Tutorial para uso do *software*

(Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires)

Brigido Vizeu Camargo e Ana Maria Justo

Laboratório de Psicologia Social da Comunicação e Cognição - UFSC – Brasil

www.laccos.com.br

Florianópolis, 21 de novembro de 2018

Sumário

O IRaMuTeQ é um *software* licenciado por GNU GPL (v2) que permite fazer análises estatísticas sobre *corpora*[1] textuais e sobre tabelas indivíduos/palavras (Loubère & Ratinaud, 2014). Ele ancora-se no *software* R (www.r-project.org) (Aquino, 2014) e na linguagem *python* (www.python.org).

Figura 1- Interface inicial do software IRaMuTeQ

Para instalar o *software* gratuitamente em seu computador, basta fazer o *download* do *software* R em www.r-project.org e instalá-lo; e em seguida fazer o *download* do *software* IRaMuTeQ em www.iramuteq.org, e instalá-lo também. É necessário que antes de instalar o IRaMuTeQ se instale o R, pois este software utilizará o *software* R para processar suas análises. No caso do sistema operacional Windows pode-se usar o kit IRaMuTeQ disponível no LACCOS/ UFSC – Brasil (https://drive.google.com/drive/u/1/folders/0B1sJtjYHLc94QnAxQkM5RmM3MVU).[2]

[1] Plural de "*corpus*".
[2] O kit IRaMuTeQ do LACCOS oferece três pastas: "*Softwares"*, "Referências" e "Corpora" (com 4 *corpora* em língua portuguesa para exercício).

Instalação do software para sistema operacional *Windows* usando o "*Kit IRaMuTeQ"*

(Necessário estar conectado à internet)

O kit oferece uma pasta denominada *Softwares*, ela contém 5 arquivos, três programas (LibreOffice_6.0.4_Win_x64, R-3.5.1-win, setup_iramuteq-0.7-alpha2) e dois complementos (Rgraph.R e lexique_pt).

1- Instale o *software* "LibreOffice (LibreOffice_6.0.4_Win_x64)

Ele é o equivalente gratuito do pacote Microsoft Office. Dois tipos de arquivos deste pacote Libre Office nos interessam: o Documento Writer, que cria arquivos de texto tipo "odt"; e o Planilha Calc que cria arquivos tipo planilha "ods". O primeiro é usado para digitar os *corpora* e ler relatórios e resultados, e o segundo para entrar dados sob a forma de matrizes de associação de palavras e também para ler, e exportar resultados**. Não abra estes arquivos** ou qualquer outro gerado pelo IRaMuTeQ **com aplicativos da Microsoft (Word, Excel, WordPad ou Bloco de notas),** pois eles produzem bugs com o Unicode (UTF-8), o usado pelo *software*.

2- Instale o *software* R (R-3.5.1-win)

Após instalar o *software* "R" abra-o no menu "Pacotes" selecione "Atualizar pacotes".

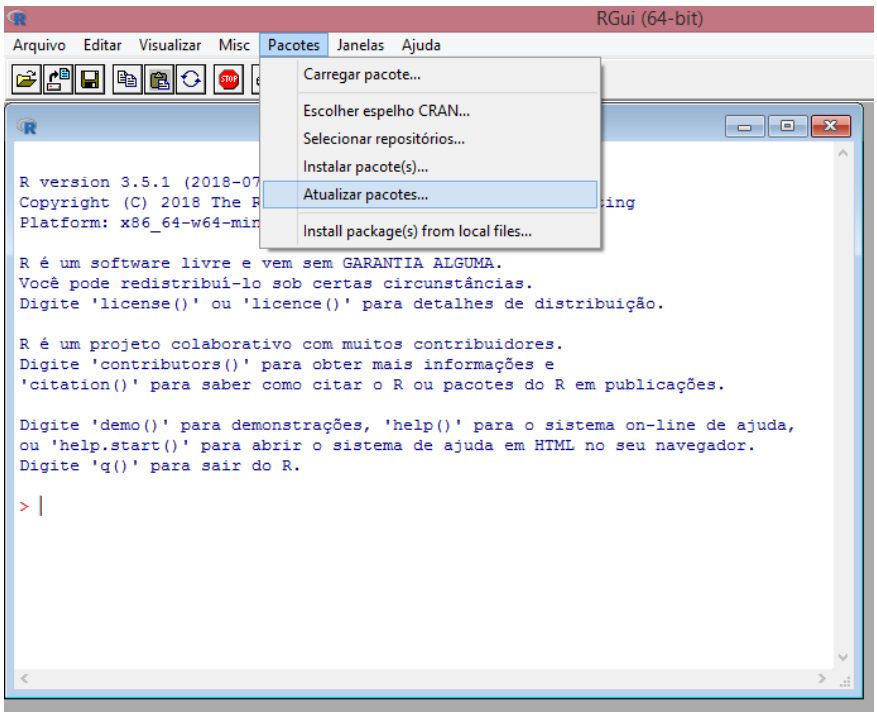


Figura 2- Atualização dos pacotes nas interfaces R

Escolha o espelho do estado mais próximo de seu local.

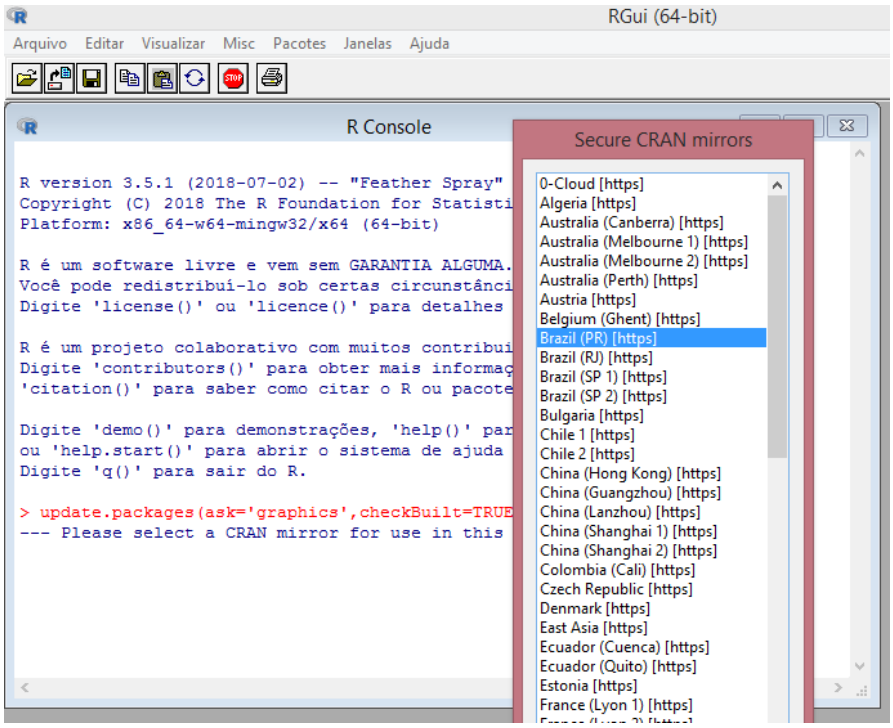


Figura 3- Escolha do espelho para atualização dos pacotes nas interfaces R

Quando aparecer uma caixa com todos os nomes, apenas clique em OK. Aguarde até que tenha a certeza de que o processo tenha finalizado. Dependendo da velocidade de seu computador e da internet pode demorar.

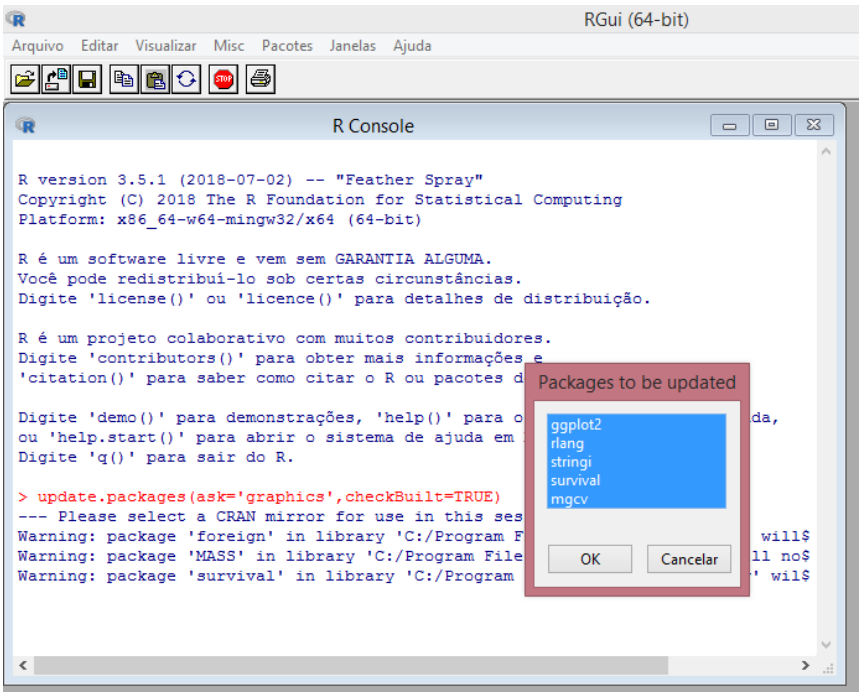


Figura 4- Baixar os pacotes nas interfaces R

3- Instale o *software* IRaMuTeQ (setup_iramuteq-0.7-alpha2)

Após instalar o *software* abra-o, se automaticamente ele fizer atualizações de pacotes do R, deixe o processo ir até o final (geralmente há uma demora nesta etapa).

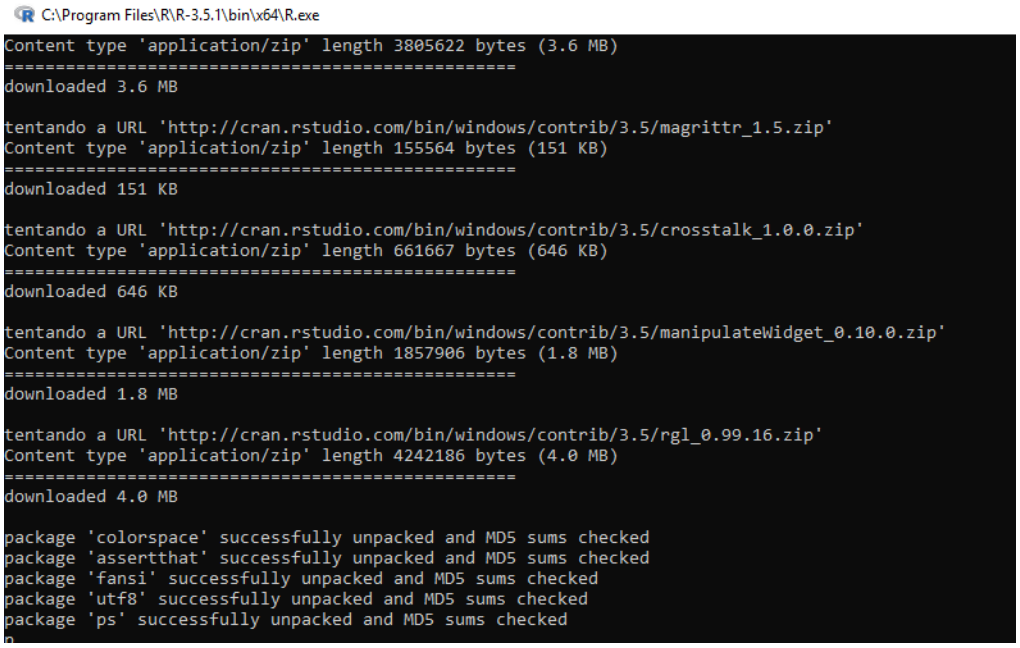


Figura 5- Tela de atualização das bibliotecas do R para o IRaMuTeQ

Pode aparecer uma tela apontando que a instalação está incompleta, listando alguns pacotes, clique em OK e aguarde o tempo necessário para atualização completa dos arquivos do *software* R.

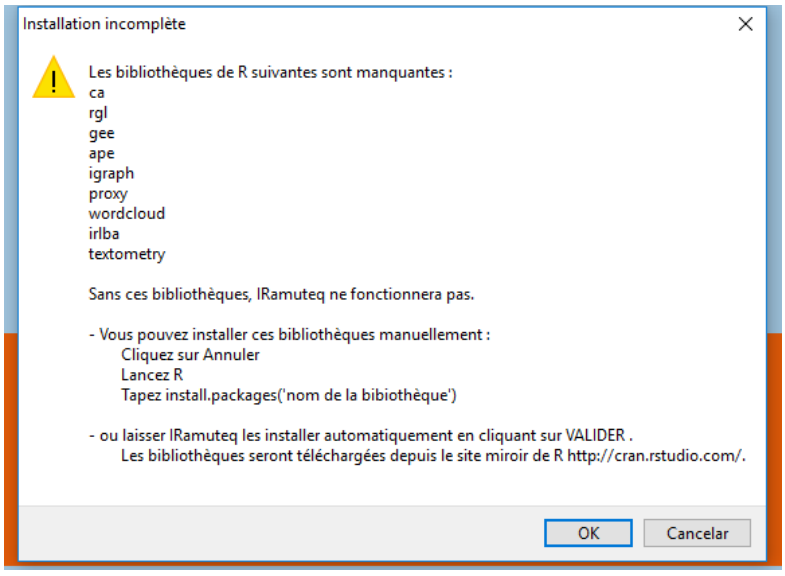


Figura 6- Atualização das bibliotecas na instalação do IRaMuTeQ

Se nenhuma das duas coisas acontecer, abra novamente o *software* IRaMuTeQ. Clique em "Edição" + "Preferências". Vá em "Verifique instalação de pacotes R", e clique em "Verificar" (figura 4). Caso ainda haja bibliotecas a instalar a tela da figura 3 aparecerá neste momento, espere o tempo necessário.

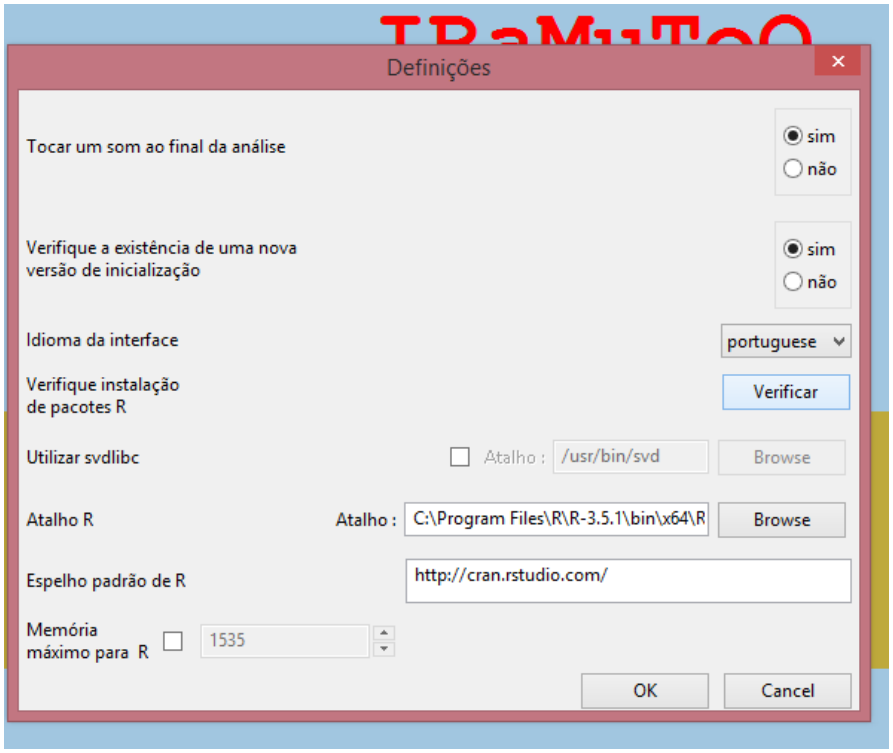


Figura 7- Verificação da instalação das bibliotecas no IRaMuTeQ

Importante: para finalizar a instalação do IRaMuTeQ copie o arquivo complementar “Rgraph.R” e cole na pasta “Rscripts” substituindo o arquivo já instalado. O caminho para encontrar a referida pasta geralmente é: “C: Arquivos de programas (x86) /iramuteq / Rscripts”.

Para atualizar o dicionário em português é só copiar o arquivo “lexique.pt” e colar na pasta do dicionário substituindo o arquivo já instalado no IRaMuTeQ. O caminho para esta pasta geralmente é: “C: Arquivos de programas (x86) /iramuteq /dictionnaires”.

Introdução

Trata-se de um *software* que viabiliza diferentes tipos de análise de dados textuais, desde aquelas bem simples, como a lexicografia básica, que abrange sobretudo a lematização[3] e o cálculo de frequência de palavras; até análises multivariadas como classificação hierárquica descendente de segmentos de texto, análise de correspondências e análises de similitude (Camargo & Justo, 2013). Por meio desse *software*, a distribuição do vocabulário pode ser organizada de forma facilmente compreensível e visualmente clara com representações gráficas pautadas nas análises utilizadas (Loubère & Ratinaud, 2014).

No IRaMuTeQ essas análises podem ser realizadas tanto a partir de um grupo de textos a respeito de uma determinada temática (*corpus* textual) reunidos em um único arquivo de texto; como a partir de matrizes com indivíduos em linha e palavras em coluna, organizadas em planilhas, como é o caso dos bancos de dados construídos a partir de testes de evocações livres.

Parte 1: Análise de *corpus* textual

A análise textual é um tipo específico de análise de dados (Lebart & Salem, 1988), na qual tratamos de material verbal transcrito, ou seja, de textos. Essa análise tem várias finalidades, sendo possível analisar textos, entrevistas, documentos, redações etc. Pode-se a partir da análise textual descrever um material produzido por um produtor, seja individual ou coletivamente, como também podemos utilizar a análise textual com a finalidade comparativa, relacional, comparando produções diferentes em função de variáveis específicas que descrevem quem produziu o texto. Para que se possa compreender a análise textual, é necessário inicialmente delimitar alguns conceitos importantes.

As noções de *corpus*, texto e segmento de texto

Corpus

O *corpus* é construído pelo pesquisador. É o conjunto de textos que se pretende analisar. Por exemplo, se um pesquisador decide analisar as matérias sobre beleza que

[3] Processo que reduz as palavras com base em suas raízes (formas reduzidas).

saíram numa revista no período de cinco anos; o conjunto destas matérias constituirá um *corpus*. O *corpus* é construído pelo pesquisador.

Texto

A definição destas unidades é feita pelo pesquisador e depende da natureza da pesquisa. No exemplo anterior cada matéria sobre beleza seria um texto. Se a análise for aplicada a um conjunto de entrevistas, cada uma delas será um texto. Caso a análise diga respeito às respostas de "n" participantes a uma questão aberta, cada resposta será um texto e teremos "n" textos. Quando se tratar de pesquisas documentais, atas de reuniões, cartas, etc.; cada exemplar destes documentos será um texto.

Um conjunto de unidades de textos constitui um *corpus* de análise. O *corpus* adequado à “Classificação Hierárquica Descendente” deve ser um conjunto textual centrado em um tema. O material monotemático evita que a análise de textos sobre vários itens previamente estruturados, ou diversos temas, resulte na reprodução da estruturação prévia dos mesmos.

No caso de entrevistas, onde há falas que produzem textos mais extensos, desde que o grupo seja homogêneo, é suficiente entre 20 e 30 textos (Ghiglione e Matalon, 1993). Se o delineamento é comparativo, sugere-se pelo menos 20 textos para cada grupo.

Em se tratando de respostas a questões abertas de um questionário, recomenda-se compor o *corpus* com respostas a uma mesma questão, para garantir que elas se refiram a um mesmo tema. Caso as questões digam respeito a temas ou aspectos diferentes, é necessário realizar uma análise para cada questão. Como mencionado anteriormente, a análise é sensível à estruturação do estímulo que produz o material textual, e isto é uma importante fonte de invalidação das conclusões. Quando as respostas apresentarem uma média em torno de três ou quatro linhas, é necessário um número bem maior de respostas para a constituição de um *corpus* de análise.

Os textos são separados por linhas de comando também chamadas de "linhas com asteriscos" ou metadados. No caso de entrevistas, p. ex., como cada uma delas é um texto, e eles necessariamente devem começar com uma linha de comando, esta linha informa o número de identificação do entrevistado (do produtor do texto que se segue) e algumas características (variáveis) que são importantes para o delineamento da pesquisa (como: sexo, faixa etária, afiliação a determinados grupos, nível social e cultural, etc.). Isto depende de cada pesquisa e o número de modalidades de cada uma destas variáveis depende do delineamento da pesquisa e do número de entrevistas realizadas.

Segmentos de texto

Os segmentos de texto (ST), na maior parte das vezes, tem o tamanho aproximado de três linhas, dimensionadas pelo *software* em função do tamanho do *corpus*. Os segmentos de textos são os ambientes das palavras. Podem ser construídos pelo pesquisador, ou automaticamente pelo *software*. São as principais unidades de análise textual deste tipo de *software*.

Embora seja o pesquisador que demarca os textos, nem sempre é ele que controla a divisão do *corpus* em segmentos de texto (ST). Numa análise padrão (*standart*), após reconhecer as indicações dos textos (pelas linhas com asteriscos) é o *software* que divide o material em ST. Mas o pesquisador pode configurar a divisão dos segmentos, p. ex.: no caso de uma grande quantidade de respostas curtas a uma pergunta aberta de um questionário, aconselha-se que cada texto seja definido como um único ST.

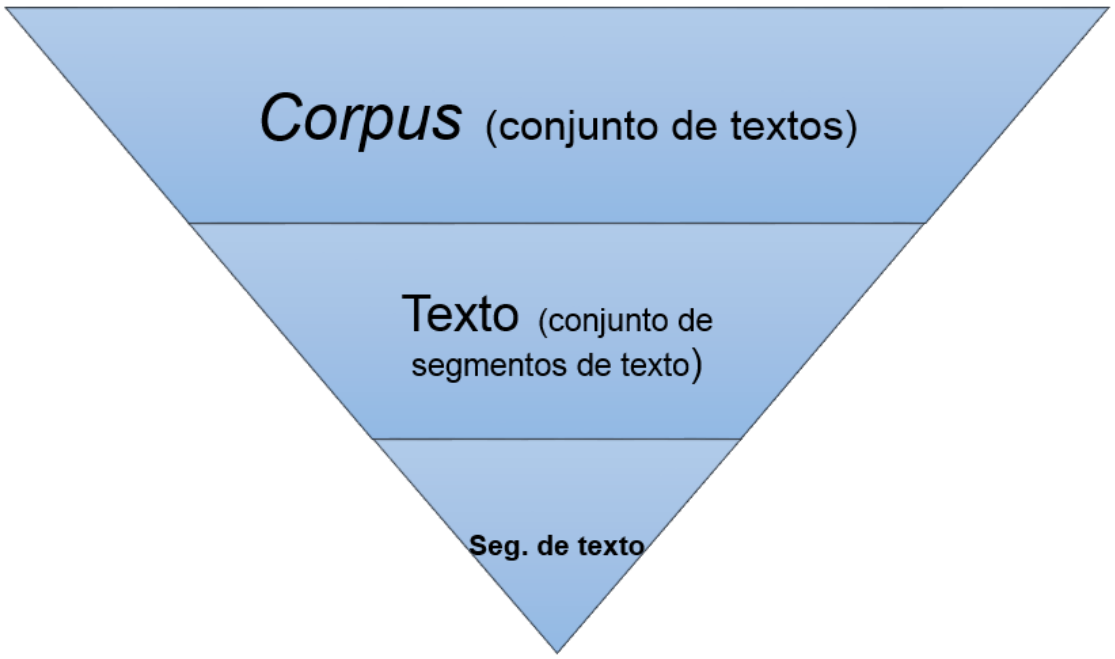


Figura 8- Noções de corpus, texto e segmento de texto.

Preparação de um *corpus* textual para análise

O primeiro passo para realizar a análise é construir o *corpus* a ser analisado, que deve ser feito de acordo com os seguintes procedimentos:

1- Colocar todos os textos (entrevistas, artigos, textos, documentos ou respostas a uma única questão) em **um único arquivo de texto** no *software* **Libre Office** (http://pt-br.libreoffice.org/) ou Open Office (http://www.openoffice.org/), deixando a primeira linha em branco. ***Jamais abra estes arquivos*** *ou qualquer outro gerado pelo IRaMuTeQ* ***com aplicativos da Microsoft*** *(Word, Excel, WordPad ou Bloco de notas), pois eles produzem bugs com o Unicode (UTF-8), o usado pelo software em questão.*

2- **Separar os textos com linhas de comando** (com asteriscos). Por exemplo, para cada entrevista ser reconhecida pelo *software* como um texto, elas devem começar por uma linha deste tipo. **Observação:** Deixe uma linha em branco antes da primeira linha de comando.

Exemplo de uma linha com asteriscos:

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

Digitar quatro asteriscos (sem espaço em branco antes deles), um espaço branco depois, um asterisco e o nome da variável (sem espaço branco entre eles), um traço em baixo da linha (*underline*) e o código da modalidade da variável (também sem espaço branco entre eles), um espaço em branco e depois o asterisco da segunda variável, e assim por diante.
Esta linha exemplo indica que o material textual que a segue (conteúdo verbalizado nos grupos focais) refere-se ao **grupo (gru)** nº 01 (utiliza-se dois dígitos, pois a amostra tem mais de 10 indivíduos e menos de 100), cujo **contexto (ctx)** de discussão foi o de beleza (onde 1= beleza; 2= saúde), composto por participantes jovens **(ida)** (onde 1=adultos jovens e 2= adultos maduros) do **sexo (sex)** feminino (onde 1= masculino; 2= feminino).
Imediatamente após esta linha com asterisco teclar ENTER, e sem tabulação, e linha em branco, digite ou coloque o texto correspondente a esta linha de comando.

3- **Existem duas maneiras de preparar as linhas de um *corpus*.** A primeira, a original ou monotemática, onde cada linha é seguida por um texto sem separações. Uma segunda maneira, a chamada temática, onde cada linha pode conter dois textos, correspondentes a duas ou mais temáticas, com a inclusão de linhas subordinadas a principal. A análise de *corpora* com divisões temáticas (temas diferentes) nos informa sobre as relações entre o conteúdo de um tema com o outro tema; e pode ser usada como uma análise preliminar de natureza mais exploratória (para se ter uma visão de todo) da coleta do material textual, mas deve-se fazer as análises monotemáticas, pois são elas que aprofundam a compreensão do significado do material estudado.

Extrato exemplo de um *corpus* da maneira original (monotemática)

**** *gru_01 *ctx_1 *ida_1 *sex_2
A palavra que me veio agora na cabeça foi cartão de visitas. Pelas imagens do vídeo é a ideia de que o corpo é a primeira impressão, é o cartão de visitas é o que vai apresentar, então tem que estar com os cabelos bem cuidados, bem_vestido, com o corpo em forma. Para mim veio também essa questão de identidade mesmo. Porque é assim que a gente se mostra. Não só a forma de se vestir, mas a forma como você se expressa, o jeito como anda. Tudo é uma questão de como as pessoas te percebem. E como acaba virando meio que um consumismo. As pessoas querem parecer bem e por isso elas compram bastante, investem. Corpo como um objeto. Essa questão da identidade é uma coisa também imposta socialmente. Tu tens que ser bonita, tu tens que ser magra, tu tens que estar sempre bem_vestida. Uma coisa que me veio foi esta tentativa de padronização, padrão de beleza. Uma coisa que marcou no vídeo é que 95 por cento das imagens eram de pessoas bonitas, magras, homens malhados. E aí tinha 3 gordos, só 3 gordos. Foi bem para esse lado mesmo. Isso também me marcou e também se eu estou realmente satisfeita com o meu corpo ou se eu tento o que as pessoas estão satisfeitas, porque todo mundo é assim. CONTINUA
**** *gru_02 *ctx_1 *ida_1 *sex_2
O corpo hoje é visto como uma ferramenta. Uma ferramenta de trabalho de até mesmo para a vida social ou para o trabalho, ele tem meio que se enquadrar num padrão social. E nesse padrão social o que se vê? A mídia passando, através. O que mais me chamou atenção foi a menina tentando imitar a Barbie, que influencia já desde pequena a ter que ser magra, ter que ter peito, ter que ter bunda, ter cabelos compridos, tem que ter cabelos lisos, tem que ter cabelos crespos, tem um padrão determinado. E onde muitos não pensam. E isso influi também na autoestima. Se eu não conseguir chegar nesse ponto, não estou bem, não estou legal, estou fora do certo. E isso influi muito também no comércio, capitalismo, é academia, nutricionista, moda. Porque eu não visto uma calça 36, veste um número maior. CONTINUA
**** *gru_03 *ctx_2 *ida_1 *sex_2
Eu pensei enquanto estrutura para tudo. Independentemente de ser desde estrutura mínima, quase microscópica, mas também na questão de estrutura para as nossas expressões, para as nossas interações, o que possibilita isso é a gente ter um corpo. Não tem como, não adiantaria de nada, eu acho, seria completamente diferente se a gente tivesse um corpo que não nos possibilitasse essa interação, esse movimento. A nossa própria constituição seria totalmente diferente. CONTINUA

Extrato exemplo de um *corpus* da maneira temática

**** *ind_01 *grup_1 *sex_1 *ren_1 *paph_3 *papf_1 *papp_2
-*tema_1
A hipertensão eu acho o seguinte, ela aparece, é silenciosa, se a pessoa não tiver o cuidado de saber que é hipertenso, através da consulta médica, é muito ruim, porque aquilo se agrava e a pessoa sofre muito. Eu, por exemplo, eu tive a experiência de praticamente não saber que era hipertenso, e eu tinha desconforto com a parte cardíaca, mas eu não tive infarto, não tive nada, eles fizeram vários exames e no exame cardiológico mais elaborado foi descoberto que eu tinha uma coronária obstruída e foi necessário fazer uma ponte safena, então eu fiz a cirurgia e eu tenho tido controle da pressão e está sendo muito bem feito. CONTINUA
-*tema_2
Há 1 ano eu fui atendido aqui no posto e o cardiologista mudou a medicação e eu achei que foi muito importante, porque me deu uma sensação de melhora, a pressão melhorou, e está bem controlada, eu recentemente fiz um check_up, fiz uns exames por causa da cirurgia que eu fiz e pela idade, eles analisaram, mas eu não tive a consulta ainda, mas eu estou em um nível adequado para a minha situação. CONTINUA
**** *ind_02 *grup_1 *sex_1 *ren_3 *paph_2 *papf_2 *papp_1
-*tema_1
É uma doença escondida, é lenta, quando a gente vê, já está com ela lá em cima. Tem se cuidar para não ter um avc ou coisas piores, e o resto, a vida leva a gente. Isso é um ponto que coloquei na minha cabeça, o restante eu esqueço, se eu vou morrer ou não, eu não sei, um dia eu sei que vou, não sei se com ela ou sem ela, mas vou levar ela comigo. Pressão alta todo mundo tem, quem que não tem pressão alta hoje, o que a gente planta depois a gente colhe, CONTINUA
-*tema_2
Eu cortei a bebida destilada, eu gostava de whisky, não bebi mais, a cerveja não cortei, a cerveja eu sou obrigado a tomar, a gente tem um escape para o corpo, mas estou numa situação boa agora. Eu não me cuido muito com a alimentação, mas o principal é o sal, não cheguei a 0, mas quase 0, agora a alimentação eu como de tudo, eu faço comida, se eu decido fazer uma feijoada, eu faço uma feijoada, eu vou e faço e como. CONTINUA

Observação: *Após preparar o corpus, recomenda-se que se leia o mesmo atentamente, especialmente no que se refere às linhas de comando. Esta verificação precisa ser realizada pelo pesquisador para que o texto possa ser processado.*

4- **Corrigir e revisar todo o arquivo**, para que os erros de digitação, ortográficos e gramaticais não sejam tratados como palavras diferentes.

5- **A pontuação deve ser observada**, no entanto sugere-se não deixar parágrafos em cada texto (devido à dificuldade entre nós no uso correto dos mesmos).

6- No caso de entrevistas ou questionários, as perguntas e o material verbal produzido pelo pesquisador (intervenções e anotações) devem ser suprimidos para não entrar na análise. Ao suprimir recupere os referentes.

7- Não justifique o texto, não use negrito, nem itálico ou outro recurso semelhante.

8- É desejável certa **uniformidade em relação às siglas**, sugere-se usá-las como siglas e em minúsculo. Ex: oms no lugar de organização mundial de saúde.

9- **As palavras compostas hifenizadas** quando digitadas com hífen são entendidas como duas palavras (o hífen vira espaço em branco), **una-as com um traço *underline***. Ex: "alto-mar" fica "alto_mar"; “terça-feira” fica “terça_feira”; e “bate-papo” fica “bate_papo”.

10- Todos os **verbos que utilizem pronomes** devem estar **na forma de próclise**, pois o dicionário não prevê as flexões verbo-pronominais. Ex: No lugar de “tornei-me”, a escrita deve ser “me tornei”.

11- **Evite uso de diminutivos** pelas características do dicionário.

12- **Números** devem ser mantidos em sua forma algarísmica. Ex: usar “2013” no lugar de “dois mil e treze”; “70” no lugar de “setenta”.

13- **Não usar em nenhuma parte do arquivo** dos textos os seguintes caracteres: aspas ("), apóstrofo ('), hífen (-), cifrão ($), percentagem (%), reticências (...), e nem asterisco (*). Este último é usado somente nas linhas que antecedem cada texto (linhas de comando).

14- **O arquivo com o *corpus*** preparado no *software* Libre Office ou no Open Office deve ser salvo em uma **nova pasta** criada no desktop, somente para a análise, como **“Texto: Escolha a codificação”** (arquivo do tipo “txt”). No Libre Office esta opção abre uma primeira janela e devemos escolher “Utilizar o formato texto – Escolha a codificação”, e uma segunda janela onde as opções “Conjuntos de caracteres” e “Quebra de parágrafo” devem ser respectivamente “Unicode (UTF- 8)” e “LF”.

15- Sugere-se que se arquive o *corpus* no formato “odt” e não no formato “txt”.

Tipos de análise de *corpus* textual IRaMuTeQ

O menu do IRaMuTeQ de análises textuais oferece cinco possibilidades: I) Estatísticas (análises lexicográficas), II) Especificidades e AFC, III) Classificação (método de Reinert), IV) Análise de similitude e V) Nuvens de palavras.

I) Estatísticas (análises lexicográficas)

Identifica e reformata as unidades de texto, transformando textos em ST, identifica a quantidade de palavras, frequência média e *hápax* (palavras com frequência igual a um), pesquisa o vocabulário e reduz as palavras com base em suas raízes (formas reduzidas) ou lematiza[4], cria do dicionário de formas reduzidas, identifica formas ativas e suplementares (Lebart & Salem, 1988).

II) Especificidades e Análise Fatorial de Correspondência (AFC)

Associa textos com modalidades de uma única variável de caracterização, ou seja, possibilita a comparação (contraste) da produção textual destas modalidades. Oferece uma análise fatorial de correspondência (Cibois, 1990; Lebart & Salem, 1988) para variáveis com no mínimo 3 modalidades.

III) Classificação (método de Reinert)

Os ST são classificados em função dos seus respectivos vocabulários, e o conjunto deles é repartido em função da presença ou ausência das formas reduzidas (Reinert, 1990).

No exemplo já utilizado dos três segmentos de texto (ST) lematizados: 1- o corpo ser como uma ferramenta, 2- ser uma ferramenta de trabalho ou mesmo para a vida social, 3- não precisar ter um corpo, mas ele ter que se enquadrar no padrão; a tabela seria a seguinte:

ST	corpo	enquadrar	ferramenta	padrão	precisar	social	trabalho	vida
1	1	0	1	0	0	0	0	0
2	0	0	1	0	0	1	1	1
3	1	1	0	1	1	1	0	0

[4] Lematizar significa transformar as várias flexões (de número, de gênero, etc.) ou lexemas de uma palavra no seu lema ou base comum. Exemplos: as palavras “corpo” e “corpão” tornam-se “corpo”; as palavras “preciso”, “precisamos”, “precisou” são reduzidas a “precisar”. Neste *software* os substantivos são reduzidos ao masculino singular, os verbos ao infinitivo e os adjetivos ao masculino singular.

A partir de matrizes cruzando ST e formas reduzidas (em repetidos testes do tipo x^2), aplica-se o método de CHD e obtém-se uma classificação definitiva. A CHD objetiva reagrupar as linhas dessa tabela em função da sua similaridade entre si, por meio de diversos testes qui-quadrado, particionando o *corpus* em classes. Uma ilustração aproximada disto é apresentada na figura 9.

Formas/ ST	a	b	c	d	e	f	g	h	i	Totais
1	1	1	1	1	0	0	0	0	0	4
2	0	0	0	0	1	1	1	1	1	5
3	0	0	1	0	1	0	1	0	0	3
4	1	0	1	0	1	0	0	0	1	4
5	0	0	1	0	1	0	1	0	0	3
6	1	1	1	1	0	0	0	0	1	5
7	0	0	0	0	1	1	1	1	0	4
8	1	0	1	0	1	0	0	0	0	3
9	0	0	1	0	1	0	1	0	1	4
10	0	0	0	0	1	1	1	1	0	3
Totais	4	2	8	2	8	2	6	2	4	38

Formas/ ST	a	b	c	d	e	f	g	h	i	
1	1	1	1	1	0	0	0	0	0	Classe 1
6	1	1	1	1	0	0	0	0	1	
8	1	0	1	0	1	0	0	0	0	Classe 3
4	1	0	1	0	1	0	0	0	1	
9	0	0	1	0	1	0	1	0	1	Classe 2
3	0	0	1	0	1	0	1	0	0	
5	0	0	1	0	1	0	1	0	0	
10	0	0	1	0	1	0	1	0	0	
2	0	0	0	0	1	1	1	1	1	Classe 4
7	0	0	0	0	1	1	1	1	0	

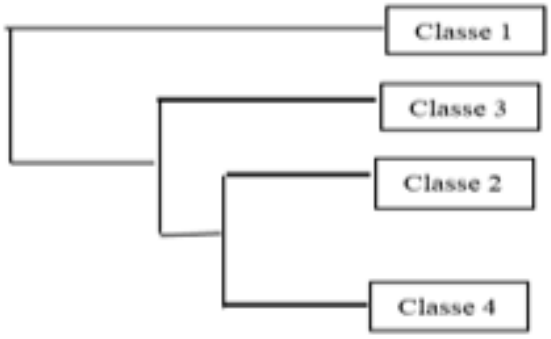


Figura 9- Classificação hierárquica descendente de segmentos de texto.

Esta análise visa obter classes de ST que, ao mesmo tempo, apresentam vocabulário semelhante entre si, e vocabulário diferente dos segmentos das outras

classes. A partir dessas análises o *software* organiza a análise dos dados em **um dendrograma** que ilustra as relações entre as classes.

O *software* executa cálculos e fornece resultados que nos permite a descrição de cada uma das classes, principalmente, pelo vocabulário presente nos segmentos de texto característicos e pelas suas “palavras” com asterisco (variáveis). Além disto, o *software* fornece uma outra forma de apresentação dos resultados, através de uma análise fatorial de correspondência feita a partir da CHD. Com base nas classes escolhidas, o *software* calcula e fornece os ST mais característicos de cada classe.

Estas classes de segmentos de texto, em nível do *software* são compostas por uma classificação segundo a presença ou ausência de determinado vocabulário. Em nível interpretativo, a significação das classes depende do marco teórico de cada pesquisa. Reinert (1990), ao estudar a literatura francesa considerou cada classe como uma noção de "mundo", enquanto um quadro perceptivo-cognitivo com certa estabilidade temporal associado a um ambiente complexo. Em pesquisas no campo da linguística estas classes foram interpretadas como campos lexicais (Cros, 1993) ou contextos semânticos.

Em pesquisas no campo da psicologia social, particularmente aquelas interessadas em estudar o conhecimento do senso comum, tendo em vista o estatuto que elas conferem às manifestações linguísticas, estas classes podem indicar representações sociais ou campos de imagens sobre um dado objeto, ou somente aspectos de uma mesma representação social (Veloz, Nascimento-Schulze e Camargo, 1999). Na maior parte das vezes não há coincidência entre o número de classes e o número de representações sociais envolvidas. O que vai definir se elas indicam representações sociais ou apenas uma representação social é o seu conteúdo, e sua relação com fatores ligados ao plano geral de cada pesquisa, geralmente expresso na seleção diferenciada dos participantes segundo sua afiliação grupal, suas práticas sociais anteriores, etc.

IV) Análise de similitude

Esse tipo de análise baseia-se na teoria dos grafos e é utilizada frequentemente por pesquisadores das representações sociais. Esta teoria estuda as relações de objetos de um dado conjunto. Sua fórmula é: G (V, E), onde G significa grafo e é composto de vértices (V) e de várias ligações entre dois vértices (E) (Degenne & Vergès, 1973; Flament, 1981).

Este tipo de análise permite identificar as coocorrências entre as palavras e seu resultado traz indicações da conexidade entre as palavras, auxiliando na identificação da estrutura do conteúdo de um *corpus* textual (Flament, 1981). Permite também

identificar as partes comuns e as especificidades em função das variáveis descritivas identificadas na análise (Marchand & Ratinaud, 2012).

V) Nuvem de palavras

Agrupa as palavras e as organiza graficamente em função da sua frequência. Elas são apresentadas com tamanhos diferentes: as palavras maiores são aquelas com maior frequência (ou outro indicador escolhido) no *corpus*, e as menores apresentam frequências inferiores. As primeiras são colocadas no centro do gráfico.

É uma análise lexical bem simples. Porém ela é graficamente interessante, pois fornece uma ideia inicial do conteúdo do material textual.

Processando a análise no *software* IRaMuTeQ

Para a exposição e exercício de análises de *corpora* textuais (parte 1) estão disponíveis três *corpora* no kit IRaMuTeQ (na pasta “Corpora”): “corpo”, “aids” e “hipertensão”. Estes *corpora* serão utilizados para esta parte do tutorial.

Três exemplos de *corpora* textuais

Corpus monotemático com textos longos (grupos focais): RS do corpo

Quando temos um *corpus* com apenas um tema e composto de textos longos, preparamos ele de forma monotemática (apenas com linhas de comando), como no exemplo do ***corpus* “corpo”**. Este *corpus* é formado pela transcrição da discussão de 16 grupos focais após visualizarem materiais audiovisuais, de um experimento de laboratório realizado numa dissertação (Justo, 2011). Os grupos foram formados por estudantes e funcionários de uma universidade. A instrução inicial foi: “Discutam, em grupo, este vídeo sobre corpo que acabaram de ver”.

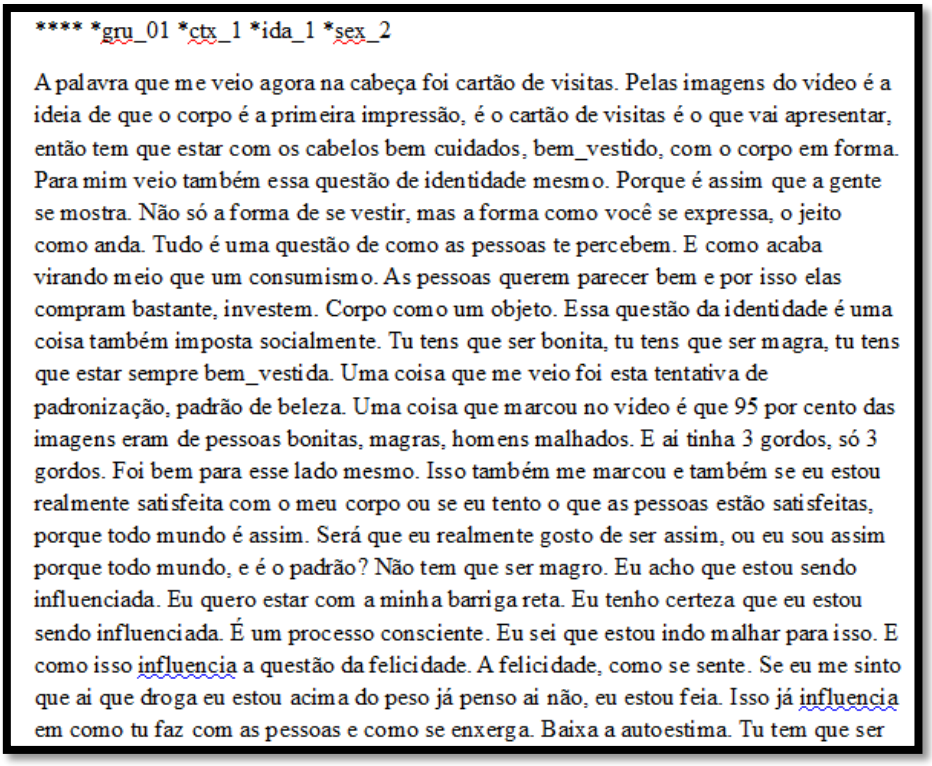
As variáveis que formam a linha de comando ou linha de metadados são:

*gru (Grupo): com 16 modalidades ou grupos focais.

*ctx (Contexto de discussão de cada grupo: 1= beleza e 2= saúde).

*ida (Idade dos participantes: 1= adultos jovens e 2= adultos maduros).

*sex (Sexo dos participantes: 1= masculino e 2= feminino).



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

A palavra que me veio agora na cabeça foi cartão de visitas. Pelas imagens do vídeo é a ideia de que o corpo é a primeira impressão, é o cartão de visitas é o que vai apresentar, então tem que estar com os cabelos bem cuidados, bem_vestido, com o corpo em forma. Para mim veio também essa questão de identidade mesmo. Porque é assim que a gente se mostra. Não só a forma de se vestir, mas a forma como você se expressa, o jeito como anda. Tudo é uma questão de como as pessoas te percebem. E como acaba virando meio que um consumismo. As pessoas querem parecer bem e por isso elas compram bastante, investem. Corpo como um objeto. Essa questão da identidade é uma coisa também imposta socialmente. Tu tens que ser bonita, tu tens que ser magra, tu tens que estar sempre bem_vestida. Uma coisa que me veio foi esta tentativa de padronização, padrão de beleza. Uma coisa que marcou no vídeo é que 95 por cento das imagens eram de pessoas bonitas, magras, homens malhados. E aí tinha 3 gordos, só 3 gordos. Foi bem para esse lado mesmo. Isso também me marcou e também se eu estou realmente satisfeita com o meu corpo ou se eu tento o que as pessoas estão satisfeitas, porque todo mundo é assim. Será que eu realmente gosto de ser assim, ou eu sou assim porque todo mundo, e é o padrão? Não tem que ser magro. Eu acho que estou sendo influenciada. Eu quero estar com a minha barriga reta. Eu tenho certeza que eu estou sendo influenciada. É um processo consciente. Eu sei que estou indo malhar para isso. E como isso influencia a questão da felicidade. A felicidade, como se sente. Se eu me sinto que aí que droga eu estou acima do peso já penso aí não, eu estou feia. Isso já influencia em como tu faz com as pessoas e como se enxerga. Baixa a autoestima. Tu tem que ser

Figura 10- Extrato inicial do corpus "corpo".

Corpus monotemático com ST curtos (resposta a uma única questão aberta): RS da aids

Quando temos grande volume de respostas curtas de uma questão aberta de um questionário (mínimo em torno de 80), a técnica da classificação hierárquica ascendente (CHD) também pode ser utilizada, como no exemplo do ***corpus* "aids".** Este *corpus* é formado pela digitação das respostas a uma das questões do questionário utilizado numa dissertação (Antunes, 2012). Participaram 312 estudantes do ensino médio. A questão foi: "O que você pensa a respeito da Aids? ", respondida por 300 estudantes.

As variáveis que formam a linha de comando são:

*ind (Indivíduo participante): com 312 modalidades ou 312 estudantes.

*sex (Sexo): 1= masculino e 2= feminino.

*esc (Tipo de escola): 1= pública e 2= particular.

*pes (Pessoa soropositiva): 1= conhece e 2= não conhece.

*conh (Conhecimento sobre a transmissão do HIV): 1= bom e 2= pouco.

*ati (Atitude frente ao soropositivo): 1= favorável, 2= neutra e 3= desfavorável.

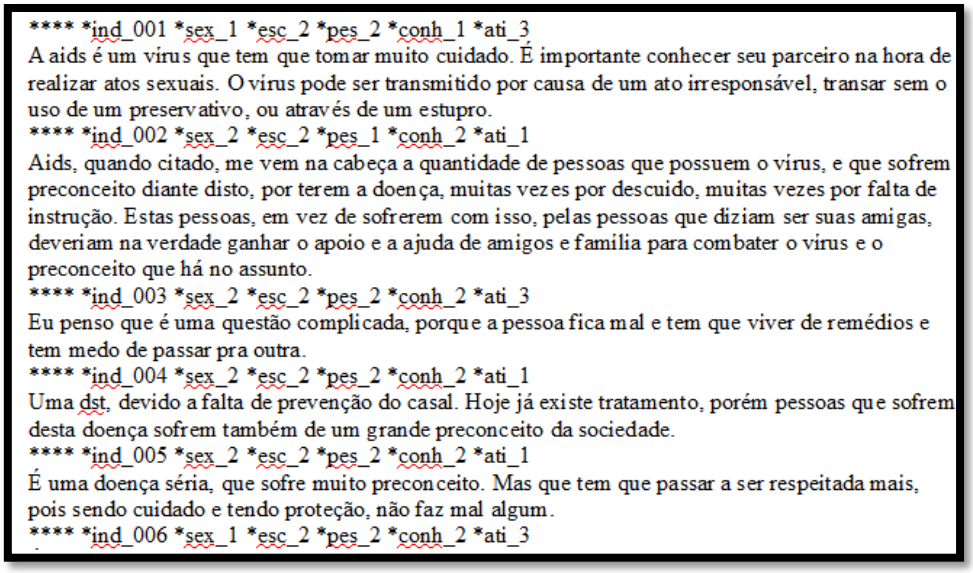

**** *ind_001 *sex_1 *esc_2 *pes_2 *conh_1 *ati_3
A aids é um virus que tem que tomar muito cuidado. É importante conhecer seu parceiro na hora de realizar atos sexuais. O virus pode ser transmitido por causa de um ato irresponsável, transar sem o uso de um preservativo, ou através de um estupro.
**** *ind_002 *sex_2 *esc_2 *pes_1 *conh_2 *ati_1
Aids, quando citado, me vem na cabeça a quantidade de pessoas que possuem o virus, e que sofrem preconceito diante disto, por terem a doença, muitas vezes por descuido, muitas vezes por falta de instrução. Estas pessoas, em vez de sofrerem com isso, pelas pessoas que diziam ser suas amigas, deveriam na verdade ganhar o apoio e a ajuda de amigos e familia para combater o virus e o preconceito que há no assunto.
**** *ind_003 *sex_2 *esc_2 *pes_2 *conh_2 *ati_3
Eu penso que é uma questão complicada, porque a pessoa fica mal e tem que viver de remédios e tem medo de passar pra outra.
**** *ind_004 *sex_2 *esc_2 *pes_2 *conh_2 *ati_1
Uma dst, devido a falta de prevenção do casal. Hoje já existe tratamento, porém pessoas que sofrem desta doença sofrem também de um grande preconceito da sociedade.
**** *ind_005 *sex_2 *esc_2 *pes_2 *conh_2 *ati_1
É uma doença séria, que sofre muito preconceito. Mas que tem que passar a ser respeitada mais, pois sendo cuidado e tendo proteção, não faz mal algum.
**** *ind_006 *sex_1 *esc_2 *pes_2 *conh_2 *ati_3

Figura 11- Extrato inicial do corpus "aids".

Corpus temático com ST longos de entrevistas com dois temas complementares: RS da hipertensão

Quando temos um *corpus* composto de mais de um tema, preparamos ele de forma temática (com sub-linhas de comando), como no exemplo do *corpus* "hipertensão". Este *corpus* é formado pela transcrição de entrevistas em profundidade feitas com 60 pessoas adultas, ligadas de alguma forma com o tema da hipertensão arterial, numa tese (Antunes, 2017). A instrução inicial foi: "Eu queria que você me falasse o que pensa sobre a hipertensão arterial, suas causas, suas características, consequências, experiências, tratamento e outras coisas importantes para você sobre este assunto". A maior parte do material transcrito foi dividida entre dois temas (ou duas sub-linhas de comando), onde o *tema_1 trata da hipertensão e o *tema_2 do seu tratamento.

As variáveis que formam a linha de comando principal são:

*ind (Indivíduo participante): com 60 modalidades ou 60 adultos.

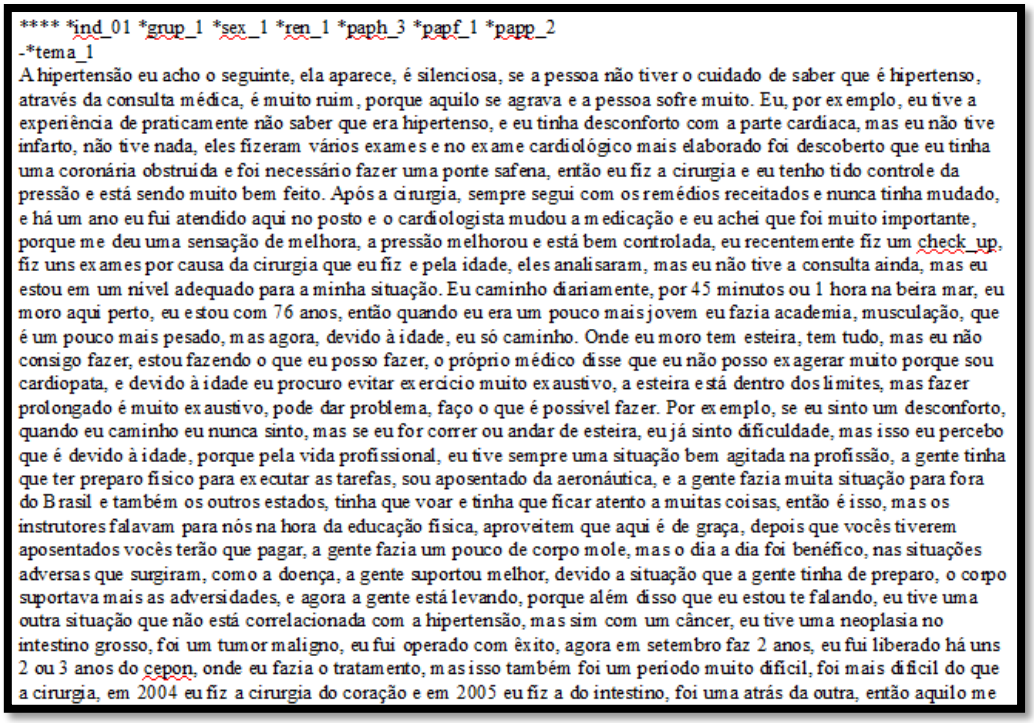
*grup (Grupo): 1= prof. saúde, 2- hipertensos e 3= familiares.

*sex (Sexo): 1= masculino e 2= feminino.

*ren (Renda): 1= 1 a 3 sm, 2= 4 a 6 sm, 3= 7 a 10 sm e 4= +10 sm.

*paph (Papel do hipertenso) e *papf (do familiar) no tratamento: 1= alimentação, 2= medicação e 3= exercício.

*papp (Papel do prof. saúde): 1= orientação, 2= rotina e 3= medicação.



**** *ind_01 *grup_1 *sex_1 *ren_1 *paph_3 *papf_1 *papp_2
-*tema_1
A hipertensão eu acho o seguinte, ela aparece, é silenciosa, se a pessoa não tiver o cuidado de saber que é hipertenso, através da consulta médica, é muito ruim, porque aquilo se agrava e a pessoa sofre muito. Eu, por exemplo, eu tive a experiência de praticamente não saber que era hipertenso, e eu tinha desconforto com a parte cardíaca, mas eu não tive infarto, não tive nada, eles fizeram vários exames e no exame cardiológico mais elaborado foi descoberto que eu tinha uma coronária obstruída e foi necessário fazer uma ponte safena, então eu fiz a cirurgia e eu tenho tido controle da pressão e está sendo muito bem feito. Após a cirurgia, sempre segui com os remédios receitados e nunca tinha mudado, e há um ano eu fui atendido aqui no posto e o cardiologista mudou a medicação e eu achei que foi muito importante, porque me deu uma sensação de melhora, a pressão melhorou e está bem controlada, eu recentemente fiz um check_up, fiz uns exames por causa da cirurgia que eu fiz e pela idade, eles analisaram, mas eu não tive a consulta ainda, mas eu estou em um nível adequado para a minha situação. Eu caminho diariamente, por 45 minutos ou 1 hora na beira mar, eu moro aqui perto, eu estou com 76 anos, então quando eu era um pouco mais jovem eu fazia academia, musculação, que é um pouco mais pesado, mas agora, devido à idade, eu só caminho. Onde eu moro tem esteira, tem tudo, mas eu não consigo fazer, estou fazendo o que eu posso fazer, o próprio médico disse que eu não posso exagerar muito porque sou cardiopata, e devido à idade eu procuro evitar exercício muito exaustivo, a esteira está dentro dos limites, mas fazer prolongado é muito exaustivo, pode dar problema, faço o que é possível fazer. Por exemplo, se eu sinto um desconforto, quando eu caminho eu nunca sinto, mas se eu for correr ou andar de esteira, eu já sinto dificuldade, mas isso eu percebo que é devido à idade, porque pela vida profissional, eu tive sempre uma situação bem agitada na profissão, a gente tinha que ter preparo físico para executar as tarefas, sou aposentado da aeronáutica, e a gente fazia muita situação para fora do Brasil e também os outros estados, tinha que voar e tinha que ficar atento a muitas coisas, então é isso, mas os instrutores falavam para nós na hora da educação física, aproveitem que aqui é de graça, depois que vocês tiverem aposentados vocês terão que pagar, a gente fazia um pouco de corpo mole, mas o dia a dia foi benéfico, nas situações adversas que surgiram, como a doença, a gente suportou melhor, devido a situação que a gente tinha de preparo, o corpo suportava mais as adversidades, e agora a gente está levando, porque além disso que eu estou te falando, eu tive uma outra situação que não está correlacionada com a hipertensão, mas sim com um câncer, eu tive uma neoplasia no intestino grosso, foi um tumor maligno, eu fui operado com êxito, agora em setembro faz 2 anos, eu fui liberado há uns 2 ou 3 anos do cepon, onde eu fazia o tratamento, mas isso também foi um período muito difícil, foi mais difícil do que a cirurgia, em 2004 eu fiz a cirurgia do coração e em 2005 eu fiz a do intestino, foi uma atrás da outra, então aquilo me

Figura 12- Extrato inicial do corpus "hipertensão".

Inicialmente, abra o *software* para trabalhar em sua interface, e importe o *corpus.* Na barra de ferramentas superior clique em *Arquivo* e *Abrir um corpus textual*, conforme indica a Figura 13.

Figura 13- Importação do corpus de análise.

Localize e selecione o *corpus* que deseja analisar e clique em *Abrir* (Figura 14). Sugere-se que o *corpus* tenha um nome curto e que a pasta que o contenha tenha o mesmo nome.

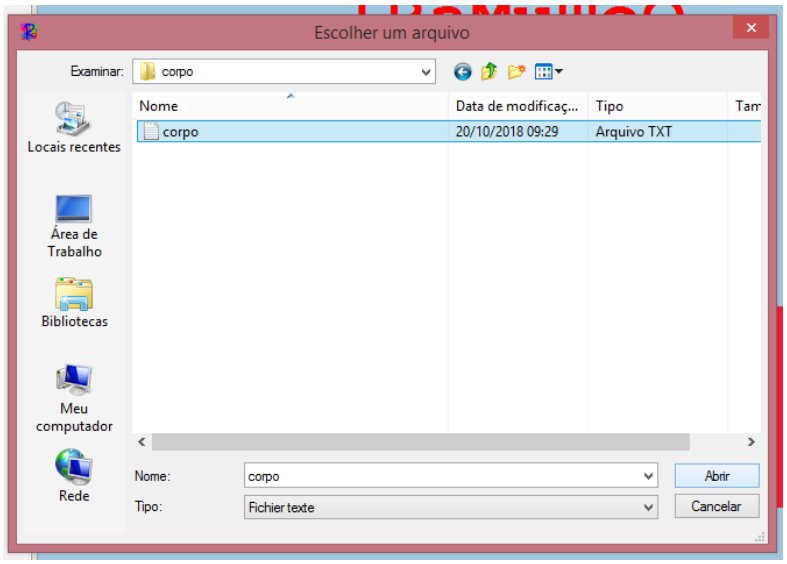


Figura 14- Importação do corpus de análise denominado "corpo".

No momento em que o *software* importar o *corpus*, uma nova janela será aberta. Nessa janela (Figura 15) podem ser observadas algumas configurações do *software* para analisar os dados textuais. A maior parte das configurações, na aba *Geral*, pode ser mantidas conforme o padrão, com exceção de duas que precisam ser modificadas. A primeira refere-se à codificação (*Definir caracteres*) do texto, que deve ser a segunda opção de cima para baixo: "*uft-8 – all languages*".

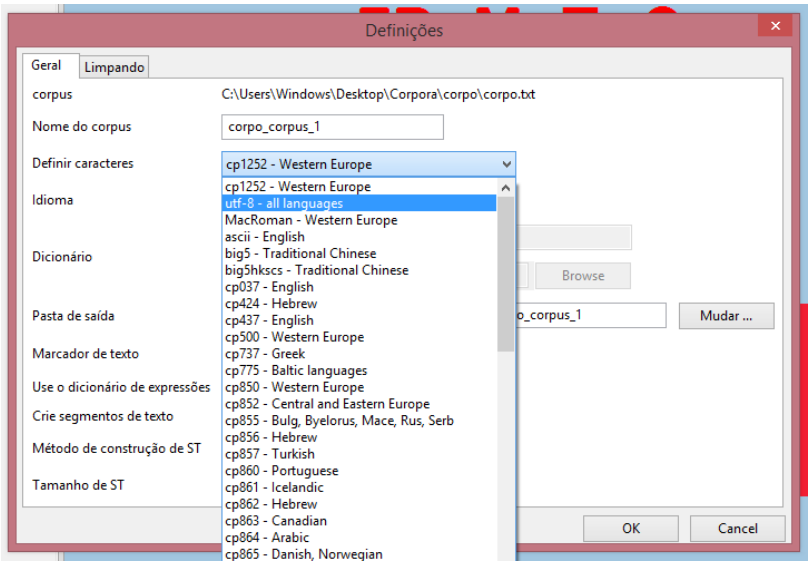


Figura 15- Configurações de análise – codificação do corpus.

A outra configuração é a da língua (*Idioma*). Conforme a Figura 16, selecione a língua: português no caso de o texto estar nesta língua, ou escolha a língua correspondente ao caso.

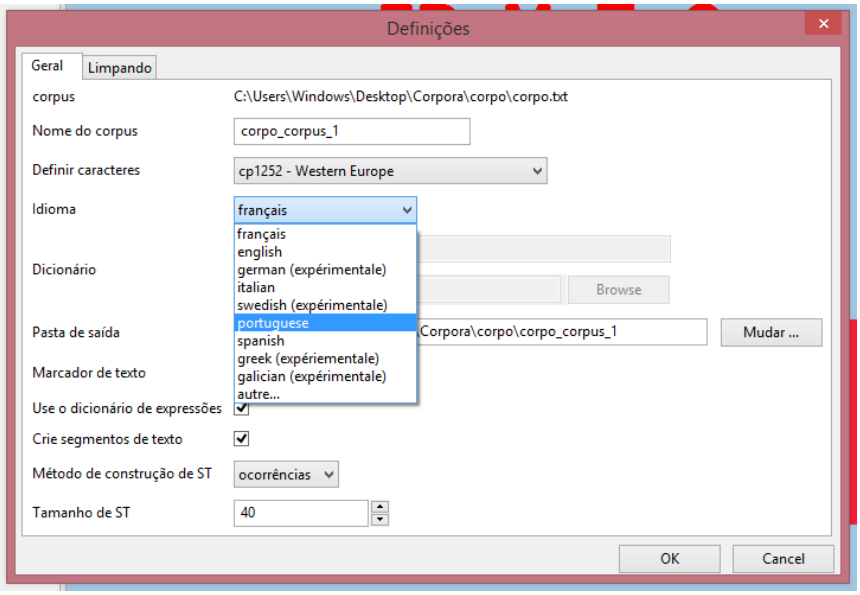


Figura 16- Configurações de análise – língua.

Na aba limpando pode-se fazer algumas correções no *corpus* antes de analisá-lo. Mas é melhor que isto seja feito antes desta etapa, conforme as instruções já indicadas para a formatação de *corpora* textuais.

Clique em OK e aguarde alguns segundos para que se processe importação dos dados. Em seguida, na grande janela da direita aparecerá uma breve descrição do *corpus*, como indicado na figura 17, onde se pode verificar, o número de textos (40) e de segmentos de texto (994), ocorrências (34724), formas (palavras diferentes) (3305), e *Hápax* (palavras com frequência igual a 1) (1618).

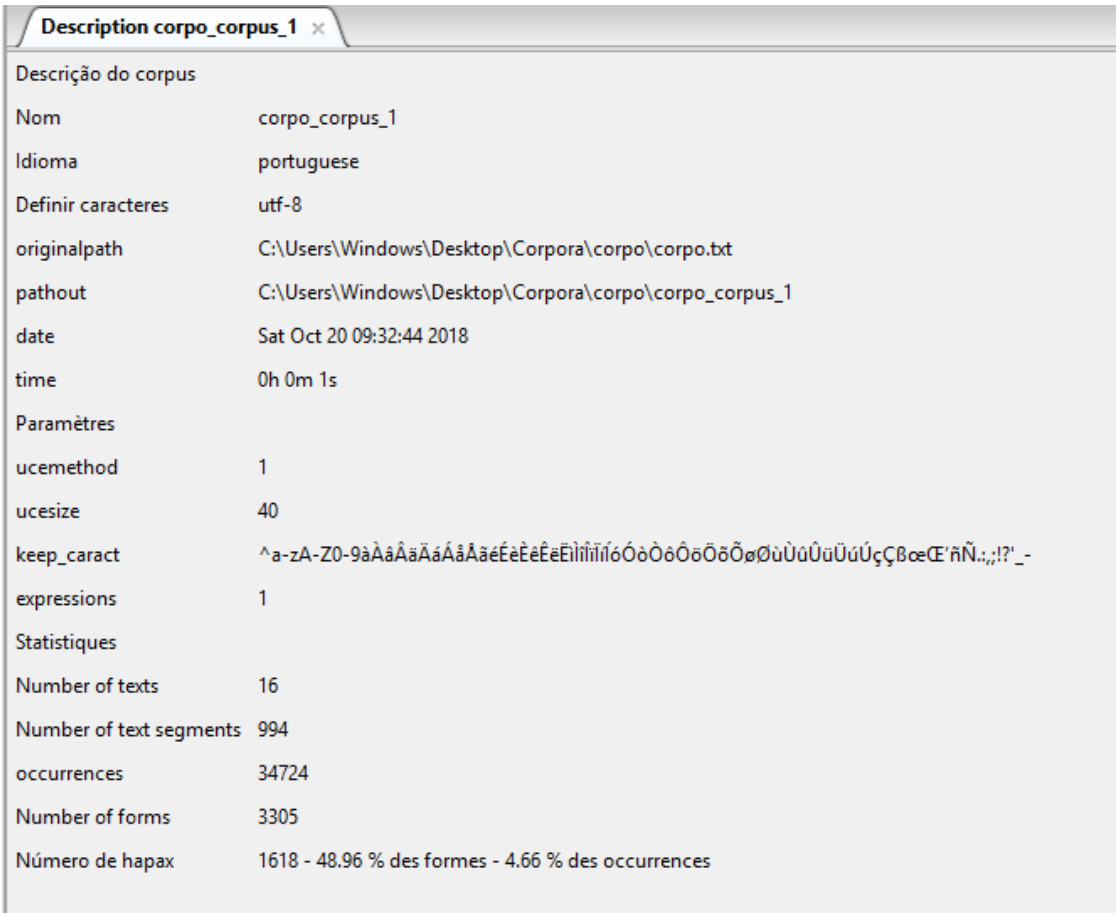


Figura 17- Resultados preliminares do corpus denominado "corpo".

Tendo sido realizada a importação do *corpus*, as análises já podem ser iniciadas. Para realizá-las, na barra de ferramentas superior, selecione *Análise de texto,* e aparecerão as possibilidades de análise (Figura 18).

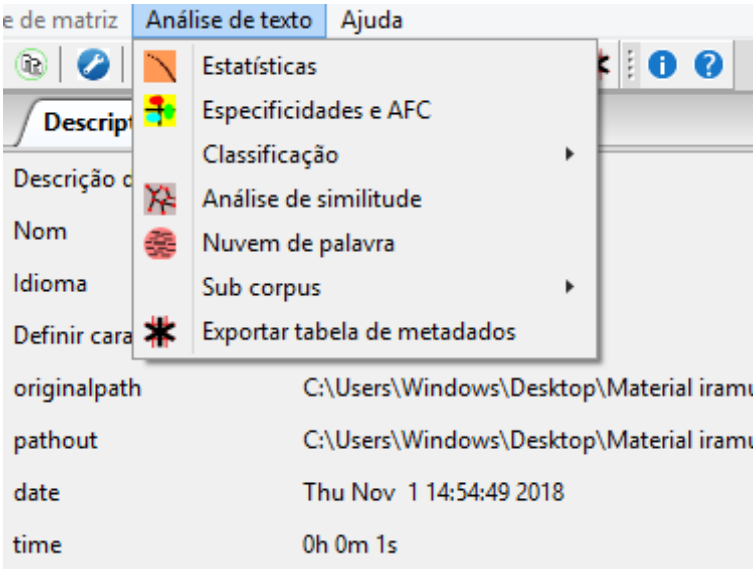


Figura 18- Escolha da análise textual.

Toda a vez que for escolhida uma análise, surgirá uma nova janela (Figura 19) perguntando se você deseja manter a *Lematização*. Deixe selecionado SIM, pois assim o *software* utilizará o dicionário de formas reduzidas para processar a análise. Nessa janela você também poderá editar as formas ativas e suplementares, se assim desejar, clicando em *Propriedades Chave.*

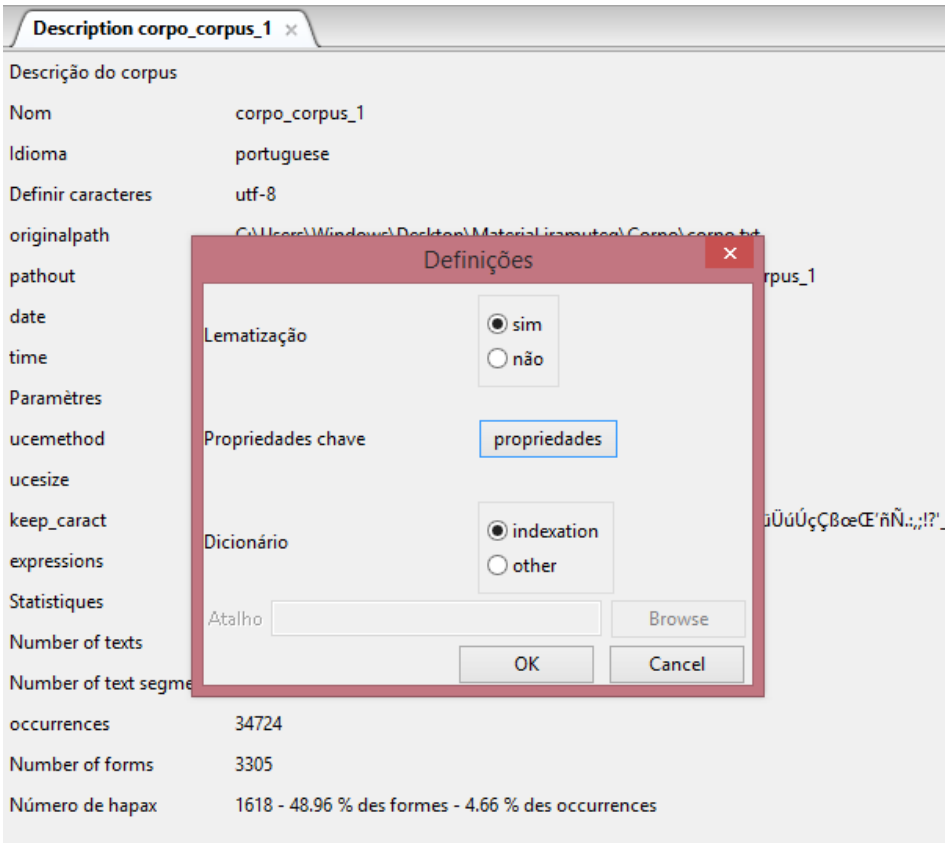


Figura 19- Lematização e edição das formas ativas (palavras consideradas nos cálculos).

Conforme a figura 20, o pesquisador pode selecionar quais as classes gramaticais ele deseja considerar ativas na análise **(0= palavras são eliminadas; 1= palavras são ativas; 2= palavras são suplementares)**. Uma fez feita essa alteração nas preferências da lematização, ela se manterá nas análises subsequentes para um mesmo *corpus*. O pesquisador pode alterá-las novamente no momento que desejar. Após escolher as classes gramaticais clique em *Ok*, e novamente em *Ok* que a análise será realizada.

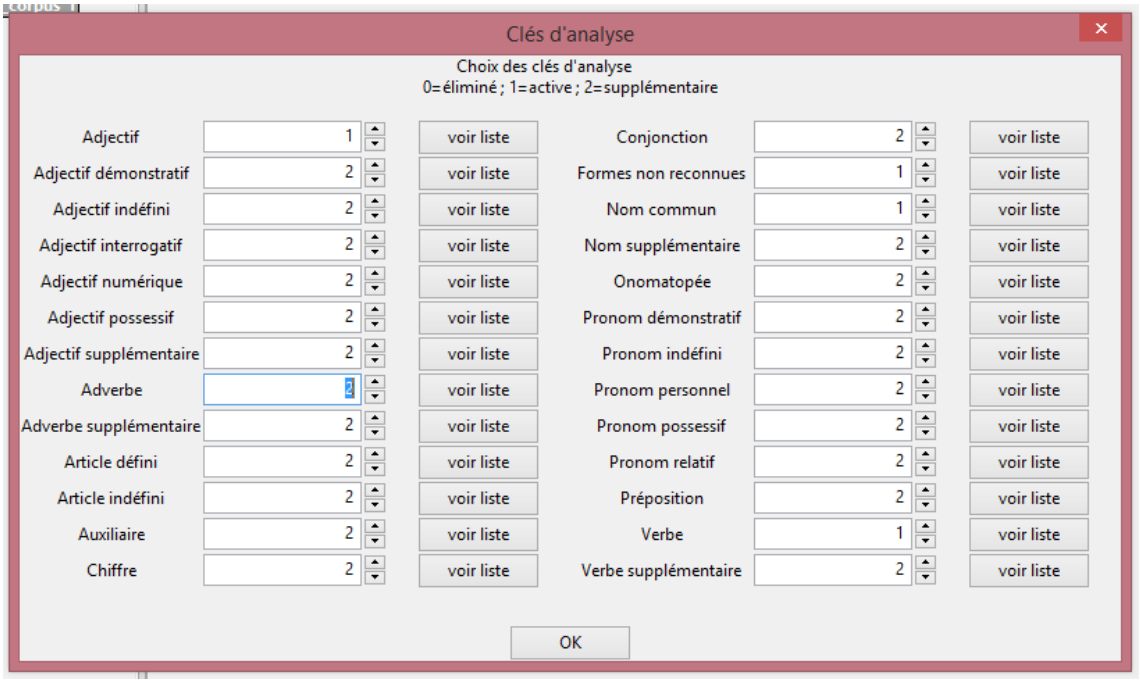


Figura 20- Escolha das formas ativas (palavras consideradas nos cálculos).

Sugere-se que se utilize os parâmetros padrão, conforme a ilustração da figura 20, com uma única alteração: passar o advérbio de 1 (ativa) para 2 (suplementar). Esta parametragem traz uma boa limpeza para pesquisa onde o conteúdo do texto é o mais importante. A lógica é trabalhar com os elementos de linguagem "plenos" como ativos: adjetivos, formas não reconhecidas, nomes (substantivos), verbos; e com as demais formas como suplementares.

Análise: Estatísticas

Na primeira opção de análise, "Estatísticas", o *software* fornece o número de textos e segmentos de textos, ocorrências, frequência média das palavras, bem como a frequência total de cada forma; e sua classificação gramatical, de acordo com o dicionário de formas reduzidas. Na interface dos resultados você poderá visualizar o diagrama de Zipf (Figura 21), que apresenta o comportamento das frequências das palavras no *corpus*, num gráfico que ilustra no eixo vertical (y) a posição das frequências das palavras por ordem decrescente, e no eixo horizontal (x) as frequências das formas, ambas em escalas logarítmicas (Lebart & Salem, 1988).

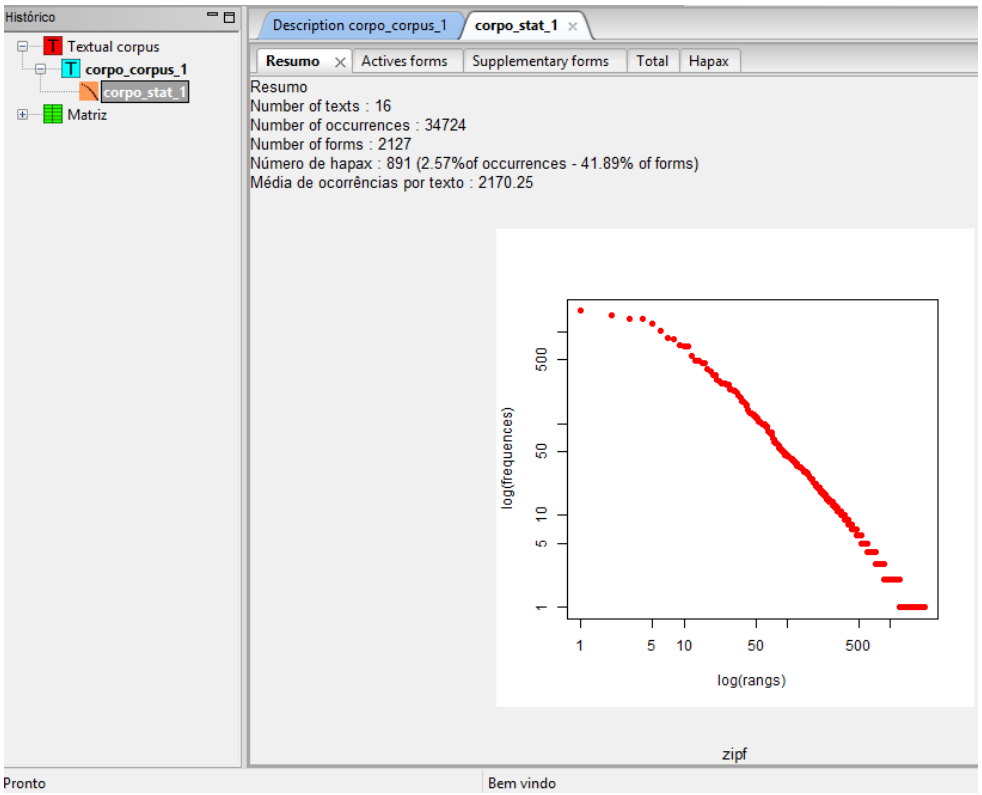


Figura 21- Diagrama de Zipf do corpus "corpo".

Ao clicar na aba "Formas ativas", o *software* exibe o dicionário do *corpus* (Figura 22), listando as frequências de cada forma (palavra) e suas respectivas categorias gramaticais. O *software* classifica as palavras com a seguinte codificação, a qual será utilizada ao longo de todas as análises:

adj = adjetivo
adj_num = adjetivo numeral
adj_sup = adjetivo colocado em forma suplementar
adv = advérbio
adv_sup = advérbio colocado em forma suplementar
art_def = artigo definido
conj = conjunção
nom = nome
nom_sup = nome colocado em forma suplementar
nr = não reconhecida
ono = onomatopéia
pro_ind = pronome indefinido
pre = preposição
ver = verbo
verbe_sup = verbo colocado em forma suplementar

As formas não reconhecidas (nr) são consideradas ativas na análise. Elas podem indicar erros de digitação do *corpus* ou palavras muito específicas que não se encontram no dicionário do IRaMuTeQ.

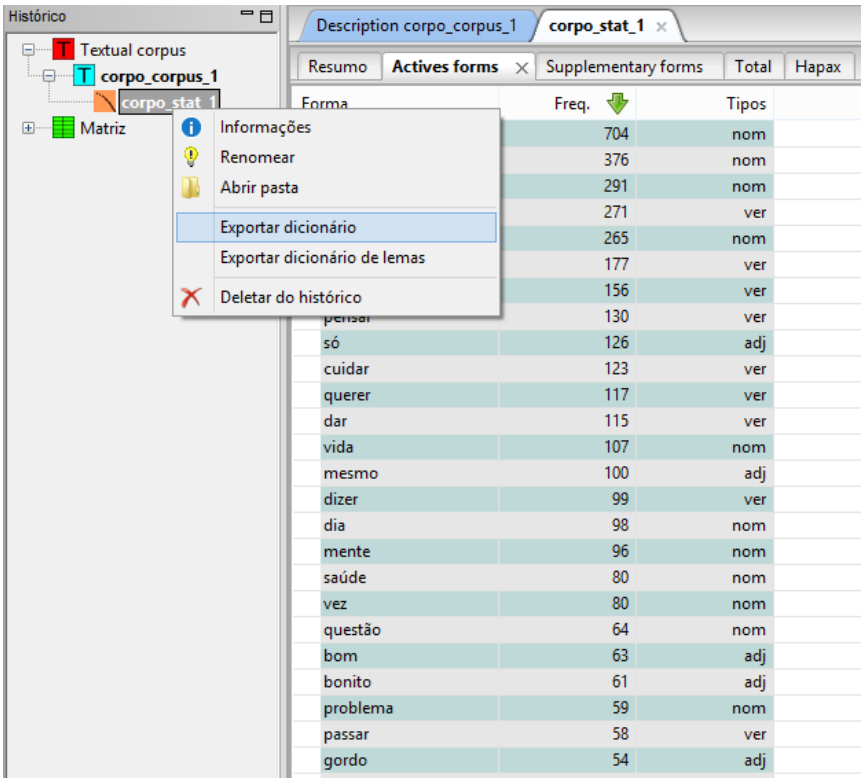


Figura 22- Dicionário de formas ativas do corpus "corpo".

Na coluna que se apresenta à esquerda, na interface do *software* exibida pela figura 22, você identifica essa análise como: "nomedocorpus_stat_n[5]". Colocando o cursor sobre esse nome, você pode clicar com o botão direito do mouse sobre o mesmo e selecionar algumas opções, dentre elas, exportar o dicionário (aquele que apresenta as formas sem a lematização), o qual será salvo como uma planilha ("dictionary.csv") na pasta em que foi salvo o *corpus* inicial, dentro de uma sub-pasta denominada: "nomedocorpus_stat_n". Você também pode exportar o dicionário de lemas (aquele que apresenta as formas com a lematização).

Análise: Especificidades e AFC

Para as análises das "Especificidades e AFC" você deverá escolher a variável categorial em função da qual deseja comparar suas modalidades. Vamos utilizar para isto o *corpus* "aids". Selecione a variável "*ati" (atitude em relação a pessoas soropositivas ao HIV) na janela que aparece na interface (Figura 23) e clique em *Ok*.

[5] O "n" geralmente é igual a 1, pois cada vez que iniciamos toda a análise do *corpus* de uma mesma pasta o *software* enumera casa análise.

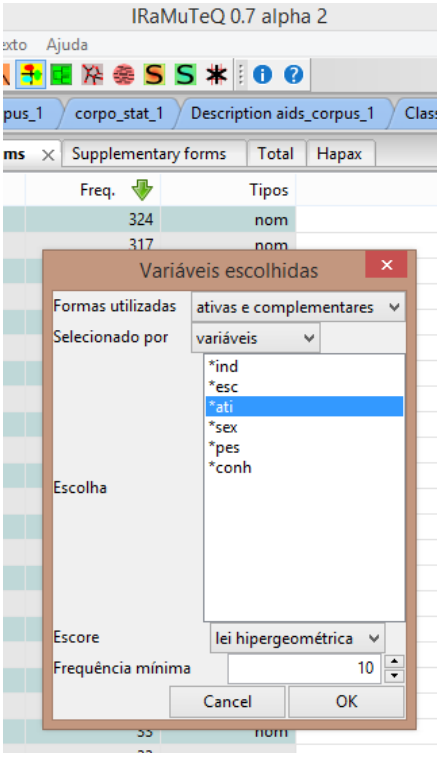


***Figura 23- Escolha de uma variável do corpus "aids"* para análise de especificidades.**

Aguarde alguns instantes e os resultados aparecerão na janela principal, conforme a figura 24.

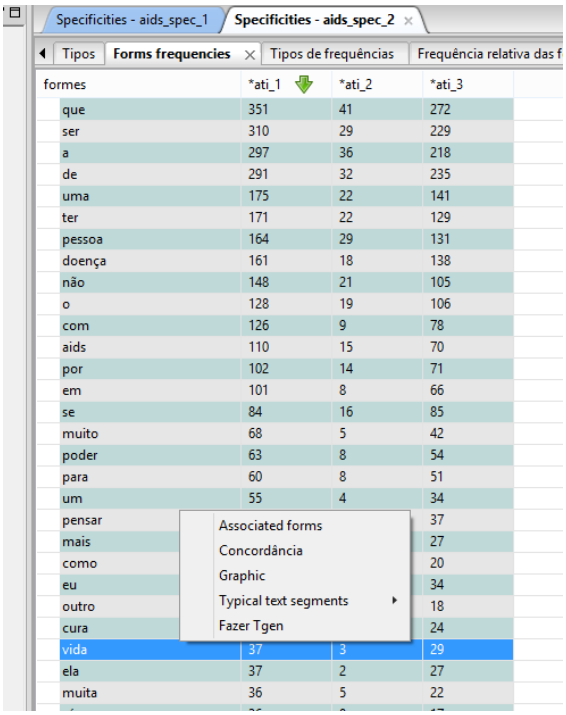


Figura 24- Resultados das especificidades da variável "atitude" do corpus "aids".

Ao clicar com o botão direito do mouse sobre qualquer uma das palavras apresentadas na tabela (Fig. 24) aparece um menu específico para aquela palavra selecionada. Ele contém cinco opções, e duas delas apresentam um maior interesse: “Formas associadas” e “Concordância”. A primeira apresenta uma lista de variações da forma reduzida escolhida (no caso de “vida”: “vida” e “vidas”). A segunda oferece os ambientes (segmentos de texto) onde encontra-se cada ocorrência da forma (da palavra “vida”), podendo seu contexto ser recuperado.

Quando se tem uma variável com mais de duas modalidades, o resultado da comparação entre estas modalidades pode ser representado num plano fatorial (Análise Fatorial de Correspondências). Para isto é necessário clicas na aba superior direta denominada “AFC”. Esta representação permite a distribuição num espaço bidimensional dos elementos textuais em função das modalidades desta variável (ver Figura 25).

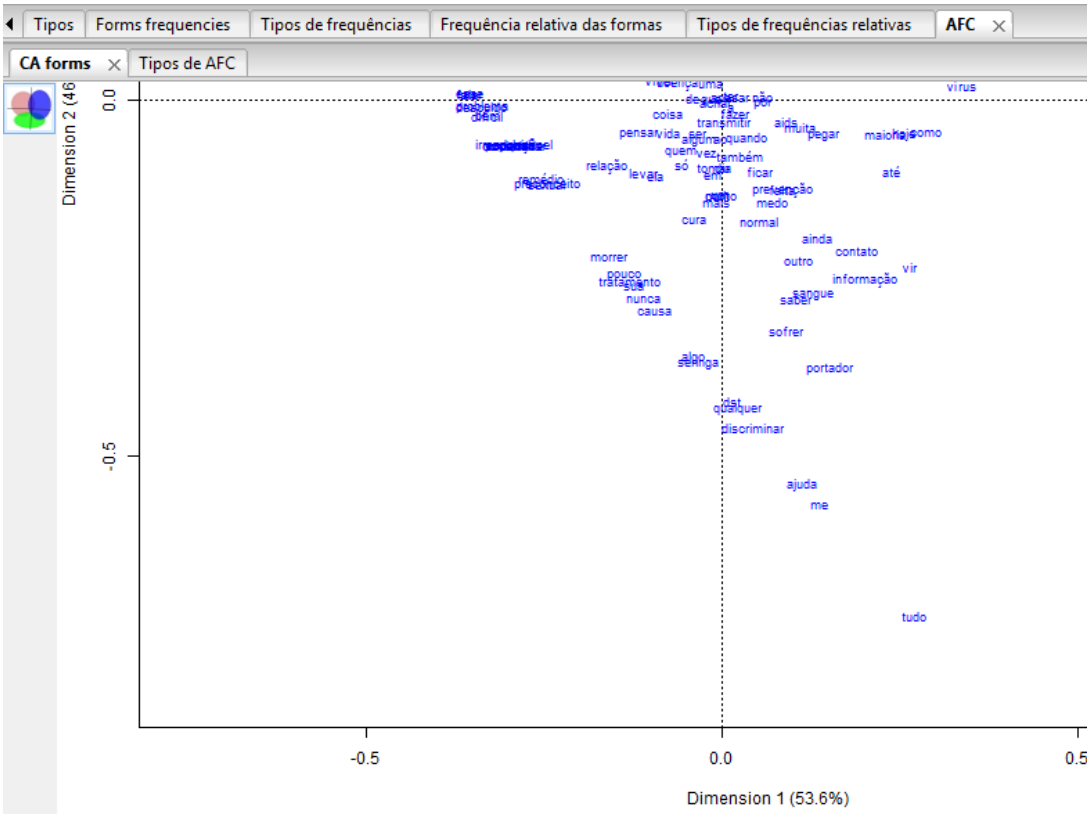


Figura 25- AFC com base em uma variável com três modalidades (“atitude”) do corpus “aids”.

Normalmente é necessário alterar alguns parâmetros do gráfico para uma melhor visualização. Conforme a figura 26, o primeiro é o seu tamanho (indicado em pixels[6]. A largura deve ser maior que a altura (p. ex. de 800x800 passa-se para 1200x800). Também é interessante escolher o tamanho do texto. Caso o gráfico apresente muitas palavras, deve-se limitar uma quantidade de pontos a serem apresentados (p. ex. “Pegue os 50 primeiros pontos”). E por fim, escolhe a opção: “Evitar sobreposição”.

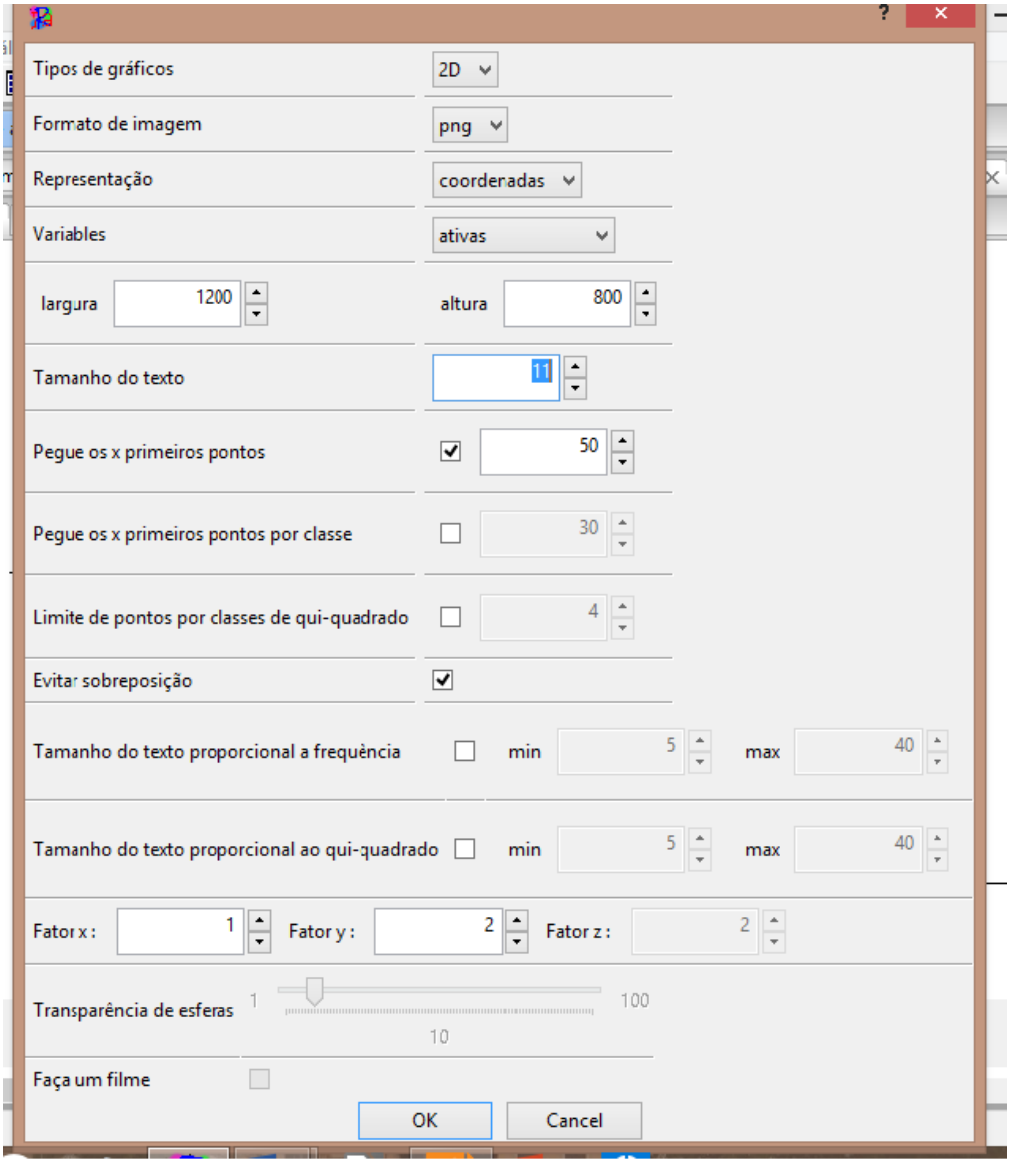


Figura 26- Interface para alterar os parâmetros do gráfico da AFC do corpus “aids”.

A figura 27 mostra o mesmo gráfico da figura 25 depois de alterados os paramentos conforme indica a figura 26. O plano fatorial 1 (eixo horizontal) x 2 (eixo

[6] Pixel é uma unidade de imagens digitais. Um ponto digital que somado a outros milhares forma uma imagem digital completa.

vertical) distribui as formas (palavras) partir de uma variável sobre a atitude de jovens a respeito de pessoas soropositivas ao HIV. Esta variável apresenta três modalidades: favorável (cor vermelha), neutra (verde) e desfavorável (cor azul).

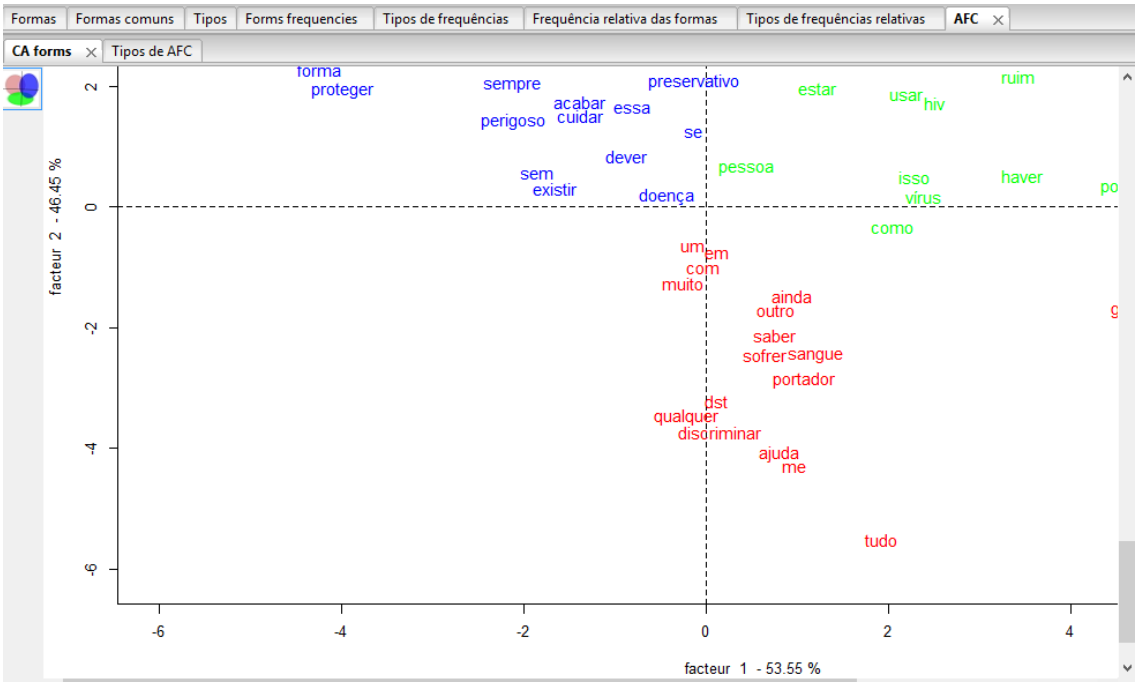


Figura 27- AFC com base na variável "atitude frente ao soropositivo" do corpus "aids".

Análise: Classificação (Método de Reinert)

Ao escolher "Classificação (Método de Reinert) ", conforme a figura 28, você pode optar por três possibilidades de classificação hierárquica descendente (CHD):

1- Dupla sobre RST: não utilizada, pois usualmente tem baixo aproveitamento do *corpus.* É uma análise dupla sobre reagrupamento de textos.
2- Simples sobre ST: que equivale a uma análise sobre os segmentos de texto delimitados pelo *software* (análise *standart*) ou pré-determinada pelo *software*, recomendada quando se dispõe de textos longos, como dois dos nossos *corpora* que utilizaremos aqui: o "corpo" e o denominado "hipertensão". Aqui a segmentação dos textos é feira com base no método de Reinert.
3- Simples sobre textos – que realiza a análise considerando os textos como segmentos de texto (ST), sem dividi-los em segmentos. Recomendada para respostas curtas a questões abertas de questionários e de entrevistas com roteiros, como o *corpus* denominado "aids".

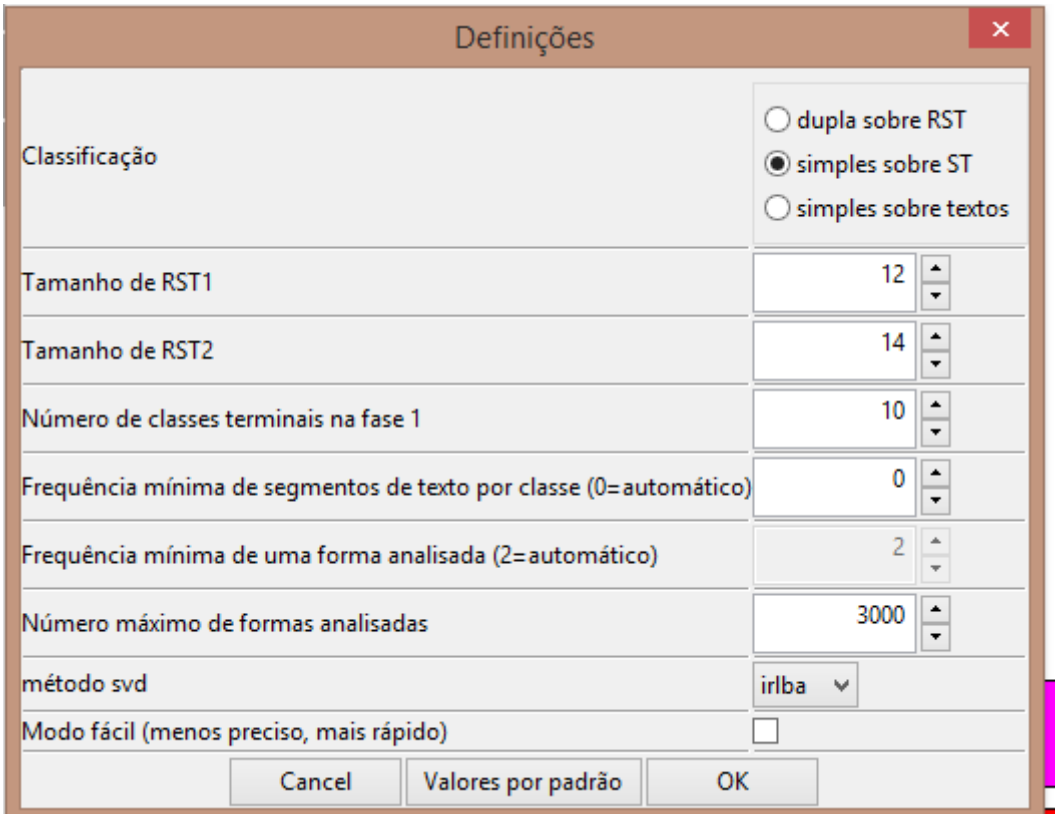


Figura 28- Interface para a escolha do tipo de classificação hierárquica descendente (CHC)

Escolha uma das modalidades de classificação. Geralmente a escolha da classificação simples sobre ST não exige nenhuma modificação nas configurações.

No caso de uma CHD anterior que não reteve no mínimo 75% dos ST, aumente os diminua o valor do “Número de classes terminais da fase 1” e refaça a CHD para obter uma retenção satisfatória. Quanto ao “método svd”, são escolhas do algoritmo para a segmentação do texto; por padrão emprega-se o “irlba”, não o “svdR”. A escolha do “Modo fácil” oferece uma análise mais rápida, mas menos precisa, pois suprime a segunda fase de cada partição.

Nesse último caso, o da classificação simples sobre textos, é necessária uma configuração anterior, conforme indica a figura 28. Logo ao importar o *corpus*, além de indicar a codificação e a língua, selecione “parágrafos” como método de construção dos ST.

Quando temos uma grande quantidade de respostas curtas a uma questão aberta de questionário, temos que configurar a CHD de forma específica (veja a figura 29). Ao importar um *corpus* deste tipo, além de identificar a codificação e a língua, selecione "parágrafos" como método de construção de segmentos de texto (ST). E quando for realizar a CHD escolha a classificação "simples sobre textos", para que o

software não segmente o texto de cada resposta. Assim o segmento de texto será considerado o próprio texto ou resposta curta a pergunta de um questionário.

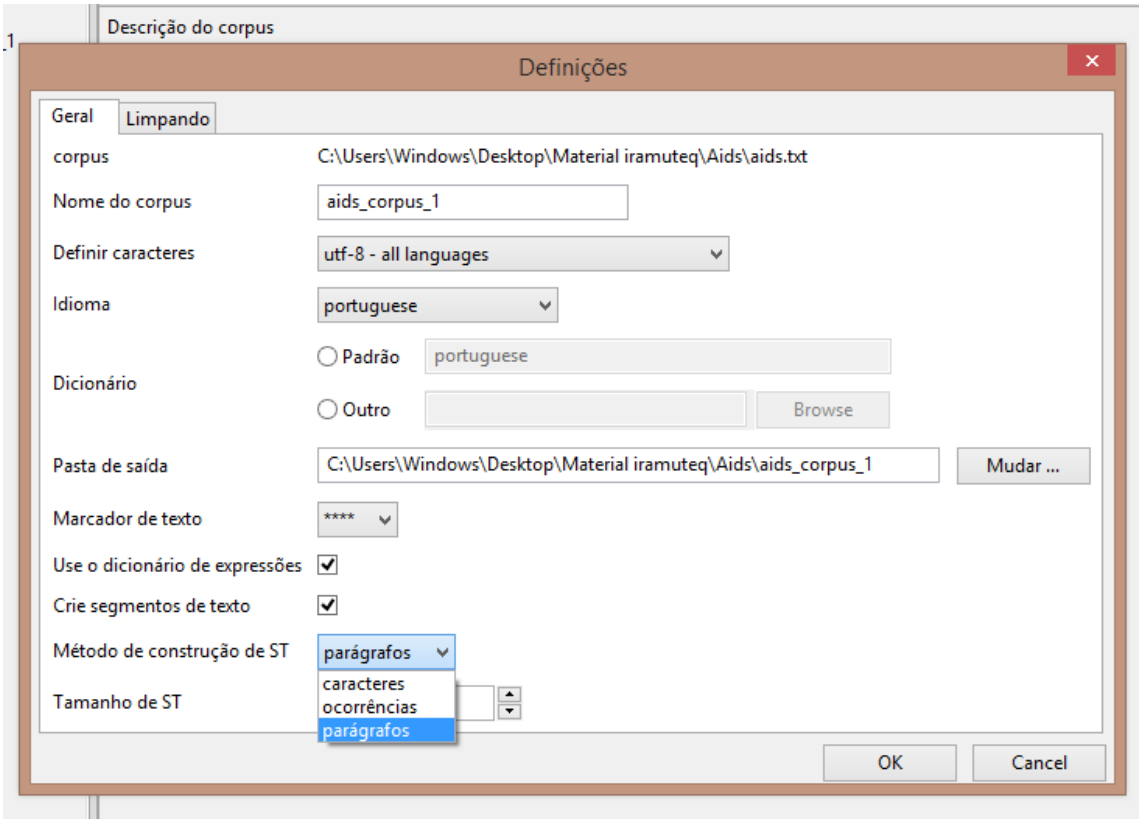
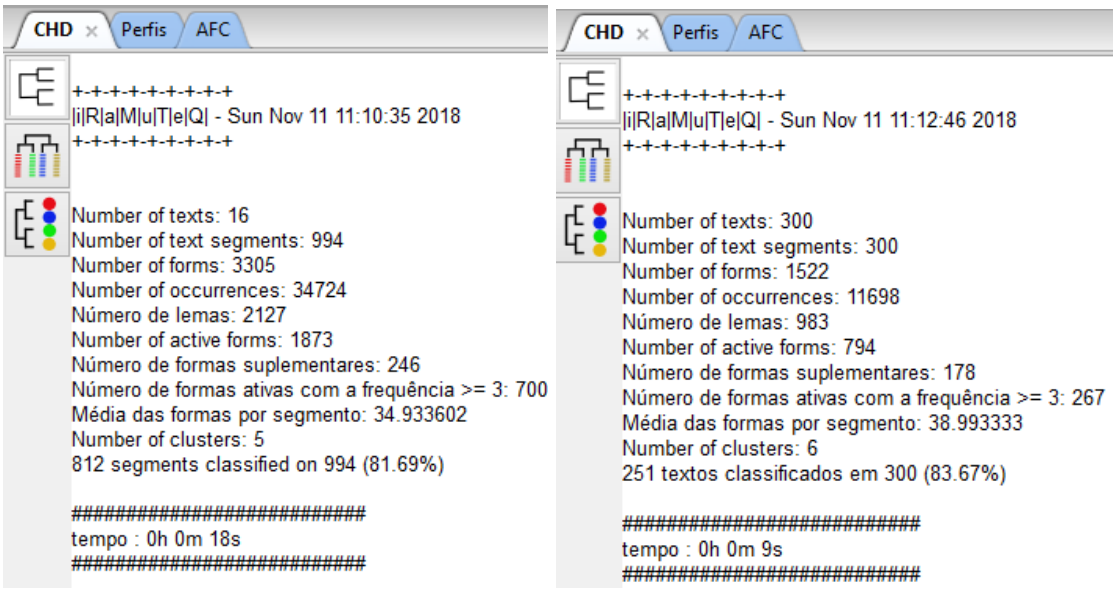


Figura 29- Configuração prévia do método de construção dos ST para a classificação simples sobre textos.

Na interface de resultados aparecerão alguns dados importantes da CHD (Fig. 30 e 31), na parte superior esquerda do dendrograma.



Figuras 30 e 31- Principais pontos da CHD a serem considerados em relação aos corpora: "corpo" e "aids" (o primeiro a esquerda e o segundo a direita).

Nessa parte da descrição dos resultados, as principais características da análise a serem consideradas são as seguintes:

1- Número de textos (*Number of texts*) = respectivamente 16 (o *software* reconhece a separação do *corpus* "corpo" em 16 unidades de texto ou 16 transcrições de sessões de grupos focais); e 300 (referindo-se as 300 respostas a uma única questão obtidas com a aplicação de um questionário).
2- Número de segmentos de textos (*Number of text segments*) = 994 (o *software* desmembrou o texto em 994 segmentos); e 300 (o pesquisador configurou o *software* para não desmembras uma resposta, então o número de textos (respostas curtas) ficou igual ao número de ST.
3- Número de formas distintas (*Number of forms*) = para o primeiro *corpus* foi 3.305 e para o segundo 1.522 formas.
4- Número de ocorrências (*Number of occurrences*) = respectivamente 34.724 e 11.698.
5- Número de lemas = 2.127 e 983.
6- Número de formas ativas (*Number of actives forms*): 1.876 e 794.
7- Número de classes (*Number of clusters*) = 5 para o *corpus* "corpo" e 6 para o *corpus* "aids".
8- Retenção de segmentos de texto: **812 segmentos classificados de um total de 994 (81,69%)** para o *corpus* "corpo" e **251 classificados de 300 (83,67%)** para o *corpus* "aids".

É importante relembrar que as análises do tipo CHD, para serem úteis à classificação de qualquer material textual, requerem uma retenção mínima de 75% dos segmentos de texto. Caso a CHD ofereça uma classificação com retenção inferior a esta, refaça a CHD alterando o número inicial de classes (Figura 28) para um valor diferente do padrão (10), e em caso de persistência do problema de retenção deve-se abandonar a CHD e empregar outros recursos, como p. ex., a análise de especificidades, etc.

Dendrogramas

O dendrograma que segue a descrição dos principais resultados (Figura 32) é apresentado de forma horizontal e lê-se da esquerda para a direita. Ele apresenta as

partições ou iterações que foram executadas na classificação dos segmentos de texto do *corpus*. Estas partições geram *sub-corpora* que correspondem as classes. No caso do *corpus* “corpo” elas são 5.

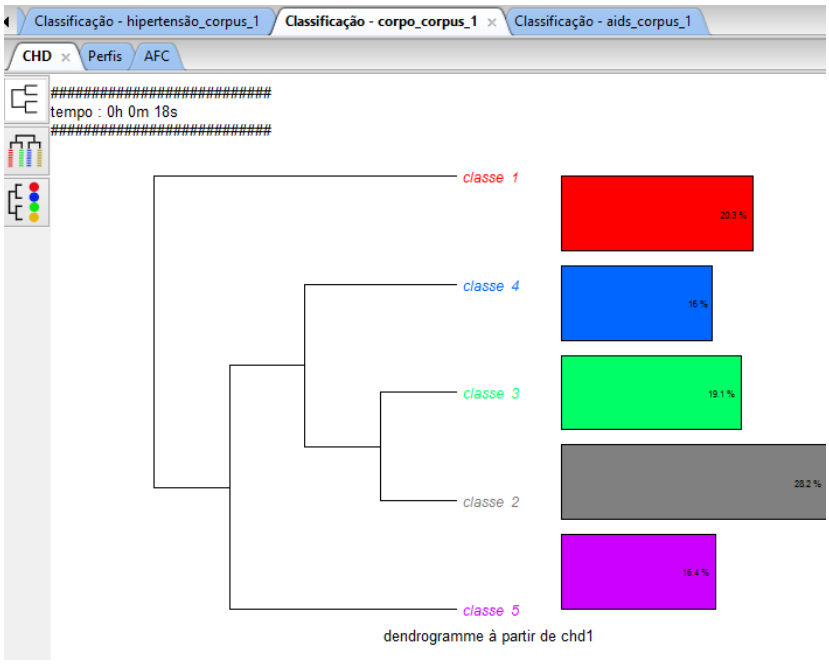


Figura 32- Dendrograma da classificação (CHD) do corpus “corpo” (forma horizontal).

No exemplo da figura 32, num primeiro momento, o *corpus* “Corpo" foi dividido (1ª partição ou iteração) em dois *sub-corpora*, separando a classe 1 do restante do material. Num segundo momento o *sub-corpora* maior foi dividido, originando a classe 5 (2ª partição ou iteração). Num terceiro momento há uma partição (a 3ª) gerando a classe 3, e uma última partição (a 4ª) separa as classes 2 e 3. A classificação (CHD) parou aqui, pois as 5 classes mostraram-se estáveis, ou seja, compostas de unidades de segmentos de texto com vocabulário semelhante. O número de partições é igual ao número de classes menos um.

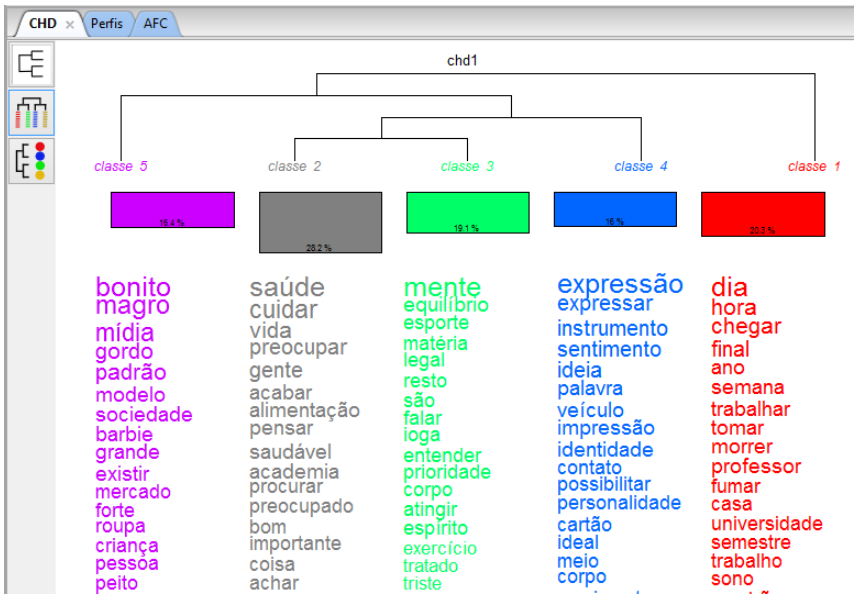


Figura 33- Dendrograma da classificação (CHD) do corpus "corpo" (forma vertical).

A figura 33 ilustra uma segunda forma de apresentação do mesmo dendrograma apresentado na figura 32. Neste caso a leitura é de cima para baixo. São indicadas as formas ativas (palavras) contidas nos segmentos de textos associados a cada classe. Esta imagem e a anterior são automaticamente salvas na sub-pasta "corpo_alceste_n" que se encontra na sub-pasta "corpo_corpus_n" da pasta em que foi salvo o *corpus* inicial (denominada corpo). Elas são imagens do tipo ".png"[7].

Além do dendrograma, essa interface de resultados também possibilita que se identifique o conteúdo lexical presente nos segmentos de texto (ST) de cada uma das classes (para acessá-lo, basta clicar na segunda aba (Perfis) e uma representação fatorial da CHD (para acessá-la, basta clicar na aba AFC).

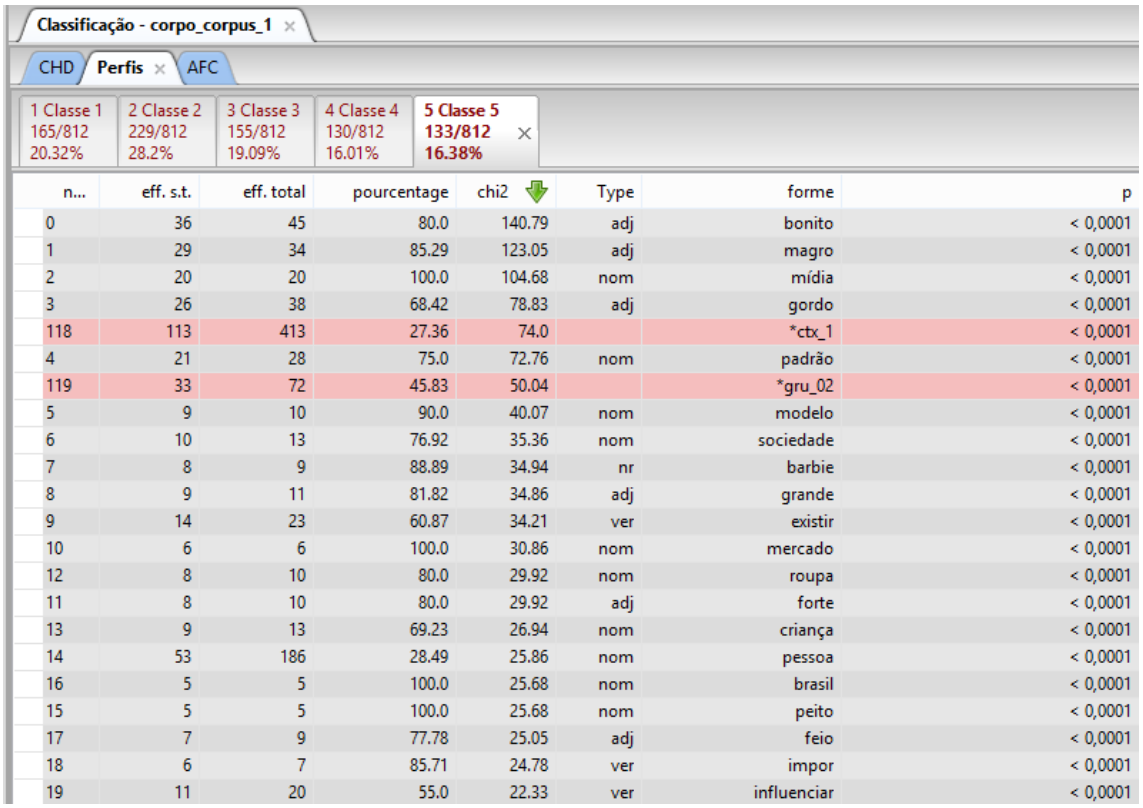
Perfis das classes da classificação (CHD)

Segundo a figura 34, na aba "Perfis", para cada classe encontram-se em colunas dados referentes ao seu conteúdo:

1- *n.* (número que ordena as palavras na tabela);
2- *eff. st* (número de segmentos de texto que contêm a palavra na classe);
3- *eff. total* (número de segmentos de texto no *corpus* que contém, ao menos uma vez, a palavra);
4- *pourcentage* (percentagem de ocorrência dos segmentos de texto que contem a palavra nessa classe, em relação a sua ocorrência no *corpus*);

[7] A extensão ".png" significa "*Portable Network Graphics*", um dos formatos de dados utilizado para imagens.

5- *chi2* (X^2 de associação dos segmentos de texto que contem a palavra com a classe);
6- *Type* (classe gramatical da palavra presente no segmento de texto identificada no dicionário de formas);
7- *forme* (identifica a palavra);
8- *P* (identifica o nível de significância da associação do segmento de texto contendo a palavra com a classe).



n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	36	45	80.0	140.79	adj	bonito	< 0,0001
1	29	34	85.29	123.05	adj	magro	< 0,0001
2	20	20	100.0	104.68	nom	mídia	< 0,0001
3	26	38	68.42	78.83	adj	gordo	< 0,0001
118	113	413	27.36	74.0		*ctx_1	< 0,0001
4	21	28	75.0	72.76	nom	padrão	< 0,0001
119	33	72	45.83	50.04		*gru_02	< 0,0001
5	9	10	90.0	40.07	nom	modelo	< 0,0001
6	10	13	76.92	35.36	nom	sociedade	< 0,0001
7	8	9	88.89	34.94	nr	barbie	< 0,0001
8	9	11	81.82	34.86	adj	grande	< 0,0001
9	14	23	60.87	34.21	ver	existir	< 0,0001
10	6	6	100.0	30.86	nom	mercado	< 0,0001
12	8	10	80.0	29.92	nom	roupa	< 0,0001
11	8	10	80.0	29.92	adj	forte	< 0,0001
13	9	13	69.23	26.94	nom	criança	< 0,0001
14	53	186	28.49	25.86	nom	pessoa	< 0,0001
16	5	5	100.0	25.68	nom	brasil	< 0,0001
15	5	5	100.0	25.68	nom	peito	< 0,0001
17	7	9	77.78	25.05	adj	feio	< 0,0001
18	6	7	85.71	24.78	ver	impor	< 0,0001
19	11	20	55.0	22.33	ver	influenciar	< 0,0001

Figura 34- Exibição das formas associadas à classe 5 do corpus "corpo".

Na figura 34 as formas (palavras) ativas são apresentadas em cinza e as variáveis ilustrativas ou os metadados em rosa. Como a figura está organizada em ordem decrescente segundo o valor do X^2, as formas suplementares (em azul) não aparecem.

Outros recursos para interpretar o perfil de cada classe e cada forma

Ainda na aba "Perfis", o conteúdo de cada uma das classes pode ser explorado pelo pesquisador a partir de mais recursos disponibilizados pelo *software*, conforme ilustra a figura 35. Para ter acesso a esses recursos basta **clicar com o botão direito no mouse sobre qualquer palavra** pertencente à classe que você deseja explorar. A

parte superior da lista oferece recursos para melhor explorar a presença da palavra selecionada nos ST da classe, enquanto sua parte inferior fornece informações referentes à respectiva classe.

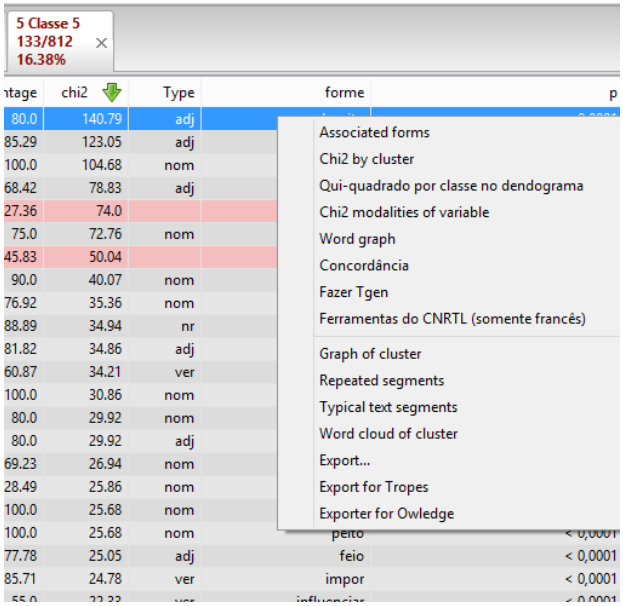


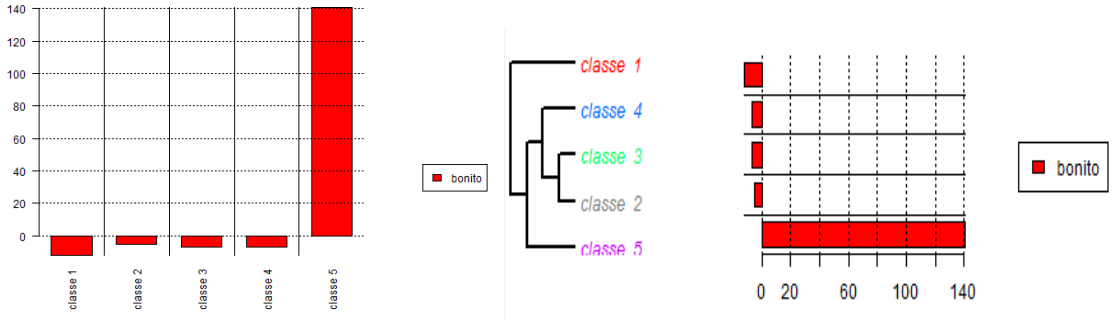
Figura 35- Recursos para interpretar cada uma das formas e das classes.

A parte superior do menu ilustrado pela figura 35 oferece as seguintes opções:

1- Formas ou palavras associadas: mostra a frequência de cada forma (lexema) que originou o lema indicado no perfil da classe (o termo selecionado “bonito” corresponde a 17 ocorrências de “bonito”, 15 de “bonita” e 6 de “bonitas”.
2- Qui-quadrado por classe: fornece um gráfico que exibe a associação da forma, no exemplo “bonito”, a cada classe.
3- Qui-quadrado por classe no dendrograma: fornece um gráfico equivalente, mas diferente do anterior.
4- Qui-quadrado das modalidades de uma variável (metadado): ao clicar numa variável (linha rosa) esta escolha fornece um gráfico da associação das suas modalidades com cada uma das classes.
5- Gráfico da palavra (forma): oferece um gráfico de similitude representando as ligações entre a forma escolhida e as outras formas da classe.
6- Concordância: mostra os ST onde a forma ou palavra ocorre (em função da classe, em função de todas as classes do dendrograma ou da totalidade do *corpus*).

7- Fazer Tgen**:** construir reagrupamentos de formas ou lemas que serão considerados como um conjunto ou um todo.

8- Ferramentas do CNRTL: conecta-se a base de dados do *"Centre National Ressources Textuelles et Lexicales"* (http://www.cnrtl.fr/); é necessária uma conexão com a internet e o interesse é somente para *corpus* em francês.



Figuras 36 e 37- Gráficos do qui-quadrado da forma por classes sem ou com dendrograma (itens 2 e 3 da parte superior do menu).

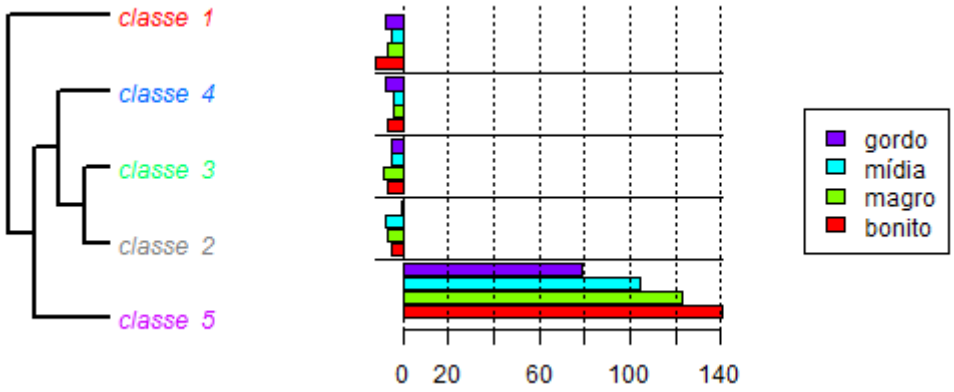


Figura 38- Gráfico do qui-quadrado das formas "gordo, mídia, magro e bonito" por classes e com dendrograma (item 3 da parte superior do menu).

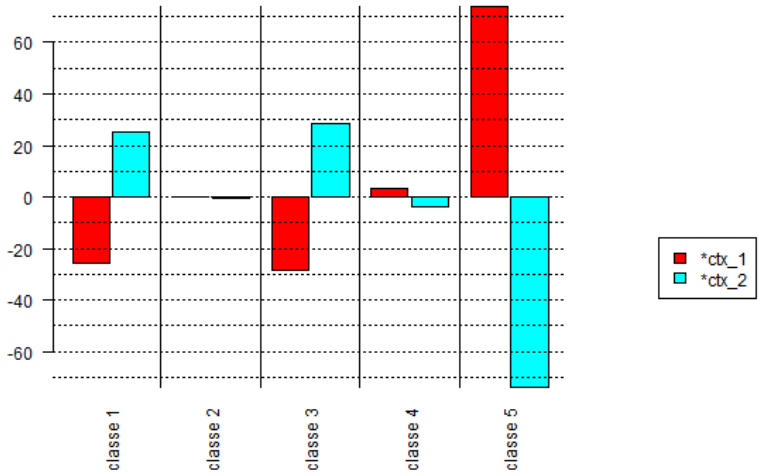


Figura 39- Gráfico do qui-quadrado das modalidades da variável "contexto de discussão dos grupos focais" por classes (item 4 da parte superior do menu).

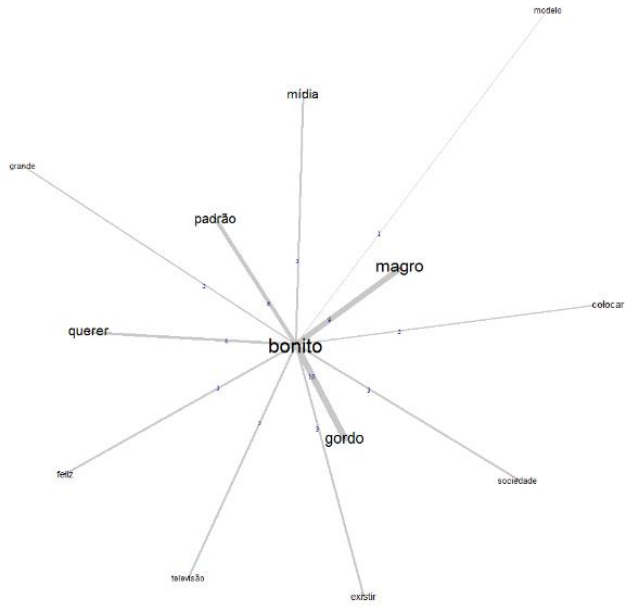


Figura 40- Gráfico de similitude da forma "bonito" da classe 5 do corpus "corpo" (item 5 da parte superior do menu).

A lista completa dos segmentos de texto da classe 5 que contém a palavra "bonito" pode ser salva, como arquivo ".html", clicando no botão correspondente indicado na parte inferior direita da figura 41.

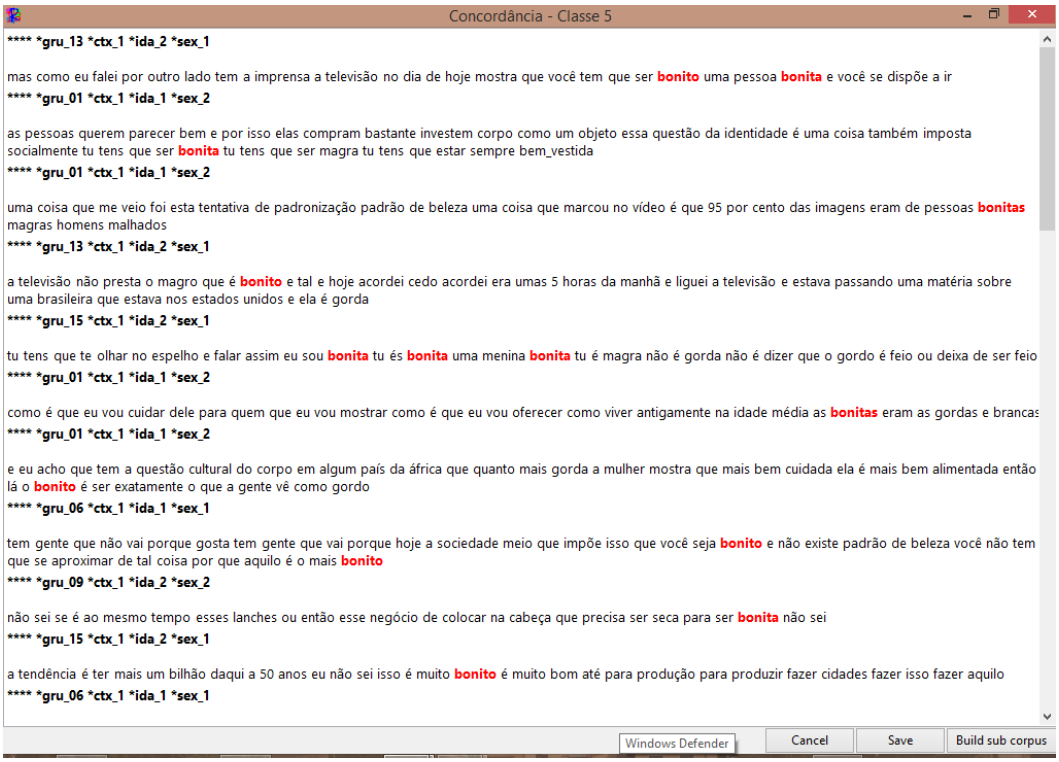


Figura 41- Segmentos de texto onde ocorre a forma "bonito" na classe 5 do corpus "corpo" (item 6 da parte superior do menu).

A parte inferior do menu ilustrado pela figura 35 oferece as seguintes opções:

1- Gráfico da classe: similitude representando as ligações entre as formas da classe.
2- Segmentos repetidos: os mais frequentes da classe.
3- Segmentos de texto típicos: por ordem decrescente do valor do qui-quadrado de associação com a classe.
4- Nuvem de palavras da classe.
5- Exporta: os segmentos de texto da classe escolhida são colocados num arquivo ".txt", ele poderá ser um *corpus* para novas análises.
6- Exporta para os aplicativos Tropes e Owledge.

As figuras seguintes ilustram resultados das opções da parte inferior do menu. Quanto ao gráfico da figura 42 é necessário apoiar no botão "EXPORT" para ele salvar na pasta de análise, já o gráfico da figura 45 é salvo automaticamente em uma das sub-pastas da pasta de análise "corpo_corpus_n".

As configurações da figura 42 foram as seguintes: selecionou-se as palavras com frequência igual ou superior a 6 (aproximadamente 1/4 do total de palavras); na aba "Configurações gráficas" assina-se a opção "Escores nas bordas", coloca-se em branco a opção "Edge curved" (aresta curva), seleciona-se "Comunidades" e "halo",

altera-se o “Tamanho do texto” de 10 para 8; na aba “Ajustes gráficos” apenas assinala-se “chi2” (qui-quadrado) no lugar de “eff.” (efetivo) na opção “Texto do vértice proporcional a frequência”.

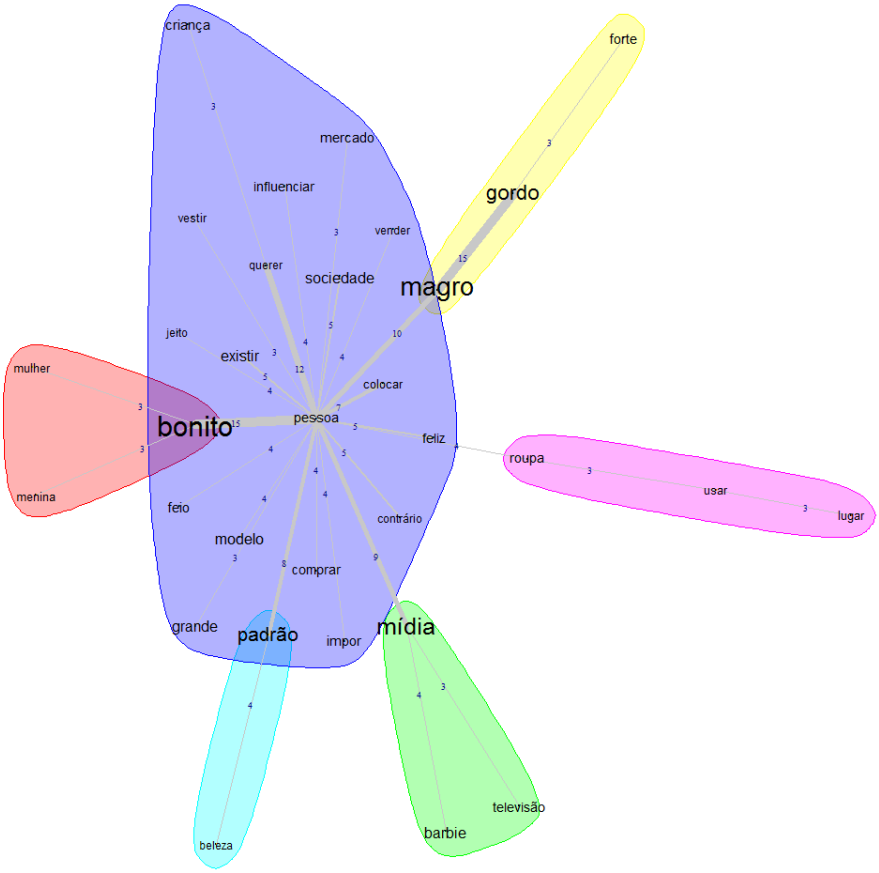


Figura 42- Gráfico de similitude (árvore máxima) das formas ou palavras da classe 5 do corpus “corpo” (item 1 da parte inferior do menu).

formes		
acho que	32	2
eu acho	31	2
que é	30	2
eu acho que	29	3
tem que	27	2
não é	26	2
a gente	24	2
as pessoas	24	2
o que	23	2
a mídia	20	2
que a	17	2
que eu	17	2
o corpo	16	2
é o	16	2
é uma	16	2
a pessoa	15	2

Figura 43- Segmentos repetidos mais frequentes da classe 5 do corpus "corpo" (item 2 da parte inferior do menu).

Figura 44- Segmentos de texto característicos da classe 5 do corpus "corpo" (item 3 da parte inferior do menu).

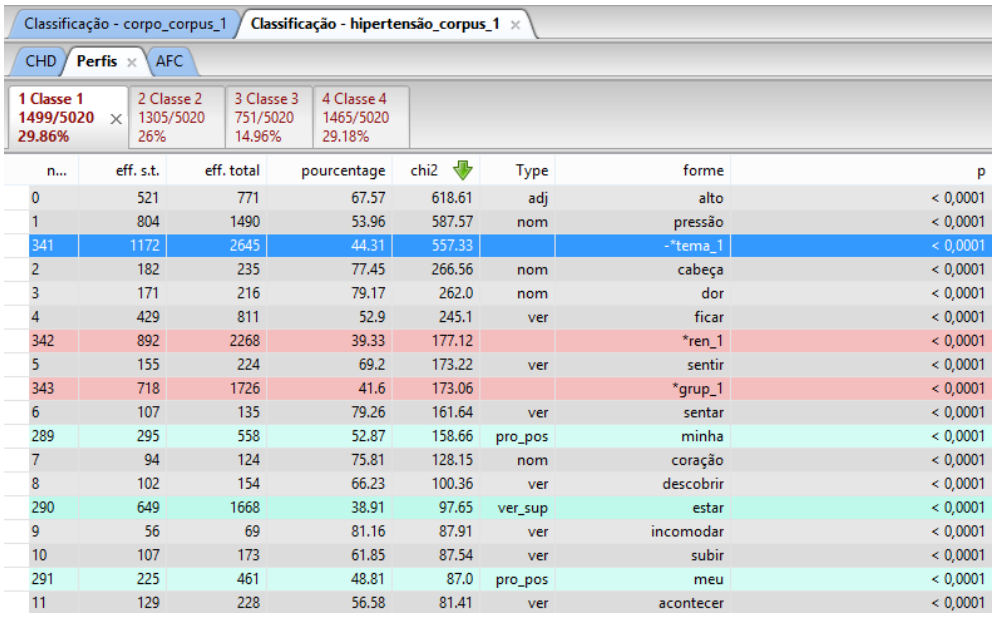
A figura 44 traz uma informação muito importante, trata-se do ambiente das palavras que ilustram cada classe, ou seja, os segmentos de textos associados a ela.

Figura 45- Nuvem de palavras características da classe 5 do corpus "corpo" (item 4 da parte inferior do menu).

As configurações da figura 45 foram as seguintes: selecionou-se as palavras com frequência igual ou superior a 6 (aproximadamente metade do total de palavras); na opção "Tamanho do texto" no mínimo mudou-se de 5 para 10 e máximo de 50 para 100; na "Cor do fundo" o branco foi substituído por azul claro.

Algumas indicações para classificação (CHD) de *corpus* temáticos

Relembrando, o *corpus* "hipertensão" envolve dois sub-temas: o "*tema_1" trata da hipertensão e o "*tema_2" do seu tratamento.



Classificação - corpo_corpus_1 | Classificação - hipertensão_corpus_1

CHD | Perfis | AFC

1 Classe 1 1499/5020 29.86%	2 Classe 2 1305/5020 26%	3 Classe 3 751/5020 14.96%	4 Classe 4 1465/5020 29.18%

n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	521	771	67.57	618.61	adj	alto	< 0,0001
1	804	1490	53.96	587.57	nom	pressão	< 0,0001
341	1172	2645	44.31	557.33		-*tema_1	< 0,0001
2	182	235	77.45	266.56	nom	cabeça	< 0,0001
3	171	216	79.17	262.0	nom	dor	< 0,0001
4	429	811	52.9	245.1	ver	ficar	< 0,0001
342	892	2268	39.33	177.12		*ren_1	< 0,0001
5	155	224	69.2	173.22	ver	sentir	< 0,0001
343	718	1726	41.6	173.06		*grup_1	< 0,0001
6	107	135	79.26	161.64	ver	sentar	< 0,0001
289	295	558	52.87	158.66	pro_pos	minha	< 0,0001
7	94	124	75.81	128.15	nom	coração	< 0,0001
8	102	154	66.23	100.36	ver	descobrir	< 0,0001
290	649	1668	38.91	97.65	ver_sup	estar	< 0,0001
9	56	69	81.16	87.91	ver	incomodar	< 0,0001
10	107	173	61.85	87.54	ver	subir	< 0,0001
291	225	461	48.81	87.0	pro_pos	meu	< 0,0001
11	129	228	56.58	81.41	ver	acontecer	< 0,0001

Figura 46- Perfis das classes do corpus temático "hipertensão".

Este tipo de *corpus* promove a interferência dos sub-temas na elaboração das classes. Observa-se na figura 46 que o "*tema_1" está associado com a classe 1. Embora não apareça na figura, ele também está associado com a classe 3, enquanto o "*tema_2" associou-se com as classes 2 e 4.

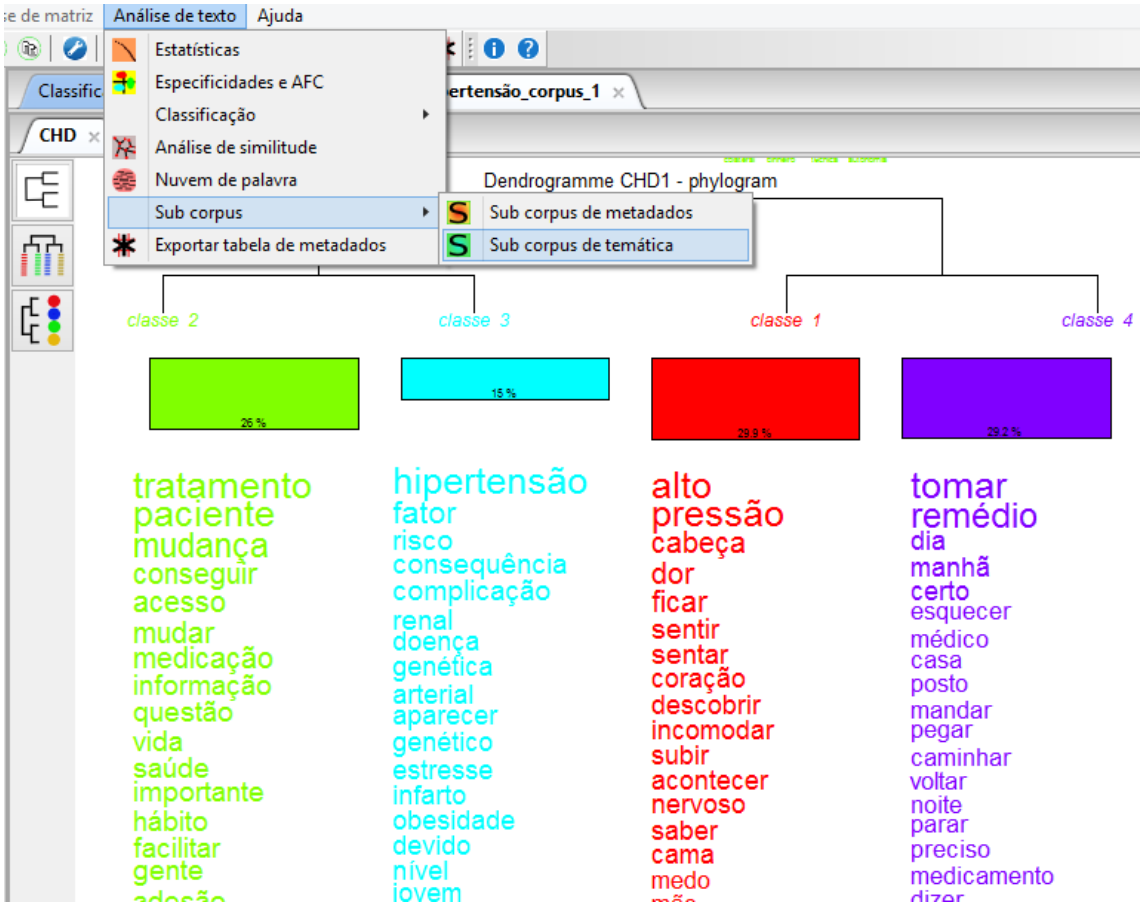


Figura 47- Dendrograma das classes do corpus temático "hipertensão" e menu para separação do em corpora monotemáticos.

A figura 47 indica que o sub-tema "hipertensão" é tratado de duas maneiras, no ramo esquerdo (classe 3, azul) com palavras mais especializadas, e no ramo direito (classe 1, vermelha) com palavras mais comuns. O que está em jogo aqui também é o tipo de participante, respectivamente o "profissional de saúde" e a "pessoa hipertensa". A mesma coisa acontece com o sub-tema "tratamento da hipertensão", no ramo esquerdo (classe 2, verde) as palavras são mais especializadas, e no ramo direito (classe 4, roxa) as palavras são mais próprias ao senso comum. Houve uma primeira partição em função dos participantes serem profissionais ou usuários, e no interior destas partições resultantes houve a interferência dos dois sub-temas: doença e seu tratamento.

Aconselha-se, quando tratamos de *corpora* temáticos, a realizar classificações (método de Reinert) de cada sub-tema. Para isto, conforme também ilustra a figura 47, no menu "Análise de texto" devemos criar os *corpora* temáticos ("Sub corpus de temática").

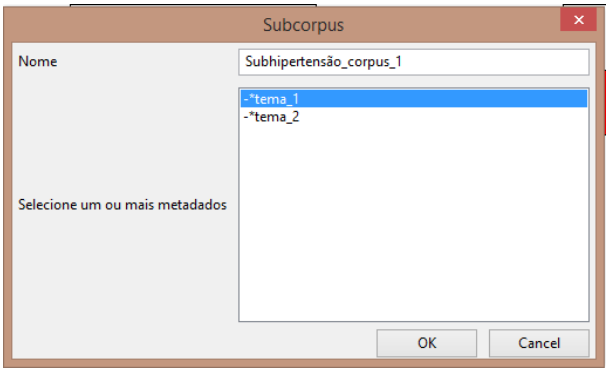


Figura 48- Criação de corpora monotemáticos para novas classificações (escolha do "*tema_1": hipertensão).

Ao clicar no tema escolhido (Figura 48), o *software* criará automaticamente uma nova pasta denominada "Subhipertensão_corpus_n_1" na pasta "hipertensão_corpus_n" (Figura 49). Nela temos arquivos do tipo ".db"[8] que permitem todas as análises textuais.

[8] Arquivos do tipo *"Data Base".*

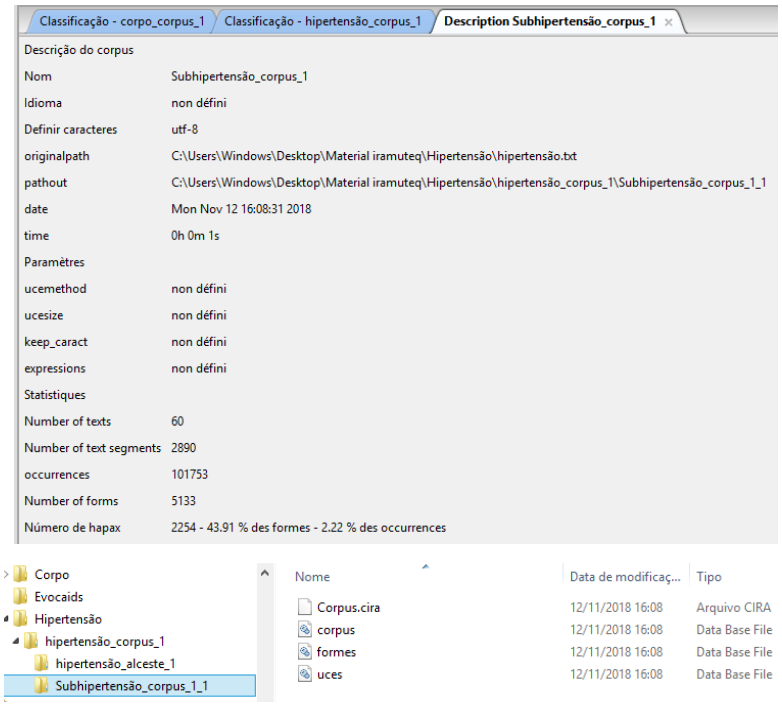
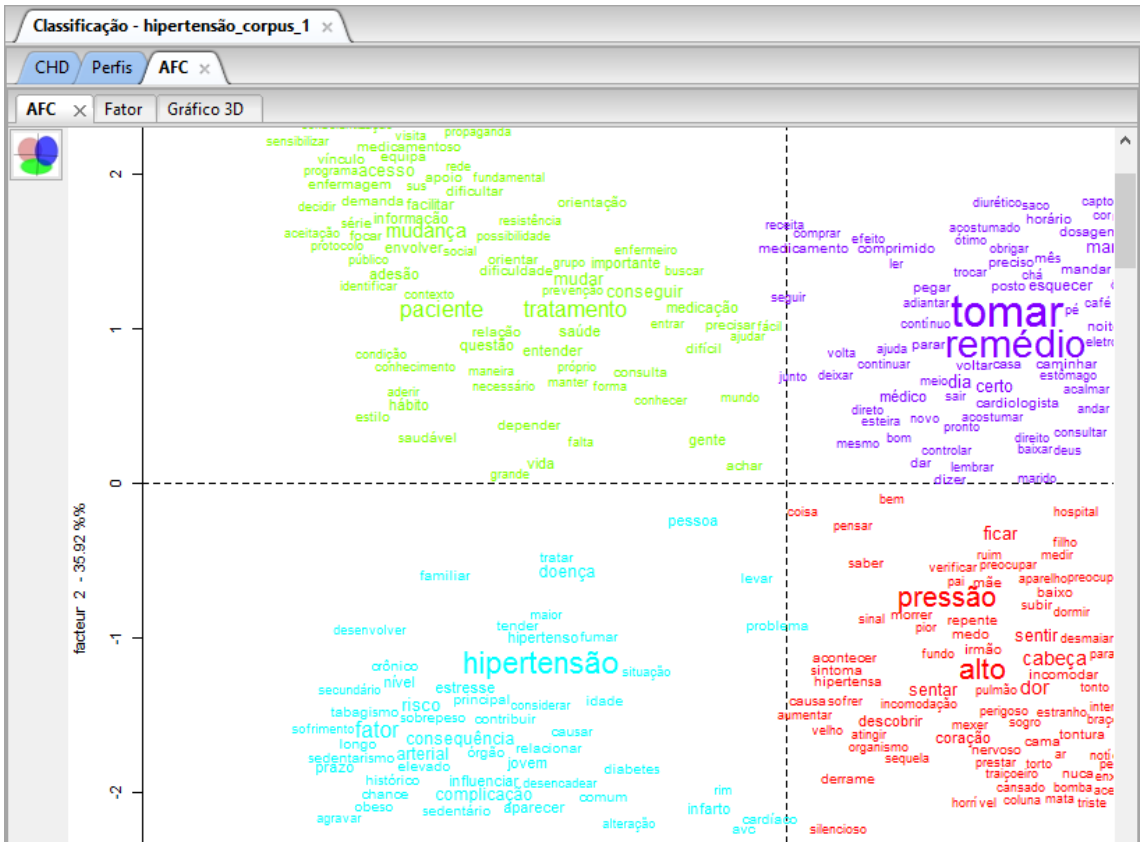


Figura 49- Resultados preliminares do corpus relativo ao sub-tema "hipertensão".

Análise fatorial de correspondência?

Retomando o *corpus* temático hipertensão, na interface "Classificação", a última aba é a "AFC" (análise fatorial de correspondência). Embora haja uma polêmica se seria realmente uma AFC (Reinert, 1995) ou uma análise pós-fatorial (Cibois, 1983), pois trata-se de uma análise feita a partir da classificação hierárquica descendente (CHD), neste tutorial considera-se que estas representações em planos fatoriais são uma outra forma de visualizar os conteúdos e relações entre as classes.



AFC | **Fator** | Gráfico 3D

formes	Valeurs propres	Pourcentages	Pourcentage cumules
facteur 1	0.27352	42.82591	42.82591
facteur 2	0.22942	35.92095	78.74686
facteur 3	0.13574	21.25314	100

Figura 50- Representação em planos fatoriais e contribuições dos fatores da análise classificatória do corpus "hipertensão"

Na parte inferior da figura 50, observa-se que o número de fatores é igual ao número de classes da CHD menos 1. O plano fatorial 1x2 retém 78,75% da variância. E é este plano que está representado na figura. No entanto, há um número excessivo de palavras, e com tamanho pequeno, o que prejudica a visualização. Ao clicar no botão superior esquerdo (que representa um gráfico) tem-se acesso as configurações gráficas ilustradas na figura 51.

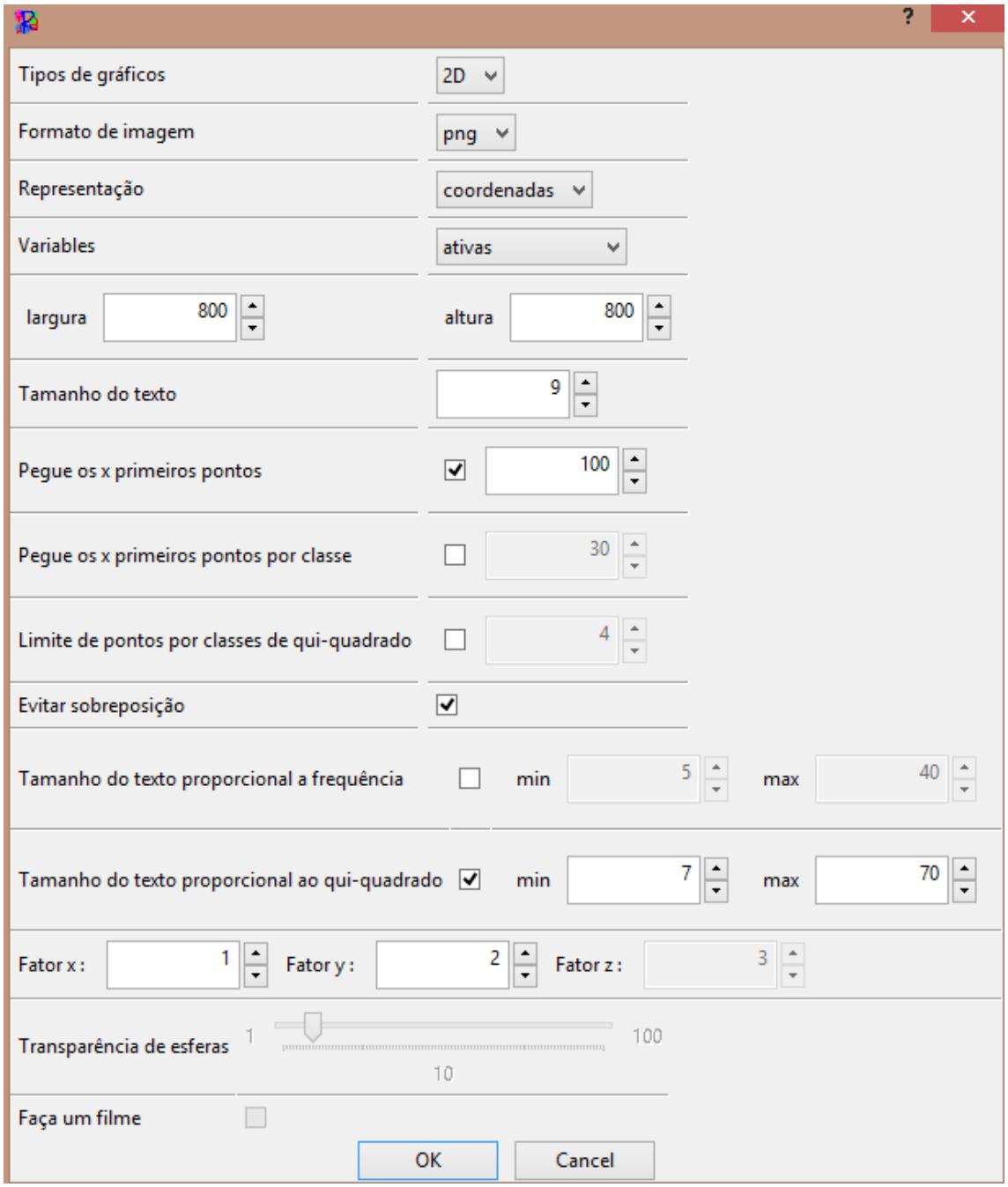


Figura 51- Interface de configurações gráficas da AFC.

Conforme a figura 51, com as seguintes alterações: “pegar os 100 primeiros pontos”, marcar para “evitar sobreposição”, marcar e alterar o “tamanho do texto proporcional ao qui-quadrado” de mínimo 5 e máximo 40 para mínimo 7 e máximo 70; obtém-se o gráfico da figura 52.

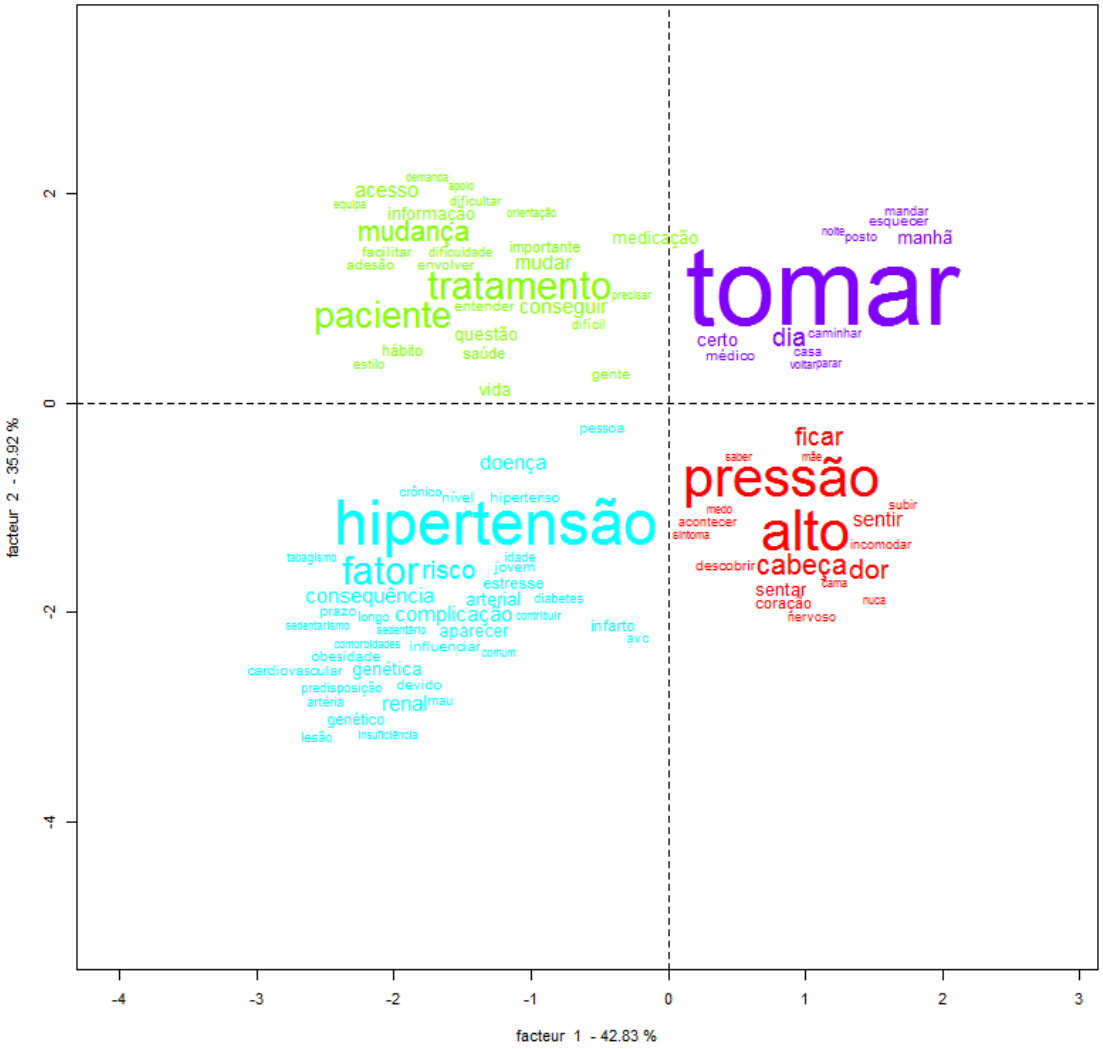


Figura 52- Gráfico do plano fatorial 1x2 da AFC do corpus "hipertensão".

No gráfico 52, observa-se no fator 1 (eixo horizontal) uma contraposição entre as classes 2 (verde) e 3 (azul) do lado esquerdo e as classes 4 (roxa) e 1 (vermelha) do lado direito; o que indica um contraste entre profissionais de saúde e pacientes, em relação a esta doença crônica. Também há uma contraposição no fator 2 (vertical) entre as classes 2 e 4 na parte superior e as classes 3 e 1 na inferior, trata-se aqui da temática "tratamento da hipertensão" na parte superior e da temática "hipertensão" na parte inferior.

Mais recursos (menu do lado direito da interface do *software*)

Na coluna da esquerda na interface do IRaMuTeQ, **clicando com o botão direito do mouse sobre a análise denominada "nomedocorpus_alceste_n"**, no

caso “hipertensão_alceste_n”, um menu oferece mais recursos, conforme a figura 53.você pode ter acesso a mais alguns resultados da análise. A figura exibe a lista destes recursos.

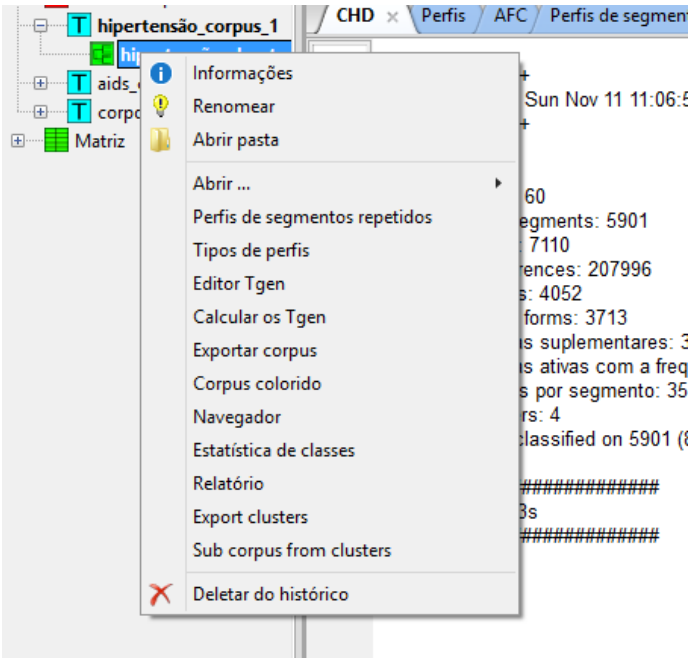


Figura 53- Menu do lado direito da interface do software IRaMuTeQ.

As opções deste menu são as seguintes:

1- Informações: fornece algumas informações sobre o *corpus*.
2- Abrir ...: fornece as formas ou palavras significativamente ausentes de determinada classe (“antiperfis”).
3- Perfis de segmentos repetidos (Figura 54 e 55): permite a configuração do número de palavras e fornece os segmentos repetidos por classe.

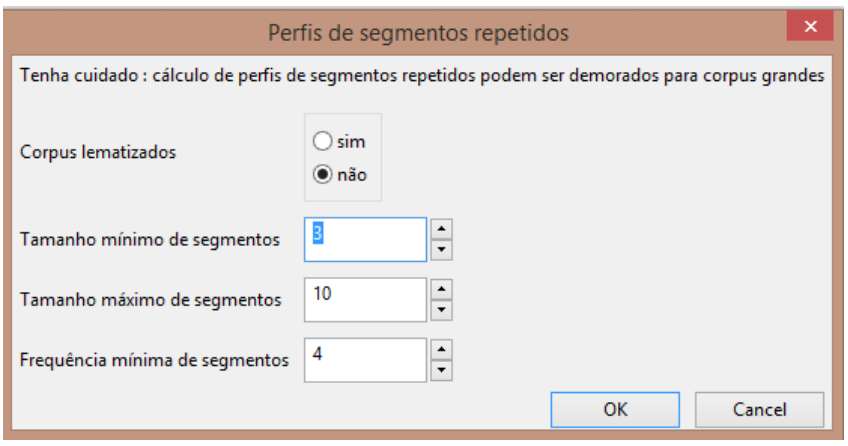


Figura 54- Configuração do número mínimo e máximo de palavras por segmento de texto.

CHD | Perfis | AFC | Profils des segements répétés

classe 1 | classe 2 | classe 3 | classe 4

n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	172	236	72.88	187.02		a pressão alta	< 0,0001
1	125	150	83.33	186.39		dor de cabeça	< 0,0001
2	73	104	70.19	71.88		com a pressão alta	< 0,0001
3	96	152	63.16	70.32		com a pressão	< 0,0001
4	41	47	87.23	67.43		causa da pressão	< 0,0001
5	41	47	87.23	67.43		por causa da pressão	< 0,0001
6	68	103	66.02	56.63		por causa da	< 0,0001
7	75	121	61.98	51.87		que a pressão	< 0,0001
8	27	29	93.1	50.83		causa da pressão alta	< 0,0001
9	27	29	93.1	50.83		por causa da pressão a...	< 0,0001
10	54	79	68.35	49.51		pressão alta e	< 0,0001
11	32	38	84.21	48.74		pressão alta é	< 0,0001
12	29	33	87.88	48.43		tinha pressão alta	< 0,0001
13	38	49	77.55	47.96		pressão está alta	< 0,0001
14	59	92	64.13	45.18		a pressão está	< 0,0001
15	20	20	100.0	43.35		dor na nuca	< 0,0001
16	41	57	71.93	42.99		da pressão alta	< 0,0001
17	34	44	77.27	42.55		a pressão está alta	< 0,0001
18	21	22	95.45	41.56		uma dor de cabeça	< 0,0001
19	27	32	84.38	41.3		não sente nada	< 0,0001
20	22	24	91.67	40.12		uma dor de	< 0,0001
21	24	28	85.71	37.99		que a pressão alta	< 0,0001

Figura 55- Resultado dos segmentos de texto associados a classe 1 da CHD do corpus "hipertensão".

4- Tipos de perfis: calcula os perfis das categorias gramaticais, gerando um arquivo tipo planilha (".csv").
5- Editor Tgen: edita os reagrupamentos de formas ou lemas.
6- Calcular os Tgen: oferece cálculos sobre estes reagrupamentos.
7- Exportar *corpus*: permite exportar o *corpus* já dividido em segmentos de texto associados a variável descritiva (metadado) e a classe que eles pertencem.

**** *ind_01 *grup_1 *sex_1 *ren_1 *paph_3 *papf_1 *papp_2

a hipertensão eu acho o seguinte ela aparece é silenciosa se a pessoa não tiver o cuidado de saber que é hipertenso através da consulta médica é muito ruim porque aquilo se agrava e a pessoa sofre muito

eu por exemplo eu tive a experiência de praticamente não saber que era hipertenso e eu tinha desconforto com a parte cardíaca mas eu não tive infarto não tive nada

eles fizeram vários exames e no exame cardiológico mais elaborado foi descoberto que eu tinha uma coronária obstruída e foi necessário fazer uma ponte safena então eu fiz a cirurgia e eu tenho tido controle da pressão e está sendo muito bem feito

após a cirurgia sempre segui com os remédios receitados e nunca tinha mudado e há um ano eu fui atendido aqui no posto e o cardiologista mudou a medicação e eu achei que foi muito importante

porque me deu uma sensação de melhora a pressão melhorou e está bem controlada eu recentemente fiz um check_up fiz uns exames por causa da cirurgia que eu fiz e pela idade eles analisaram

mas eu não tive a consulta ainda mas eu estou em um nível adequado para a minha situação eu caminho diariamente por 45 minutos ou 1 hora na beira mar eu moro aqui perto

eu estou com 76 anos então quando eu era um pouco mais jovem eu fazia academia musculação que é um pouco mais pesado mas agora devido à idade eu só caminho

onde eu moro tem esteira tem tudo mas eu não consigo fazer estou fazendo o que eu posso fazer o próprio médico disse que eu não posso exagerar muito porque sou cardiopata

Figura 56- Corpus "hipertensão" colorido.

8- *Corpus* colorido (Figura 56): o *corpus* total é apresentado sob a forma de arquivo tipo ".html" com seus segmentos de texto com as cores das suas respectivas classes da CHD (os ST não classificados aparecem em preto).

9- Navegador: apresenta uma matriz com todas as formas ou palavras e os valores do qui-quadrado de ligação com cada classe (figura 57).

Navigation

	formes	classe 1	classe 2	classe 3	classe 4
0	pessoa	-2.757	51.259	105.311	-176.567
1	pressão	587.57	-344.046	-19.441	-9.282
2	tomar	-173.749	-125.033	-187.875	1211.891
3	gente	-18.585	108.231	-0.167	-28.914
4	remédio	-125.961	-129.107	-139.208	993.399
5	achar	-0.101	50.847	-1.891	-30.037
6	coisa	23.405	0.004	-0.637	-18.529
7	ficar	245.099	-121.03	-105.045	8.383
8	alto	618.61	-231.36	-14.206	-54.843
9	vez	-0.694	5.893	-18.389	3.464
1	hipertensão	-38.725	-0.126	678.511	-191.28
1	saber	77.419	-30.65	-0.415	-9.064
1	bem	22.142	-22.888	-10.898	6.096
1	cuidar	7.397	-17.455	-0.336	3.054
1	só	-2.981	1.25	-1.417	2.539
1	médico	-7.834	-6.257	-38.509	102.008
1	dar	4.533	-30.543	-20.168	45.054
1	dia	-27.298	-45.388	-19.835	232.684
1	dizer	13.123	-54.942	-20.187	49.419
1	vida	-104.178	128.924	86.517	-63.646
2	problema	33.567	-27.158	52.456	-42.08
2	tratamento	-117.602	386.75	-11.623	-28.969
2	medicação	-84.234	145.048	-47.244	9.074
2	comer	1.771	-6.605	-0.899	3.549
2	mesmo	-2.79	-4.735	-3.637	27.85
2	sal	-5.252	-0.321	-0.046	9.136
2	doença	-26.435	16.473	170.704	-80.839

Figura 57- Interface da opção "Navegador" da CHD do corpus "hipertensão".

10- Estatísticas de classes: gera um arquivo denominado "stat_par_classe.csv" (planilha) na sub-pasta "nome do corpus_alceste_n", ele contém números totais de ocorrências; de formas diferentes; de hápax; de segmentos de texto; e um índice do número de hápax pelo número de formas.

11- Relatório: cria um documento em .txt, denominado "RAPPORT" dentro da pasta que contém o *corpus*, em uma sub-pasta denominada "nomedocorpus_alceste_n"; ele pode ser visualizado em qualquer editor de texto, contém a descrição lexical de cada uma das classes formadas pela CHD.

```
+-+-+-+-+-+-+-+-+
|i|R|a|M|u|T|e|Q| - Sun Nov 11 11:06:55 2018
+-+-+-+-+-+-+-+-+


Number of texts: 60
Number of text segments: 5901
Number of forms: 7110
Number of occurrences: 207996
Número de lemas: 4052
Number of active forms: 3713
Número de formas suplementares: 330
Número de formas ativas com a frequência >= 3: 1838
Média das formas por segmento: 35.247585
Number of clusters: 4
5020 segments classified on 5901 (85.07%)

#########################
tempo : 0h 0m 33s
#########################

classe 1 - 1499 uce sur 5020 - 29.86%
    0|  521|   771| 67.57|  618.61|     adj|              alto   < 0,0001
    1|  804|  1490| 53.96|  587.57|     nom|           pressão   < 0,0001
    2|  182|   235| 77.45|  266.56|     nom|           cabeça   < 0,0001
    3|  171|   216| 79.17|   262.0|     nom|              dor   < 0,0001
    4|  429|   811|  52.9|   245.1|     ver|            ficar   < 0,0001
    5|  155|   224|  69.2|  173.22|     ver|           sentir   < 0,0001
    6|  107|   135| 79.26|  161.64|     ver|           sentar   < 0,0001
    7|   94|   124| 75.81|  128.15|     nom|          coração   < 0,0001
    8|  102|   154| 66.23|  100.36|     ver|        descobrir   < 0,0001
    9|   56|    69| 81.16|   87.91|     ver|        incomodar   < 0,0001
   10|  107|   173| 61.85|   87.54|     ver|           subir   < 0,0001
   11|  129|   228| 56.58|   81.41|     ver|       acontecer   < 0,0001
   12|   48|    57| 84.21|   81.32|     adj|         nervoso   < 0,0001
   13|  283|   631| 44.85|   77.42|     ver|           saber   < 0,0001
   14|   36|    38| 94.74|   76.95|     nom|            cama   < 0,0001
   15|   62|    87| 71.26|   72.47|     nom|            medo   < 0,0001
```

Figura 58- Extrato inicial do "Relatório" da classificação do corpus "hipertensão".

12- Exportar classes: cria arquivos textuais "classe_x_export.txt" para cada classe que possibilita considerar uma classe como um novo *corpus* para análises textuais.

13- Sub-*corpus* de classes: permite a criação de uma pasta "Subnomedocorpus_corpus_x_y" que permite agrupar as classes que se escolhe (ver figuras 59 e 60) e contém três arquivos tipo ".db" (Data Base).

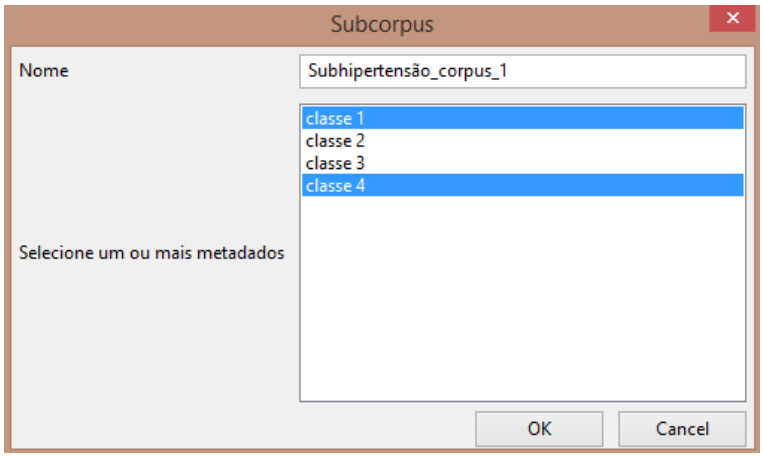


Figura 59- Interface de escolha das classes para a construção do sub-corpus.

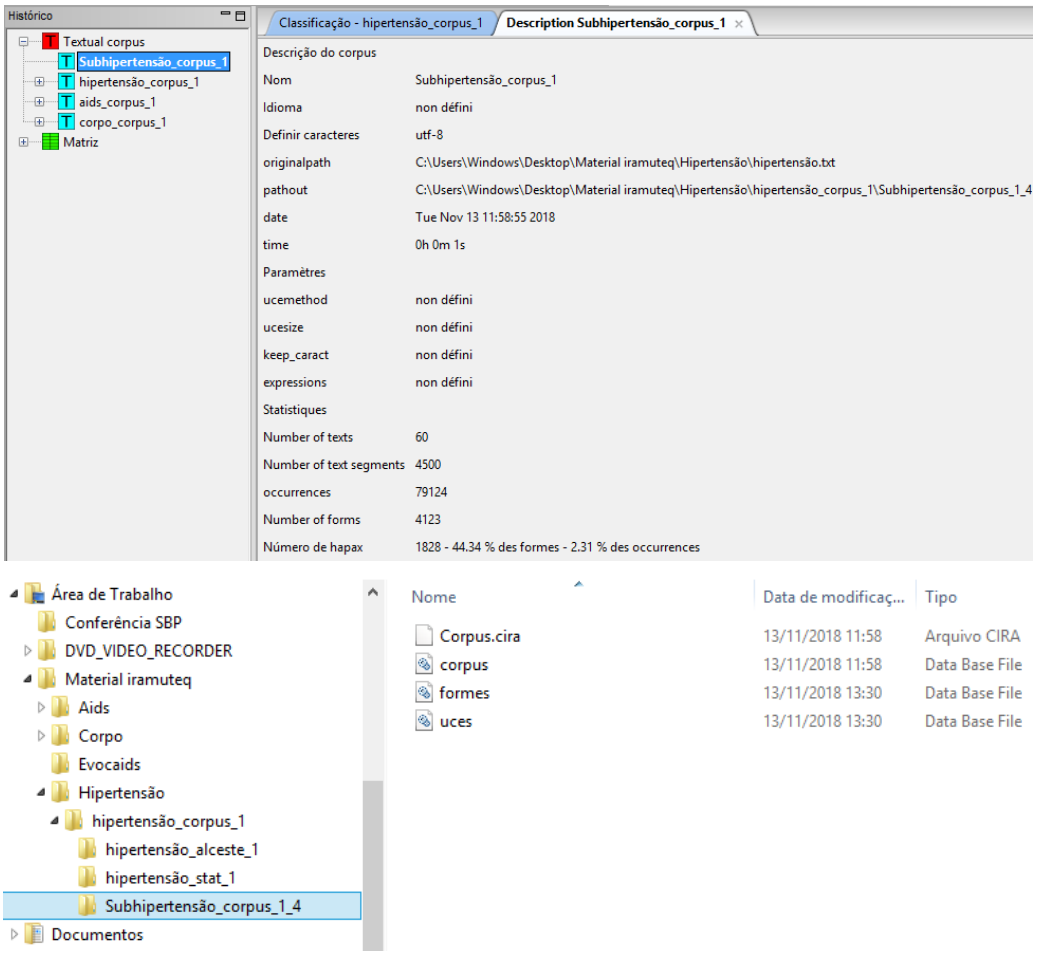


Figura 60- Interface de características do corpus resultante do agrupamento das classes 1 e 4 da classificação do corpus inicial "hipertensão".

Análise: Similitude

Para esta parte vamos utilizar o *corpus* de respostas a um questionário, denominado "aids". Ao escolher a análise de similitude, uma nova janela se abrirá (figura 61), possibilitando que sejam escolhidos alguns parâmetros para a construção da árvore de similitudes. Do lado esquerdo escolhe-se as palavras que comporão o gráfico; do lado direito, na aba "Configurações gráficas" há uma série de opções, e também podemos acessar outra aba, a "Ajustes gráficos".

A figura 62 mostra uma árvore máxima caso não se configure nada. No canto superior esquerdo dessa janela, logo abaixo da aba "Gráfico", temos dois botões. O primeiro deles, com traços vermelhos e pontos pretos permite que se modifique as configurações da análise, abrindo novamente a janela para edição dos parâmetros. O segundo botão, no qual está escrito "EXPORT" e tem uma seta, depois do gráfico terminado, salva-o em um arquivo imagem ("graph_simi_n.png"), na pasta "aids_simitxt_n".

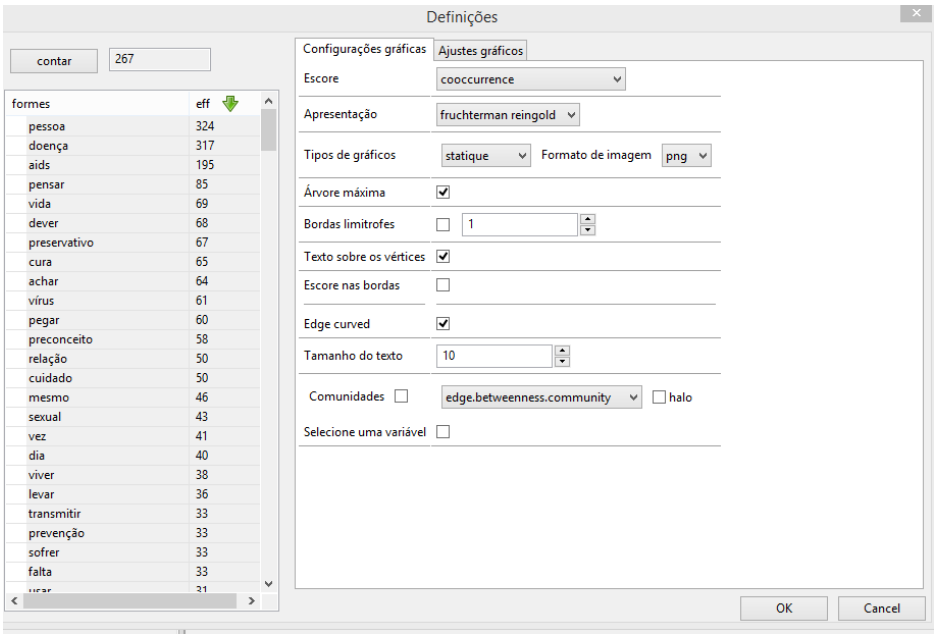


Figura 61- Interface de configurações para análise de similitude do corpus "aids".

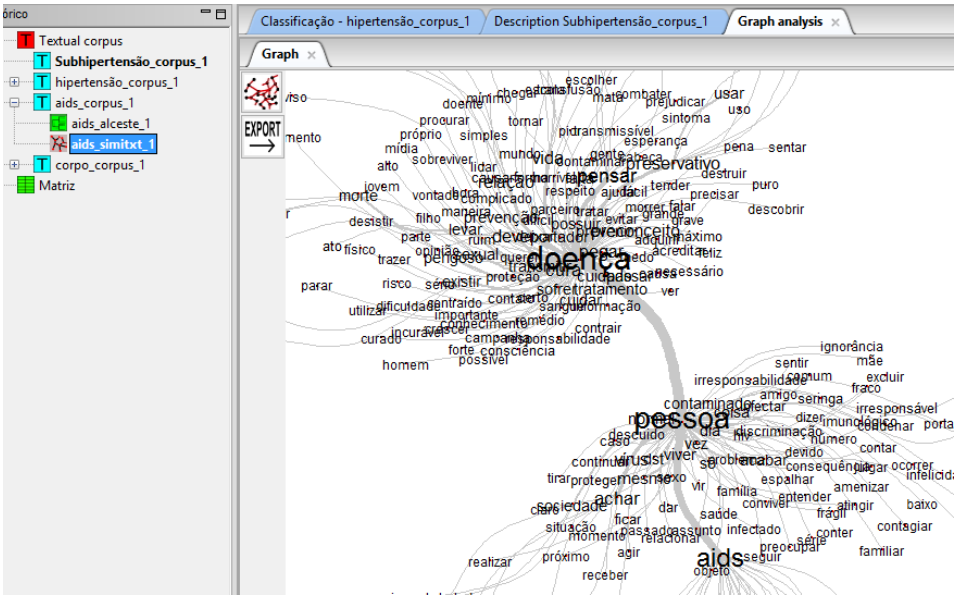


Figura 62- Árvore máxima de similitude do corpus "aids" sem configurações.

Na aba *Ajustes Gráficos*, por sua vez, é possível fazer edições gráficas (tamanho do texto, tamanho das arestas, cores, etc.).

A figura 63 mostra algumas configurações escolhidas. Primeiramente, na seleção de palavras sugere-se não selecionar as palavras com frequências muito altas, como no caso: "pessoa, doença, aids e pensar"; já que elas estão ligadas as questões ou consignas da coleta dos dados que geraram este material textual.

Não se deve selecionar também as palavras com frequências baixas, em benefício da visibilidade e comunicabilidade do gráfico. No caso do corpo “aids” com exceção destas palavras selecionou-se aquelas com frequência igual ou superior a 15 (cerca de 1/4 do total de palavras desta lista).

Na aba “Configurações gráficas” assinou-se a opção “Escores nas bordas”, colocou-se em branco a opção “Edge curved” (aresta curva), selecionou-se “Comunidades” e “halo”, alterou-se o “Tamanho do texto” de 10 para 8. A escolha de “Comunidades” e “halo” permite que as palavras mais associadas fiquem agrupadas, envoltas por nuvens coloridas. E a escolha de “Escores nas bordas” deixa visível no gráfico os valores relativos às relações entre as palavras (no exemplo as coocorrências).

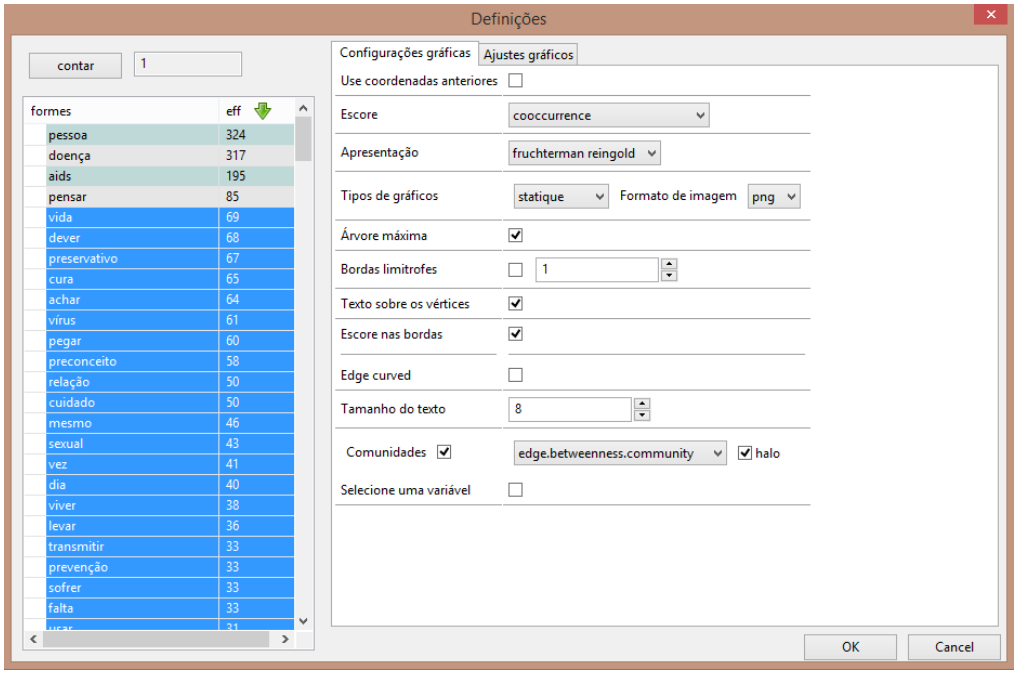


Figura 63- Aba das “Configurações gráficas” da análise de similitude do corpus “aids”.

A figura 64 indica as opções da aba “Ajustes gráficos”, para se gerar o gráfico da figura 65: assinalou-se na opção “Texto do vértice proporcional a frequência” “chi2” (qui-quadrado) no lugar de “eff.” (efetivo), diminuindo o valor mínimo de 10 para 8 e o máximo de 25 para 20.

Todas estas configurações permitiram gerar o gráfico da figura 65, ilustração das relações entre as principais palavras ou formas que compõem os segmentos de texto ou resposta a um questionário sobre aids.

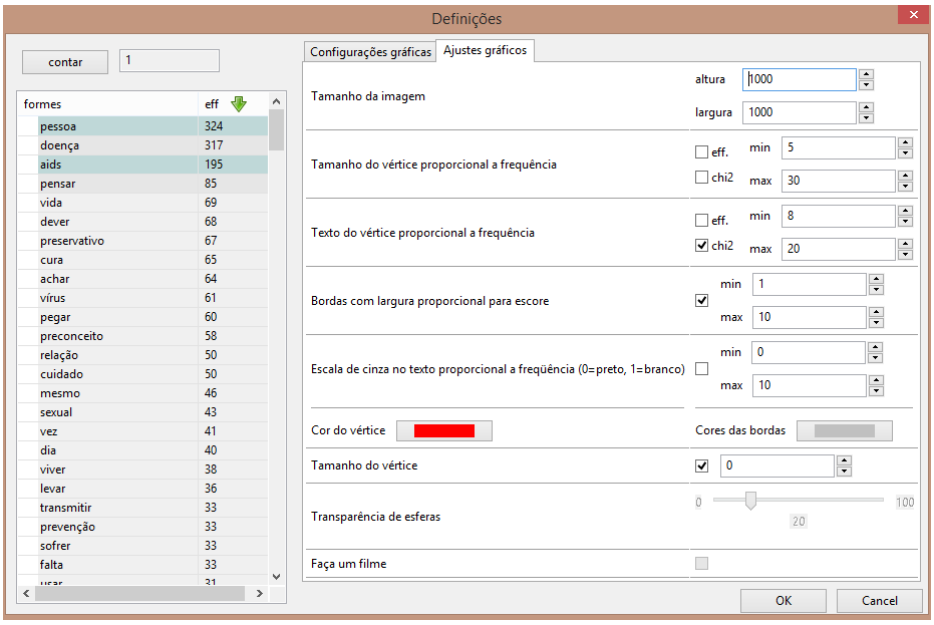


Figura 64- Aba dos "Ajustes gráficos" da análise de similitude do corpus "aids".

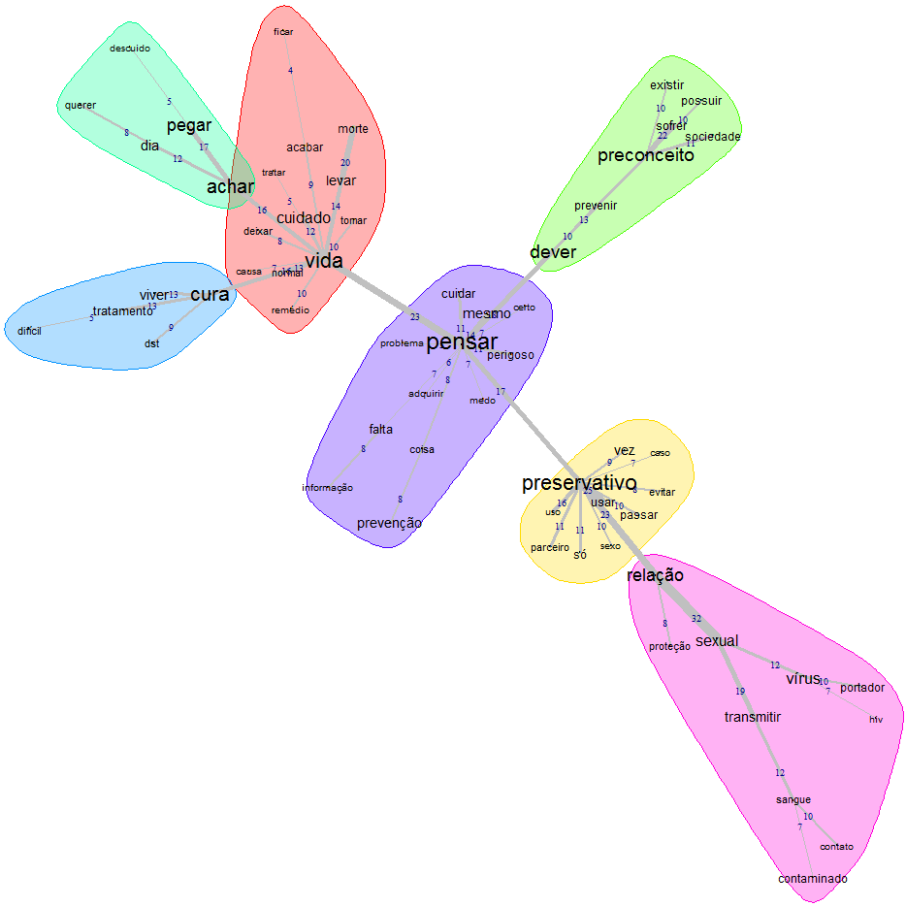


Figura 65- Árvore máxima de similitude do corpus "aids" com configurações.

Ao marcar “Selecione uma variável” (Figura 66) é possível escolher uma variável categorial para participar da análise de similitude, podendo identificar diferenças entre grupos. Aqui escolheu-se a variável “*esc” (Tipo de escola). Na aba “Configurações gráficas” preserve as configurações do gráfico anterior, exceto as opções “Comunidades” e “halo” que não são marcadas. Na aba “Ajustes gráficos” deixe como estava anteriormente.

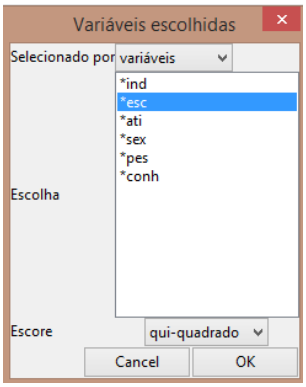


Figura 66- Janela quando se marca “Selecione uma variável” para indicar o papel das modalidades desta variável na análise de similitude do corpus “aids”.

Tendo escolhido a variável descritiva (metadado) clique em “OK” e aguarde enquanto a análise se finaliza.

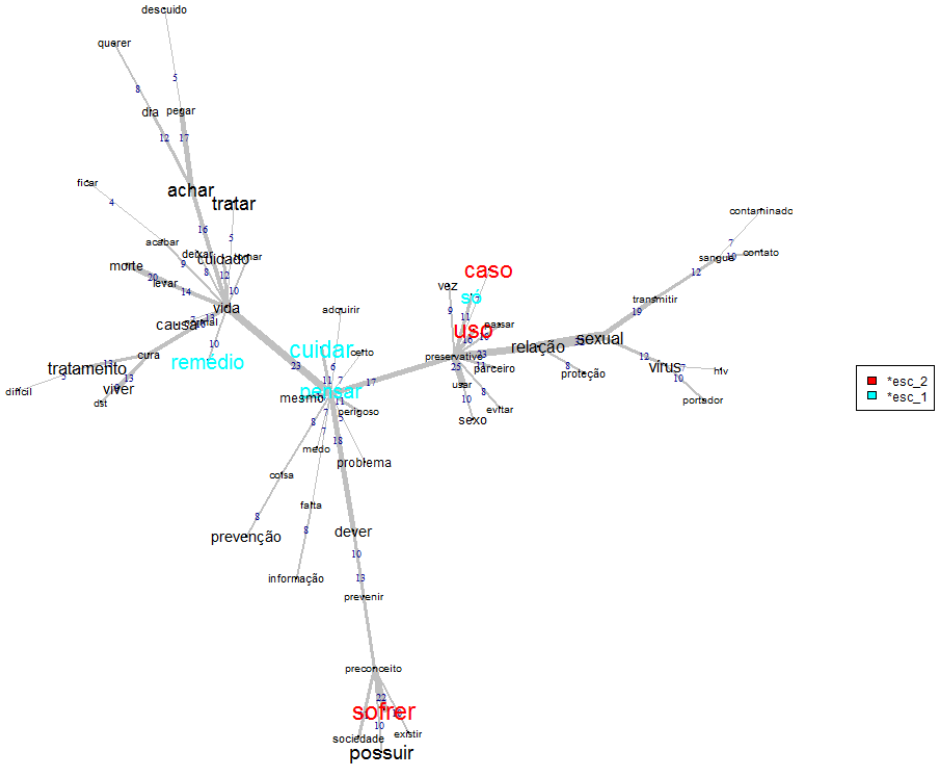


Figura 67- Árvore máxima de similitude do corpus “aids” segundo a variável “tipo de escola”.

Outra possibilidade é ajustarmos manualmente o gráfico de similitude. Para isto na aba “Configurações gráficas” escolha o “Tipo de gráfico” dinâmico, assinale a opção “Escores nas bordas”, coloque em branco a opção “Edge curved” (aresta curva), não selecione “Comunidades” e nem “halo”. Na aba “Ajustes Gráficos” na opção “Tamanho do vértice proporcional a frequência” selecione qui-quadrado, deixando o mínimo de 5 e o máximo de 30). Para “Cor do vértice” escolha um azul claro e na opção “Cores dos vértices” deixe a cor cinza. O resultado será a figura 68.

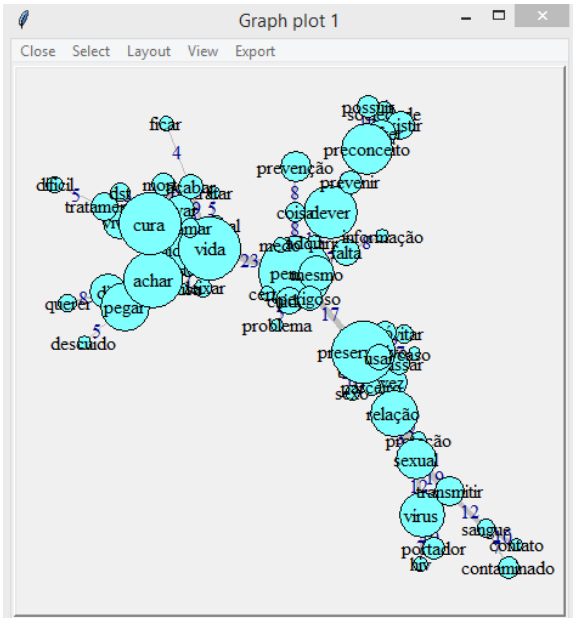


Figura 68- Gráfico dinâmico da arvore máxima de similitude do corpus “aids” (primeira etapa).

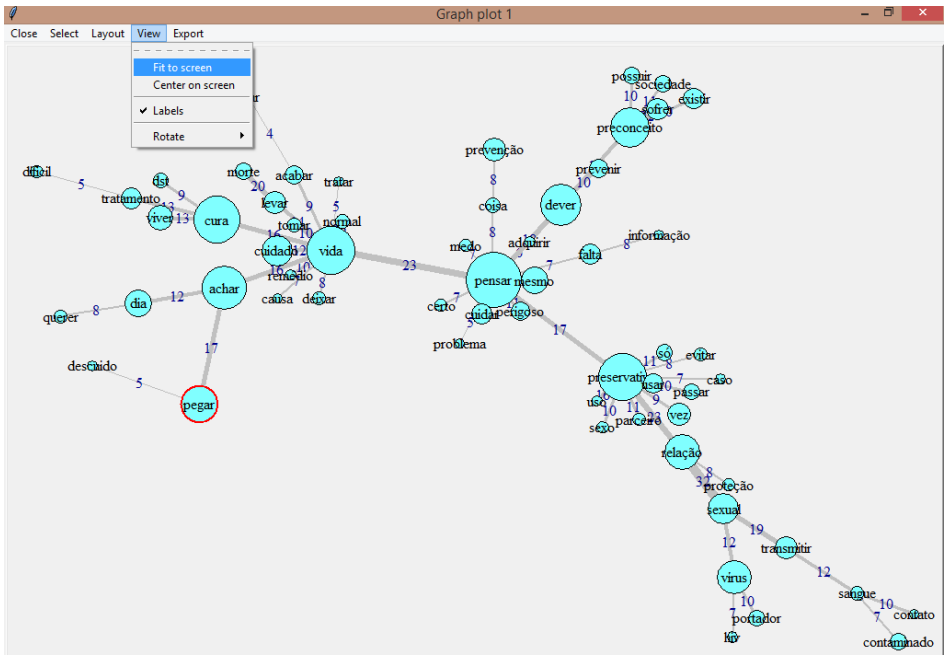
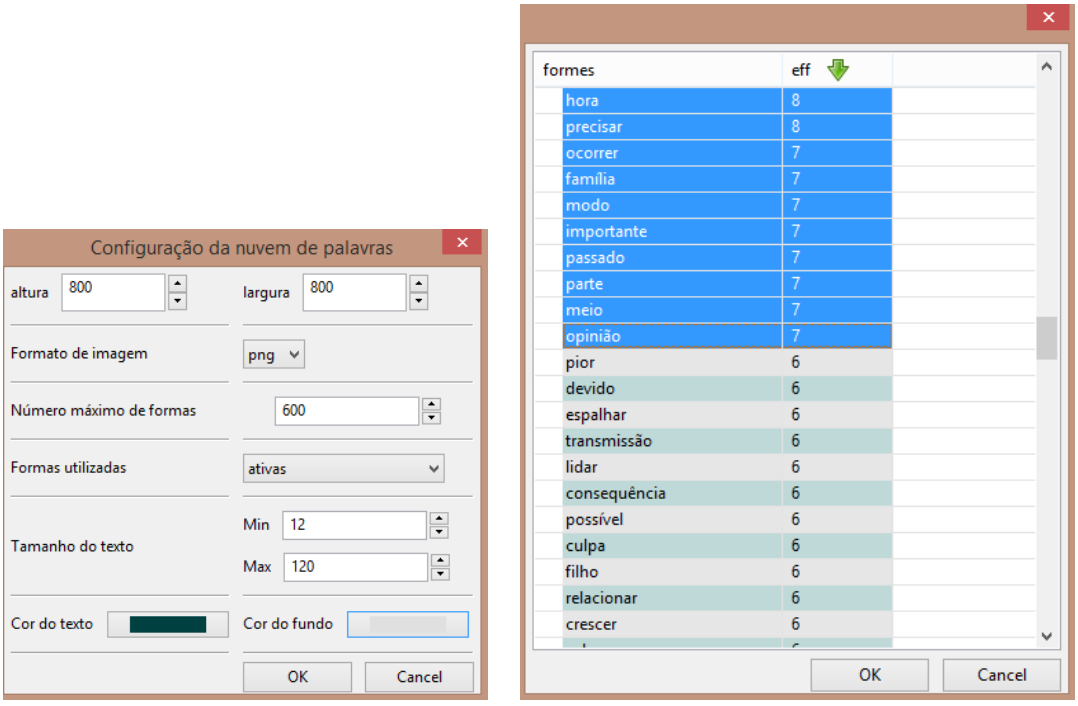


Figura 69- Gráfico dinâmico da arvore máxima de similitude do corpus “aids” após o ajuste da tela (segunda etapa).

A figura 69 ilustra o gráfico anterior (Figura 68) após a escolha da opção “Ajustar a tela” (*Fit to screen*) do menu. Quando se seleciona uma palavra (ou vértice) segurando o botão do mouse, podemos modificar o posicionamento desta palavra, como ilustra a palavra “pegar”, circulada em vermelho. Ela foi alterada. Assim podemos fazer com todas. Para salvar o gráfico resultante sugere-se usar o procedimento de imprimir a tela (*print screen*) e depois salvar como uma figura. Outra opção é no menu “Exportar”, criar um arquivo “Postscript”, para ser importado por um *software* gráfico compatível.

Análise: Nuvem de palavras

Esta é uma análise mais simples, que trabalha com a representação gráfica em função da frequência das palavras. Ao escolher a nuvem de palavras, uma nova janela se abrirá (Figura 70). Na opção “Tamanho do texto” mude o mínimo de 5 para 12 e máximo de 50 para 120. Escolha para “Cor do texto” um verde escuro, e para a “Cor do fundo” um cinza claro.



Figuras 70 e 71- Configurações do gráfico de nuvens e escolha das formas do corpus “aids”.

A figura 71 mostra a interface para a “Escolha das formas”: selecione aproximadamente metade das formas deste *corpus*, ou seja, as palavras com frequência igual ou superior a 7. Tendo realizado estas configurações clique em “OK”. Em seguida aparecerá o resultado gráfico, a nuvem de palavras do *corpus* “aids”, conforme indica a figura 72.

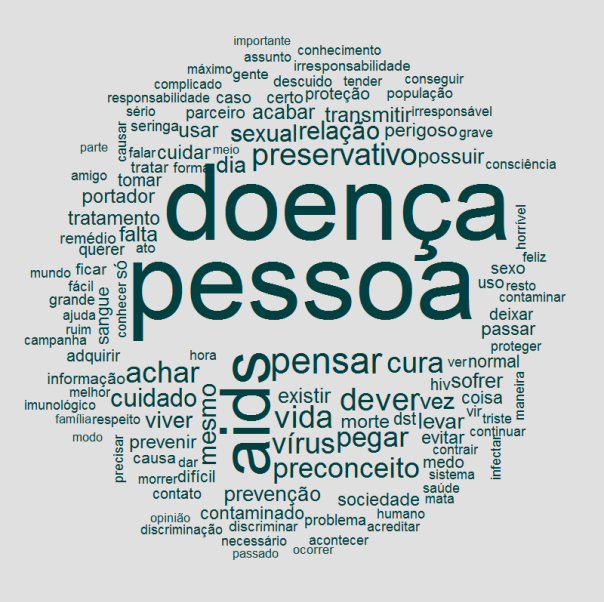
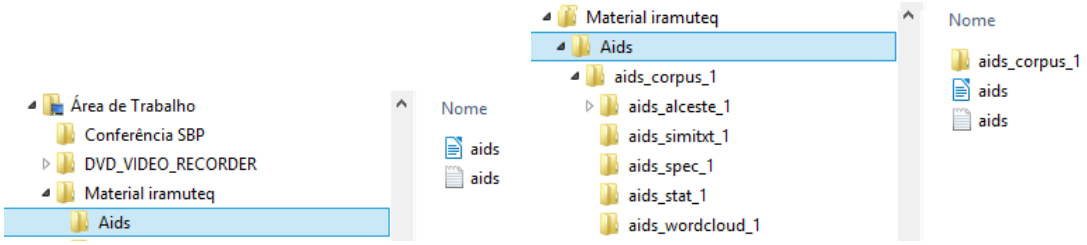


Figura 72- Nuvem de palavras do corpus "aids".

Este gráfico da nuvem de palavras automaticamente é salvo na pasta de análises, na sub-pasta "nomedocorpus_wordcloud_n", sob a forma de arquivo de imagem denominado "nuage_n".

Todos os resultados das análises, incluindo as figuras e os gráficos estarão localizados também dentro da pasta na qual foi salvo inicialmente o *corpus* de análise. Cada análise (estatísticas, especificidades, CHD, similitude e nuvem de palavras) terá uma sub-pasta com os documentos relativos à mesma.



Figuras 73 e 74- Impressão da tela com conteúdo da pasta "Aids" antes e depois das análises do software IRaMuTeQ.

A figura 73 indica que na pasta "Aids" temos apenas dois arquivos, o "aids.odt" e o "aids.txt", gerado a partir do primeiro ao empregar a opção "Salvar como": Texto – Escolha a codificação (.txt). E a figura 74, mostra que, depois das análises, além destes dois arquivos originais foi criada uma pasta "aids_corpus_1". E no seu interior foram

criadas mais 5 sub-pastas, respectivamente: “aids_alceste_1, aids_simitxt_1, aids_spec_1, aids_stat_1 e aids_wordcloud_1”.

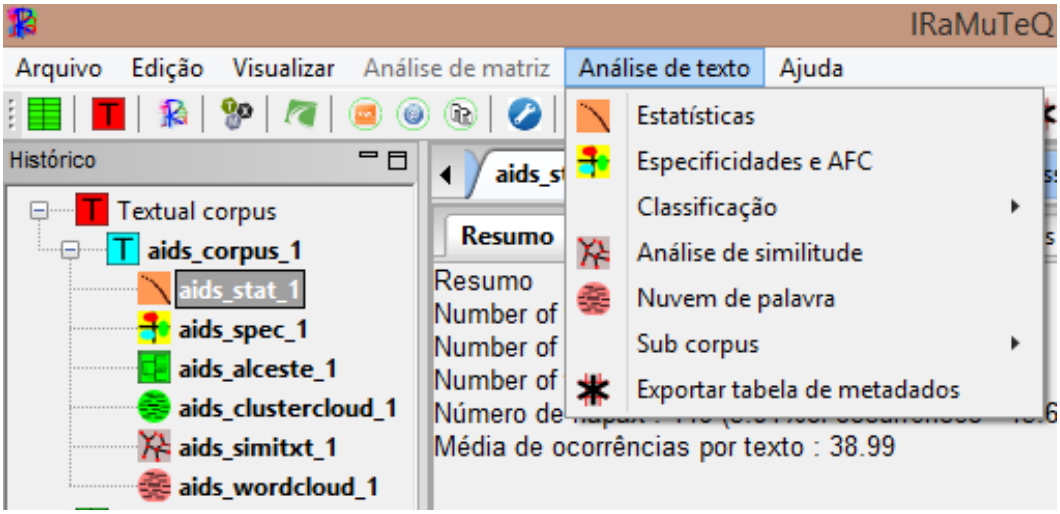
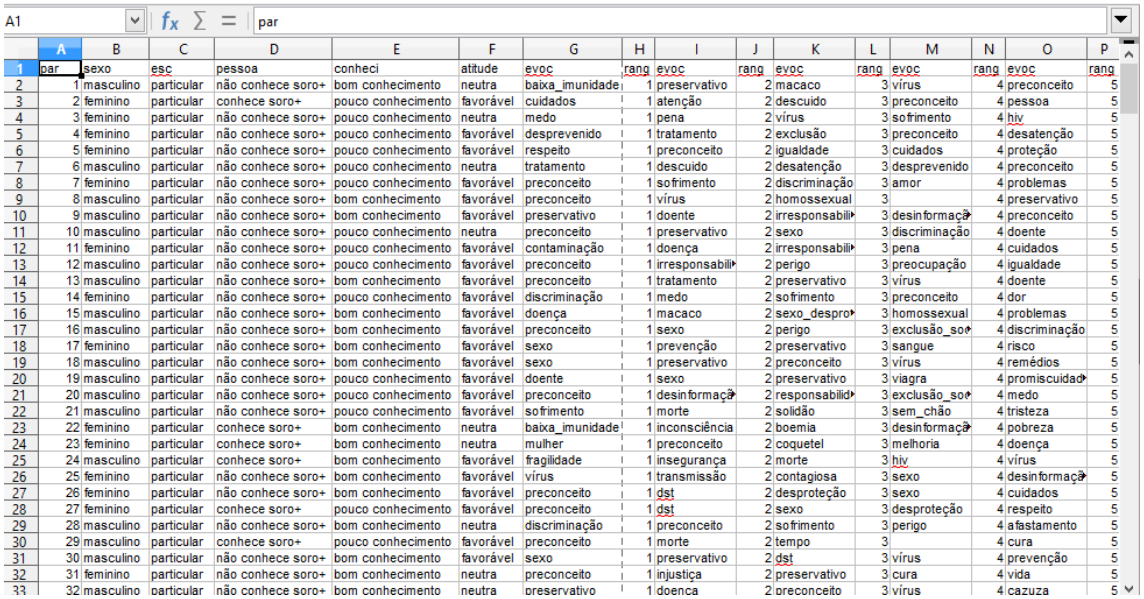


Figura 75- Correspondência entre o histórico das análises do corpus “aids” realizadas pelo IRaMuTeQ e o menu de análise de texto.

Na figura 75, pode-se observar que cada análise realizada acionando o menu “Análise de texto” corresponde a criação de sub-pastas no histórico do *software* (a direita). Estas sub-pastas contém outras pastas e arquivos gerados pelas análises. Na pasta de análise, aqui no caso a “aids_corpus_1”, criada no interior da pasta onde inicialmente colocou-se o *corpus*, é criado um arquivo denominado “Corpus.cira”. Este arquivo permite abrir todas as análises, quando escolhemos no menu principal do *software* “Arquivo” e “Abrir um corpus textual”.

Parte 2: Análise de matrizes

O IRaMuTeQ permite que se trabalhe com matrizes que envolvam variáveis categoriais e listas de palavras, tais como aquelas obtidas de tarefas de associações ou evocações livres (Sá, 1996). Nesse caso, o *software* viabiliza contagem de frequência, cálculo de qui-quadrado, análise de similitude e análise prototípica. Para isso, trabalha-se em um banco de dados montado a partir de um arquivo do Libre Office “Planilha Calc”, conforme ilustra a figura 73.



par	sexo	esc	pessoa	conheci	atitude	evoc	rang	evoc	rang	evoc	rang	evoc	rang	evoc	rang
1	masculino	particular	não conhece soro+	bom conhecimento	neutra	baixa_imunidade	1	preservativo	2	macaco	3	vírus	4	preconceito	5
2	feminino	particular	conhece soro+	pouco conhecimento	favorável	cuidados	1	atenção	2	descuido	3	preconceito	4	pessoa	5
3	feminino	particular	não conhece soro+	pouco conhecimento	neutra	medo	1	pena	2	vírus	3	sofrimento	4	hiv	5
4	feminino	particular	não conhece soro+	pouco conhecimento	favorável	desprevenido	1	tratamento	2	exclusão	3	preconceito	4	desatenção	5
5	feminino	particular	não conhece soro+	pouco conhecimento	favorável	respeito	1	preconceito	2	igualdade	3	cuidados	4	proteção	5
6	masculino	particular	não conhece soro+	pouco conhecimento	neutra	tratamento	1	descuido	2	desatenção	3	desprevenido	4	preconceito	5
7	feminino	particular	não conhece soro+	pouco conhecimento	favorável	preconceito	1	sofrimento	2	discriminação	3	amor	4	problemas	5
8	masculino	particular	não conhece soro+	bom conhecimento	favorável	preconceito	1	vírus	2	homossexual	3		4	preservativo	5
9	masculino	particular	não conhece soro+	bom conhecimento	favorável	preservativo	1	doente	2	irresponsabili	3	desinformaçã	4	preconceito	5
10	masculino	particular	não conhece soro+	pouco conhecimento	neutra	preconceito	1	preservativo	2	sexo	3	discriminação	4	doente	5
11	feminino	particular	não conhece soro+	pouco conhecimento	favorável	contaminação	1	doença	2	irresponsabili	3	pena	4	cuidados	5
12	masculino	particular	não conhece soro+	pouco conhecimento	favorável	preconceito	1	irresponsabili	2	perigo	3	preocupação	4	igualdade	5
13	masculino	particular	não conhece soro+	bom conhecimento	favorável	preconceito	1	tratamento	2	preservativo	3	vírus	4	doente	5
14	feminino	particular	não conhece soro+	pouco conhecimento	favorável	discriminação	1	medo	2	sofrimento	3	preconceito	4	dor	5
15	masculino	particular	não conhece soro+	bom conhecimento	favorável	doença	1	macaco	2	sexo_despro	3	homossexual	4	problemas	5
16	masculino	particular	não conhece soro+	bom conhecimento	favorável	preconceito	1	sexo	2	perigo	3	exclusão_so	4	discriminação	5
17	feminino	particular	não conhece soro+	bom conhecimento	favorável	sexo	1	prevenção	2	preservativo	3	sangue	4	risco	5
18	masculino	particular	não conhece soro+	bom conhecimento	favorável	sexo	1	preservativo	2	preconceito	3	vírus	4	remédios	5
19	masculino	particular	não conhece soro+	pouco conhecimento	favorável	doente	1	sexo	2	preservativo	3	viagra	4	promiscuidad	5
20	masculino	particular	não conhece soro+	pouco conhecimento	favorável	preconceito	1	desinformaçã	2	responsabilid	3	exclusão_so	4	medo	5
21	masculino	particular	não conhece soro+	pouco conhecimento	favorável	sofrimento	1	morte	2	solidão	3	sem_chão	4	tristeza	5
22	feminino	particular	conhece soro+	bom conhecimento	neutra	baixa_imunidade	1	inconsciência	2	boemia	3	desinformaçã	4	pobreza	5
23	feminino	particular	conhece soro+	bom conhecimento	neutra	mulher	1	preconceito	2	coquetel	3	melhoria	4	doença	5
24	masculino	particular	conhece soro+	bom conhecimento	favorável	fragilidade	1	insegurança	2	morte	3	hiv	4	vírus	5
25	feminino	particular	não conhece soro+	bom conhecimento	favorável	vírus	1	transmissão	2	contagiosa	3	sexo	4	desinformaçã	5
26	feminino	particular	não conhece soro+	bom conhecimento	favorável	preconceito	1	dst	2	desproteção	3	sexo	4	cuidados	5
27	feminino	particular	conhece soro+	pouco conhecimento	favorável	preconceito	1	dst	2	sexo	3	desproteção	4	respeito	5
28	masculino	particular	não conhece soro+	bom conhecimento	neutra	discriminação	1	preconceito	2	sofrimento	3	perigo	4	afastamento	5
29	masculino	particular	conhece soro+	pouco conhecimento	favorável	preconceito	1	morte	2	tempo	3		4	cura	5
30	masculino	particular	não conhece soro+	bom conhecimento	favorável	sexo	1	preservativo	2	dst	3	vírus	4	prevenção	5
31	feminino	particular	não conhece soro+	bom conhecimento	neutra	preconceito	1	injustiça	2	preservativo	3	cura	4	vida	5
32	masculino	particular	não conhece soro+	bom conhecimento	neutra	preservativo	1	doenca	2	preconceito	3	vírus	4	cazuza	5

Figura 76- Matriz ou planilha denominada “evocaids”.

Exemplo de matriz

A figura 76 ilustra uma matriz denominada “evocaids” para a exposição e exercício de análises deste tipo de material. Esta matriz será utilizada nesta parte do tutorial e se encontra no kit IRaMuTeQ (na pasta “Corpora”).

Ela tem origem numa dissertação já utilizada aqui (Antunes, 2012), é composta das respostas a outra questão, a saber: “Escreva as cinco primeiras palavras que lhe vem à cabeça quando você lê a palavra ‘aids’”. Esta tarefa de associação foi respondida por 300 estudantes do ensino médio.

As colunas da direita (da A até a F) referem-se a variáveis descritivas ou metadados. Elas são as seguintes: “par” (Indivíduo participante) com 300 modalidades ou 300 estudantes; “sexo” com duas modalidades (masculino e feminino); “esc” (Tipo de escola) com duas modalidades (pública e particular); “pessoa” com duas modalidades (não conhece e conhece pessoa soropositiva); “conheci” (Conhecimento

sobre a transmissão do HIV) com duas modalidades (bom ou pouco conhecimento); “atitude” (Atitude frente ao soropositivo) com três modalidades (favorável, neutra e desfavorável).

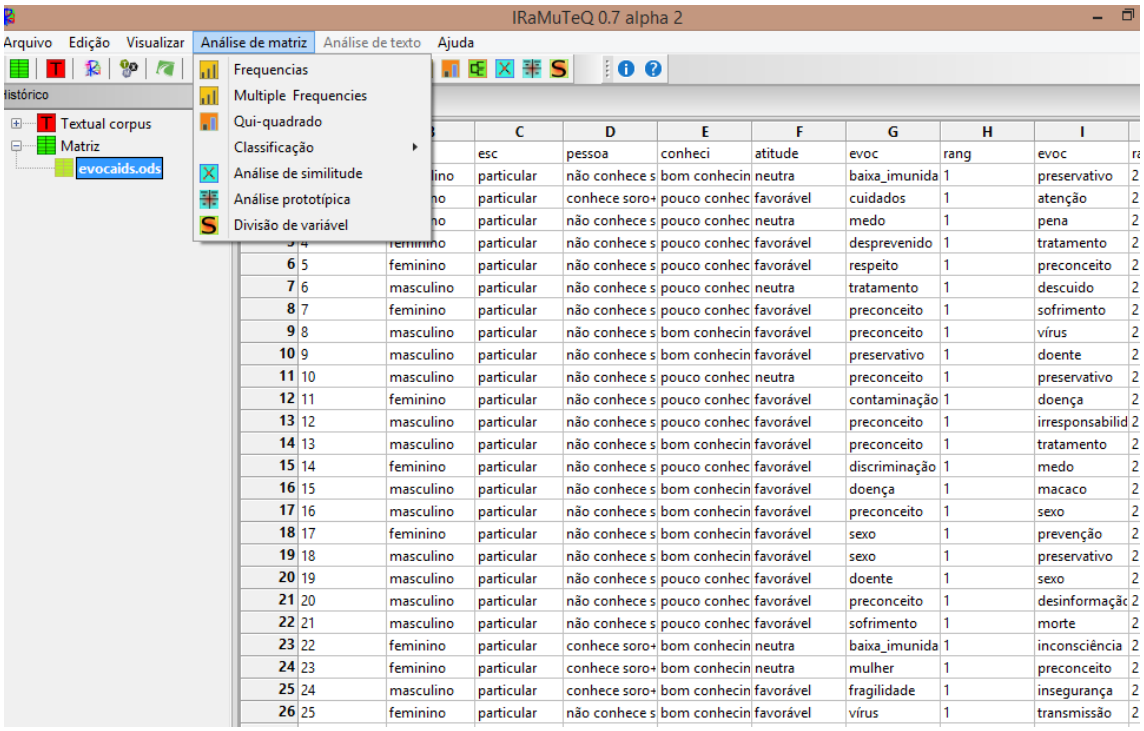
As colunas da esquerda (G a P) referem-se a cada evocação (“evoc”) seguida da respectiva ordem que foi evocada (“rang”). São 10 colunas intercalando evocações e ordem (de 1 a 5).

Aconselha-se que o banco de dados siga as seguintes indicações:

1- O tipo do arquivo de entrada seja: “.ods”; “.csv” ou “.xls” (não usar o tipo “.xlsx” ou o excel atual, pois ele é incompatível com o IRaMuTeQ). A codificação deve ser a mesma usada para as análises de texto: “UTF 8 all languages”.
2- O banco de dados não pode conter os caracteres: **: ; ‘ “.**
3- Não conter espaços no texto das células referentes às evocações (use *underline* para ligar as palavras compostas ou expressões).
4- Não conter acentos ou caracteres especiais no nome do arquivo.
5- Sugere-se que as modalidades das variáveis de caracterização sejam apresentadas sob a forma de rótulos (categorias) para facilitar a compreensão dos gráficos.
6- Caso tenha-se além da informação da ordem de aparecimento a da ordem de importância das palavras atribuída pelos participantes, esta deve ser acrescentada como variável numérica (de 1 a 5) em uma coluna logo cada coluna “rang”.
7- É necessária uma ampla revisão da matriz, uma vez que esse tipo de análise não realiza a lematização.

Após salvar o banco de dados em uma pasta exclusiva para a análise, ao abrir o *software* IRaMuTeQ, seleciona-se a aba “Arquivo”, e em seguida a opção “Abrir uma matriz”. Localiza-se o arquivo que contém o banco de dados e clica-se em “Abrir”. Para a importação dos dados, uma outra janela se abrirá e nela pode-se indicar se a primeira linha da planilha contém os nomes das colunas (opção indicada); e se a primeira coluna é um identificador (opção também indicada). A figura 77 mostra a matriz reconhecida pelo IRaMuTeQ e o menu “Análise de matriz” com suas opções.

Tipos de análise de matrizes



		C	D	E	F	G	H	I
		esc	pessoa	conheci	atitude	evoc	rang	evoc
		particular	não conhece s	bom conhecin	neutra	baixa_imunida	1	preservativo
		particular	conhece soro+	pouco conhec	favorável	cuidados	1	atenção
		particular	não conhece s	pouco conhec	neutra	medo	1	pena
		particular	não conhece s	pouco conhec	favorável	desprevenido	1	tratamento
6	5 feminino	particular	não conhece s	pouco conhec	favorável	respeito	1	preconceito
7	6 masculino	particular	não conhece s	pouco conhec	neutra	tratamento	1	descuido
8	7 feminino	particular	não conhece s	pouco conhec	favorável	preconceito	1	sofrimento
9	8 masculino	particular	não conhece s	bom conhecin	favorável	preconceito	1	vírus
10	9 masculino	particular	não conhece s	bom conhecin	favorável	preservativo	1	doente
11	10 masculino	particular	não conhece s	pouco conhec	neutra	preconceito	1	preservativo
12	11 feminino	particular	não conhece s	pouco conhec	favorável	contaminação	1	doença
13	12 masculino	particular	não conhece s	pouco conhec	favorável	preconceito	1	irresponsabilid
14	13 masculino	particular	não conhece s	bom conhecin	favorável	preconceito	1	tratamento
15	14 feminino	particular	não conhece s	pouco conhec	favorável	discriminação	1	medo
16	15 masculino	particular	não conhece s	bom conhecin	favorável	doença	1	macaco
17	16 masculino	particular	não conhece s	bom conhecin	favorável	preconceito	1	sexo
18	17 feminino	particular	não conhece s	bom conhecin	favorável	sexo	1	prevenção
19	18 masculino	particular	não conhece s	bom conhecin	favorável	sexo	1	preservativo
20	19 masculino	particular	não conhece s	pouco conhec	favorável	doente	1	sexo
21	20 masculino	particular	não conhece s	pouco conhec	favorável	preconceito	1	desinformação
22	21 masculino	particular	não conhece s	pouco conhec	favorável	sofrimento	1	morte
23	22 feminino	particular	conhece soro+	bom conhecin	neutra	baixa_imunida	1	inconsciência
24	23 feminino	particular	conhece soro+	bom conhecin	neutra	mulher	1	preconceito
25	24 masculino	particular	conhece soro+	bom conhecin	favorável	fragilidade	1	insegurança
26	25 feminino	particular	não conhece s	bom conhecin	favorável	vírus	1	transmissão

Figura 77- Importação do banco de dados do corpus "evocaids" para "Análise de matriz" e menu de análises.

As análises possíveis de serem realizadas com os bancos de dados do tipo matriz envolvem cálculos de frequências, qui-quadrado, classificação hierárquica descendente (aconselhada apenas nos casos em que o número de participantes é bastante alto), análise de similitude e análise prototípica.

Para processar as análises, basta clicar na opção "Análise de matriz" e em seguida selecionar a análise desejada (figura 77).

1- Frequências: fornece a distribuição, em frequência absoluta e relativa, por variável descritiva ou de cada palavra evocada.
2- Múltiplas frequências: fornece a distribuição, em frequência absoluta e relativa, de diversas variáveis descritivas ou de uma lista de palavras evocadas.
3- Qui-quadrado: oferece o valor do qui-quadrado e o nível de significância entre variáveis, entre palavras (formas), ou ainda entre variável e forma.
4- Classificação: realiza a classificação hierárquica descendente das palavras evocadas (limitando-se as três primeiras) e sua relação com uma das variáveis descritivas.
5- Análise de similitude: fornece árvores (grafos) de associação entre as palavras evocadas e suas relações com uma ou mais variáveis descritivas (metadados).

6- Análise prototípica: proporciona a criação de um diagrama de quatro casas para o estudo da centralidade ou não das palavras evocadas.
7- Divisão de variável: separa a matriz original em sub-matrizes em função das modalidades das variáveis descritivas (metadados), colocando-as em pastas denominadas “nomedamodalidade_matrix_n” dentro da sub-pasta de análise “nomedoarquivo_matrix_n”.

A análise de frequências é mais simples. Ela é indicada para acessar as frequências das variáveis categoriais da matriz e a “Análise de frequências múltiplas” (*Multiple Frequencies*) para se obter um relatório de frequência absoluta e relativa das palavras presentes na matriz.

Ao selecionar a análise desejada é necessário escolher sobre quais variáveis serão processados os cálculos. A figura 78 indica a escolha de todas as 5 evocações para gerar um relatório de “Frequências múltiplas”. Nesse caso, não há interesse no “rang” (ordem de evocação) mas apenas nas palavras evocadas.

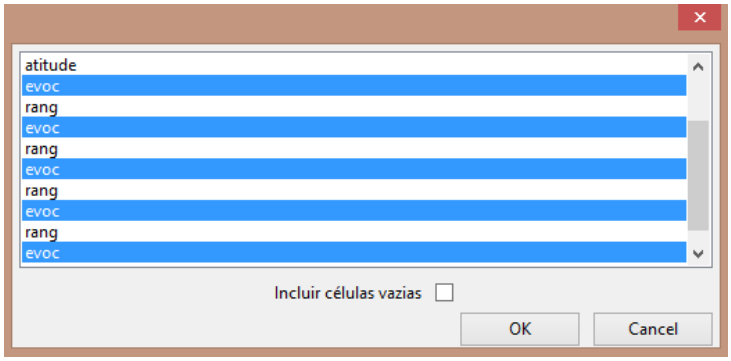
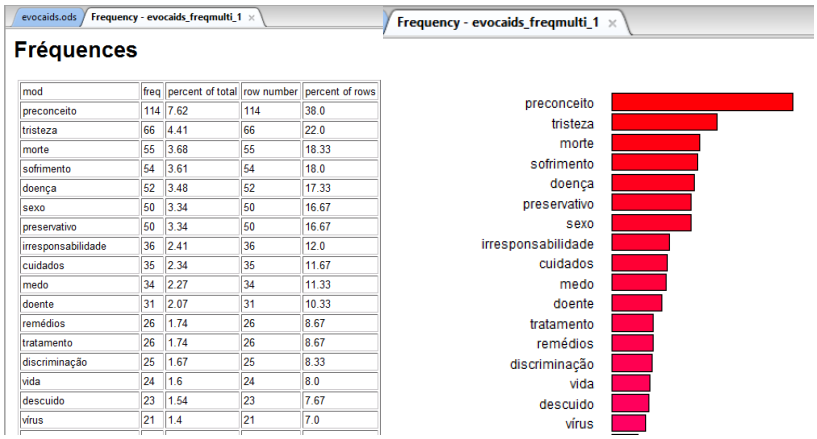


Figura 78- Seleção das evocações (“evoc”) do corpus “evocaids” para cálculo das frequências múltiplas.

A figura 79 ilustra um relatório das frequências múltiplas relativas às palavras evocadas no teste de associação livre que se está utilizando aqui.



mod	freq	percent of total	row number	percent of rows
preconceito	114	7.62	114	38.0
tristeza	66	4.41	66	22.0
morte	55	3.68	55	18.33
sofrimento	54	3.61	54	18.0
doença	52	3.48	52	17.33
sexo	50	3.34	50	16.67
preservativo	50	3.34	50	16.67
irresponsabilidade	36	2.41	36	12.0
cuidados	35	2.34	35	11.67
medo	34	2.27	34	11.33
doente	31	2.07	31	10.33
remédios	26	1.74	26	8.67
tratamento	26	1.74	26	8.67
discriminação	25	1.67	25	8.33
vida	24	1.6	24	8.0
descuido	23	1.54	23	7.67
virus	21	1.4	21	7.0

Figura 79- Fragmento inicial do resultado da análise de frequências múltiplas do corpus "evocaids".

Conforme se observa na figura 79, a análise fornece uma tabela com as palavras ordenadas por sua frequência, na segunda coluna suas percentagens em relação ao total de evocações, o número de linhas que contém cada uma destas palavras, bem como suas proporções em relação ao número total de linhas. Lembrado, cada linha representa um participante respondente.

Análise de similitude

Outra possibilidade é a análise de similitude, indicadora da estrutura do conjunto das palavras evocadas. O processamento da análise para matrizes se dá de modo análogo ao realizado para os corpora textuais; mas há uma maneira diferente de fazer esta análise em função de variáveis descritivas (ou metadados). Conforme ilustra a figura 80, numa primeira janela seleciona uma ou mais variáveis de caracterização (no caso: "sexo, esc, pessoa, conheci e ati") e as evocações ou o material textual (as "evoc").

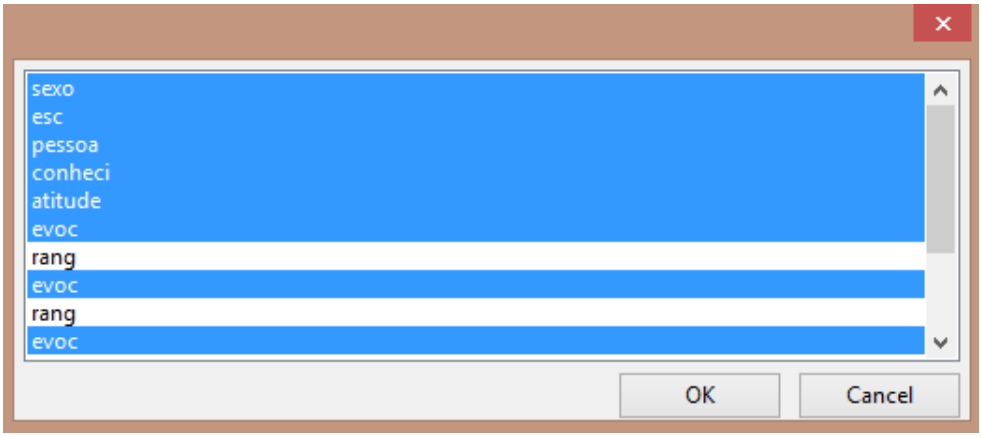


Figura 80- Seleção das variáveis descritivas (metadados) e das evocações ("evoc") do corpus "evocaids" para análise de similitude.

Após clicar em “OK” aparece a figura 81. Num primeiro momento, escolhe-se as evocações que aparecerão no gráfico. Para clareza da representação gráfica sugere-se considerar em torno de 1 / 4 das palavras ou formas com maior frequência. Aqui considerou-se as palavras com frequência igual ou superior a 10.

Do lado direito da figura 81, na aba “Configurações gráficas” assinalou-se a opção “Escores nas bordas”, colocou-se em branco a opção “Edge curved” (aresta curva), as opções “Comunidades” e “halo” não foram selecionadas neste caso. Por fim, alterou-se o “Tamanho do texto” de 10 para 8.

Na aba “Ajustes gráficos”, que não está representada na figura, alterou-se a largura do “Tamanho da imagem” de 800 para 1200 pixels; no “Tamanho do vértice proporcional a frequência”, com a opção “eff.” (efetivo) assinalada alterou-se o valor mínimo de 5 para 3 e o máximo de 30 para 12; na opção “Texto do vértice proporcional a frequência” assinalou-se a opção “chi2” (qui-quadrado) e alterou-se o mínimo de 8 para 5 e o máximo de 25 para 15; a “Cor do vértice” escolhida foi um tom de verde.

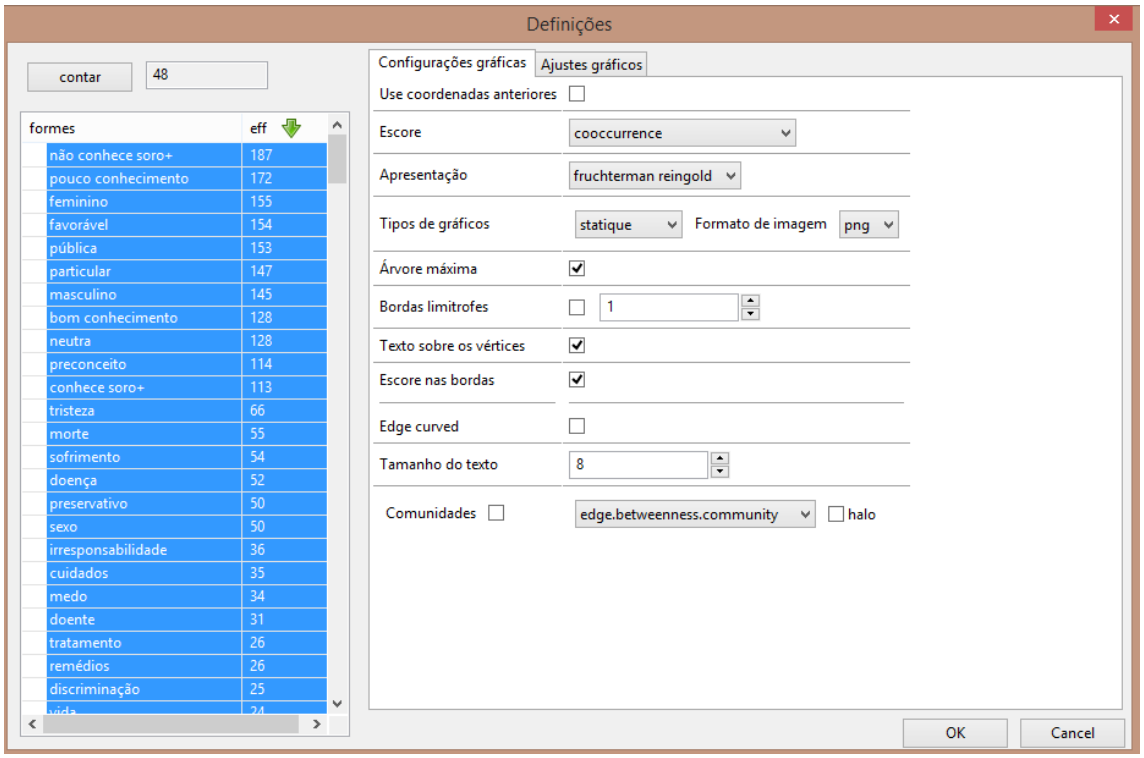


Figura 81- Escolha das formas e “Configurações gráficas” para a arvore máxima de similitude do corpus “evocaids”

O gráfico resultante da análise de similitude está indicado na figura 82, onde o tamanho dos vértices (círculos verdes) é proporcional à frequência das modalidades das variáveis descritivas (metadados) e das palavras evocadas. E as arestas indicam os valores da associação entre as modalidades e as palavras. No caso utilizou-se as como

indicador de associação as frequências de coocorrências. Por exemplo: há 77 coocorrências entre a modalidade “não conhece uma pessoa soropositiva” e a evocação da palavra “preconceito”.

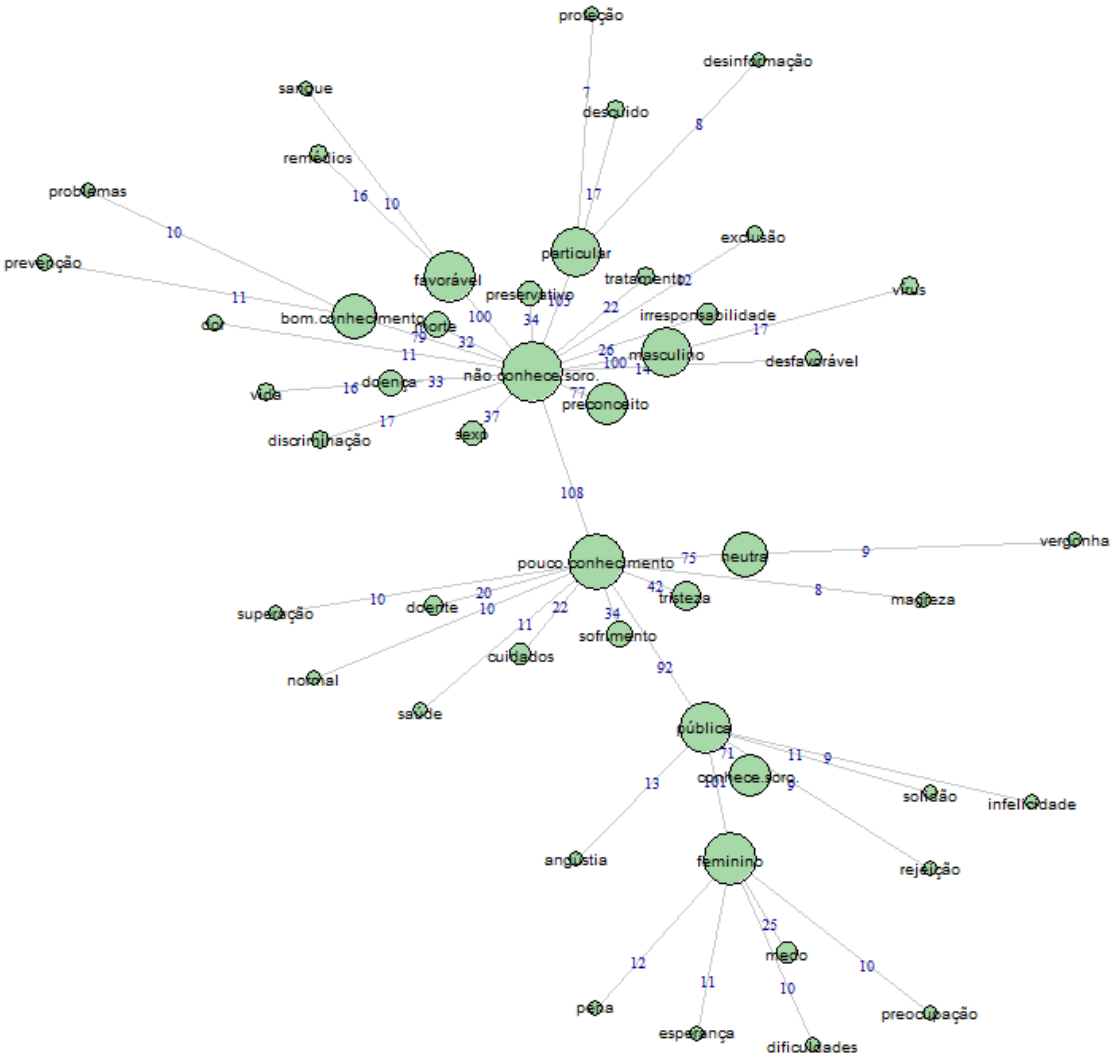


Figura 82- Árvore máxima de similitude do corpus “evocaids” com configurações.

Análise prototípica

A análise prototípica é uma técnica simples e eficaz desenvolvida especificamente pelo campo de estudo de representações sociais (Sá, 1996). Visa identificar a estrutura representacional a partir dos critérios de frequência e ordem de evocação das palavras, provenientes de tarefas de associações ou evocações livres (Wachelke & Wolter, 2011). A mesma pode ser realizada com o *software* IRaMuTeQ a partir do menu “Análise de matriz e da opção “Análise prototípica”.

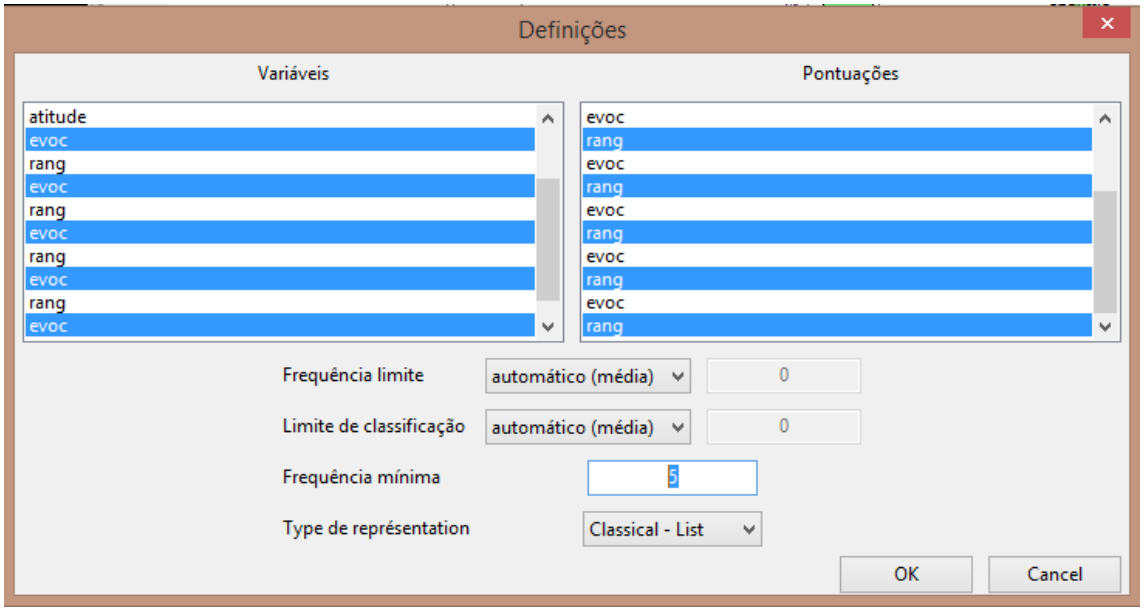


Figura 83- Seleção das evocações ("evoc") e de suas ordens ("rang") do corpus "evocaids" para análise prototípica.

Ao abrir a janela de definições (figura 83) deve-se selecionar (com um clique simples) na parte esquerda as variáveis correspondentes às evocações e na parte direita as variáveis correspondentes ao RANG (seja ele a ordem de evocação ou de importância atribuída, à escolha segundo os critérios do pesquisador). Os demais parâmetros referem-se aos critérios de cálculo da análise prototípica e podem ser mantidos os padrões automáticos, com exceção da "Frequência mínima" sugerindo-se alterar de 2 para no mínimo 5. Nos "Tipos de representação" há duas outras opções além da "Clássica- Lista" (*Classical- List*) ou do diagrama de quatro quadrantes, a saber: "Clássica- Nuvem" e a interessante "Plano" (*Plan*).

Definidos os padrões, clique em OK e em alguns segundos será apresentado o produto da análise prototípica (figura 84). Este diagrama de quatro quadrantes representa quatro tipos de elementos da representação social sua dimensão estrutural.

No exemplo em questão, trata-se de uma tarefa de evocação livre com termo indutor "Aids". O primeiro quadrante (superior esquerdo) indica as palavras que têm alta frequência (uma frequência maior que a média) e baixa ordem de evocação (aquelas que foram mais prontamente evocadas). Essas seriam as prováveis indicadoras do núcleo central de uma representação. Mas só é possível determinar se efetivamente são elementos do núcleo central por meio de outras técnicas, que podem até envolver uma nova etapa da pesquisa ou uma nova pesquisa.

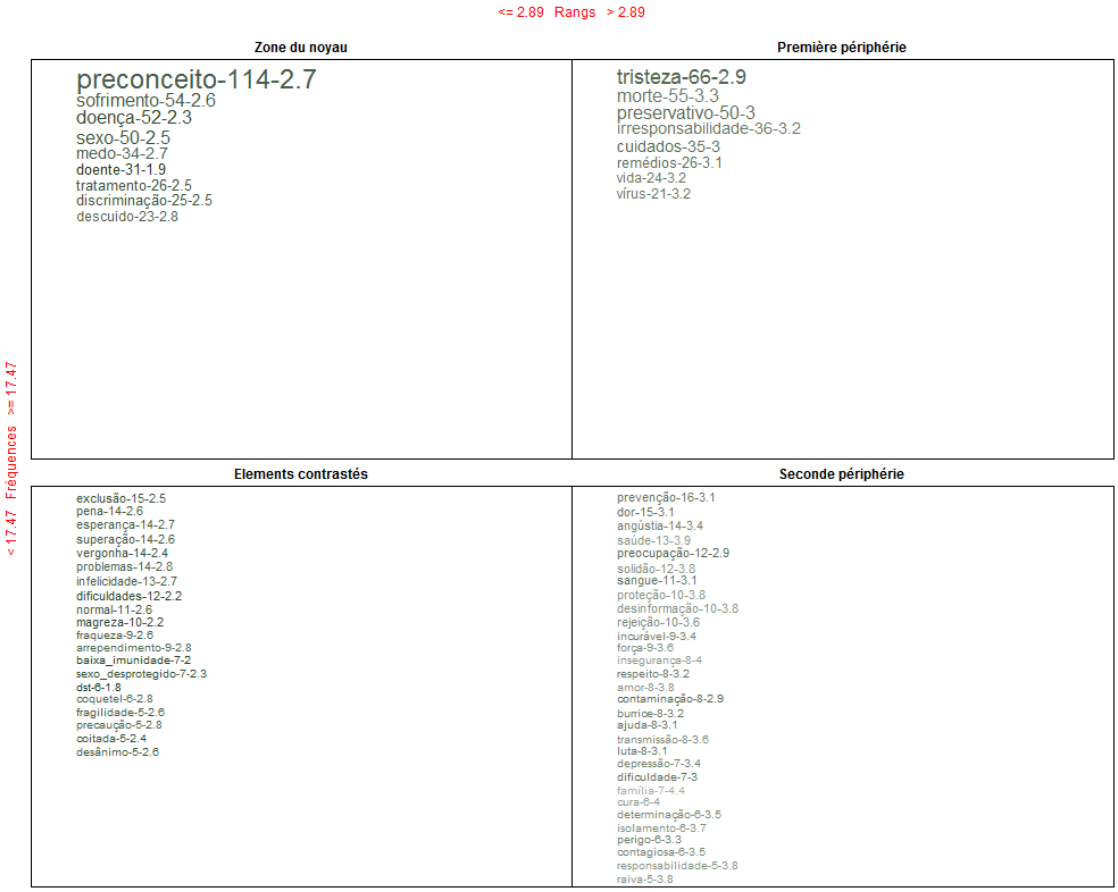


Figura 84- Diagrama dos quatro quadrantes da "Análise prototípica" do corpus "evocaids".

No segundo quadrante (superior direito), temos a primeira periferia, com as palavras que têm alta frequência, mas que tiveram ordem média maior, ou seja, não foram tão prontamente evocadas. No terceiro quadrante (inferior esquerdo), a zona de contraste contém elementos que foram prontamente evocados, porém com frequência abaixo da média. Por fim, a segunda periferia no quarto quadrante (inferior direito) indica os elementos com menor frequência e maior ordem de evocação.

Referências

Antunes, L. (2012). 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.

Antunes, L. (2017). Representações sociais da hipertensão arterial e do tratamento para profissionais de saúde, pessoas que vivem com hipertensão e seus familiares. Tese de Doutorado em Psicologia (não publicada). Programa de pós-graduação em Psicologia. Universidade Federal de Santa Catarina. Florianópolis, SC.

Aquino, J. A. (2014). Livro R para cientistas. Ilhéus: Editora da UESC.

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.

Cibois. P. (1983). Méthodes post-factorielles pour le dépouillement d'ênquete. Bul. Methodo. Socio. (1), 41-78.

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.

Degenne, A.; Vergès, P. (1973). Introduction à l'analyse de similitude. *Revue Française de Sociologie*. 14, 513-528.

Flament, C. (1981). L'analyse de similitude: Une technique pour les recherches sur les representations sociales. *Cahiers de Psychologie Cognitive.* 1, 375-395.

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.

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

Loubère, L. & Ratinaud, P. (2014). Documentation IraMuTeQ 0.6 alpha 3 - version 0.1 *[Computer software].* Recuperado em 19 fevereiro de 2014, de http://www.iramuteq.org

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

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.

Reinert, M. (1995). *Quelques aspects du choix des unités d'analyse et leur contrôle dans la méthode ALCESTE.* Manuscrito não publicado.

Sá, C. P. (1996). Núcleo central das representações sociais. Petrópolis: Vozes.

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.