Tutorial para uso do software de análise textual IRAMUTEQ

Brigido Vizeu Camargo e Ana Maria Justo
Laboratório de Psicologia Social da Comunicação e Cognição – LACCOS
Universidade Federal de Santa Catarina, Brasil (2013).

O presente tutorial tem o objetivo de oferecer ao usuário de língua portuguesa as principais indicações para o uso do IRAMUTEQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires). Ele não é completo, em vista da grande gama de procedimentos envolvida na aplicação deste *software*. O foco deste tutorial é na análise de *corpus* textual.

O IRAMUTEQ é um software gratuito e com fonte aberta, desenvolvido por Pierre Ratinaud (Lahlou, 2012; Ratinaud & Marchand, 2012) e licenciado por GNU GPL (v2), que permite fazer análises estatísticas sobre corpus textuais e sobre tabelas indivíduos/palavras. Ele ancora-se no software R (www.r-project.org) e na linguagem Python (www.python.org).

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 o IRAMUTEQ se utilizará do software R para processar suas análises.

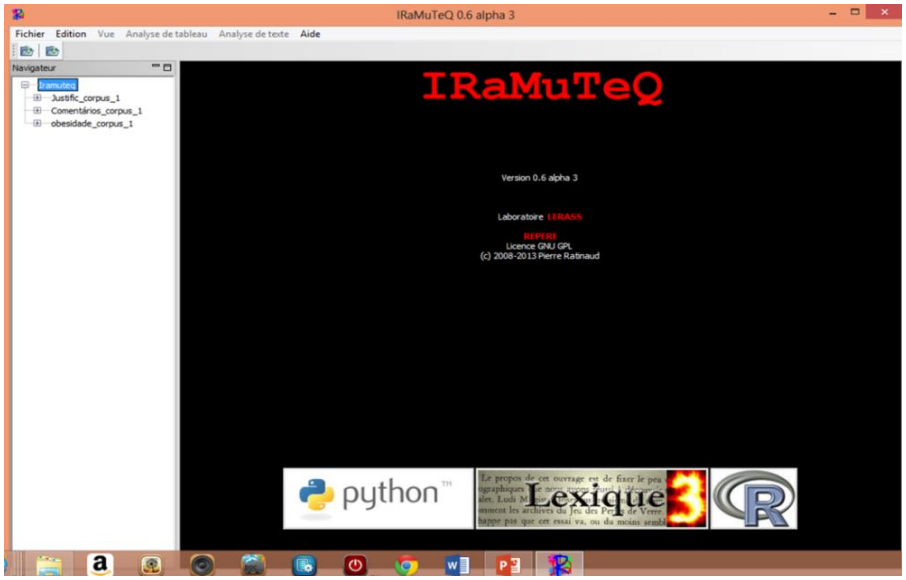

Fig. 1 Interface inicial do software IRAMUTEQ

Tipos de análises possíveis com o IRAMUTEQ

Análises sobre corpus textuais:

1) Estatísticas textuais clássicas.
2) Pesquisa de especificidades a partir de segmentação definida do texto (análise de contraste de modalidades de variáveis).
3) Classificação Hierárquica Descendente (CHD) conforme o método descrito por Reinert (1987 e 1990).
4) Análise de similitude de palavras presentes no texto.
5) Nuvem de palavras.

Análises sobre tabelas indivíduos / palavras:

1) CHD conforme algorítimo proposto por Reinert (1987).
2) CHD por matrizes de distância.
3) Análise de similitude (por exemplo, de palavras resultantes de evocações livres).
4) Nuvem de palavras.
5) Descrição e χ^2.

ANÁLISES SOBRE CORPUS TEXTUAIS

A análise textual é um tipo específico de análise de dados, na qual tratamos de material verbal transcrito, ou seja, de textos (Nascimento-Schulze & Camargo, 2000). Essa análise tem várias finalidades, sendo possível analisar textos, entrevistas, documentos, redações etc. A partir da análise textual é possível descrever um material produzido por um produtor, seja individual ou coletivamente, como também pode-se utilizar a análise textual com a finalidade 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 explicitar alguns conceitos importantes:

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

Corpus

O *corpus* é construído pelo pesquisador. É o conjunto texto que se pretende analisar. Por exemplo, numa pesquisa documental se um pesquisador decide analisar os

artigos que saíram na sessão de saúde de um jornal, em um determinado período temporal, o corpus seria o conjunto destes artigos. Outro exemplo seria um conjunto de 40 transcrições de entrevistas não diretivas sobre um tema, feitas por um pesquisador no âmbito de um estudo de casos. E ainda podemos ter, por exemplo, um *corpus* composto de 200 respostas a uma questão aberta, que faz parte de um questionário empregado como instrumento de uma pesquisa do tipo enquete.

Textos

Como já vimos nos exemplos relativos a um *corpus*, a definição destas unidades é feita pelo pesquisador e depende da natureza da pesquisa. Se a análise vai ser 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 artigos de jornais, atas de reuniões, cartas, etc.; cada exemplar destes documentos será um texto.

Um conjunto de textos constitui um *corpus* de análise. O *corpus* adequado à análise do tipo Classificação Hierárquica Descendente deve constituir-se num conjunto textual centrado em um tema. O material textual deve ser monotemático, pois a análise de textos sobre vários itens previamente estruturados ou diversos temas resulta na reprodução da estruturação prévia dos mesmos (Camargo, 2005).

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, sugerem-se pelo menos 20 textos para cada grupo.

Em se tratando de respostas a questões abertas de um questionário, cada texto será composto da adição dos trechos obtidos das respostas somente quando elas se referirem a um mesmo tema (uma mesma questão ou pergunta). Caso as questões referiram-se 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 à cinco linhas, é necessário um número bem maior de respostas para a constituição de um corpus de análise (refere-se aqui a um número mínimo em torno de uma centena de textos).

Os textos são separados por linhas de comando também chamadas de "linhas com asteriscos". No caso de entrevistas, por exemplo, como cada uma delas é um texto, elas 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 textos coletados. É desejável certo balanceamento das modalidades das variáveis da linha de comando, e parcimônia quanto ao número de variáveis utilizadas.

Segmentos de Texto

São excertos de texto, na maior parte das vezes, do tamanho de três linhas, dimensionadas pelo próprio *software* em função do tamanho do *corpus*. Os segmentos de textos que são considerados o ambiente das palavras. Seu tamanho também pode ser configurado pelo pesquisador. Numa análise padrão, após reconhecer as indicações dos textos a serem analisados, é o *software* IRAMUTEQ que divide os textos do *corpus* em segmentos de texto.

Como o pesquisador pode configurar a divisão dos segmentos de texto, no caso de uma grande quantidade de respostas curtas a uma pergunta aberta de um questionário, aconselha-se que os segmentos de texto sejam definidos enquanto textos, ou seja, enquanto a resposta dada à questão. Neste caso configura-se o software a não segmentar os textos componentes do corpus.

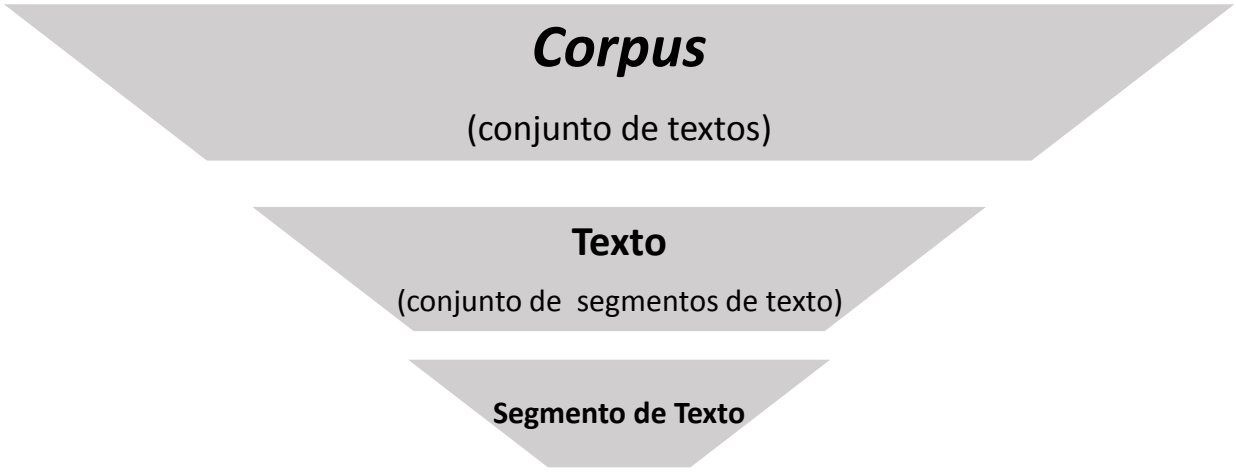

Figura 2: Noções de *corpus*, texto, segmento de texto

POSSIBILIDADES DE ANÁLISE DE DADOS TEXTUAIS NO IRAMUTEQ

O IRAMUTEQ oferece a possibilidade de diferentes formas de análise de dados textuais, desde aquelas bem simples, como a lexicografia básica (como cálculo

de frequência de palavras), até análises multivariadas (classificação hierárquica descendente, análise pós-fatorial) (Lebart & Salem, 1994; Doise, Clemence & Lorenzi-Cioldi, 1992).

I) Análises lexicográficas clássicas – Identifica e reformata as unidades de texto, identifica a quantidade de palavras, frequência média e *hapax* (palavras com frequência um), pesquisa o vocabulário e reduz das palavras com base em suas raízes (formas reduzidas), cria do dicionário de formas reduzidas, identifica formas ativas e suplementares.

II) Especificidades – Associa textos com variáveis, ou seja, possibilita a análise da produção textual em função das variáveis de caracterização. É possível modelo de análise de contrastes das modalidades das variáveis e também a apresentação em plano fatorial.

III) Método da Classificação Hierárquica Descendente (CHD) – Os segmentos de texto são classificados em função dos seus respectivos vocabulários, e o conjunto deles é repartido em função da frequência das formas reduzidas. A partir de matrizes cruzando segmentos de textos e palavras (em repetidos testes do tipo χ^2), aplica-se o método de CHD e obtém-se uma classificação estável e definitiva (Reinert,1990). Esta análise visa obter **classes** de segmentos de texto que, ao mesmo tempo, apresentam vocabulário semelhante entre si, e vocabulário diferente dos segmentos de texto das outras classes (Camargo, 2005). A partir dessas análises em matrizes o *software* organiza a análise dos dados em um **dendograma** da CHD, que ilustra as relações entre as classes. O programa executa cálculos e fornece resultados que nos permite a descrição de cada uma das classes, principalmente, pelo seu vocabulário característico (léxico) e pelas suas palavras com asterisco (variáveis). Além disto, o programa 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 programa calcula e fornece-nos os segmentos de texto mais característicos de cada classe (corpus em cor) permitindo a contextualização do vocabulário típico de cada classe. O que são estas classes de palavras e de segmentos de texto? Em nível do programa informático, cada classe é composta de vários segmentos de texto em função de uma classificação segundo a distribuição do vocabulário (formas) destes segmentos de texto. Em nível interpretativo Reinert (1990), ao estudar a literatura, utilizou a noção de "mundo", enquanto um quadro perceptivo-cognitivo com certa estabilidade temporal associado a um ambiente complexo. Em pesquisas no campo da linguística e comunicação estas classes são

interpretadas como campos lexicais (Cros, 1993) ou contextos semânticos. Em pesquisas sobre representações sociais, tendo em vista o estatuto que elas conferem às manifestações linguísticas, estas classes podem indicar teorias ou conhecimentos do senso comum ou campos de imagens sobre um dado objeto, ou ainda apenas aspectos de uma mesma representação (Veloz, Nascimento-Schulze e Camargo, 1999).

IV) Análise de similitude – Esse tipo de análise baseia-se na teoria dos grafos (Marchand & Ratinaud, 2012) e é utilizada frequentemente por pesquisadores das representações sociais (cognição social). Possibilita identificar as coocorrências entre as palavras e seu resultado traz indicações da conexidade entre as palavras, auxiliando na identificação da estrutura da representação.

V) Nuvem de palavras – Agrupa as palavras e as organiza graficamente em função da sua frequência. É uma análise lexical mais simples, porém graficamente interessante.

COMO ANALISAR DADOS TEXTUAIS NO IRAMUTEQ

Para realizar a análise o primeiro passo é configurar o *corpus* a ser analisado. Isso 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* OpenOffice.org** (http://www.openoffice.org/) **ou LibreOffice** (http://pt-br.libreoffice.org/). 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 IRAMUTEQ.

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.

Exemplo de uma linha de comando (com asteriscos):

**** *n_014 *sex_1 *posic_1 *cur_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 foi extraída de uma pesquisa realizada em comentários na internet referentes a um ensaio fotográfico com mulheres obesas. Ela indica que o material textual que a segue (comentários em determinado *site*) refere-se ao indivíduo nº 014 (utiliza-se três dígitos, pois a amostra tem mais de 99

indivíduos e menos de 1000), de sexo masculino (onde 1= masculino; 2= feminino), com posicionamento favorável em relação ao ensaio fotográfico (onde 1= favorável; 2= contra; 3=neutro); e cujo comentário teve entre 11 e 50 “curtidas” (onde 1= até 10; 2= 11 a 50; 3= mais de 50). Imediatamente após esta linha com asterisco teclar ENTER, e sem tabulação e linha em branco digite ou coloque o texto correspondente a este indivíduo.

3- Corrigir e revisar todo o arquivo, para que os erros de digitação ou outros não sejam tratados como palavras diferentes.

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

5- 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.

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

7- É desejável certa uniformidade em relação às siglas, ou as usa sempre ou coloque tudo por extenso unido por traço *underline.* Por exemplo: ou oms ou organização_mundial _de_saúde.

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

9- 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”.

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

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

12- O arquivo com o corpus preparado no *software* OpenOffice.org ou no LibreOffice deve ser salvo em uma nova pasta criada no desktop, somente para a análise, com um nome curto, como texto codificado (nome_do_arquivo.txt). No OpenOffice.org esta opção abre uma primeira janela e devemos escolher “manter formato atual”, e uma segunda janela onde as opções “Conjuntos de caracteres” e “Quebra de parágrafo” devem ser respectivamente “Unicode (UTF-8)” e “LF”.

Exemplo de extrato de um corpus

**** *n_014 *sex_1 *posic_1 *cur_2

Achei interessante o trabalho dele, pois muitas pessoas geralmente não estão satisfeitas com o corpo e acabam esquecendo a sensualidade, achando que ninguém lhe acha atraentes. Essas meninas deram seu melhor dentro das limitações delas e ficou ótimo! Amei. Parabéns ao artista e as modelos.

**** *n_016 *sex_1 *posic_1 *cur_2

Ainda bem que há pessoas que nadam contra a maré da nossa cultura de massas e nos proporciona uma visão mais abrangente do espaço e das pessoas que habitam ao nosso redor. Uma bela iniciativa do fotógrafo e uma linda lição de autoestima das modelos. CONTINUA /.../

OBS: *Após preparar o corpus, recomenda-se que se leia o mesmo atentamente, especialmente no que se refere às linhas de comando. O IRAMUTEQ não possui ferramenta para verificação e correção do corpus. Essa verificação precisa ser realizada pelo pesquisador antes de lançar o procedimento de análise dos dados.*

PROCESSANDO A ANÁLISE NO SOFTWARE IRAMUTEQ

Abra o programa para trabalhar em sua interface, e importe o *corpus.* Na barra de ferramentas superior clique em ARQUIVO (*Fichier*) e ABRIR UM CORPUS (*Ouvrir un corpus*), conforme indica a Figura 3. Selecione o *corpus* que deseja analisar e clique em abrir (*Ouvrir).*

Fig. 3 Importação do *corpus* de análise.

No momento em que o *software* importar o *corpus*, uma nova janela será aberta:

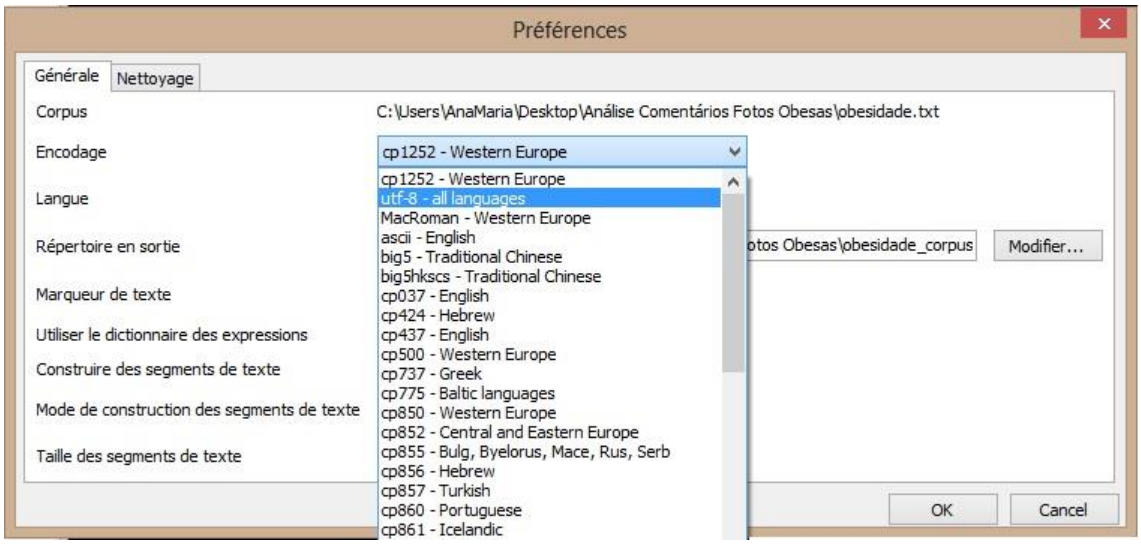

Fig. 4. Configurações de análise.

Nessa janela (Figura 4) podem ser observadas algumas configurações do software para analisar os dados textuais. A maior parte das configurações, na aba Générale, pode ser mantida conforme o padrão, com exceção de duas que precisam ser modificadas. A primeira refere-se a codificação (Encodage) do texto, que deve ser a segunda opção de cima para baixo: “uft-8 – all languages”.

A outra configuração é a da língua (Langue). Conforme a Figura 5, selecione a língua: portuguese (expérimentale) no caso do texto estar nesta língua, ou escolha a língua correspondente ao caso (francês, inglês e italiano). Atualmente trabalha-se para aprimorar o dicionário da língua portuguesa durante este ano de 2013, ele ainda é experimental como os dicionários de outras línguas (alemã, sueca, espanhola e grega).

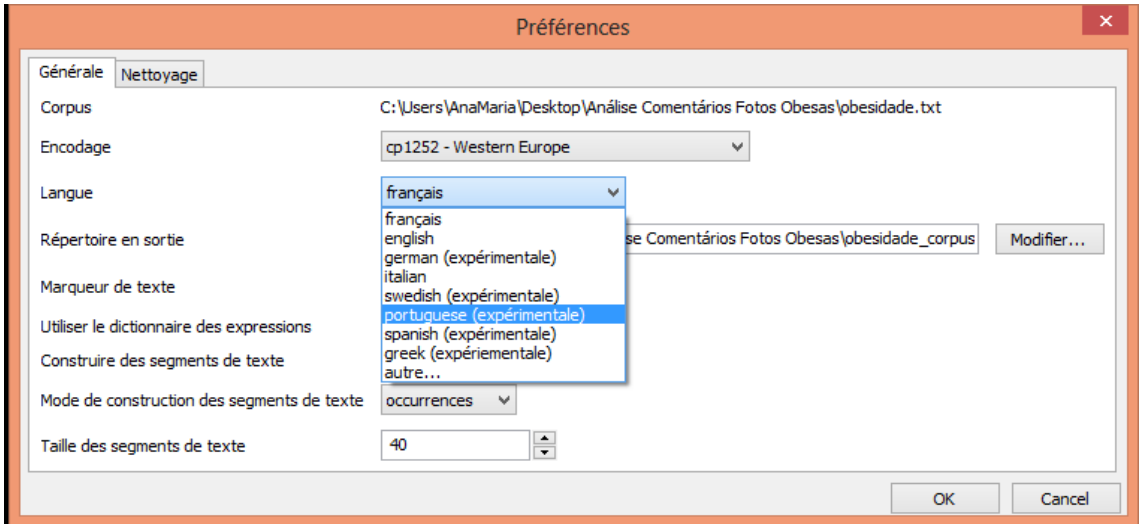

Fig. 5. Configurações de análise.

.

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 6, onde se pode verificar, o número de Textos e de Segmentos de texto, Formas identificadas, Ocorrências, e Frequência de *Hapax*.

Fig. 6 – Resultados preliminares, descrição do corpus.

Tendo sido realizada a importação do corpus, as análises já podem ser iniciadas. Para realiza-las, na barra de ferramentas superior, selecione ANÁLISE DO TEXTO (*Analyse de texte),* e aparecerão as possibilidades de análise (Figura 7).

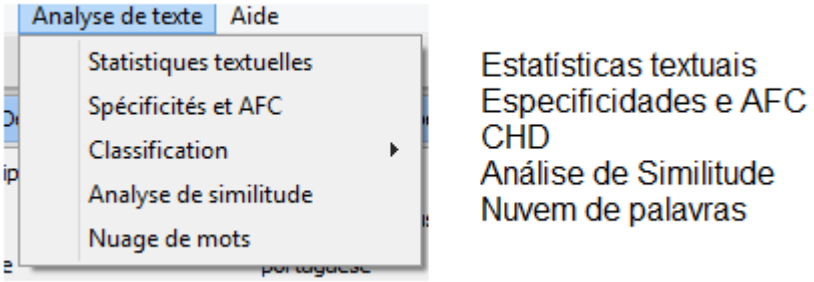

Fig. 7. Escolha da análise

Toda a vez que for escolhida uma análise, surgirá uma nova janela perguntando se você deseja manter a Lematização (*Lematisation*). Deixe selecionado SIM (*OUI*), 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 *Preferences.* É indicado que o pesquisador selecione quais as classes gramaticais 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.

ESTATÍSTICAS TEXTUAIS

Na primeira opção de análise, **Estatísticas textuais**, 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 8), que apresenta o comportamento das frequências das palavras no *corpus*, num gráfico que ilustra a distribuição de frequência X rang.

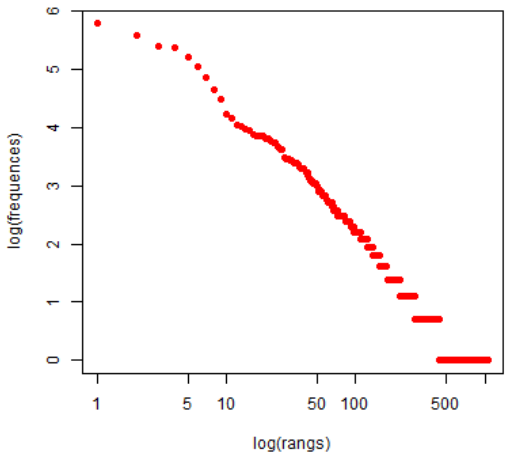

Fig. 8. Diagrama de Zipf

Na coluna que se apresenta à esquerda, na interface do *software*, você identifica essa análise como: **NOME DO CORPUS_stat_1**. 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 de formas reduzidas (*exporter le dicttionaire*), o qual será salvo na pasta em que foi salvo o corpus inicial, dentro de uma subpasta denominada: **NOME DO CORPUS_stat_1.**

ESPECIFICIDADES E AFC

Ao selecionar o modo **Especificidades e AFC**, você deverá escolher a variável categorial em função da qual deseja realizar a análise. Selecione-a na janela que aparece na interface e clique em *Ok*. Aguarde alguns instantes e os resultados aparecerão na janela principal. A identificação dos resultados encontra-se descrito na figura 9.

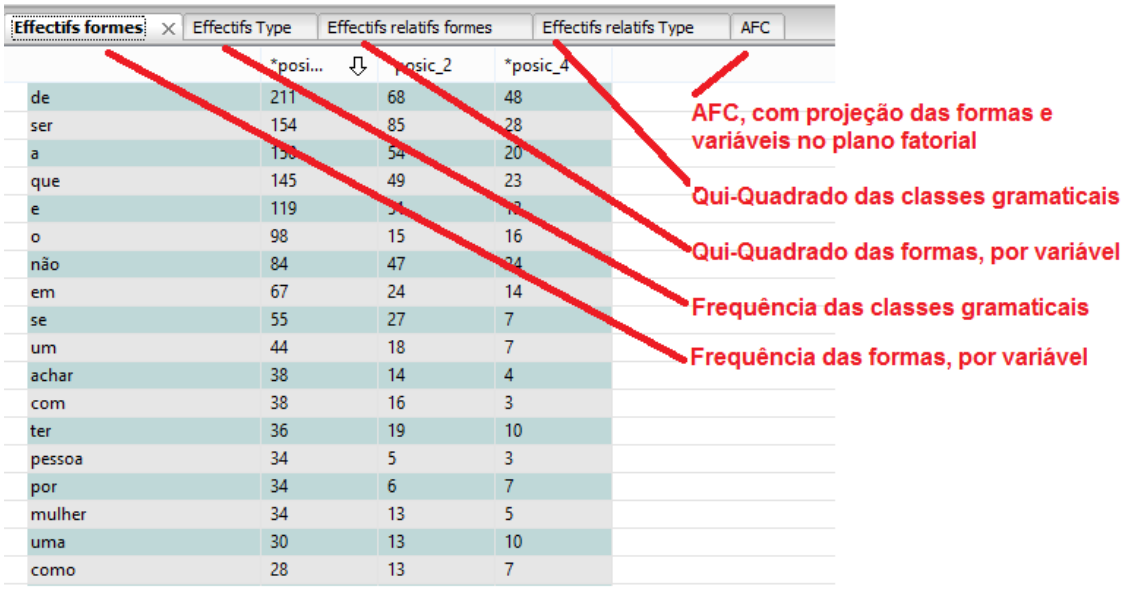

Fig. 9. Resultados, especificidades e AFC.

CLASSIFICAÇÃO HIERÁRQUICA DESCENDENTE (CHD)

Ao escolher a CHD, você pode optar por três possibilidades de análise na janela que aparecerá na interface do IRAMUTEQ.

- DOUBLE SUR SRT – não utilizada, pois usualmente tem baixo aproveitamento do c*orpus.*
- SIMPLE SUR SEGMENTS DE TEXTE – que equivale a uma análise sobre os segmentos de texto, delimitados pelo programa (Análise Standart), recomendada para respostas longas.

- SIMPLE SUR TEXTES – que realiza a análise considerando a os textos, sem dividi-los em segmentos de texto. Recomendada para respostas curtas.

Escolha uma das modalidades de classificação. Nas demais configurações (parametragens) não é necessária nenhuma modificação. Clique em *OK* e aguarde alguns segundos até que a análise seja finalizada. Na interface de resultados aparecerão alguns dados importantes à CHD (Fig. 10), seguidos do dendograma (Fig. 11):

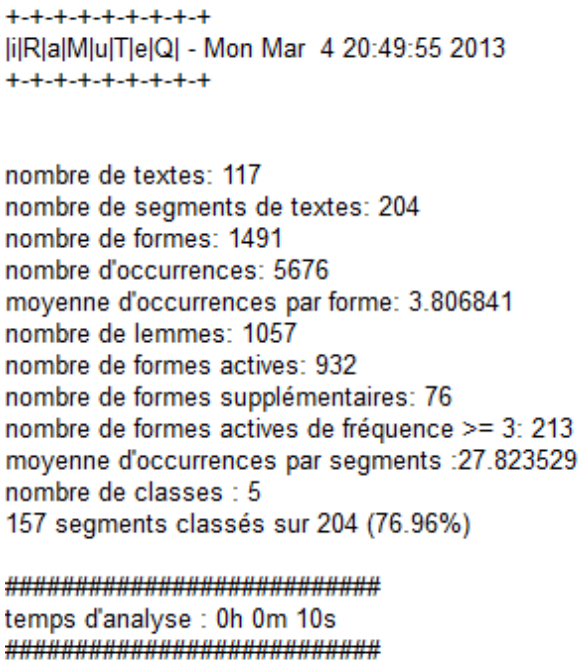

Fig. 10. Principais pontos da CHD a serem considerados

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

- Número de textos (*nombre de textes*) = 117 (o programa reconhece a separação do corpus em 117 unidades de texto iniciais).
- Número de segmentos de textos (*nombre de segments de textes*) = 204 (o programa reparte em 204 segmentos de texto)
- Número de formas distintas (*nombre de formes*) = 1491.
- Número de ocorrências (*nombre d'occurrences*) = 5676
- Frequência média das formas (*moyenne d'occurrences par forme*) = 3.80
- Número de classes (*nombre de classes*) = 5
- Retenção de segmentos de texto: **157 segments classés sur 204 (76.96%)**

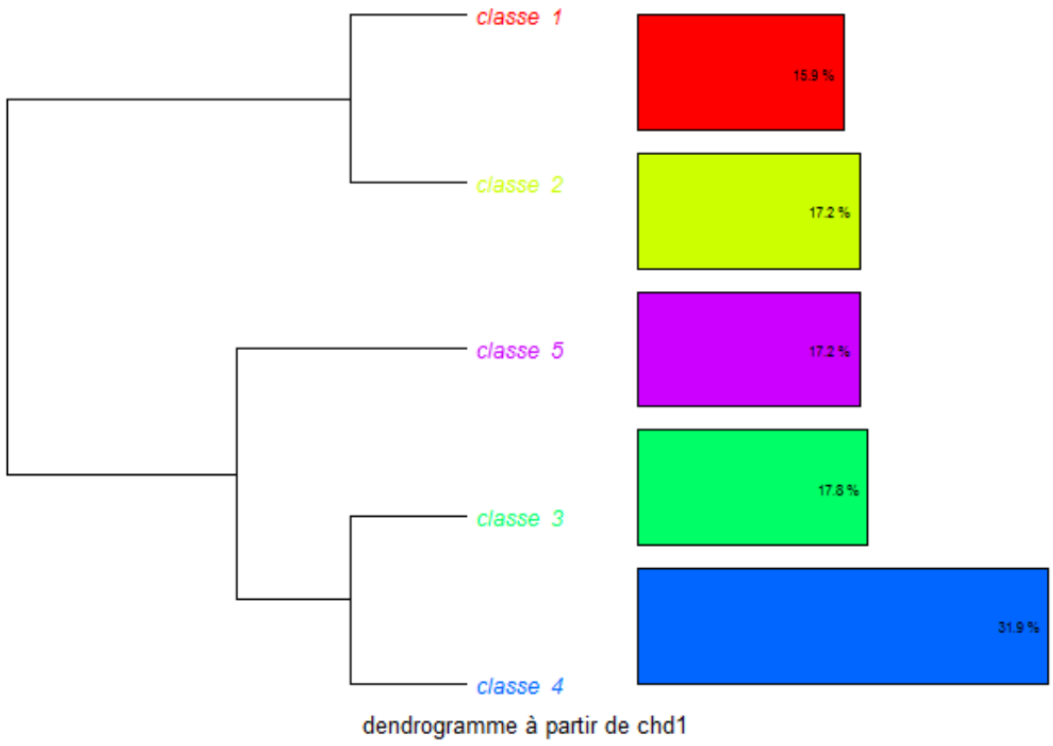

Fig. 11. Dendograma da CHD.

Na aba CHD dos resultados, é possível ter acesso ao dendograma, que apresenta as partições que foram feitas no *corpus* até que se chegasse às classes finais. Lê-se o dendograma da esquerda para a direita. No exemplo da figura 11, num primeiro momento, o *corpus* “obesidade", utilizado aqui como exemplo, foi dividido (1ª partição ou iteração) em dois *sub-corpus*. Num segundo momento um *sub-corpus* foi dividido em dois (2ª partição ou iteração), assim obteve-se a classe 5. E num terceiro momento, há mais partições, originando de um lado, as classes 1; e 2 e do outro, as classes 3 e 4. A CHD parou aqui, pois as 5 classes mostraram-se estáveis, ou seja, compostas de unidades de segmentos de texto com vocabulário semelhante.

Além do dendograma, essa interface de resultados também possibilita que se identifique o conteúdo lexical de cada uma das classes (para acessá-lo, basta clicar na aba *Profils*) e uma representação fatorial da CHD (para acessá-la, basta clicar na aba AFC).

Na aba *Profils*, para cada classe encontram-se dados referentes ao seu conteúdo: *n.* (número que ordena as palavras na tabela); *eff. st* (número de segmentos de texto que contêm a palavra na classe); *eff. total* (número de segmentos de texto no *corpus* que contém, ao menos uma vez, a palavra citada); *pourcentage* (percentagem

de ocorrência da palavra nos segmentos de texto nessa classe, em relação a sua ocorrência no *corpus*); *chi2* (χ^2 de associação da palavra com a classe); *Type* (classe gramatical em que a palavra foi identificada no dicionário de formas); *Forme* (identifica a palavra) e *P* (identifica o nível de significância da associação da palavra com a classe).

Na coluna da esquerda na interface, clicando com o botão direito do mouse sobre a análise denominada NOME DO CORPUS_alceste_1, você pode ter acesso a mais alguns resultados da análise. Dentre eles, os mais importantes são:

- **Coupus em Couleur** - o qual abrirá uma interface de navegação da internet que permitirá que você visualize os segmentos de texto característicos de cada classe, identificando-a pelas cores das classes, conforme as apresentadas no dendograma.
- **Rapport** – que criará um documento em .txt, denominado **Rapport**, dentro da pasta que contém o corpus, em uma subpasta denominada **NOME DO CORPUS__alceste_1**. Esse documento, que poderá ser visualizado em qualquer editor de texto, contém a descrição lexical de cada uma das classes formadas pela CHD, numa espécie de Relatório Simplificado da Análise.

Codificação das formas gramaticais

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

ANÁLISE DE SIMILITUDE

Ao escolher a análise de similitude, uma nova janela se abrirá, possibilitando que sejam escolhidos alguns parâmetros para a construção da árvore de coocorrências. Em *Paramètres du graph,* você pode editar a análise, trocar o índice de coocorrências por algum outro, escolher se será uma árvore máxima ou não, etc. Na

aba *Paramètres graphiques*, por sua vez, é possível fazer edições gráficas (tamanho do texto, tamanho das arestas, cores, etc). Tendo escolhido os parâmetros clique em *OK* e aguarde enquanto a análise se finaliza.

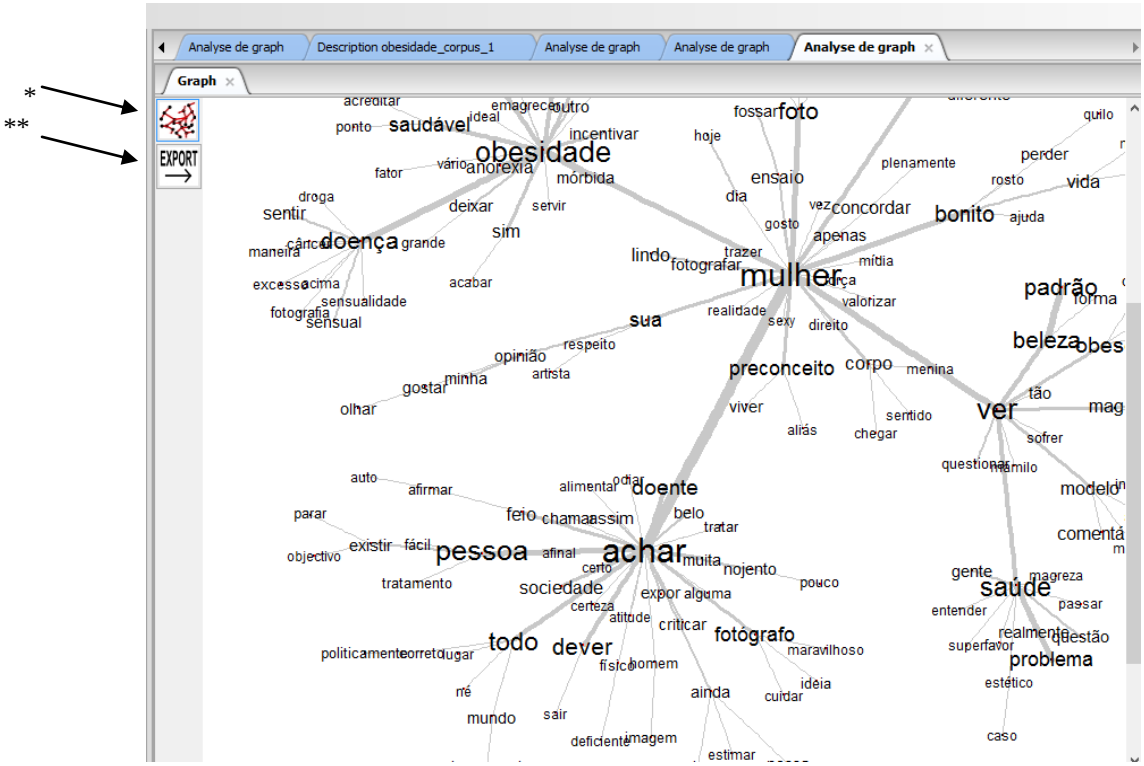

Fig.12. Resultados da Análise de similitude

Conforme se observa na Figura 12, a árvore é apresentada na interface dos resultados. No canto superior esquerdo dessa janela, aparecem dois botões. O primeiro deles (*) com traços vermelhos e pontos pretos permite que se modifique a parametragem da análise, abrindo novamente a janela para edição dos parâmetros. O segundo botão (**), no qual está escrito *EXPORT*, exportará a imagem para a pasta das análises, dentro de uma subpasta denominada NOME DO CORPUS_ simitxt_1.

NUVEM DE PALAVRAS

Ao escolher a nuvem de palavras, uma nova janela se abrirá, também possibilitando que sejam escolhidos alguns parâmetros para a análise, os quais não necessariamente precisam ser editados. Esta é uma análise mais simples, que trabalha com a representação gráfica em função da frequência das palavras. Tendo escolhido os parâmetros, clique em *OK* nas duas janelas que aparecerão e aguarde alguns instantes.

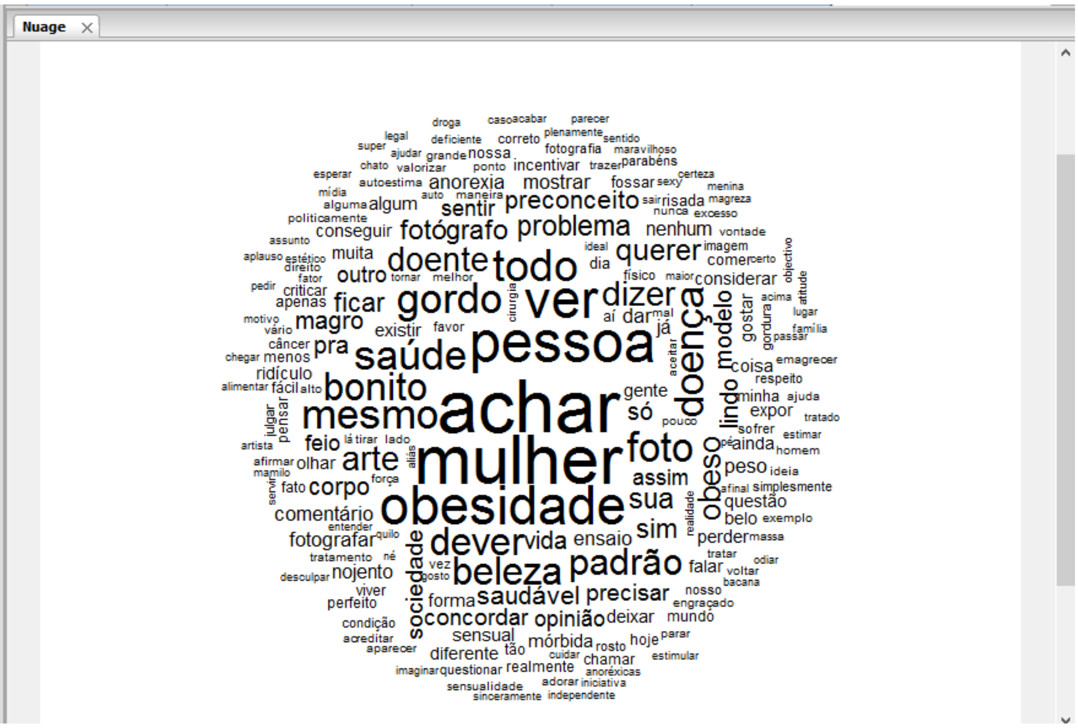

Fig. 13. Resultados da Nuvem de palavras

Na interface dos resultados (Figura 13) você poderá visualizar a nuvem de palavras, a qual também pode ser visualizada, dentro da pasta de análises, na subpasta NOME DO CORPUS_wordcloud_1, em arquivo de imagem denominado **nuage_1.**

REFERÊNCIAS

Camargo, B. V. (2005). ALCESTE: Um programa informático de análise quantitativa de dados textuais. In Moreira, A. S. P.; Camargo, B. V.; Jesuíno, J. C.; Nóbrega, S. M. (Eds.) *Perspectivas teórico-metodológicas em representações sociais* (pp. 511-539). João Pessoa: Editora da UFPB.

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.

Doise, W.; Clemence, A.; Lorenzi-Cioldi, F. (1992). *Représentations sociales et analyses de données.* Grenoble: P.U.G.

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

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

Lebart, L.; Salem, A. (1994). *Statistique textuelle.* Paris: DUNOP.

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

Nascimento-Schulze, C. M.; Camargo, B. V. (2000). Psicologia social, representações sociais e métodos. *Temas de psicologia. Ribeirão Preto*, 8 (3), 287-299.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IraMuTeQ. Em: *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles* (835–844). Presented at the 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012, Liège.

Reinert, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud, *Bulletin de méthodologie sociologique*, (13).

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

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