface (Picture 25) as well as inside the analysis folder (sub folder “Corpus Name_wordcloud_1”, as an image archive called “nuage_1”).

All the results, including the pictures and graphics will be inside the folder containing the corpus of analysis as well as each analysis run (statistics, specificities, DHC, similarity and word cloud).

PART 2: MATRIX ANALYSIS

IRAMUTEQ allows users to analyze matrices involving categorical variables and word lists (similar to the ones of free evocation tasks), conducting frequency calculus, prototypical analysis and similarity analysis. The database should be created with Open Office Calc (see Picture 26).



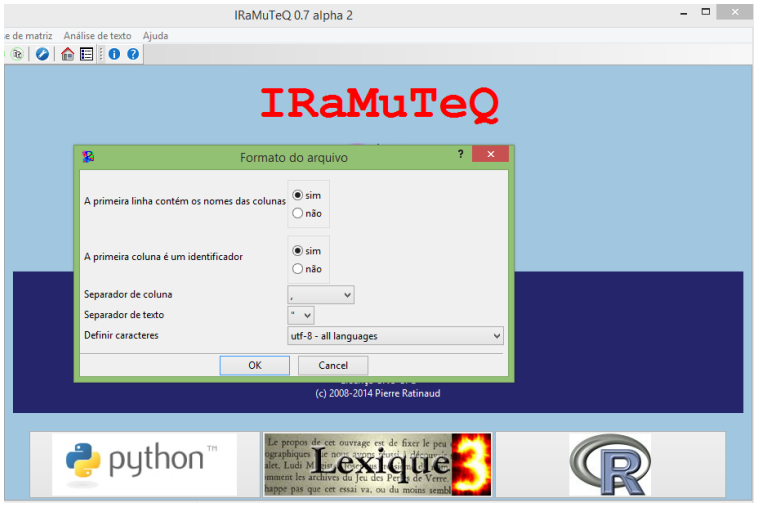
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	participante	sexo	esc	evoc	rang	evoc	rang	evoc	rang	evoc	rang	evoc	rang	
2	1	1	2	imunidade	1	preservativo	2	macaco	3	vírus	4	preconceito	5	
3	2	2	2	cuidados	1	mais_atenç	2	descuido_co	3	preconceito	4	pessoa_norr	5	
4	3	2	2	medo_de_pe	1	pena	2	vírus	3	sofrimento	4	HIV	5	
5	4	2	2	desprevenid	1	tratamento	2	exclusão	3	preconceito	4	falta_de_ate	5	
6	5	2	2	respeito	1	preconceito	2	igualdade	3	cuidado	4	proteção	5	
7	6	1	2	tratamento	1	descuidado	2	falta_de_ate	3	desprevenid	4	preconceito	5	
8	7	2	2	preconceito	1	sofrimento	2	discriminaçã	3	amor	4	problemas	5	
9	8	1	2	preconceito	1	vírus	2	homossexual	3		4	camisinha	5	
10	9	1	2	preservativo	1	doente	2	irresponsáve	3	desinformad	4	preconceito	5	
11	10	1	2	preconceito	1	camisinha	2	sexo	3	discriminaçã	4	doente	5	
12	11	2	2	contaminaçã	1	doença	2	irresponsabi	3	pena	4	cuidado	5	
13	12	1	2	preconceito	1	irresponsabi	2	perigo	3	preocupação	4	igual	5	
14	13	1	2	preconceito	1	tratamento	2	preservativo	3	vírus	4	doente	5	
15	14	2	2	discriminaçã	1	medo	2	sofrimento	3	preconceito	4	dor	5	
16	15	1	2	doença	1	macaco	2	sem_proteçã	3	gays	4	problemas	5	
17	16	1	2	preconceito	1	sexo	2	perigo	3	exclusão_so	4	discriminaçã	5	
18	17	2	2	sexo	1	prevenção	2	preservativo	3	sangue	4	risco	5	
19	18	1	2	sexo	1	camisinha	2	preconceito	3	vírus	4	remédio	5	
20	19	1	2	doente	1	sexo	2	camisinha	3	viagra	4	prostitutas	5	
21	20	1	2	preconceito	1	falta_de_info	2	responsabili	3	exclusão_so	4	medo	5	
22	21	1	2	sofrimento	1	fim_da_vida	2	solidão	3	sem_chão	4	tristeza	5	
23	22	2	2	baixa	1	inconsciênci	2	boemia	3	falta_de_info	4	pobreza	5	
24	23	2	2	mulher	1	preconceito	2	coquetel	3	melhoria	4	doença	5	
25	24	1	2	fragilidade	1	insegurança	2	morte	3	HIV	4	vírus	5	
26	25	2	2	vírus	1	transmissão	2	contágio	3	sexo	4	falta_de_info	5	
27	26	2	2	preconceito	1	DST	2	falta_de_pro	3	sexo	4	cuidado	5	
28	27	2	2	preconceito	1	DST	2	sexo	3	falta_de_pro	4	respeito	5	
29	28	1	2	discriminaçã	1	preconceito	2	sofrimento	3	perigo	4	afastamento	5	

PICTURE 26- Database model for matrices analysis

It is recommended to create the database according to the following:

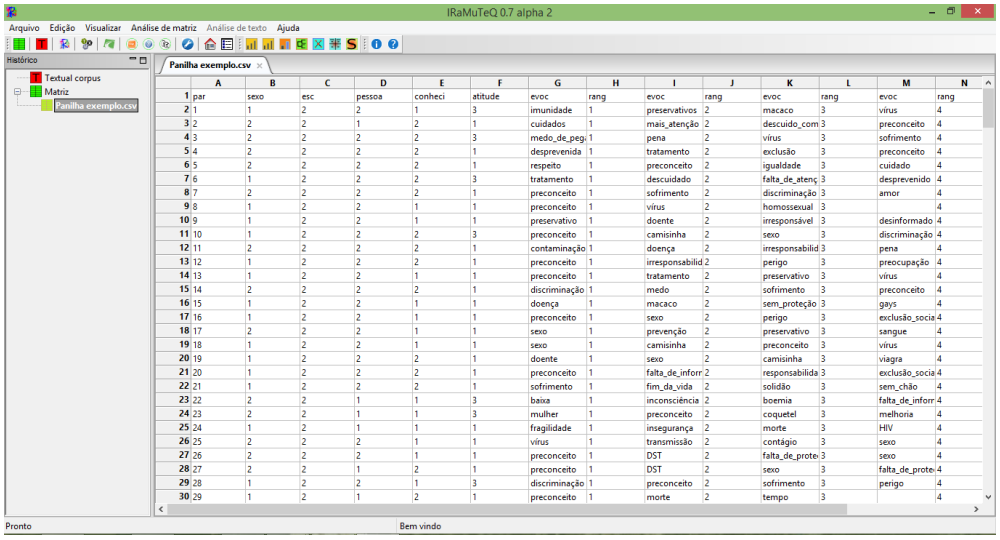
- Archive format should be: ods; csv; xls (don't use xlsx- excel because its incompatible with IRAMUTEQ). The coding must be the same used for text analysis: **UTF8 all languages**
- Avoid the following characters : ; ' "
- Don't insert blank spaces (use underscore to connect more than 2 words)
- The archive's name can't include accentuation or special characters
- The numeric variables can be presented in the archive but can't be used in the analysis (except rangs in prototypical analysis).
- If you know the order or importance of the words, this may be added in a column immediately after the word.
- A broad corpus revision is necessary, since this type of analysis doesn't do lemmatization.

Save the database inside an exclusive folder for the analysis, open IRAMUTEQ and select *File* and *Open a matrix.* Locate the archive containing the database and press *Open.* To import the data, another window will appear (see Picture 27) and you'll be able to indicate some parameters of the database: the first line of the spreadsheet must include the column names (indicated); the first column is an identifier (indicated); column delimiter (will be , in case of CSV format); text delimiter (") character coding (utf-8-all languages)



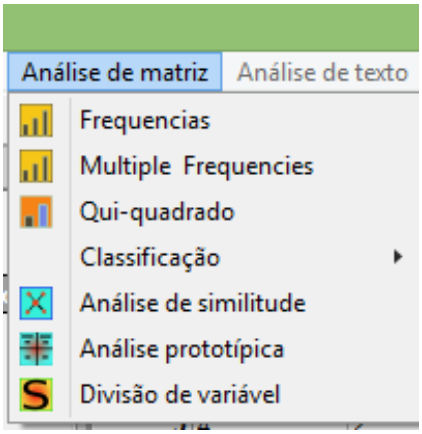
PICTURE 27- Matrix database import

Select the parameters and press ok to access the imported matrix (Picture 28). The available analysis are frequencies, descending hierarchical classification (recommended only with a high number of participants), similarity analysis and prototypical analysis.



PICTURE 28- Imported matrix

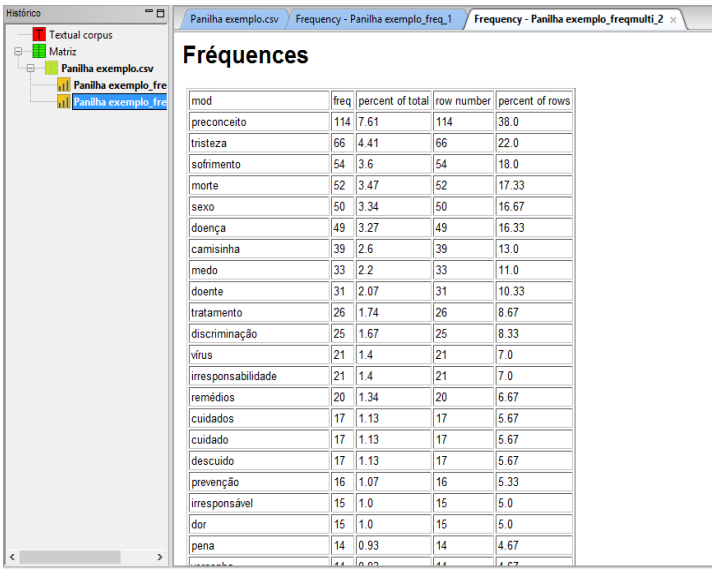
To run the analyses, click Matrix analysis icon and select (Picture 29).



PICTURE 29- Possible analysis for the matrices

Frequency analysis is the simpler one. Click *Frequency* analysis to access frequencies of the matrix's categorical variables and *Multiple Frequencies* analysis to obtain a report for the absolute and relative frequency of the words in the matrix. It's necessary to choose which variables are going to be calculated. In this case, there is no interest in Rang (evocation order) but only in words and eventually on descriptive variables inserted in the matrix.

Picture 30 shows a multiple frequency report related to evoked words and a free association test.



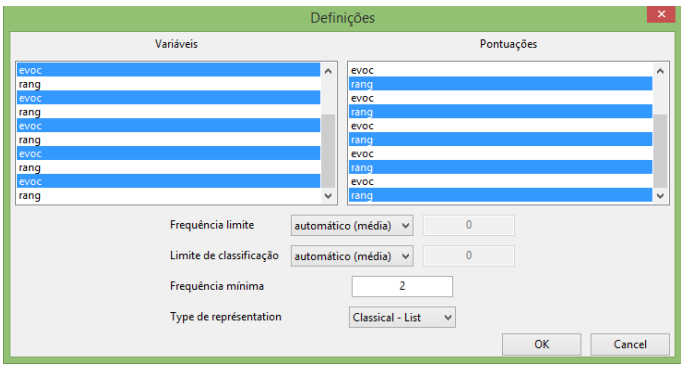
mod	freq	percent of total	row number	percent of rows
preconceito	114	7.61	114	38.0
tristeza	66	4.41	66	22.0
sofrimento	54	3.6	54	18.0
morte	52	3.47	52	17.33
sexo	50	3.34	50	16.67
doença	49	3.27	49	16.33
camisinha	39	2.6	39	13.0
medo	33	2.2	33	11.0
doente	31	2.07	31	10.33
tratamento	26	1.74	26	8.67
discriminação	25	1.67	25	8.33
vírus	21	1.4	21	7.0
irresponsabilidade	21	1.4	21	7.0
remédios	20	1.34	20	6.67
cuidados	17	1.13	17	5.67
cuidado	17	1.13	17	5.67
descuido	17	1.13	17	5.67
prevenção	16	1.07	16	5.33
irresponsável	15	1.0	15	5.0
dor	15	1.0	15	5.0
pena	14	0.93	14	4.67

PICTURE 30-Frequency analysis

As shown in Picture 30, the analysis provides a table with words ordered by frequency, also including the absolute frequency on the second column, followed by its proportion

in relation to the total of evocations. Also includes the number of lines with this word as well as the proportion related to the total number of lines. Each line represents a participant.

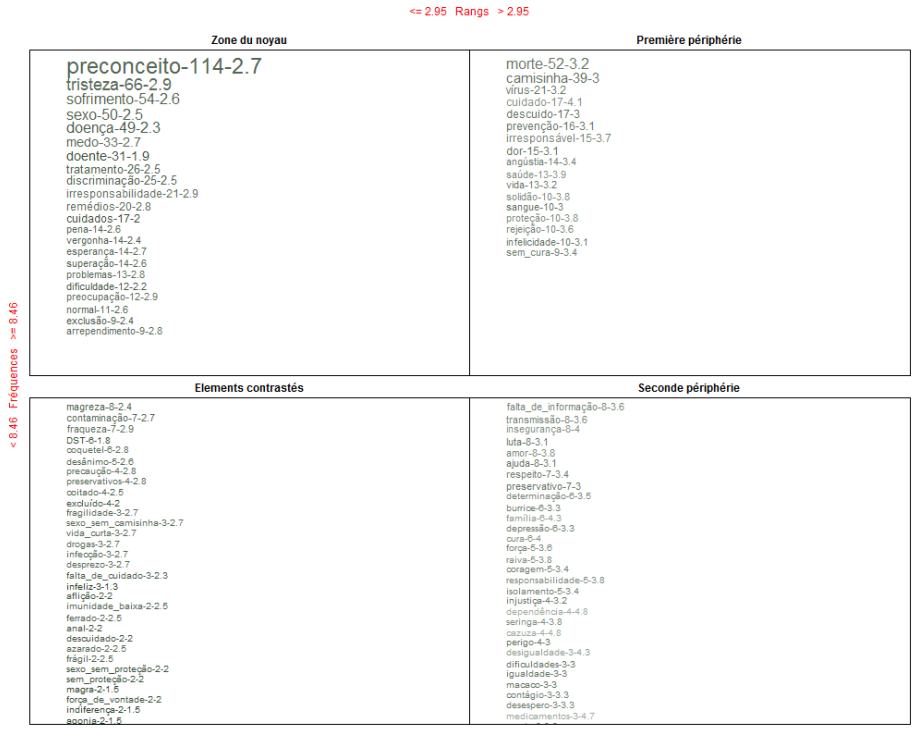
The prototypical analysis is a simple and efficient technique specifically developed by the social representations field. It is aimed at identifying the representational structure according to the frequency and word evocation order derived from a free evocation test (Wachelke & Wolter, 2011). You can conduct this analysis pressing *Matrix analysis* followed by *prototypical analysis.*



PICTURE 31- prototypical analysis configuration

In the configuration window, select the variables related to evocations (on the left side) and the variables corresponding to RANG (on the right side) - (depending on your criteria, evocation order or attributed importance). The other parameters refer to the criteria for calculate the prototypical analysis, you can keep the automatic definitions (see Picture 31).

Standards defined, press ok to access the prototypical analysis (Picture 32). This 4 quadrants diagram represents four dimensions of the social representations structure. In this example, using a free evocation task with the inductive term AIDS, the first quadrant (upper left side) indicate the words with high frequency (higher frequency than the media) and low evocation frequency (those immediately evoked). These probably correspond to the central nucleus of a representation.

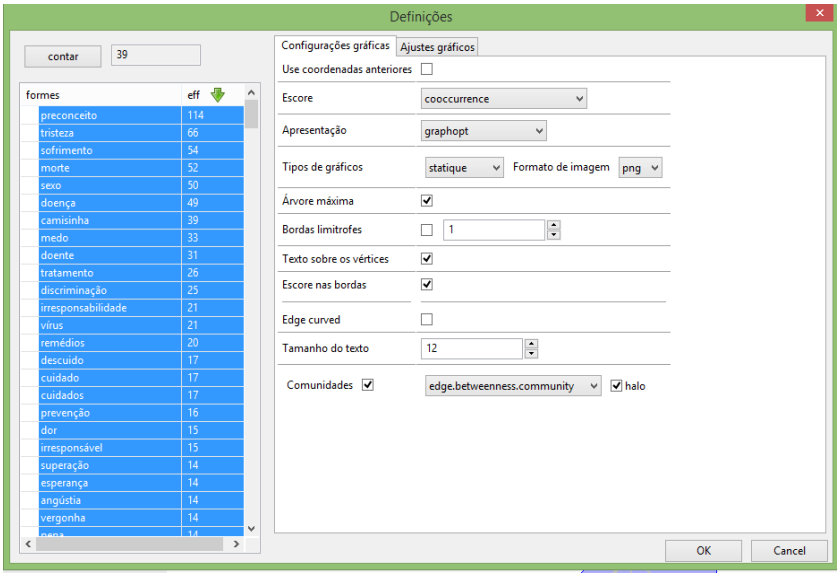


PICTURE 32- four quadrants diagram- prototypical analysis

The second quadrant (upper right side), corresponds to the first periphery, including words with high frequency but with a higher media, thus not so readily evoked. The third quadrant (bottom left side) corresponds to the contrast zone with elements readily evoked but with a lower frequency. The second periphery in the fourth quadrant (bottom right side) indicates the elements with lower frequency and higher evocation order.

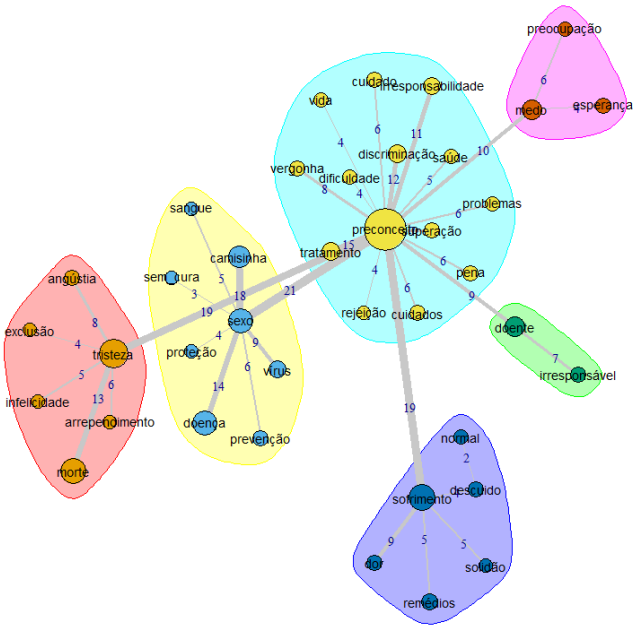
At last, the similarity analysis, also an indicator of a social representations structure, can be conducted from *Matrix analysis* and *Similarity analysis.*

The analysis is analogous to the one performed with text material (see Picture 33).



PICTURE 33- Similarity analysis definitions

A similarity analysis is presented in Picture 34, where the colorful vertices size is proportional to the words frequency and the edges indicate the words co - occurrence strength.



PICTURE 34- Similarity Analysis

References

Antunes, L. (2013). O papel dos estereótipos nas representações sociais compartilhadas por adolescentes sobre as pessoas que vivem com HIV/aids. *Dissertação de Mestrado* (não publicada). Programa de Pós-Graduação em Psicologia. Universidade Federal de Santa Catarina. Florianópolis, SC.

Camargo, B. V., Justo, A. M. (2013). *IRAMUTEQ: Um Software Gratuito para Análise de Dados Textuais.*Temas em Psicologia, 21 (2), 513-518.

Cibois, P. (1990). *L'analyse des données en sociologie.* Paris: P.U.F.

Cros, M. (1993). Les apports de la linguistique: langage des jeunes et sida. In ANRS (Agence Nationale de Recherche sur le Sida). *Les jeunes face au Sida: de la recherche à l'action* (pp. 50-61). Paris: ANRS.

Ghiglione, R.; Matalon, B. (1993). *O inquérito: Teoria e prática.* Oeiras: Celta.

Justo, A. M. (2011). Representações sociais sobre o corpo e implicações do contexto de inserção desse objeto. *Dissertação de Mestrado* (não publicada). Programa de Pós-Graduação em Psicologia. Universidade Federal de Santa Catarina. Florianópolis, SC.

Justo, A. M.; Camargo, B. V. (2014). Estudos qualitativos e uso de softwares para análises lexicais. Em: C. Novikoff; S. R. M. Santos; O. B. Mithidieri (Orgs.). *Cadernos de artigos: X SIAT e II SERPRO Lageres/UNIGRANRIO* (pp. 37-54). Duque de Caxias: UNIGRANRIO.

Lahlou, S. (2012). Text Mining Methods: An answer to Chartier and Meunier. *Papers on Social Representation*s, 20 (38), 1-7.

Lebart, L. & Salem, A. (1988). *Analyse statistique des données textuelles*. Paris: Dunod.

Marchand, P.; P. Ratinaud. (2012). L'analyse de similitude appliqueé aux corpus textueles: les primaires socialistes pour l'election présidentielle française. Em: *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012*. (687–699). Presented at the 11eme Journées internationales d'Analyse Statistique des Données Textuelles. JADT 2012. Liège, Belgique

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires *[Computer software].* Recuperado em 5 março, 2013, de http://www.iramuteq.org

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IraMuTeQ. Em: *Actes des 11eme Journées internationales d'Analyse statistique*

des Données Textuelles (835–844). Presented at the 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012, Liège.

Reinert, M. (1990). ALCESTE, une méthodologie d'analyse des données textuelles et une application: Aurélia de G. de Nerval. *Bulletin de méthodologie sociologique*, (28) 24- 54.

Veloz, M. C. T.; Nascimento-Schulze, C. M.; Camargo, B. V. (1999). Representações sociais do envelhecimento. *Psicologia: Reflexão e Crítica,* 12 (2), 479-501.

Wachelke, J. F. R. & Wolter, R. (2011). Critérios de construção e relato da análise prototípica para representações sociais*. Psicologia Teoria e Pesquisa*, 27 (4), 521-526.