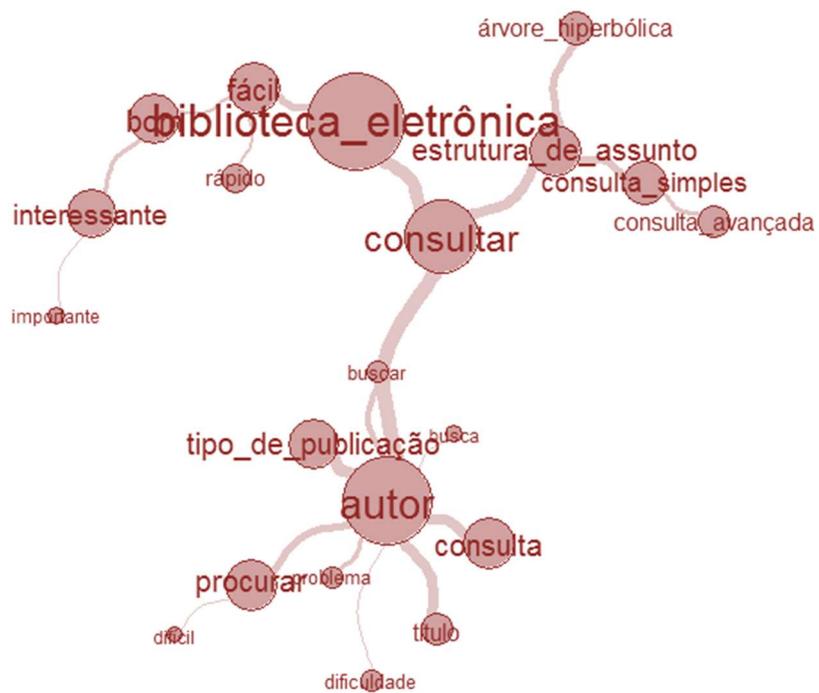


Manual do Aplicativo Iramuteq

(versão 0.7 Alpha 2 e R Versão 3.2.3)

compilação, organização e notas de

Maria Elisabeth Salviati



Planaltina, março de 2017

Sumário

I – Introdução	4
II - Instalação de aplicativos	6
1. Instalação do Open Office	6
2. Instalação do software estatístico R	6
3. Instalação do Iramuteq	7
4. Instalação do Gephi	8
III – Nomenclatura utilizada neste documento	10
IV – Construção do <i>corpus</i>	16
1. Formatação	16
2. Variáveis e temáticas	19
3. Gravação do <i>corpus</i>	20
4. Dicionário de termos	22
V – Importação do <i>corpus</i>	24
1. Aba geral	24
2. Aba limpando	26
VI – Introdução às análises	29
VII – Estatísticas textuais	33
1. Parâmetros	
2. Resultados	33
3. Análises estatísticas textuais suplementares	35
4. Análises estatísticas textuais arquivos criados	37
VIII – Especificidades e AFC	39
1. Parâmetros	39
2. Resultados	40

3. Análises de Especificidades suplementares	43
4. Análises de Especificidades e AFC arquivos criados	44
IX – Classificação pelo Método de Reinert	46
1. Parâmetros	47
2. Resultados	48
3. Análises de Classificação suplementares	55
4. Análises de Classificação arquivos criados	66
X – Análises de Similitude	69
1. Parâmetros	69
2. Resultados	73
3. Análises de Similitude arquivos criados	78
XI – Nuvem de Palavras	79
1. Parâmetros	79
2. Resultados	80
3. Análises de Nuvem de palavras arquivos criados	81
XII – Reedição de Gráficos da Análise Textual	82
1. Reedição dos Dendrogramas	82
2. Reedição da AFC – Análise Fatorial de Correspondência	84
3. Reedição de Gráficos Dinâmicos	86
Referências	92
Anexo – Exemplo de aplicação do Iramuteq	95

I – Introdução

A análise textual é um tipo específico de análise de dados, que trata especificamente da análise de material verbal transcrito, ou seja, de textos produzidos em diferentes contextos.

Ela é aplicada nos estudos de pensamentos, crenças e opiniões produzidas em relação a determinado fenômeno, tema de investigação, permitindo a quantificação de variáveis essencialmente qualitativas originadas de textos, a fim de descrever o material produzido por determinado sujeito ou sujeitos (CAMARGO & JUSTO, 2013).

Para se analisar grande volume de textos têm sido utilizados softwares específicos de análise textual tais como Alceste e Iramuteq. O uso de novas técnicas para manipular e apresentar grandes volumes de dados leva a novas possibilidades de análise – pois construir uma representação, naturalmente, é propor uma interpretação. Esses softwares possibilitam identificar o contexto em que as palavras ocorrem. Eles executam análise lexical do material textual e particionam o texto em classes hierárquicas, identificadas a partir dos segmentos de textos que compartilham o mesmo vocabulário, facilitando, assim, o pesquisador conhecer seu teor.

O software Iramuteq - Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires foi criado em 2009 por Pierre Ratinaud. É um software gratuito de código fonte aberto, licenciado por GNU GPL (v2), que utiliza o ambiente estatístico do software R. Assim como os outros softwares de fonte aberta, ele pode ser alterado e expandido por meio da linguagem Python (www.python.org). Ele é utilizado no estudo das Ciências Humanas e Sociais e utiliza o mesmo algoritmo do software Alceste para realizar análises estatísticas de textos, porém, incorpora, além da CHD - Classificação Hierárquica Descendente, outras análises lexicais que auxiliam na análise e interpretação de textos.

Seu funcionamento consiste em preparar os dados e escrever *scripts* que são então analisados pelo software estatístico. Os resultados são finalmente exibidos pela interface.

Este manual é resultado de trabalho desenvolvido com o objetivo reunir, organizar, testar e apresentar as informações coletadas de diversas fontes para conduzir análise de dados textuais por intermédio do Iramuteq.

Destina-se aos pesquisadores que trabalham com análise qualitativa de conteúdo textual, como entrevistas, documentos, artigos de revistas, jornais, notícias etc., fontes usadas tradicionalmente em Ciências Humanas e Sociais.

Procurou-se abranger com maior completeza possível as diversas opções oferecidas pelo aplicativo, que foram testadas e originaram um conjunto de dicas e notas de observação para o usuário usufruir de melhor desempenho dos procedimentos de análise. Incluiu-se, também, um pequeno glossário de termos técnicos, úteis para melhor compreensão dos procedimentos descritos.

Cabe salientar que o usuário deve ter em mente que a análise automática apresenta resultados genéricos que indicam caminhos a serem explorados e interpretados manualmente.

II – Instalação de aplicativos

1. Instalação do Open Office

O Open Office é o editor de texto e planilhas padrão para ser utilizado em conjunto com o Iramuteq.

Do pacote do Open Office serão utilizados o Writer para digitação de texto, leitura dos relatórios e resultados das análises efetuadas e o Calc para digitação de planilhas, leitura e exportação de resultados.

- a) Fazer download do arquivo: Apache_OpenOffice_4.1.3_Win_x86_install_pt-BR no endereço:

https://sourceforge.net/projects/openofficeorg.mirror/files/4.1.3/binaries/pt-BR/Apache_OpenOffice_4.1.3_Win_x86_install_pt-BR.exe/download

- b) Instalação do pacote do Open Office, aplicativos Writer e Calc.

Nota: Para criação e/ou abertura do *corpus* e outros arquivos do Iramuteq, veja o capítulo Preparação do *corpus*.

2. Instalação do software estatístico R

O R é um software gratuito para elaboração de gráficos e computação estatística. O programa oferece uma variedade de técnicas gráficas e estatísticas, modelos lineares e não lineares, testes estatísticos clássicos, análises de séries, classificação, etc.

Um dos destaques do programa é a facilidade para produzir diagramas com qualidade de publicações profissionais, incluindo símbolos matemáticos e fórmulas quando necessário.

O R é um pacote integrado de softwares para facilitar o trabalho com dados, cálculos e exibição de gráficos. As ferramentas incluem um gerenciador de dados e armazenamento, operadores para cálculos em exibições (matrizes em particular), uma coleção de ferramentas para análise de dados, facilidades para exibição de dados e recursos de programação para os usuários avançados.

O Iramuteq executa as análises utilizando as Bibliotecas do R. Por isso, antes de instalar o Iramuteq é necessário instalar o R e as bibliotecas, executando os passos abaixo.

- Fazer o download do software R versão 3.2.3 para Windows em <https://cran.r-project.org/bin/windows/base/old/3.2.3/>
- Instalá-lo (arquivo: "R-3.2.3-win.exe") de preferência na pasta de Arquivos de Programas.
- Durante a instalação, escolha corretamente se 32 ou 64 bits.
- Executar o R e instalar as bibliotecas (pacotes) do R específicas para o IRAMUTEQ conforme as seguintes instruções¹:
 - a. Executar o R;
 - b. Escolher no menu principal Pacotes/Instalar pacotes;
 - c. Escolher o país (França/Paris2)²;
 - d. Escolher na lista apresentada em ordem alfabética, o primeiro pacote a ser instalado (ape) e clicar Ok. Se o sistema perguntar para criar uma pasta para armazenar a biblioteca, optar por criá-la e deixar o R escolher a pasta padrão. O sistema realiza o download da biblioteca e a instala. Veja resultado dessa operação na tela de operação (Console);
 - e. Depois de terminada a instalação do primeiro, instalar todos os demais (ca, gee, igraph, irlba, proxy, rgl, textometry, wordcloud) repetindo os passos d e e;
 - f. Fechar o R.

3. Instalação do Iramuteq

¹ Para fazer essa instalação é necessário que o computador esteja conectado na Internet.

² Instale as bibliotecas do país indicado (França/Paris2), pois pode acontecer de não existir as bibliotecas do Iramuteq nas outras opções.

- a) Fazer o download do software IRAMUTEQ em <https://sourceforge.net/projects/iramuteq/files/iramuteq-0.6-alpha3/>
- b) Instalá-lo (arquivo: “setup_iramuteq-0.6-alpha3.exe”) na mesma pasta onde foi instalado o R. Se não for instalado na mesma pasta, o Iramuteq não reconhece as bibliotecas do R, mesmo se o caminho for informado em Preferências do Iramuteq.
- c) Entrar no Iramuteq e aguardar a instalação das bibliotecas do R automaticamente³.

Nota: Se a instalação do R foi correta, bem como das bibliotecas, o Iramuteq encontra todas as bibliotecas e fica pronto para proceder às análises. Caso tenha algum problema na instalação do R ou das bibliotecas, o Iramuteq não conseguirá trazê-las e continuará solicitando a sua atualização toda vez que se entrar no Iramuteq ou, ainda, não executará as análises porque as bibliotecas não existem. Pode ser, ainda, que ele não encontre o R e informe que o R não está instalado. Nesse caso, não adianta acrescentar o caminho do R em Edição/Preferências porque ele continuará sem encontrar as bibliotecas. A solução desse problema normalmente envolve a desinstalação do Iramuteq e do R e nova instalação nas pastas corretas.

4. Instalação do Gephi

O Gephi é a ferramenta para manipulação de grafos mais utilizada. É um software livre colaborativo mantido por um consórcio sediado na França, com inúmeras aplicações em áreas como ciências biológicas, economia e em projetos ligados à história e à análise textual.

Estruturas que podem ser representadas por grafos estão em toda parte e muitos problemas de interesse prático podem ser formulados como questões sobre certos grafos. Por exemplo, a estrutura de ligações da Wikipédia pode ser representada por um dígrafo: os vértices são os artigos da Wikipédia e existe uma aresta do artigo A para o artigo B se e somente se A contém um link para B. Dígrafos são também usados para representar máquinas de estado finito. O desenvolvimento de algoritmos para manipular grafos é um tema importante da ciência da computação.

³ Quando a versão do R for a 3.2.3 as bibliotecas do R são carregadas automaticamente. Porém, se o usuário instalar outra versão pode ser necessário solicitar a instalação das Bibliotecas.

É, no entanto, na área dos estudos textuais que as técnicas de visualização em grafos têm sido mais aplicadas, sobretudo em estudos voltados para grandes volumes de textos.

O Gephi é uma plataforma interativa de visualização e exploração de todos os tipos de redes e sistemas complexos, grafos dinâmicos e hierárquicos. É um software escrito em Java e roda tanto no Windows, Linux e Mac. Ele usa um mecanismo 3D de renderização para exibir gráficos em tempo real e acelerar a exploração. Você pode usá-lo para explorar, analisar, espacializar, filtrar, clusterizar, manipular e exportar todos os tipos de gráficos.

Os grafos gerados pelo Iramuteq oriundo das análises de similitude e nuvem de palavras poderão ser exportados no formato graphml e então abertos no Gephi para edição.

Para instalá-lo é necessário executar os passos abaixo.

- 1) Fazer download e instalar o software Gephi (<https://gephi.org/>), release: 0.9.1 / 14 February 2016;
- 2) Instalar o arquivo gephi-0.9.1-windows.exe.

III - Nomenclatura utilizada neste documento

1. Noção de *corpus*, texto e segmentos de texto

- **Corpus**: é um conjunto de textos construídos pelo pesquisador e que forma o objeto de análise. Por exemplo, em uma pesquisa documental da área médica o *corpus* poderia ser um conjunto de artigos que foram publicados em determinado período na sessão de saúde de um jornal. Outro exemplo seria um conjunto de 40 transcrições de entrevistas sobre um tema objeto de um estudo de casos. E ainda pode ser, por exemplo, um *corpus* composto de 200 respostas a uma questão aberta, que faz parte de um questionário de uma pesquisa do tipo enquete (CAMARGO & JUSTO, 2013b).

O vocabulário utilizado pelo aplicativo Iramuteq difere daquele utilizado no software Alceste, embora seja baseado nele. Por razões de propriedade intelectual, a terminologia da Análise de Classificação pelo método Reinert foi substituído conforme na Figura 1:

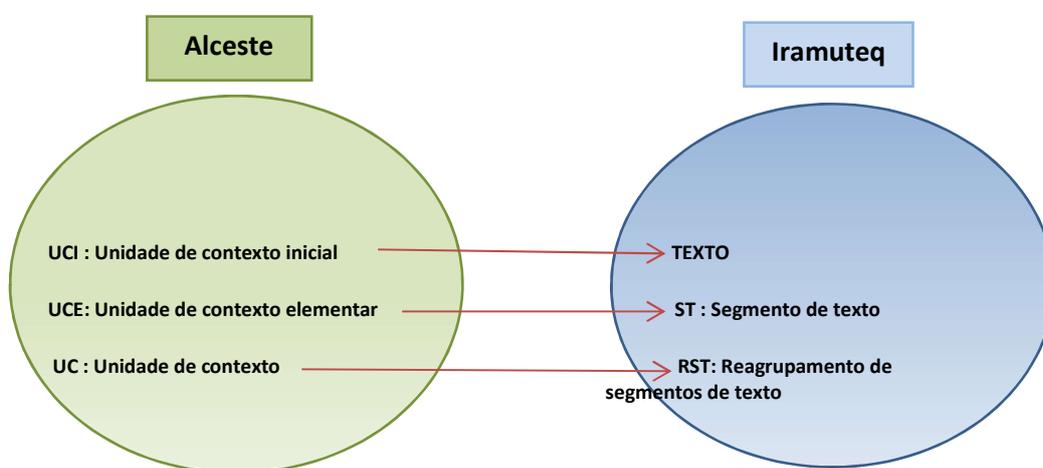


Fig. 1 – Equivalência de Nomenclatura entre Aplicativos

- **Texto**: a definição 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, portanto, constitui um *corpus* de análise. Para submissão do *corpus* à análise do tipo Classificação Hierárquica Descendente, ele 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, desde que o grupo seja homogêneo, é suficiente entre 20 e 30 textos (GHIGLIONE & MATALON, 1993). Ainda, 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, caso contrário, a análise resulta apenas na identificação dos diversos temas, invalidando a obtenção de resultados mais precisos. Outra solução seria utilizar a variável temática.

Quando as respostas apresentarem em média de três a 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).

- **Segmentos de textos:** são fragmentos 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 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, o software IRAMUTEQ divide os textos do *corpus* em segmentos de texto (CAMARGO & JUSTO, 2013b).

- **Reagrupamento de segmentos de texto:** agrupamento de UCE sucessivas dentro de uma mesma UCI, até que o número de palavras diferentes analisadas (contidas nessa unidade de contexto) seja superior ao limiar fixado na análise. (...) Esse limiar é fixado proporcionalmente ao número de palavras analisadas. As UC devem ter uma homogeneidade semântica e devem

ser de tamanho comparável. Ou seja, são unidades estatísticas de pesos idênticos. Elas estão na base de todas as estatísticas efetuadas (REINERT, 1998, p. 18).

Equivale à UC (unidade de contexto) do aplicativo Alceste. (NASCIMENTO & MENANDRO, 2006).

AFC: Análise Fatorial de Correspondência. Cruzamento entre o vocabulário (considerando a frequência de incidência de palavras) e as classes, gerando uma representação gráfica em plano cartesiano, na qual são vistas as oposições entre classes ou formas (NASCIMENTO & MENANDRO, 2006).

Classe: pode ser definida como um agrupamento constituído por várias UCE de vocabulário homogêneo (NASCIMENTO & MENANDRO, 2006).

- **CP1252:** Windows-1252 ou CP1252 é uma codificação de caracteres do alfabeto latino, usado por padrão nos componentes herdados do Microsoft Windows em Inglês e algumas outras línguas ocidentais.

- **Default:** é um termo técnico utilizado em computação e em vários contextos de informática. Termo de origem inglesa que significa "padrão", algo "predefinido", valor ou ação que é assumido por omissão. Pode ser utilizado tanto para referir-se a um valor pré-definido que o sistema computacional assume como para uma ação tomada pelo sistema a menos que seja instruído de outra maneira. <[https://pt.wikipedia.org/wiki/Default_\(computação\)](https://pt.wikipedia.org/wiki/Default_(computação))> Acesso em 18/04/2016.

- **Dicionário de lemmes:** contém o lema (palavra sem flexão), suas variantes e sua frequência de ocorrência. (MANUAL, s.d.).

- **Frequência:** Em estatística, a frequência (ou frequência absoluta) de um evento i é o número T_{li} de vezes que o evento ocorreu em um experimento ou estudo. [https://pt.wikipedia.org/wiki/Frequ%C3%Aancia_\(estat%C3%ADstica\)](https://pt.wikipedia.org/wiki/Frequ%C3%Aancia_(estat%C3%ADstica)) Acesso em 19/04/2016.

- **Grafo:** A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Para tal são empregadas estruturas chamadas de grafos, $G(V,E)$, onde V é um conjunto não vazio de objetos denominados vértices e E é um conjunto de pares não ordenados de V , chamado arestas. Sua aplicação é ampla. No caso presente, os grafos auxiliam na caracterização e visualização gráfica do *corpus*, permitindo a interpretação do conteúdo textual. <http://pt.wikipedia.org/wiki/Teoria_dos_Grafos> Acesso em : 14/04/2016.

- **Hápax:** Um hápax ou hápax legómenon é uma palavra que aparece registrada somente uma vez em um dado idioma. No Iramuteq, esta expressão designa uma palavra que se utilizou ou registrou apenas uma vez em um *corpus*. <<https://pt.wikipedia.org/wiki/Hápax>> Acesso em 18/04/2016.

- **Lei de Zipf:** a frequência e a posição das palavras (na lista ordenada por frequência decrescente) estão relacionadas por uma lei de potência. Esta lei empírica é hoje conhecida como a “Lei de Zipf”, em homenagem ao professor de linguística da Universidade de Harvard, George Kingsley Zipf (1902–1950), o primeiro a investigar de forma sistemática fenômenos da estrutura estatística em conjuntos de dados linguísticos e demográficos. Em análises de textos, ela permite estimar as frequências de ocorrência das palavras de um determinado texto científico e tecnológico e a região de concentração de termos de indexação, ou palavras-chaves, que um pequeno grupo de palavras ocorre muitas vezes e um grande número de palavras é de pequena frequência de ocorrência (GUEDES & BORSCHIVER, 2005; BORTOLOSSI; QUEIROZ; SILVA, 2011).

- **Lei hipergeométrica:** na teoria das probabilidades e estatística, a distribuição hipergeométrica é uma distribuição de probabilidade discreta que descreve a probabilidade de se retirar x elementos do tipo A numa sequência de n extrações de uma população finita de tamanho N , com K elementos do tipo A e $N-K$ elementos do tipo B, sem reposição. <https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_hipergeom%C3%A9trica> Acesso em 19/04/2016.

- **Lematização:** é o processo, efetivamente, de deflexionar uma palavra para determinar o seu lema (as flexões chamam-se lexemas) Por exemplo, as palavras *gato*, *gata*, *gatos*, *gatas* são todas formas do mesmo lema: *gato*. Igualmente, as palavras *tiver*, *tenho*, *tinha*, *tem* são do mesmo lema *ter*. E *bom*, *melhor* e *ótimo* são lexemas do lema *bom*. No Iramuteq

existem regras próprias de lematização. Os verbos são convertidos ao infinitivo, os substantivos ao singular e os adjetivos ao masculino singular. O Iramuteq realiza a lematização a partir dos dicionários, sem realizar a desambiguação. <<http://portuguese.stackexchange.com/questions/1417/o-que-%C3%A9-lematiza%C3%A7%C3%A3o>> Acesso em 18/04/2016.

- **Método IRLBA:** o método Fast Truncated SVD, PCA and Symmetric Eigendecomposition for Large Dense and Sparse Matrices são eficientes em rapidez e memória para a decomposição singular e de valores próprios, e análise de componentes principais de grandes matrizes esparsas ou densas. <<https://cran.r-project.org/web/packages/irlba/index.html>>

- **Método SDV:** Em álgebra linear, a decomposição em valores singulares ou singular value decomposition (SVD) é a fatoração de uma matriz real ou complexa, com diversas aplicações importantes em processamento de sinais e estatística. <https://pt.wikipedia.org/wiki/Decomposi%C3%A7%C3%A3o_em_valores_singulares>

- **Permilagem:** Na matemática, a expressão de um número por mil ou permilagem é uma maneira de o expressar como uma fração de 1.000 (1/1000), ou como a décima parte de 1%. <https://pt.wikipedia.org/wiki/Por_mil> Acesso em 19/04/2016.

- **Pixel:** (plural pixels ou píxeis) (aglutinação de Picture e Element, ou seja, elemento de imagem, sendo Pix a abreviatura em inglês para Pictures) é o menor elemento num dispositivo de exibição (como por exemplo um monitor), ao qual é possível atribuir-se uma cor. De uma forma mais simples, um pixel é o menor ponto que forma uma imagem digital, sendo que o conjunto de milhares de pixels forma a imagem inteira. <https://pt.wikipedia.org/wiki/Pixel> Acesso em 22/04/2016.

Num monitor colorido, cada Pixel é composto por um conjunto de 3 pontos: verde, vermelho e azul. Nos melhores monitores cada um destes pontos é capaz de exibir 256 tonalidades diferentes (o equivalente a 8 bits) e combinando tonalidades dos três pontos é então possível exibir pouco mais de 16.7 milhões de cores diferentes (exatamente 16.777.216). Em resolução de 640 x 480 temos 307.200 pixels, a 800 x 600 temos 480.000 pixels, a 1024 x 768 temos 786.432 pixels e assim por diante.

Para que o gráfico todo possa ser visualizado na tela do computador, deve-se escolher largura e altura compatíveis com a resolução de vídeo em uso no computador. Para consultar a resolução em uso, clique com o botão do mouse sobre a área de trabalho e em seguida escolha Resolução de tela no menu apresentado.

- **Qui-quadrado:** a distribuição χ^2 ou chi-quadrado é uma das distribuições mais utilizadas em estatística inferencial, principalmente para realizar testes de χ^2 . Este teste serve para avaliar quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. Isto é, ele nos diz com quanta certeza os valores observados podem ser aceitos como regidos pela teoria em questão. Muitos outros testes de hipótese usam, também, a distribuição χ^2 . <<https://pt.wikipedia.org/wiki/Chi-quadrado>> Aceso em 19/04/2016.

- **UTF-8 (8-bit Unicode Transformation Format):** é um tipo de codificação Unicode de comprimento variável criado por Ken Thompson e Rob Pike. Pode representar qualquer carácter universal padrão do Unicode, sendo também compatível com o ASCII. <<https://pt.wikipedia.org/wiki/UTF-8>> Acesso em 14/04/2016.

IV – Construção do *corpus*

1. Formatação

Para edição do *corpus* utilize o aplicativo OpenOffice Writer (para texto) e Calc (para planilhas). Os aplicativos editores de texto que fazem parte do Windows: Wordpad, Bloco de Notas e Office (Word, Excel) e mesmo o Open Office, trabalham com o padrão de codificação CP1252⁴. De forma que quando se grava um texto com formato txt por esses aplicativos eles estarão obedecendo a esse padrão, a menos que o usuário, no momento de gravação, altere a codificação para outro formato, por exemplo, UTF-8.

Existem muitos códigos de representação de letras (alfabetos diversos), números e símbolos no computador. UTF-8 é apenas um deles. O Unicode é um padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente.

O formato UTF-8 é melhor se o usuário trabalhar com aplicativos que não aceitam a codificação CP1252, principalmente quando se trata de software livre que não obedece ao padrão Windows.

Na elaboração deste manual foram testados tanto o formato CP1252 como UTF-8 e eles não apresentaram qualquer erro com os aplicativos utilizados.

Para o usuário não ter problemas com codificação, ele tem que ter em mente que esses formatos existem e são conflitantes entre si. Quando se usa o padrão CP1252, ele tem que ser informado na importação do *corpus* no Iramuteq. Além disso, toda vez que abrir esse documento com qualquer editor de texto e ele solicitar o padrão de codificação, deve-se informar também a codificação correta utilizada, caso contrário, ele abrirá o arquivo com erro de codificação. Como consequência, os sinais especiais utilizados por nossa língua como til, cedilha, acentuação e crase não serão representados e, por isso, substituídos pelo símbolo ❖, indicando caráter desconhecido. O mesmo ocorre se ele utilizar o padrão UTF-8.

Vários arquivos gerados pelo Iramuteq estão no formato CP1252 e se abertos como UTF-8 apresentarão problema de codificação.

⁴ Veja definição no capítulo III Nomenclatura utilizada neste documento.

O formato CP1252 - no Open Office está designado como Europa Ocidental (Windows-1252/WinLatin 1), como já explicado anteriormente, não precisa ser informado no momento da gravação, basta escolher o formato texto (.txt).

Porém, a gravação de texto, utilizando o formato UTF-8, deve seguir codificação especial, portanto, antes de gravá-lo, siga as orientações do item 3, Gravação do *corpus*.

As entrevistas ou questões abertas devem ser formatadas sem as questões. As questões podem ser consideradas como variáveis temáticas (veja item a seguir). Deve-se obedecer às seguintes regras:

- a) Sinais proibidos: aspas; apóstrofo; cifrão; porcentagem; asterisco; reticências; travessão; negrito, itálico, grifo e outros sinais similares; recuo de parágrafo, margens ou tabulações do texto; justificação do texto.
- b) Pontuação permitida: ponto; dois pontos; vírgula; interrogação e exclamação.
- c) Formatação de texto todo corrido, sem mudança de linha.
- d) Uso de maiúsculas só para nomes próprios.
- e) Palavras compostas devem ser unidas por *underline*, mesmo aquelas unidas ortograficamente pelo hífen. Ex.: recém_casado; anti_inflamatório; Distrito_Federal.
- f) Padronização das siglas e nomes próprios para obedecer sempre mesma grafia.
- g) Revisão gramatical do português, corrigindo-se grafia e concordância.
- h) Complementação de todas as frases incompletas: cada frase deve encerrar um sentido completo e não deve possuir palavras subtendidas. Complementar com as palavras necessárias, sem modificar o sentido. Se necessário, reexaminar o texto original para escolher as palavras adequadas. Caso haja impossibilidade de completar determinadas frases, elas deverão ser eliminadas.

i) Eliminação de expressões sem necessidade, tais como: Ahh, Uhmm, né, tá.

j) Eliminação de frases não condizentes com o assunto tratado.

A Figura 2 a seguir, mostra o trecho original de uma entrevista e a Figura 3, logo abaixo traz o mesmo trecho no corpus preparado.

Com que frequência você utiliza, olha temos as opções: diária, semanal, mensal, esporádica.

Isso varia muito. Se eu to fazendo um projeto, às vezes, diária... mas às vezes também...

Na média, você poderia dizer o que?

Eu poderia colocar semanal, mas, obviamente, não é toda semana que eu consulto.

Entendi. É na média né?!

Mas outras vezes é diária.

Eu tenho aqui diversos tipos de consulta: por assunto, por autor, tipo, título e consulta avançada. Você já utilizou todas essas formas, ou não?

Não. Eu uso a por assunto e por autor.

É?

Normalmente.

Fig. 2 – Trecho original transcrito de uma entrevista realizada

A frequência de uso da biblioteca eletrônica varia muito. Se eu estou fazendo um projeto, às vezes, a frequência de uso é diária. Eu poderia colocar frequência de uso semanal, mas, obviamente, não é toda semana que eu consulto. Mas outras vezes a frequência de uso é diária. Não utilizo todas as formas de consulta. Eu uso a consulta por assunto e por autor, normalmente.

Fig. 3 – Mesmo trecho anterior preparado para o corpus de análise automática

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”.

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

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

2. Variáveis e Temáticas

Os textos são separados por linhas de comando também chamadas de “linhas com asteriscos”, Figura 4. No caso de entrevistas, por exemplo, cada uma delas deve iniciar com uma linha de comando. Esta linha informa o número de identificação do entrevistado, 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.

Já no caso de tabelas efetuadas no Office Calc, cada linha representa um indivíduo, isto é, formato de planilhas e bancos de dados (arquivos ods).

Os textos são introduzidos por quatro estrelas (****) seguidos de uma série de (variáveis ilustrativas) introduzidas com uma * (estrela) separada por um espaço.

Um texto terá que ter obrigatoriamente, pelo menos uma variável.

Ex.: **** *subj_001 *sex_1 *ida_21 *escol_2

Fig. 4 – Formato da linha estrela

suj_001: sujeito 001

sex_1: sexo conforme tabela seguinte: 1 (masculino) e 2 (feminino)

ida_21: idade 21 (a idade é colocada após o *underscore*)

escol_2: escolaridade conforme tabela seguinte: 1 (elementar); 2 (médio); 3 (superior)

Apesar das estrelas indicarem variáveis é possível apresentá-las no interior dos textos, para indicar uma temática. Nesse caso, no início da linha coloca-se um hífen e uma estrela (-*), Figura 5.

Ex.:

```
**** *var1_1 *var2_2
-*tematica1
texto texto
texto
```

Fig. 5 – Formatação de temática

Em um *corpus* com temática, todos os parágrafos do texto devem pertencer à mesma temática. A seguinte construção não é possível, Figura 6.

```
**** *var_1
texto texto texto texto texto (texto sem temática)
-*temática
texto texto texto texto texto texto texto texto (texto com temática)
```

Fig. 6 – Formatação de temática com erro

As variáveis e temáticas não devem conter outros caracteres além de a-z, A-Z, 1-9, traço underline (_) e hífen (para não confundir com as temáticas).

3. Gravação do *corpus*

Os textos ou tabelas que vão constituir o *corpus* de análise devem ser criados, preferencialmente, pelo software OpenOffice e gravados com o formato .txt e, ainda, codificados em Unicode UTF-8. Veja a Figura 7 a seguir.

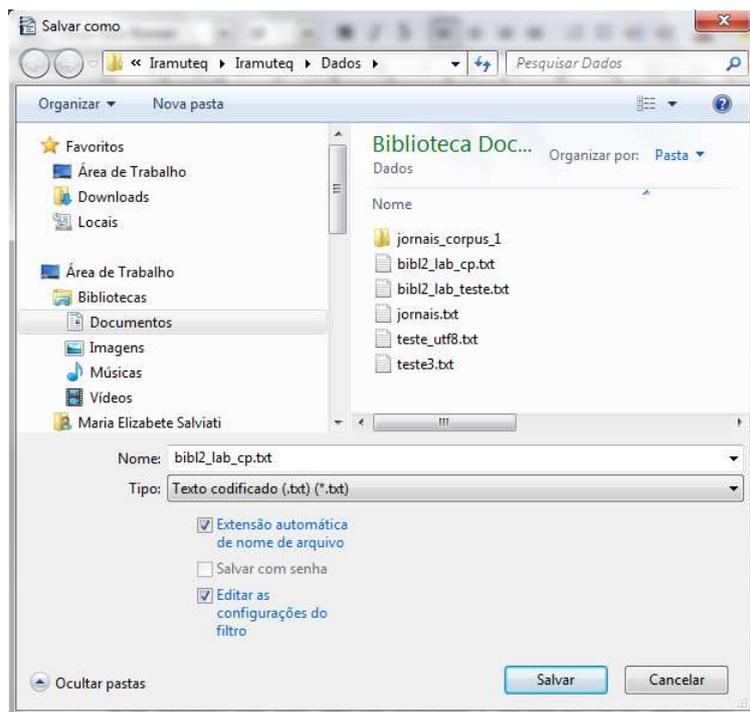


Fig. 7 – Janela do OpenOffice Writer para gravação do *corpus*

Conforme a Figura 2, o *corpus* depois de formatado deve ser salvo com a opção “Salvar como” e então escolher o tipo “Texto codificado (.txt), deixando marcada a opção de “Extensão automática de nome de arquivo” e marcando também a opção “Editar as configurações do filtro”. Em seguida, clicar em Salvar e, na nova janela, em Manter o formato atual.

O sistema mostra o menu para escolha da codificação desejada, Figura 8.

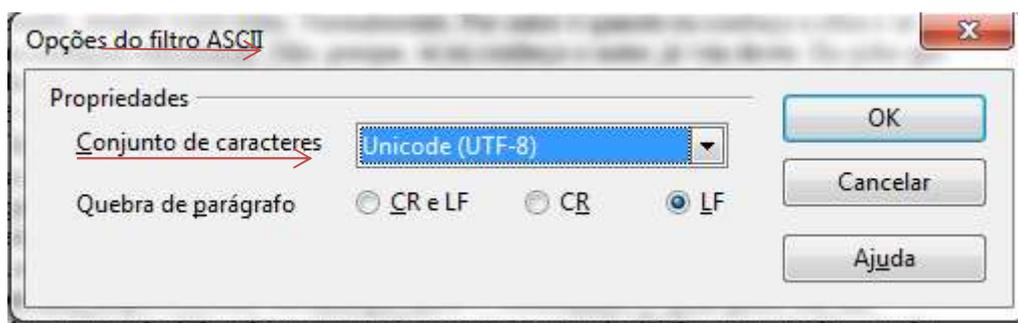


Fig. 8 – Janela do OpenOffice Writer para gravação do *corpus*

Nele, deve-se escolher Unicode UTF-8 e a Quebra de parágrafo por LF e clicar em Ok.

Toda vez que o *corpus* for aberto com o OpenOffice Writer, vai aparecer essa mesma janela da Figura 8 e sempre deverão ser escolhidas essas opções.

Cuidado com a mistura de codificações diferentes, pois ocasionam erros no Iramuteq.

Observou-se, no entanto, que alguns arquivos de formato txt gerados pelo Iramuteq não estão no formato UTF-8 e sim no CP1252 que equivale no Writer e Calc ao Europa Ocidental (Windows 1252/WinLatin 1). Abrindo-se como UTF-8 as letras acentuadas, ç e til serão substituídos por sinal de interrogação. Se isso ocorrer, feche-o sem salvar e abra novamente com a opção correta.

4. Dicionário de termos

O LACCOS - Laboratório de Psicologia da Comunicação e Cognição da Universidade Federal de Santa Catarina em parceria com o Centro Internacional de Estudos em Representações Sociais e Subjetividade – Educação, da Fundação Carlos Chagas (CIERS –ed/FCC); e com o grupo de pesquisa Valores, Educação e Formação de Professores da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) estão aprimorando o dicionário experimental em língua portuguesa, garantindo análises mais estáveis (CAMARGO & JUSTO, 2013).

Mesmo estando o dicionário atualizado, fatalmente ele não encontrará todas as palavras existentes no seu texto. Por isso, deve-se contribuir sempre com a atualização do dicionário. Sempre que realizar uma análise, deve-se encerrar o aplicativo e procurar a pasta em que se encontra a análise de seu *corpus* (nome do *corpus* seguido por “**_corpus_1**”).

Nela foi criada uma subpasta cujo nome é o mesmo do *corpus* seguido por “**_stat_1**” e que contém uma planilha denominada **Total**. Esta planilha deve ser aberta com o Calc (clique com o botão direito do mouse sobre a planilha e escolher Abrir com Office Calc⁵). Esta planilha contém todas as palavras encontradas no texto, seguidas pela frequência de ocorrência ⁶ e classe gramatical. As palavras não existentes no dicionário estarão com a identificação de “nr” (não reconhecidas).

Para atualizar o dicionário acrescentando essas palavras, proceda da seguinte maneira:

- 1) Selecione as três colunas da planilha e peça a ordenação de A/Z pela última coluna e clique Ok.

⁵ Abrir com a codificação Europa Ocidental (Windows 1252/WinLatin 1)

⁶ Para significado de Frequência, veja o capítulo III – Nomenclatura utilizada neste documento.

- 2) Selecione todas as palavras classificadas como nr (não reconhecidas), copie e cole em outra planilha do Calc.
- 3) Verifique o resto da lista se existe outra palavra mal classificada. Nesse caso, copie-a e cole-a também na nova planilha.
- 4) Substitua o nr ou a classificação errada pela correta conforme tabela abaixo.

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

- 5) Envie cópia da nova planilha gravada com o nome "SeuNome"_data (formato ddmmaaaa para o e-mail brigido.camargo@gmail.com. Escreva como assunto do e-mail o mesmo nome do arquivo anexo⁷.

⁷ Dúvidas, consulte o arquivo anexo "AJUDE NO APRIMORAMENTO DO DICIONÁRIO PORTUGUÊS".

V - Importação do *Corpus*

Para importar o texto já estruturado em formato txt, inicie o Iramuteq, e escolha **Arquivo / Abrir um *corpus* textual** no menu horizontal. Em seguida, escolha o arquivo e clique em Abrir. Vai aparecer a seguinte janela de parâmetros de importação, Figura 9:

The image shows a software dialog box titled 'Limçando' with a 'Geral' tab. The dialog contains the following fields and options:

- corpus:** C:\Users\bete\Documents\Iramuteq\Iramuteq\Dados\bibl.txt
- Nome do corpus:** bibl_corpus_3
- Definir caracteres:** utf-8 - all languages (dropdown menu)
- Idioma:** portuguese (dropdown menu)
- Dicionário:** Padrão: portuguese; Outro: [empty field] with a 'Browse' button.
- Pasta de saída:** C:\Users\bete\Documents\Iramuteq\Iramuteq\Dados\bibl_corpus_3 with a 'Mudar ...' button.
- Marcador de texto:** **** (dropdown menu)
- Use o dicionário de expressões:**
- Crie segmentos de texto:**
- Método de construção de ST:** ocorrências (dropdown menu)
- Tamanho de ST:** 40 (input field with up/down arrows)

Fig. 9 – Importação de *corpus*: configuração Geral

1. Aba Geral

Em “Nome do *Corpus*”, escreva o nome do *corpus* de análise.

Em “Definir caracteres”, escolha o formato UTF-8 (all languages) ou CP1252 conforme o padrão de gravação do *corpus* utilizado.

Se o usuário escolher a codificação diferente daquela de gravação do *corpus*, o sistema acusa erro (Problème d'encodage) e não realiza a importação. Se o *corpus* já estiver com erro de codificação, o Iramuteq realiza a importação, mas o dicionário de termos apresentará problemas de representação de caracteres especiais.

Em “Idioma” é necessário selecionar a linguagem do texto a ser analisado. O Iramuteq tem dicionários de análise completo das seguintes línguas: francês, inglês, italiano, português e espanhol. As línguas: alemão, sueco, grego e gálico ainda estão em experimentação.

Em “Dicionário”, deixe a opção Padrão (o sistema seleciona o dicionário que tem a mesma língua do texto).

Em “Pasta de saída”, escolha o diretório deseja salvar a pasta com as análises feitas. O nome dessa pasta que será criada é Nome do arquivo_ *corpus*_X, mas poderá ser alterada juntamente com o local de gravação.

Em “Marcador de texto” deixe por padrão os quatro asteriscos, que delimita o *corpus*.

Em “Use o dicionários de expressões”, deixe selecionado para utilizar o dicionário de expressões. Os dicionários de expressões permitem tratar as palavras compostas como um todo evitando assim a divisão na análise.

Em “Criar segmentos de texto”, deixe selecionado para que o software separe o texto em segmentos de texto (processo em que o software realiza uma separação do *corpus* textual em segmentos de texto para a análise fracionada e identificação de ambientes léxicos).

Em “Método de construção de segmentos de texto”, selecione “ocorrências” para segmentar os *corpus* textuais pela frequência das palavras, “caracteres” para segmentar a partir dos caracteres e “parágrafos” para segmentar por parágrafos. Utilize o padrão (por ocorrências) exceto se houver exigência devido ao *corpus* para escolha dos outros métodos.

- **Caracteres:** os segmentos são determinados pelo critério de tamanho ou do número de caracteres que contêm.

- **Ocorrências:** os segmentos são determinados por critério de tamanho e pelo número de ocorrências que contêm.

Nota: Os segmentos de texto são construídos a partir de um critério de tamanho e de pontuação. O Iramuteq busca a melhor razão entre tamanho/pontuação pela seguinte ordem de prioridade “.”, “?” “!” em primeiro, depois em segundo “;” e os dois pontos “:” em terceiro a

vírgula e por último o espaço. O objetivo é ter os segmentos com tamanhos homogêneos respeitando, o mais possível, a estrutura da língua.

- **Parágrafo:** a mudança de linha LF já indicada no OpenOffice Write é utilizada como um marcador de fim de texto. A pontuação nem sempre é tomada em conta.

Em “Tamanhos de segmentos de texto” selecione o número que definirá o tamanho em caracteres dos segmentos de texto. Por padrão deixe 40 (média de 6 a 7 palavras). Porém, se quiser que os pedaços de texto de análise sejam menores ou maiores, aumente ou diminua o número respectivamente. O tamanho dos segmentos de texto deve estar de acordo com o tipo de segmento escolhido.

2. Aba Limpando

Nessa aba é possível fazer algumas correções no *corpus* antes de submetê-lo à análise. No entanto, o ideal é que o *corpus* já esteja formatado conforme exigências da análise (veja Construção do *corpus*/Formatação).

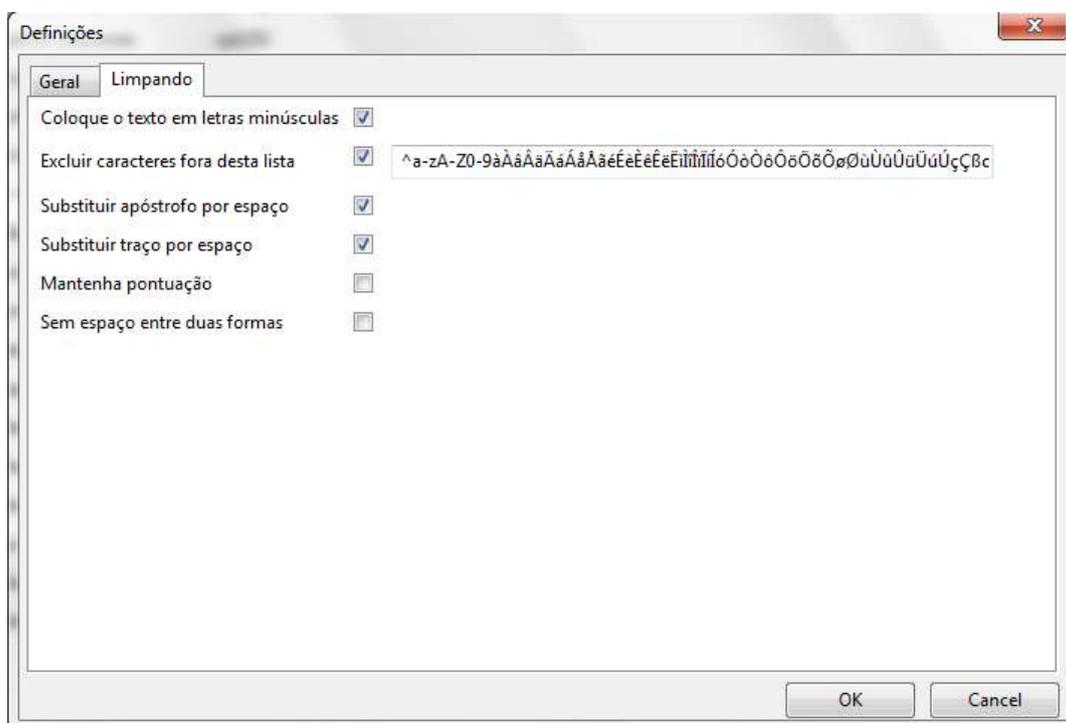


Fig. 10 – Importação de *corpus*: configuração Limpando

No entanto, é possível executar os seguintes procedimentos de limpeza:

Em “**Coloque o texto em letras minúsculas**”, converte todas as maiúsculas em minúsculas. Se estiver desativado, o Iramuteq considera “Gato” e “gato” como duas formas distintas.

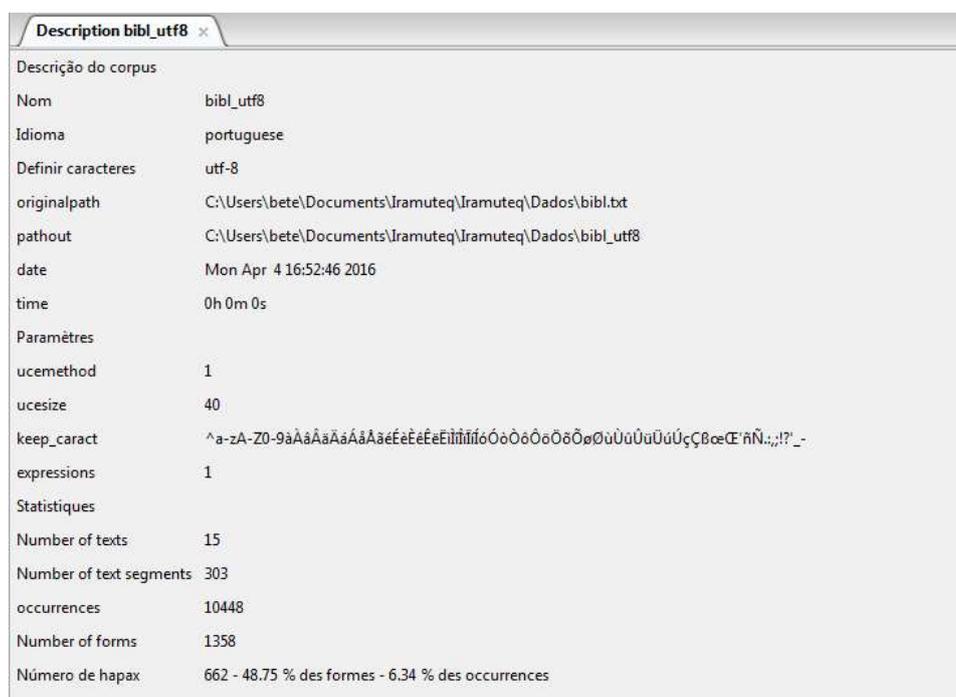
Em “**Excluir os caracteres fora dessa lista**”, selecione essa opção caso o *corpus* possa incluir caracteres não passíveis de reconhecimento. Por default, o Iramuteq considera todos os caracteres alfanuméricos, “ç”, til e os sinais de acentuação. No entanto, é possível eliminar ou acrescentar caracteres nessa lista, levando-se em conta que, para bom funcionamento do software, não se deve forçar a conservação do símbolo de aspas duplas.

Em “**Substituir apóstrofo por espaço**”, deixe selecionado para substituir as aspas simples por espaços em caso de haver esses sinais dentro do *corpus*.

Em “**Substituir traço por espaço**”, deixe selecionado para substituir traços por espaços.

Em “**Mantenha a pontuação**”, selecione essa opção se desejar conservar a pontuação do texto, exceto o ponto e vírgula que não deverá possuir dentro do *corpus*.

Em “**Sem espaço entre duas formas**”, selecione somente se desejar que os espaços entre as palavras sejam removidos.



Description bibl_utf8	
Descrição do corpus	
Nom	bibl_utf8
Idioma	portuguese
Definir caracteres	utf-8
originalpath	C:\Users\bete\Documents\Iramuteq\Iramuteq\Dados\bibl.txt
pathout	C:\Users\bete\Documents\Iramuteq\Iramuteq\Dados\bibl_utf8
date	Mon Apr 4 16:52:46 2016
time	0h 0m 0s
Paramètres	
ucemethod	1
ucesize	40
keep_caract	^a-zA-Z0-9àÁâÃäÅæÈéÊëËèÉèËìíîïóÔõÖöØøÙúÛüÜýÇβœƒřŃ.;:;!?'_-
expressions	1
Statistiques	
Number of texts	15
Number of text segments	303
occurrences	10448
Number of forms	1358
Número de hapax	662 - 48.75 % des formes - 6.34 % des occurrences

Fig. 11 – Resultado da Importação de *corpus*

Ele contém a configuração utilizada para importação do *corpus* e algumas estatísticas dessa importação como: número de textos; número de segmentos de texto; ocorrências; número de formas e número de hápax com as respectivas porcentagens nas formas e nas ocorrências.

Na abertura de um corpo ou matriz, o Iramuteq cria um relatório, Figura 11 e uma pasta no mesmo local que o arquivo aberto (ele nomeia o NomeDoCorpus_X). Este é o lugar onde ele vai armazenar os resultados a não ser que o usuário modifique.

VI – Introdução às Análises

Uma vez indexado o *corpus* na aba esquerda da janela do Iramuteq é adicionado o Nome arquivo_ *corpus* _X. As análises disponíveis podem ser vistas clicando-se na opção Análise de texto.

Existem 5 tipos de análises textuais que o Iramuteq processa: Estatísticas textuais, Especificidades e AFC, CHD, Análise de Similitude, Nuvem de Palavras, cada uma tem características e reflexões específicas.

As análises que podem ser realizadas dependem do tipo de dados: *corpus* textuais ou tabelas. Para *corpus* textual são possíveis as seguintes análises: Estatísticas textuais; Classificação Hierárquica Descendente; Análises de similitude; Nuvem de palavras; Análise de especificidades; e Análise fatorial de correspondência.

Já para tabelas são possíveis: Classificação Hierárquica Descendente; CHD por matrizes de distância; Análises de similitude; Nuvem de palavras; Descrição e qui-quadrado.

Para iniciar as análises, clique em “Analyse de texte”.

Tratamento comuns nas análises

Qualquer que seja a análise solicitada, o sistema apresentará o Menu de preferência das análises, Figura 12 que contém as seguintes opções:

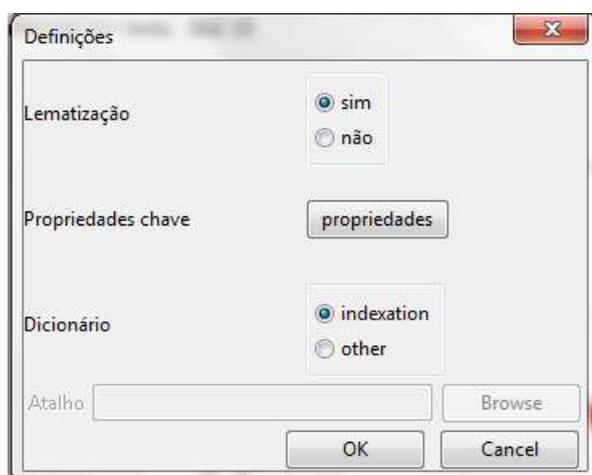


Fig. 12 – Preferências das análises

Lematização⁸: esta opção permite escolher se se deseja que o *corpus* seja lematizado ou não. A lematização é importante para o cálculo de ocorrências da palavra no *corpus*. Por exemplo, as diversas ocorrências do verbo consultar: *consultar*, *consulto*; *consultei*; *consultamos*; *consultaram* seria contabilizada como cinco ocorrências de *consultar* e não uma ocorrência de cada forma diferente.

Propriedades chave: permite escolher a classe gramatical das palavras a serem analisadas. Essa opção é importante para se eliminar palavras que não são chave para indexação do *corpus*, tais como preposições e artigos e outras que o usuário julgar necessárias. A maior parte das análises sobre o *corpus* vão diferenciar as palavras ativas, suplementares e eliminadas. Somente as formas ativas participam em certas análises como as classificações.

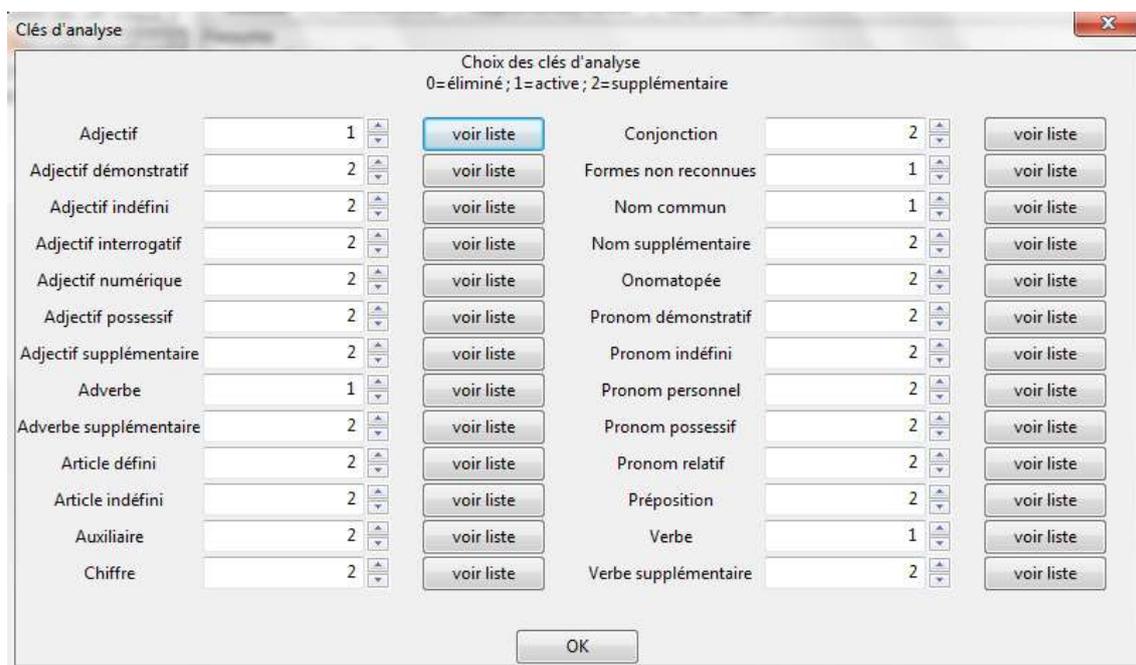


Fig. 13 – Preferências das Propriedades Chaves

- Clicando-se em Propriedades, acessa-se uma tabela que mostra todos os tipos gramaticais e a escolha *default* do sistema que o usuário pode alterar se considerar necessário. As palavras ativas estão marcadas com (1), as suplementares com (2) e as eliminadas com (0), Figura 13. Esses dados são armazenados no arquivo *key.cfg* dentro da pasta *Iramuteq*.

⁸Para significado de Lematização, veja o capítulo III – Nomenclatura utilizada neste documento.

- Escolha Voir Liste, Figura 14 para ver quais são as palavras da categoria escolhida e assim, escolher se deseja modificar as ativas, suplementares e eliminadas.

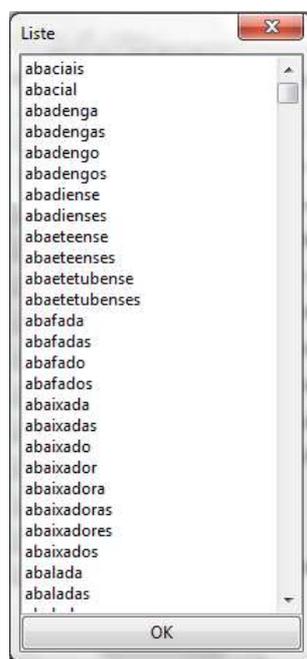


Fig. 14 – Preferências das Propriedades Chaves/Voir Liste

- Deixe selecionada a opção Indexação em Dicionário, para que o Iramuteq utilize o dicionário padrão da língua, possibilitando que seja feito o cruzamento entre as formas do dicionário e a do *corpus* analisado.

Dicas:

O conteúdo do texto pode apresentar resultados diferentes com relação ao desempenho do Iramuteq. Textos técnicos, artigos de revistas e jornais, por exemplo, possuem vocabulário diferente de entrevistas e questionários. O vocabulário contido no primeiro caso é técnico, limitado e mais fácil de interpretação pelo Iramuteq. Porém, os textos gerados por entrevistas e questionários, embora possam conter termos técnicos, misturam vocabulário informal e jargões, mesmo quando realizado com sujeitos de nível universitário. Sua classificação pelo Iramuteq pode gerar resultados insatisfatórios.

Algumas deficiências no dicionário de termos em português do Iramuteq também prejudicam as análises: palavras com tipo gramatical errado; verbos auxiliares classificados como verbos principais; palavras com conotações diferentes entre o dicionário e o texto; falta de lematização de palavras; e palavras inexistentes no dicionário. A maioria das palavras foi lematizada no dicionário apenas quanto à forma plural. Vários termos derivados foram

esquecidos. As palavras inexistentes nos dicionários, embora possam ser classificadas como ativas, não são contabilizadas corretamente por não serem lematizadas.

Além disso, observou-se que diferenças de grafia para mesma palavra, bem como os sinônimos para designar palavras importantes no *corpus* prejudicam os cálculos de frequência e qui-quadrado, impactando diretamente nos resultados das análises.

Por isso recomendam-se os seguintes procedimentos:

- a) Ao rodar as Estatísticas (capítulo VII a seguir) pode acontecer de algumas palavras sem importância recebam um escore alto e prejudiquem a análise do *corpus*. Nesse caso, é interessante rodar novamente o *corpus* eliminando-se os tipos gramaticais dessas palavras nessas listas (quando possível). Recomenda-se utilizar o 0, principalmente, para os artigos e preposições. Não é bom deixá-los como 2, pois muitas análises consideram as palavras suplementares e os artigos e preposições não tem qualquer significado para análise textual. No caso de outros tipos, pode-se escolher Voir Liste para ver quais são e determinar se devem permanecer como 0, 1 ou 2.
- b) Deve-se também examinar a lista de palavras ativas e suplementares geradas nas Estatísticas (capítulo VII). Observar os sinônimos, palavras com erro de grafia. Para resolver esse problema, corrija e importe o novo *corpus*. Em seguida rode novamente.
- c) Os erros e falhas do dicionário deveriam ser resolvidos de imediato. Por isso, seria importante que cada pesquisador tivesse acesso à atualização do dicionário localmente. Assim, o *corpus* poderia ser analisado novamente após sua atualização e, conseqüentemente, gerar análises mais fidedignas.

VII - Estatísticas textuais

Esta análise executa estatísticas simples sobre o “*corpus*” textual. Ele executa os seguintes procedimentos:

- identificação e reformatação das unidades de texto;
- identificação da quantidade de palavras, frequência média e hápax⁹;
- pesquisa no vocabulário e redução das palavras com base em suas raízes (formas reduzidas);
- criação do dicionário de formas reduzidas do *corpus*; e
- identificação das formas ativas e suplementares.

Em resultados, o Iramuteq 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.

Para executá-la, escolha **Análise de texto / Estatísticas** no menu horizontal do Iramuteq.

1. Parâmetros

Não há parâmetros para esta análise.

2. Resultados

Traz informações nas abas: Resumo; Actives Formes; Supplementary Formes; Total; e Hapax.

Na **aba Resumo**, encontra-se o diagrama de Zipf, Figura 15, que apresenta um gráfico que ilustra a distribuição de frequência das palavras no *corpus*. Traz também as seguintes informações:

- a) **Número de textos:** é o número de textos (registros) contidos no *corpus*. No caso, por exemplo, de artigos de jornal, seria o número de artigos. Já no caso de entrevistas seria o número de entrevistas realizadas.

⁹ Para significado de Hápax, veja o capítulo III – Nomenclatura utilizada neste documento.

- b) **Número de ocorrências:** é o número total de palavras contidas no *corpus*; difere se o *corpus* é lematizado ou não, porque quando ele é lematizado não inclui as variações das palavras.
- c) **Número de formas:** é o número de formas presentes no *corpus* (palavras ativas e suplementares). Também difere se este é lematizado ou não.
- d) **Número de hápax:** número de palavras que aparecem apenas uma vez em todo o *corpus*.
- e) **Média de ocorrências por texto:** (número de ocorrências) / (número de textos).

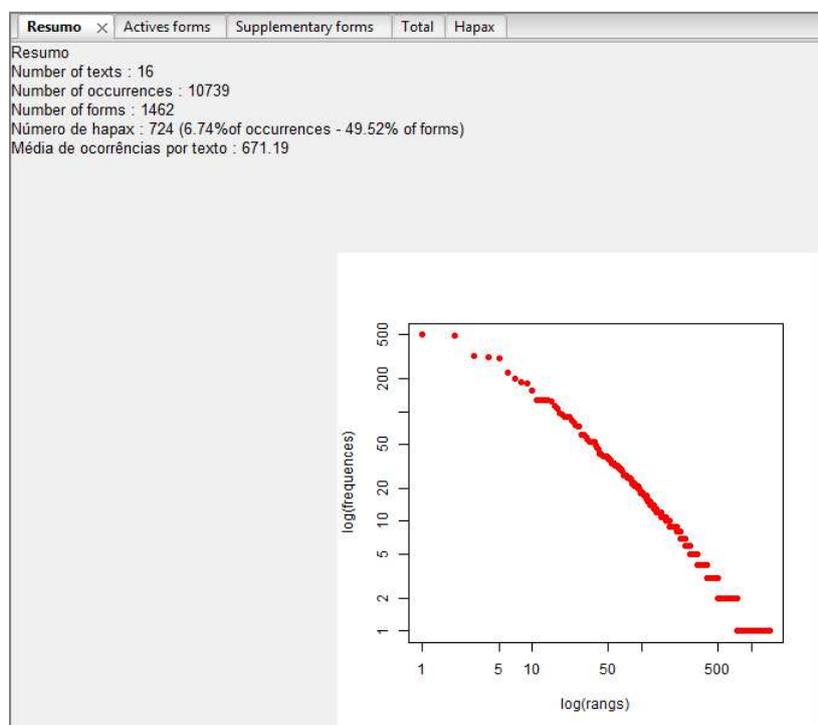


Fig. 15 – Resultado da Análise de Estatísticas Textuais

No gráfico temos presente no eixo das abscissas os logaritmos dos “pesos” (posição das frequências das palavras por ordem decrescente) e no eixo das ordenadas o das frequências das formas¹⁰.

¹⁰ Para significado de Lei de Zipf, veja o capítulo III – Nomenclatura utilizada neste documento.

Na **aba Actives Formes** estão listadas por ordem de ocorrência todas as palavras principais encontradas no *corpus*. São os verbos, adjetivos, advérbios, substantivos e também aquelas não encontradas no dicionário. A lista contém as palavras ativas (coluna Forma) com sua frequência de ocorrência (coluna Frequência) e suas categorias gramaticais coluna Tipo.

Na **aba Supplementary Formes** estão listadas por ordem de ocorrência todas as palavras suplementares encontradas no *corpus*. São aquelas cuja classe gramatical é suplementar (classificadas como 2 em Preferências das Propriedades Chaves, Figura 13).

Na **aba Total** estão todas as formas encontradas no *corpus*.

Na **aba Hapax** estão todas as formas cuja frequência foi igual a 1. Embora as listas estejam ordenadas por frequência de ocorrência, clicando-se no cabeçalho da tabela em Forma, Frequência ou Tipo, obtém-se a alfabetização pela ordem solicitada. Observe que para alfabetar é necessário clicar no cabeçalho da tabela não nas abas.

3. Análises estatísticas textuais suplementares

As opções seguintes permitem explorar a análise realizada, obtendo-se mais informação e/ou relatórios sobre as estatísticas realizadas no passo anterior.

a) Sobre cada forma

Um menu de opções suplementares fica acessível a partir de um clique com o botão direito do mouse sobre uma das formas, Figura 16.

Escolha a opção **Associated forms** para acessar os termos que estão associados a ela (plurais, palavras derivadas etc); e em **Concordância** para verificar onde elas ocorrem dentro dos segmentos de texto. Ambas as opções oferecem opção de salvar um arquivo html contendo os termos e/ou os segmentos de texto onde ocorrem.

Forma	Freq.	Tipos
a_z	1	nr
aberto	1	adj
abril	1	nom
abrir	6	ver
acabar	9	ver
acarretar	1	ver
acento	2	nom
acessa	2	nr
acessad		nr
acessam		nr
acessando	2	nr
acessar	6	nr
acesso	23	nom

Fig. 16 – Análise de Estatísticas Textuais suplementares: sobre a forma

b) Sobre análise

Na caixa da esquerda, onde estão listadas as análises efetuadas no *corpus* em questão, clique com o botão direito do mouse sobre Nome do *corpus_stat_x* para abrir um menu de opções, Figura 17. Nela pode-se escolher a opção desejada. São as seguintes:

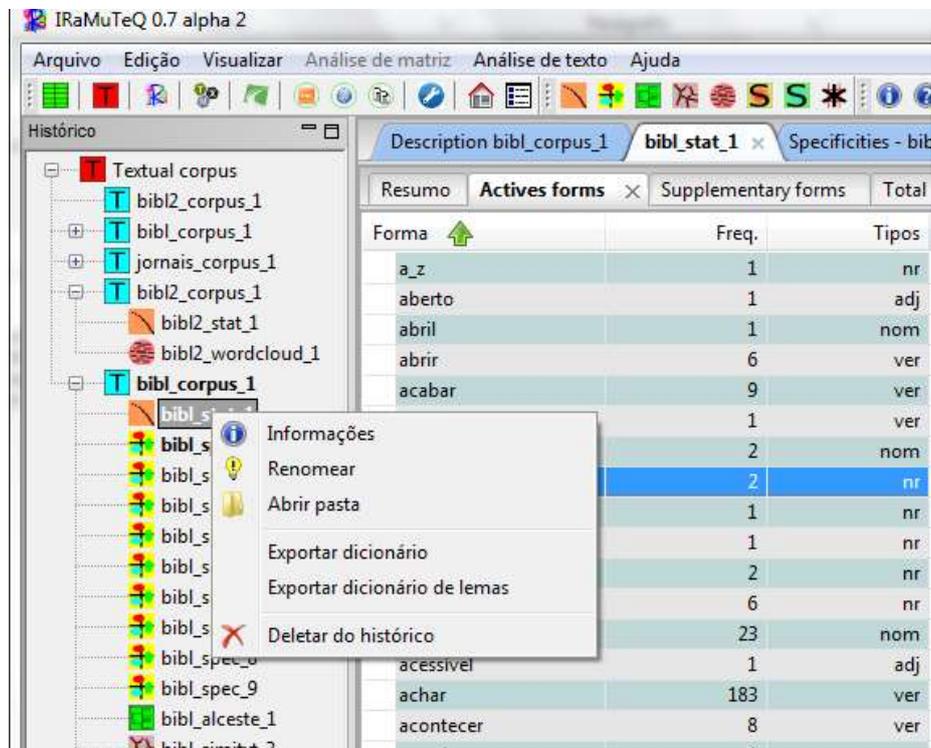


Fig. 17 – Análise de Estatísticas Textuais suplementares: sobre a análise

- **Informações:** mostra as configurações adotadas na análise.
- **Renomear:** permite renomear a pasta onde são armazenados arquivos e relatórios.
- **Abrir pasta:** abre a pasta onde se encontram armazenados arquivos e relatórios.
- **Exportar dicionário:** exportar o dicionário de formas que é colocado na mesma pasta dos resultados, possuindo o nome dictionary e extensão csv. Deve ser aberto com o Office Calc. Este dicionário já está no formato utilizado pelo Iramuteq para executar as análises.
- **Exportar dicionário de lemas:** exportar o dicionário de lemas¹¹ que é colocado na mesma pasta dos resultados, possuindo o nome lemmes e extensão csv. Deve ser aberto com o Office Calc e apresenta o termo principal e todas as variantes encontradas na mesma linha.
- **Deletar do histórico:** permite deletar o *corpus* e as análises já efetuadas.

4. Análises estatísticas textuais arquivos criados

O Iramuteq acrescenta uma nova pasta no diretório criado dentro da pasta do *corpus* e inclui os seguintes arquivos:

- **total.csv:** todas as formas com sua ocorrência e tipo gramatical.
- **formes_supplémentaires.csv:** todas as formas suplementares (receberam o código 2 durante a indexação), sua ocorrência e tipo gramatical. Por *default* são os pronomes, artigos, conjunções, preposições e outras formas suplementares.
- **formes_actives.csv:** todas as formas consideradas principais (receberam o código 1 durante a indexação), sua ocorrência e tipo gramatical. Por *default* são os substantivos, adjetivos, verbos e advérbios.
- **glob.txt:** resultados da análise.

¹¹ Para significado de Lema, veja o capítulo III – Nomenclatura utilizada neste documento.

- **hapax.csv**: todas as formas com frequência igual a 1, com tipo gramatical.
- **zipf.png**: o gráfico apresentado em Global em formato png que pode ser editado por programa gráfico que aceite abrir arquivos png.
- **Analyse.ira**: arquivo de uso do Iramuteq e permite abrir uma análise.

VIII - Especificidades e AFC

A análise de Especificidades associa textos com variáveis, ou seja, possibilita a análise da produção textual em função das variáveis de caracterização. Associam-se, ao *corpus*, variáveis que o pesquisador deseja analisar, de forma que a base de dados é dividida de acordo com a variável selecionada. Por exemplo, a comparação entre homens e mulheres em um questionário aplicado.

A Análise Fatorial de Correspondência¹² é uma representação gráfica dos dados para ajudar a visualização da proximidade entre classes ou palavras.

Os procedimentos executados nesta análise englobam o cálculo das frequências e os valores de correlação qui-quadrado de cada palavra do *corpus*, a partir da frequência pré-definida; e a execução da análise fatorial de correspondências (AFC) numa tabela de contingência que cruza as formas ativas e as variáveis.

1. Parâmetros

Depois de se escolher as opções no menu das Preferências, no qual deve-se seguir as orientações do tópico Tratamento comuns nas análises, no Capítulo VI, aparece o menu para escolha de variáveis, Figura 18.

- **Formas utilizadas:** permite que se execute a análise sobre as formas ativas e suplementares ou somente sobre as ativas ou somente sobre as suplementares.

- **Selecionado por:** permite escolher variáveis ou modalidades (valor assumido por determinada variável).

¹² Para significado de Análise Fatorial de Correspondências, veja o capítulo III – Nomenclatura utilizada neste documento.

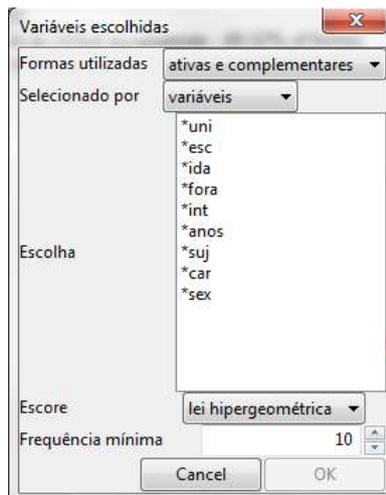


Fig. 18 – Análise de Especificidades e AFC: parâmetros

- **Escolha:** na caixa de escolha, deve-se marcar as variáveis e as modalidades desejadas. A seleção de uma variável é obrigatória, se validarmos sem selecionarmos previamente a variável uma mensagem de erro aparecerá.

Cada variável deve ter no mínimo 2 modalidades.

Pode-se escolher uma ou mais variáveis, ou então, selecionar quais modalidades desejadas. No entanto, se escolhermos a opção modalidade, todas devem pertencer a mesma variável, senão a análise não terá qualquer sentido.

Use a tecla CTRL para selecionar mais de uma variável ou modalidade.

- **Escore:** Duas possibilidades de cálculo são possíveis: o χ^2 ou a lei hipergeométrica¹³. Ambos os índices são usados para mostrar a probabilidade de existência da correlação entre as formas/palavras e as variáveis do *corpus*.

- **Frequência mínima:** só as formas onde o número de ocorrências no *corpus* for superior a este limiar serão consideradas. O *default* é 10, mas pode-se alterá-lo.

2. Resultados

Os resultados apresentados são mostrados na Figura 19 e descritos abaixo.

¹³ Para significado de Qui-quadrado e Lei Hipergeométrica, veja o capítulo III – Nomenclatura utilizada neste documento.

- a) **Formas e tipos:** na aba de Formas apresenta a lista das formas e na aba de Tipos traz a lista de categorias gramaticais e o score por cada modalidade da variável escolhida ou por cada modalidade escolhida de determinada variável.
- b) **Formas comuns:** traz a lista de formas classificadas pela frequência em que ocorrem com as variáveis, além dos scores de cada variável.
- c) **Forms frequencies:** traz a lista de formas/lemas e a frequência em que ocorrem com as variáveis.
- d) **Tipos de frequências:** traz a lista de tipos (categorias gramaticais) e a frequência em que ocorrem com as variáveis.
- e) **Frequência relativa das formas:** dá a frequência relativa das formas em permilagem (‰)¹⁴.
- f) **Tipos de frequências relativas:** dá a frequência relativa dos tipos gramaticais em permilagem (‰).

Formas	Formas comuns	Tipos	Forms frequencies				Tipos de frequências	Frequência relativa das formas	Tipos de frequências relativas	AFC
formas		*ida_30-39	*ida_40-49	*ida_50-59	*ida_m30					
eu		201	57	106	76					
que		185	82	108	68					
de		142	82	112	65					
ser		141	55	97	57					
a		128	63	113	78					
o		118	23	70	39					
não		114	29	80	40					
ter		96	40	60	47					
em		94	41	54	32					
por		68	25	53	35					
estar		64	36	45	14					
achar		62	26	46	35					
uma		50	29	24	10					
ir		49	26	22	18					
um		43	20	23	14					
para		41	17	40	9					
aquí		37	26	14	4					
mais		37	21	20	19					

Fig. 19 – Análise de Especificidades e AFC: resultados

- g) **AFC:** Apresenta o gráfico da AFC (fatores ½). Pode-se escolher nas abas a apresentação desejada: por forma ou por tipo gramatical, Figura 20.

¹⁴ Para significado de Permilogem, veja o capítulo III – Nomenclatura utilizada neste documento.

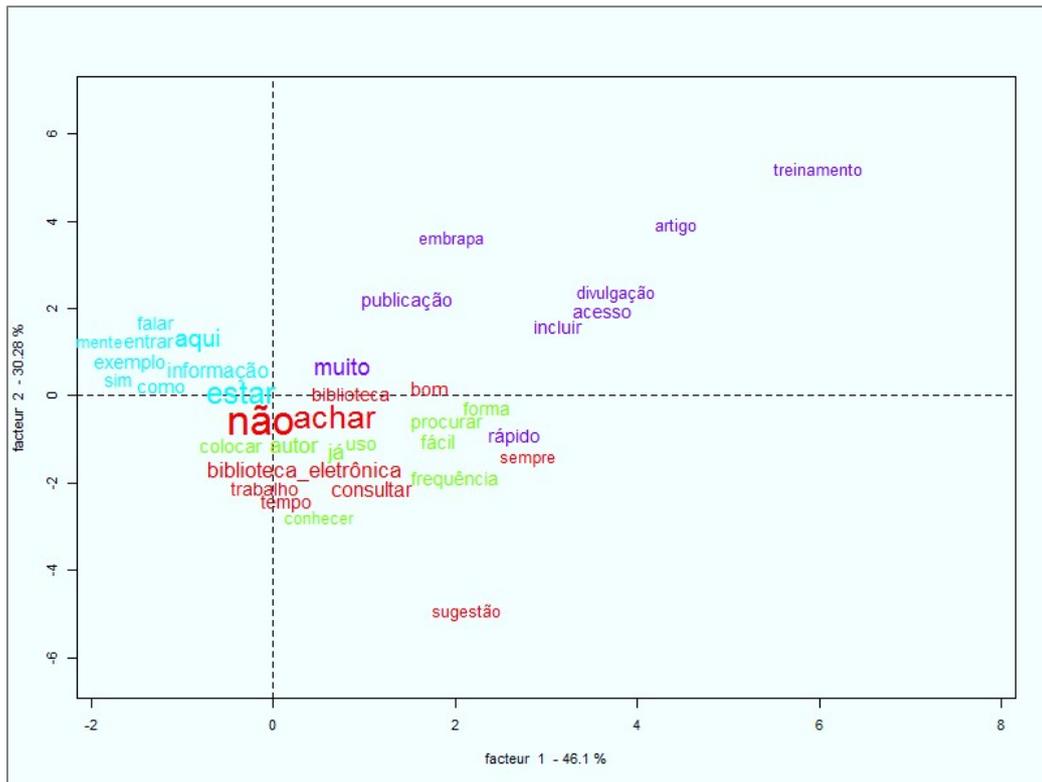


Fig. 20 – Gráfico da Análise Fatorial de Correspondência

No ícone de gráfico é possível escolher as configurações do gráfico (formato, variáveis, tamanho etc) e rerepresentá-lo, Figura 21.

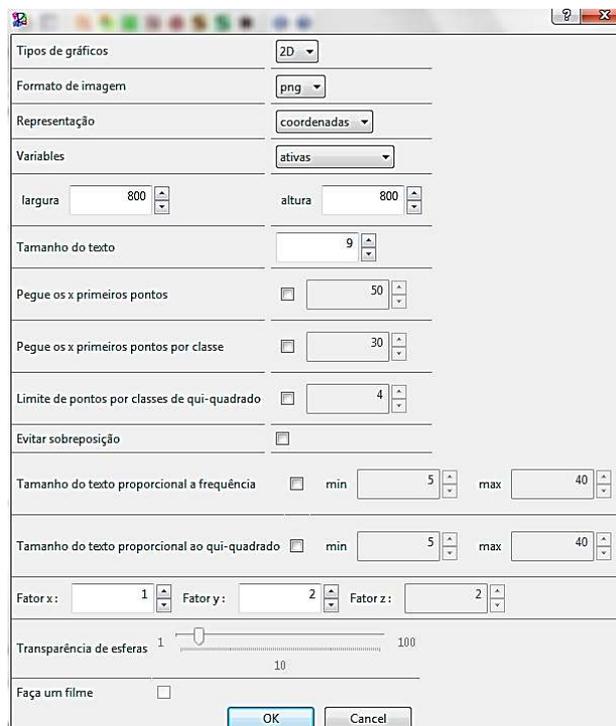


Fig. 21 – Gráfico AFC: parâmetros de configuração

Dica: é importante, principalmente, para gerar o gráfico com boa visualização na tela do computador, alterar o tamanho do gráfico que é em Pixels¹⁵. O tamanho indicado depende da resolução de tela do computador. Como a tela tem largura maior que altura, isso deve estar refletido na configuração adotada. Também é interessante escolher apenas as formas ativas (que são as mais importantes). Caso o gráfico ainda assim fique com muitas palavras de difícil visualização, pode-se limitar aos primeiros pontos (de maior frequência total ou na classe). Marque, também, a opção de evitar sobreposição.

Veja que a AFC é apresentada em 2 gráficos: o primeiro contém as formas e o segundo as modalidades das variáveis. Para analisar devemos fazer uma sobreposição.

3. Análises de Especificidades suplementares

As opções seguintes permitem explorar a análise realizada, obtendo-se mais informação e/ou relatórios sobre as estatísticas realizadas no passo anterior.

Um menu de opções suplementares, Figura 22, fica acessível a partir de um clique com o botão direito do mouse sobre uma das formas em qualquer uma das listas de especificidades; bem como nos segmentos de texto onde aparecem (os segmentos de texto são apresentados clicando-se duas vezes sobre a forma desejada); e no gráfico correspondente.

formas	eff 	*esc_esp	*esc_mes	*esc_phd	*esc_sup
não	263	0.8679	0.5589	-0.4285	-1.3016
achar	169	1.173	0.7656	-1.7665	0.9685
porque	106	-0.6287	-0.7449	0.7991	0.4451
mais	97	-0.2663	0.7231	-0.4947	-0.3265
saber	82	-0.4316	-0.5208	0.4208	0.6873
muito	68	-0.3252	0.3594	-0.685	1.3629
então	68	-0.7772	0.3594	0.3278	0.2334
coisa	63	-0.7196	-0.4114	0.8361	-0.3395
gente	61	-0.6965	-0.3677	0.5592	0.28
autor	58	0.3536	1.3758	-1.1825	-0.299
poster	57	-0.6505	0.6536	0.3017	-0.723
biblioteca eletrônica	56	1.8626	0.9208	-1.6045	-0.2832
só	56	0.126	0.2865	0.5592	-0.5698
vez	56	0.897	0.5124	-0.4257	0.4571
informação	56	0.514	-0.9683	1.2923	-0.5316
quando	56	0.05	1.1915	-1.0174	-0.5061
colocar	56	0.34	1.6287	-1.088	-0.4934
tipo_de_pul	56	0.56	-0.2862	-0.4769	-0.1459
...	25	0.2081	0.5845	0.2753	0.1045

- Associated forms
- Concordância
- Graphic
- Typical text segments
- Fazer Tgen

Fig. 22 – Lista de Especificidades das Formas: menu de opções suplementares

¹⁵ Para significado de Pixel, veja o capítulo III – Nomenclatura utilizada neste documento.

- **Associated forms:** mostra a lista de formas associadas (termos unificados pela lematização).

- **Concordância:** mostra os segmentos de texto onde a palavra aparece.

- **Graphic:** mostra o gráfico de escores da palavra para cada uma das variáveis escolhidas,

Figura 23.

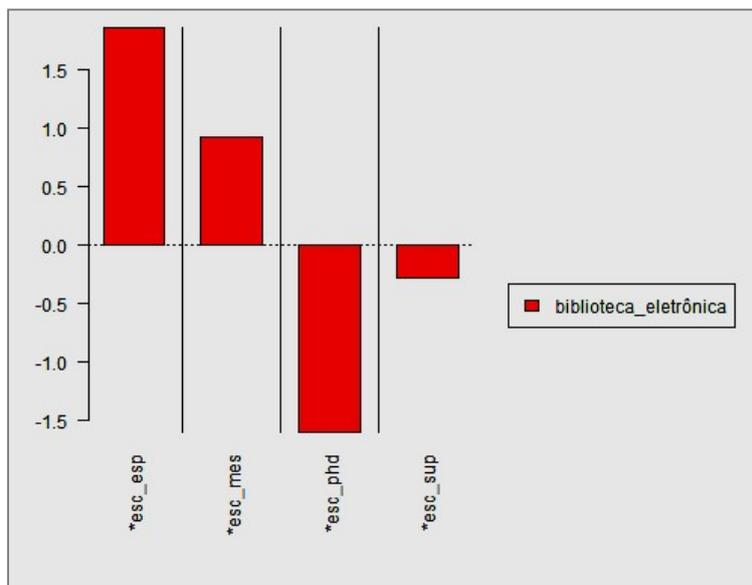


Fig. 23 – Lista de Especificidades das Formas: gráfico da forma Biblioteca_eletrônica para cada modalidade da variável escolaridade

- **Typical text segments:** mostra os segmentos de texto para a variável escolhida.

Nota: Apesar de ter sido efetuado um levantamento bastante completo das informações sobre o Iramuteq, consultando-se vários manuais e tutoriais em mais de uma língua, alguns parâmetros das análises não estão descritos em nenhum deles. Nesse caso, este manual também não os contempla.

4. Análises de Especificidades e AFC arquivos criados

São os seguintes os arquivos gerados na pasta de resultados pela Análise de Especificidades:

- **Tabletpem.csv:** Frequência dos tipos gramaticais por modalidade.

- **Tablespect.csv:** Escores das especificidades dos tipos gramaticais por modalidade.

- **Tablespecf.csv**: Escores das especificidades das formas por modalidade.

- **Tableafcm.csv**: Frequência das formas por modalidade.

- **liste_graph_afct.txt**: arquivo do Iramuteq.

- **eff_relatif_type.csv**: Frequência relativa aos tipos gramaticais por modalidade.

- **eff_relatif_forme.csv**: Frequência relativa às formas (palavras) por modalidade.

- **Analyse.ira**: Permite abrir a análise.

- **afct_row.png**: Plano fatorial dos tipos gramaticais.

- **afct_row.csv**: Detalhe dos valores das linhas da AFC sobre os tipos gramaticais.

- **afct_facteur.csv**: Valores próprios, porcentagens e porcentagens acumuladas de cada fator por tipo gramatical.

- **afct_col.png**: Plano fatorial das colunas d AFC sobre os tipos gramaticais.

- **afct_col.csv**: Detalhe dos valores da AFC sobre os tipos gramaticais.

- **afcf_row.png**: Plano fatorial das formas.

- **afcf_row.csv**: Detalhe dos valores das linhas da AFC sobre as formas.

- **afcf_facteur.csv**: Valores propostos, porcentagens acumuladas de cada fator para as formas.

- **afcf_col.png**: Plano fatorial da AFC sobre as formas.

- **afcf_col.csv**: Detalhe dos valores das colunas da AFC sobre as formas.

IX - Classificação pelo Método de Reinert

O método de Reinert propõe uma classificação hierárquica descendente segundo o método descrito por Reinert. Ele visa obter classes de segmentos de texto (ST) que, ao mesmo tempo, apresentam vocabulário semelhante entre si e vocabulário diferente das ST das outras classes. Esta análise é baseada na proximidade léxica e na ideia que palavras usadas em contexto similar estão associadas ao mesmo mundo léxico e são parte de mundos mentais específicos ou sistemas de representação. Nessa análise, os segmentos de texto são classificados de acordo com seu respectivo vocabulário e o conjunto de termos é particionado de acordo com a frequência das raízes das palavras. O sistema procura obter classes formadas por palavras que são significativamente associadas com aquela classe (a significância começa com o q -quadrado = 2).

Esta é uma das análises mais importantes do Iramuteq, nela o software, ao utilizar da lógica de correlação, utiliza as segmentações do *corpus* textual, juntamente com a lista de formas reduzidas e o dicionário embutido para apresentar um esquema hierárquico de classes. Ou seja, o Iramuteq, processa o texto de modo que possam ser identificadas classes de vocabulário, sendo assim, é possível inferir quais ideias o *corpus* textual deseja transmitir. Vale lembrar que essa análise é feita a partir de uma lógica estatística processada por computador e aplicada de forma lexical (OLIVEIRA, 2015).

Esta classificação é proposta segundo três modalidades:

- **Classificação simples sobre o texto:** os textos mantêm a sua integridade, a classificação permite o reagrupamento dos textos mais próximos;
- **Classificação simples sobre o segmento do texto:** a classificação incide sobre os segmentos de texto (ST);
- **Classificação dupla sobre os (RST):** A classificação é feita sobre duas tabelas nas quais as linhas não são mais segmentos de texto, mas reagrupamento de segmentos de texto (RST). O mesmo tratamento é, portanto, feito duas vezes, porém mudando o número de formas ativas para RST.

1. Parâmetros

Depois da escolha das opções desejadas no menu das Preferências, no qual deve-se seguir as orientações do tópico **Tratamento comuns nas análises**, no capítulo VI, são apresentados os seguintes parâmetros, Figura 24:

- **Classificação**: permite escolher o modo de classificação: simples sobre texto; simples sobre segmentos de texto e dupla sobre reagrupamentos de textos. A opção padrão é ST, exceto se o usuário tiver motivos para usar outra opção.
- **Tamanho e rst1 rst2**: para a classificação dupla sobre os RST permite variar o tamanho dos reagrupamentos dos segmentos de texto nas duas passagens.
- **Número das classes terminais da fase 1**: determina o número de classes da primeira parte da classificação. Por *default* ela é regulada em 10, mas deve ser ajustada de acordo com o tamanho e à heterogeneidade do *corpus*.

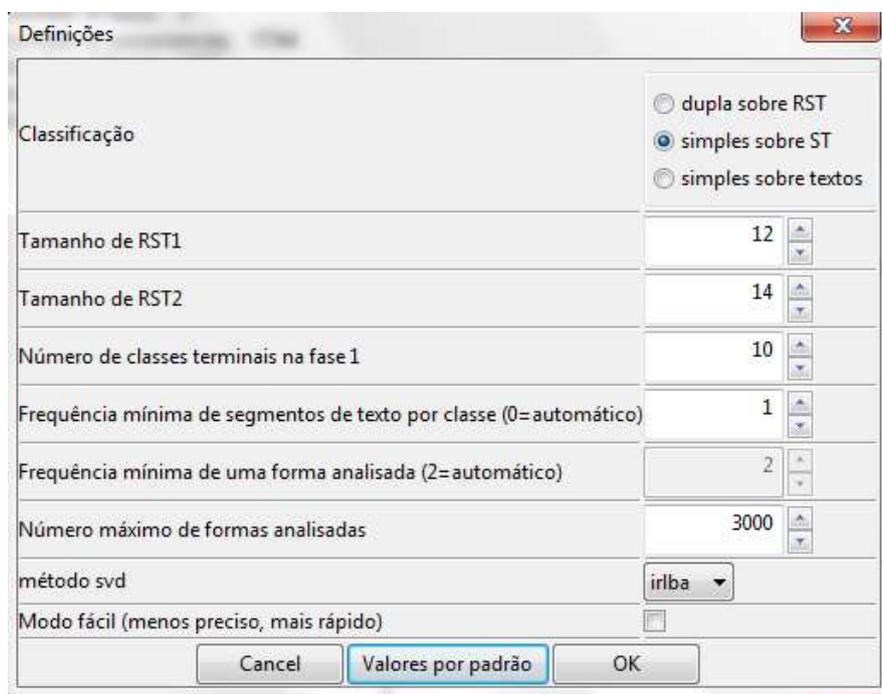


Fig. 24 – Classificação pelo Método de Reinert: parâmetros de configuração

- **Frequência mínima de segmentos de texto por classe:** determina um limiar mínimo de ST abaixo do qual as classes não serão selecionadas. Por *default* este número é igual ao número de segmentos de texto dividido pelo número de classes terminais no caso da classificação simples. E ao número de segmentos de texto, dividido duas vezes pelo número de classes terminais no caso da classificação dupla RST.

- **Frequência mínima de uma forma analisada:** esta opção não está disponibilizada.

- **Número máximo de formas analisadas:** por *default*, o Iramuteq assume as 3.000 formas ativas mais frequentes. As formas suplementares mais frequentes serão também consideradas. Uma forma deve ter no mínimo uma frequência de 3 para ser considerada. Se o *corpus* tem menos de 3.000 formas todas as formas com uma frequência igual ou superior a 3 serão tomadas em conta. Este parâmetro tem uma forte incidência sobre o tamanho das tabelas analisadas e quantidade de memória disponível no computador. Se o computador não tem capacidade para analisar um *corpus* de determinada dimensão devemos experimentar baixar este parâmetro. Se o computador possuir memória suficiente para a análise do *corpus* e este possuir mais de 3.000 formas com frequência igual ou maior de 3, nem devemos hesitar em aumentar esse parâmetro¹⁶.

- **Método SVD:** esta opção permite escolher o algoritmo usado na decomposição em valores singulares. As possibilidades são: irlba e svdR¹⁷.

- **Modo Fácil:** esta opção remove a segunda fase de cada partição. Permite uma análise um pouco mais rápida, mas menos precisa.

2. Resultados

Os resultados diretamente disponíveis apresentam um resumo da classificação (aba CHD) os perfis das classes (aba perfis) e uma análise fatorial das correspondências (AFC) realizadas sobre a tabela de contingência cruzando formas/lemas e classes (aba AFC).

Na aba CHD, primeira parte, Figura 25, o sistema apresenta uma lista de contabilizações realizadas sobre o *corpus*. Entre as quais se ressaltam:

¹⁶ Para saber o número de formas do *corpus*, consulte os resultados das Estatísticas textuais.

¹⁷ Veja mais detalhes sobre os métodos no capítulo III – Nomenclatura utilizada neste documento.

- **número de lemas:** difere do número de formas, pois os lemas são as formas lematizadas.

- **média de formas por segmento.**

- **número de segmentos classificados:** No caso presente, somente 58,87% dos segmentos foram classificados devido a escolha das categorias das palavras no menu de Preferências (primeiro menu apresentado nesta análise), bem como da escolha da forma de seleção dos segmentos de texto.

- **número de clusters:** número de classes.

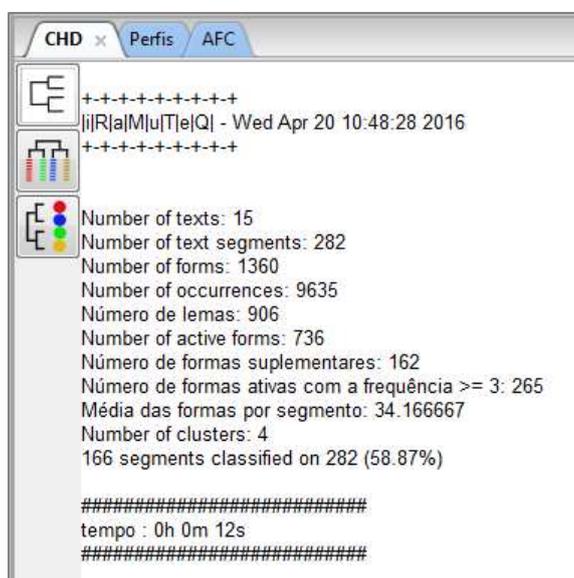


Fig. 25 – Resultado da Classificação pelo Método de Reinert: estatísticas textuais

A segunda parte é constituída pelo Dendrograma que apresenta a partição do *corpus* e a indicação do tamanho de cada classe em relação ao *corpus*, Figura 26.

Para terminar, o terceiro elemento é o Dendrograma da fase 1. As classes classificadas com “0” representam as classes que não foram consideradas, Figura 27.

Os ícones gráficos à esquerda do texto permitem gerar novamente o Dendrograma ou Filogramas, Figuras 28 e 29.

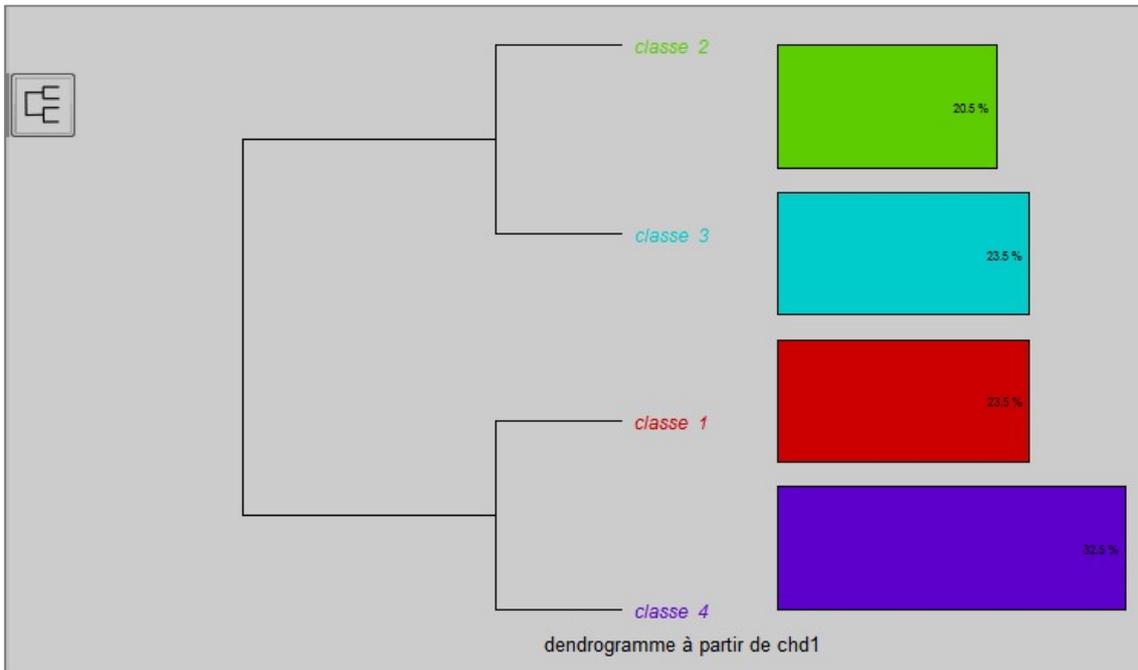


Fig. 26– Resultado da Classificação pelo Método de Reinert: Dendrograma

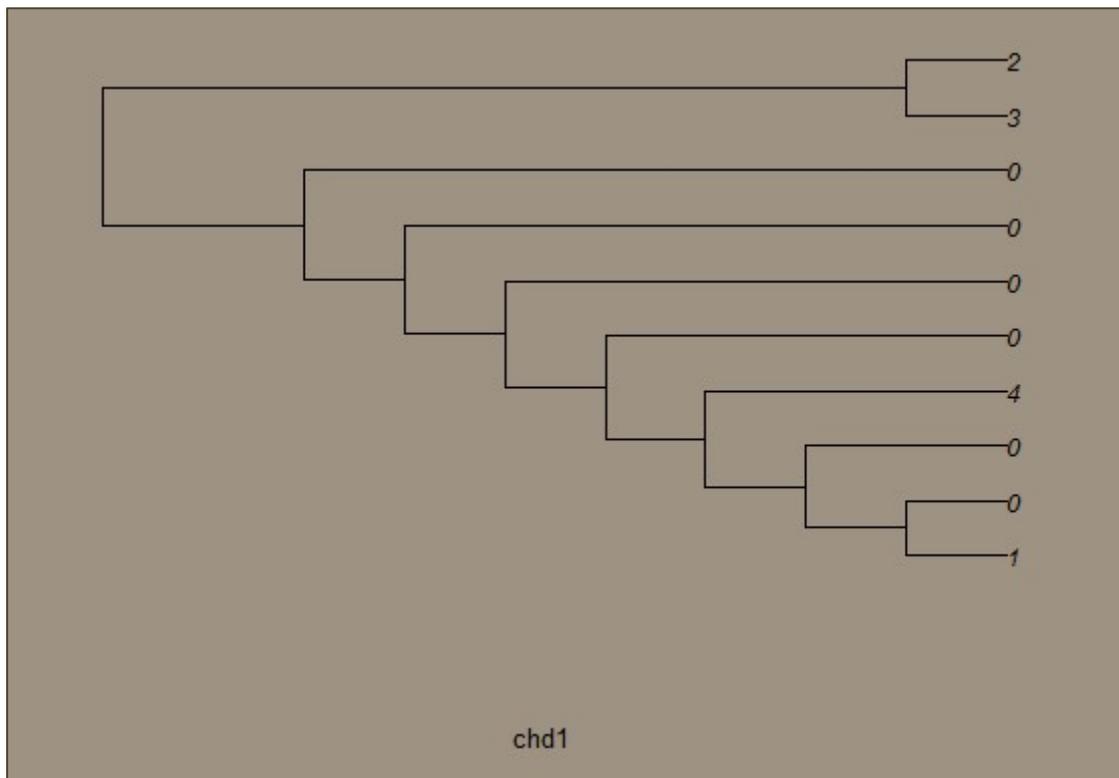


Fig. 27 – Resultado da Classificação pelo Método de Reinert: Dendrograma da Fase 1

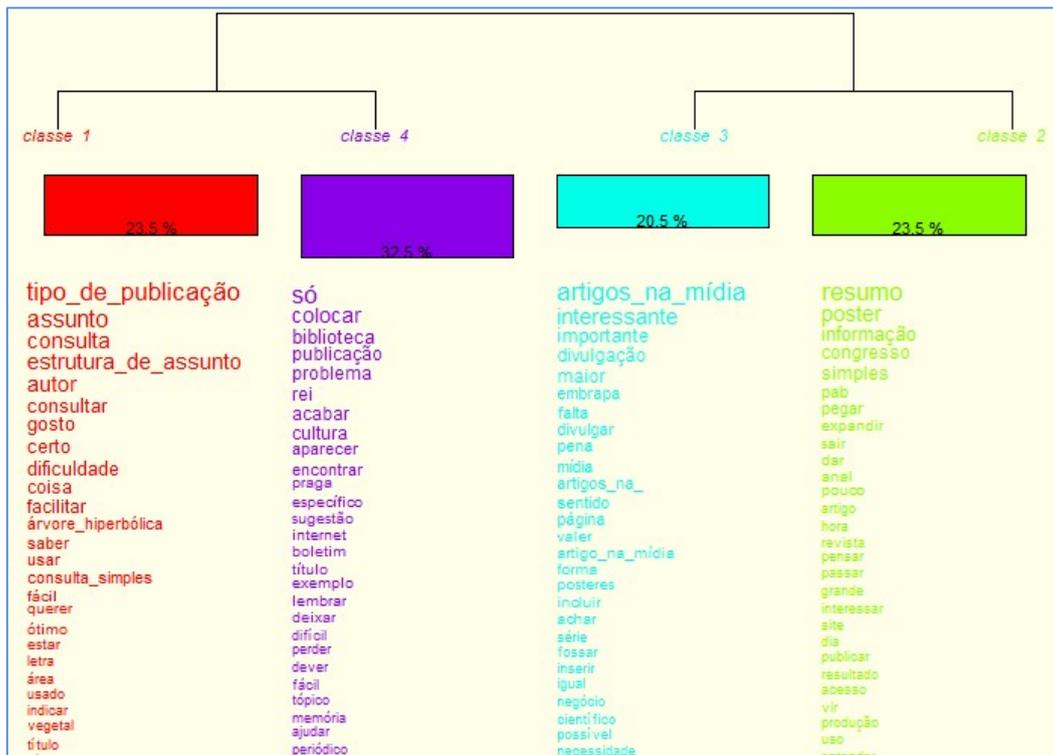


Fig. 28 – Resultado da Classificação pelo Método de Reinert: filograma

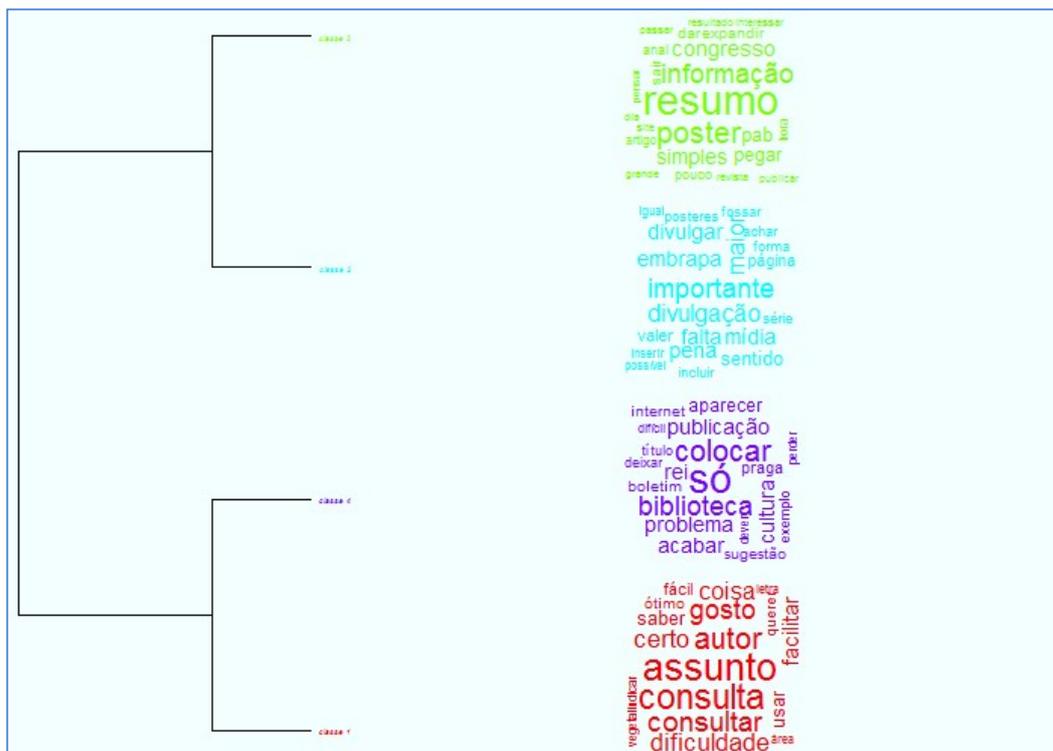


Fig. 29 – Resultado da Classificação pelo Método de Reinert: filograma

Na aba **Perfis**, Figura 30, encontram-se as tabelas de formas/lemas característicos das classes e que trazem as seguintes informações:

- **título da aba superior da tabela:** Contém o nome da classe, número de ST existentes no *corpus*; número de ST participantes da classe em questão; e a porcentagem de ST que fazem parte da classe em questão.

- **conteúdo da tabela:** os dados são apresentados em tabela de 3 cores, as formas ativas são apresentadas em cinzento, as formas suplementares em azul e as variáveis ilustrativas em rosa. No interior de cada uma destas categorias, as formas são apresentadas por ordem decrescente do valor do χ^2 e, portanto, da ligação com a classe:

- **num:** número da linha (permite ao utilizador encontrar a classificação).

- **eff. s.t:** frequência dos segmentos de texto que contém a forma na classe em questão.

- **eff. Total:** frequência total dos segmentos de texto em todo o "*corpus*" classificado que contém a forma.

- **Porcentagem:** Representa a % de segmentos de texto que contém a forma e aparecem nessa classe.

- **χ^2 :** Exprime a força de ligação entre a forma e a classe.

- **Tipo:** categoria gramatical da forma.

- **Forma:** forma ou lema próprio. Clicando-se em qualquer forma listada, obtém-se os ST (UCEs) da classe para a forma escolhida.

- **p:** nível de confiança associado (λ) ao χ^2 .

CHD		Perfis		AFC			
1 Classe 1	2 Classe 2	3 Classe 3	4 Classe 4				
39/166	39/166	34/166	54/166				
23.49%	23.49%	20.48%	32.53%				
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
10	3	3	100.0	6.34	nom	praga	0.01182
11	4	5	80.0	5.29	adj	especifico	0.02141
12	4	5	80.0	5.29	nom	sugestão	0.02141
13	4	5	80.0	5.29	nom	internet	0.02141
14	4	5	80.0	5.29	nom	boletim	0.02141
15	7	12	58.33	3.92	nom	título	0.04760
16	12	24	50.0	3.9	nom	exemplo	0.04824
17	6	10	60.0	3.66	ver	lembrar	NS (0.05578)
18	6	10	60.0	3.66	ver	deixar	NS (0.05578)
19	5	8	62.5	3.44	adj	difícil	NS (0.06364)
20	3	4	75.0	3.37	ver	perder	NS (0.06645)
21	8	15	53.33	3.25	ver	dever	NS (0.07135)
22	11	19	57.89	6.29	pro_per	ela	0.01214
23	5	7	71.43	5.04	pro_ind	qual	0.02479
24	21	48	43.75	3.87	adv	porque	0.04906
25	16	35	45.71	3.51	adv	então	NS (0.06090)
26	4	6	66.67	3.31	nom_s...	parte	NS (0.06906)
27	4	6	66.67	3.31	adv	atrás	NS (0.06906)
28	12	25	48.0	3.21	pre	com	NS (0.07322)
29	25	63	39.68	2.37	pre	por	NS (0.12395)
30	45	127	35.43	2.08	pro_per	eu	NS (0.14967)
31	34	80	42.5	6.99		*fora_m2	0.00818
32	3	3	100.0	6.34		*suj_10	0.01182
33	3	3	100.0	6.34		*esc_sup	0.01182

Fig. 30 – Resultado da Classificação pelo Método de Reinert: Perfis da Classe 4

Na aba AFC: traz o AFC derivado da classificação.

- Na aba AFC: mostra o gráfico AFC e o ícone opções do gráfico e de como o reajustar. A AFC realizada por meio da CHD representa, em um plano cartesiano, as diferentes palavras e variáveis associadas a cada uma das classes da CHD, Figura 31. Para reajustá-lo, consulte as informações na Figura 21 – Gráfico AFC: parâmetros de configuração.

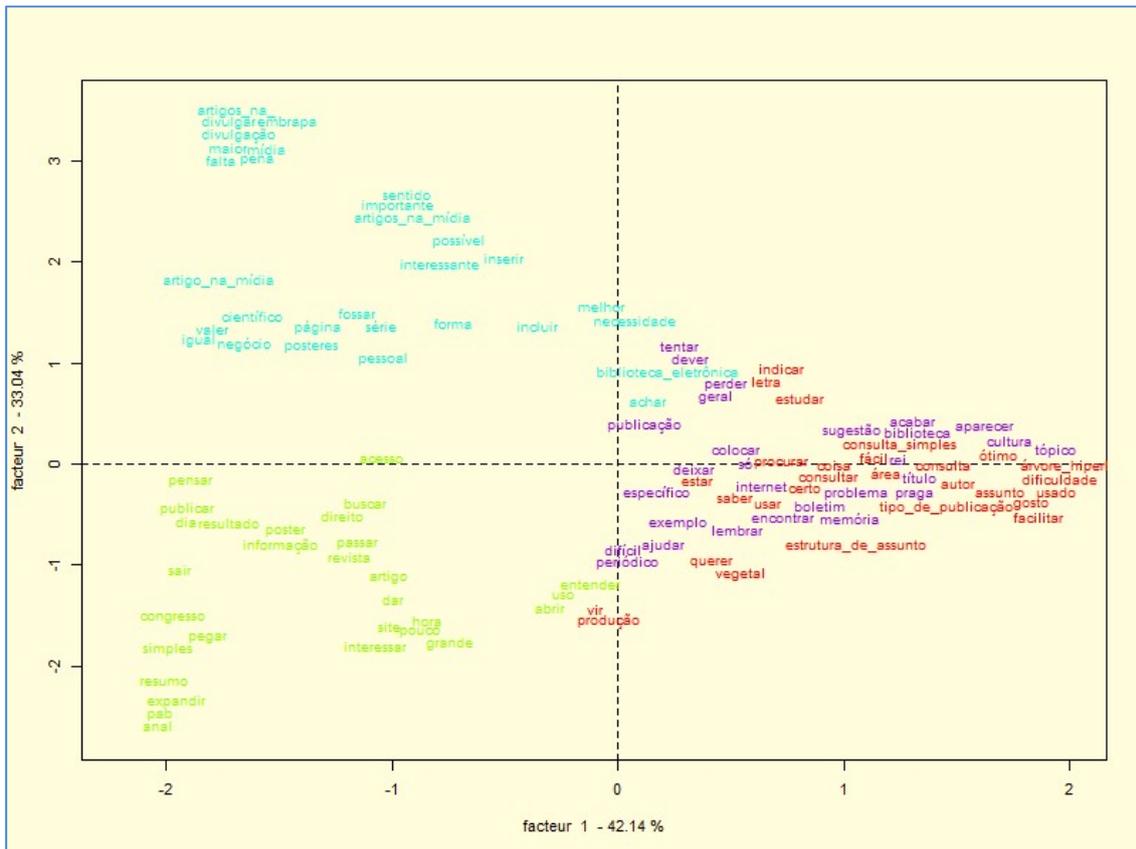


Fig. 31 – Resultado da Classificação pelo Método de Reinert: AFC

- Na aba Fator: mostra os valores próprios, percentagens e percentagens acumuladas de cada fator, Figura 32.

formes	Valeurs propres	Pourcentages	Pourcentage cumules
facteur 1	0.34755	42.13905	42.13905
facteur 2	0.27252	33.04224	75.18128
facteur 3	0.2047	24.81872	100

Fig. 32 – Resultado da Classificação pelo Método de Reinert: Fator

- Na aba Gráfico 3D: produz um gráfico de distâncias mínimas em formato tridimensional.

Notas:

Os gráficos 3D não estão funcionando nem com os arquivos UTF-8 e nem com os arquivos CP-1252. Aparece erro acusando a ocorrência de texto com caracteres não suportados na biblioteca rgl do programa R. Possivelmente são os sinais de acentuação, til e ç, porque quando se solicita por tipo gramatical (que não possui esses caracteres) o gráfico é executado. Talvez a biblioteca rgl não esteja apropriada para trabalhar com os caracteres especiais do português.

Não é possível em nenhum dos casos gravar um filme. O sistema gera vários frames do filme, porém ocorre erro na geração do arquivo gif. Ele informa que a unidade de gravação é inválida e o ImageMagick não foi encontrado. Traz, ainda, mensagens de aviso a respeito da versão do R ser a 3.2.3 e que está processando uma conversão. Talvez haja necessidade de complementação do R para gravação do filme.

3. Análises de Classificação suplementares

a) Opções disponíveis para cada classe:

Pode-se acessar o menu de análises complementares, clicando-se com o botão direito do mouse sobre a linha que contém a forma desejada na tabela de Perfis, Figura 33.

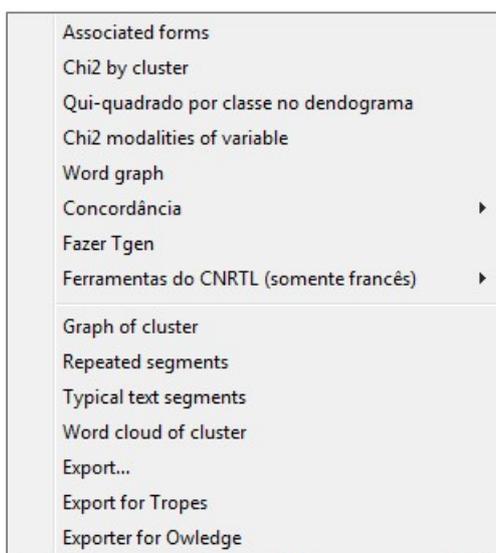


Fig. 33 – Método de Reinert: Perfis – Análise complementar

As opções de análises são as seguintes:

- **Associated forms:** mostra as frequências de cada forma agrupadas por lema em cada classe, Figura 34.



Fig. 34 – Método de Reinert: Perfis – Associated forms

- **Chi2 by cluster:** cria um gráfico, Figura 35, que mostra a associação da forma a cada uma das classes. Para ver gráfico formado por várias formas, selecione cada uma das linhas segurando a tecla CTRL.

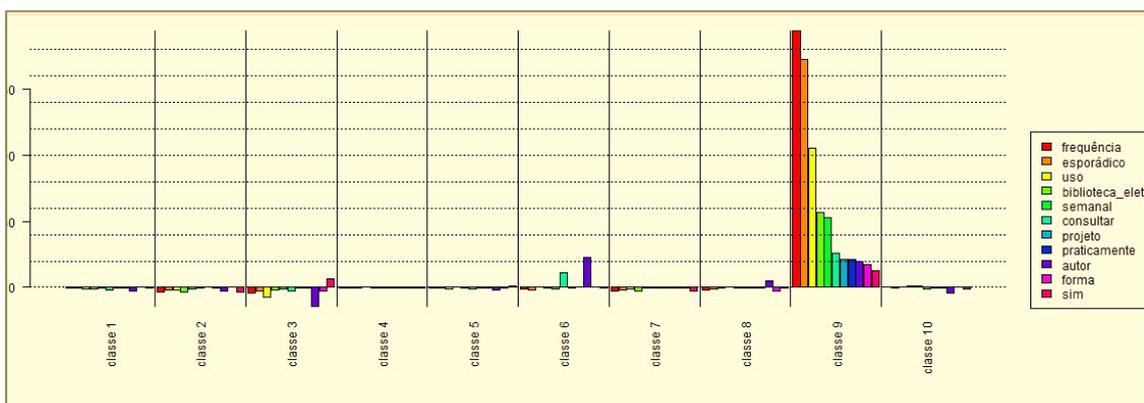


Fig. 35 – Método de Reinert: Perfis – Chi2 by cluster

- **Qui-quadrado por classe no dendrograma:** cria um gráfico que mostra a associação da forma a cada uma das classes, apresentado em formato diferente do Chi2 by cluster, Figura 36.

- **Chi2 modalities of variable:** ao se clicar em alguma variável (linhas rosa da tabela da aba Perfis), o Iramuteq cria um gráfico que representa o χ^2 de associação das diversas modalidades da variável selecionada com cada uma das classes, Figura 37.

- **Word graph:** cria um gráfico de similitude (semelhança) representando as ligações da forma selecionada com as outras formas da classe, Figura 38.

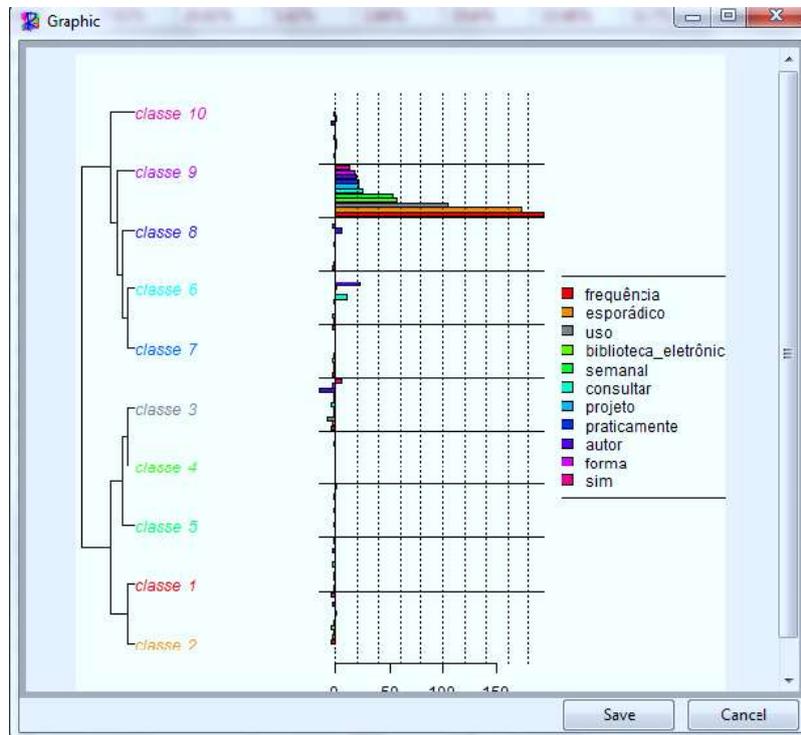


Fig. 36 – Método de Reinert: Perfis – Qui-quadrado por classe no dendrograma

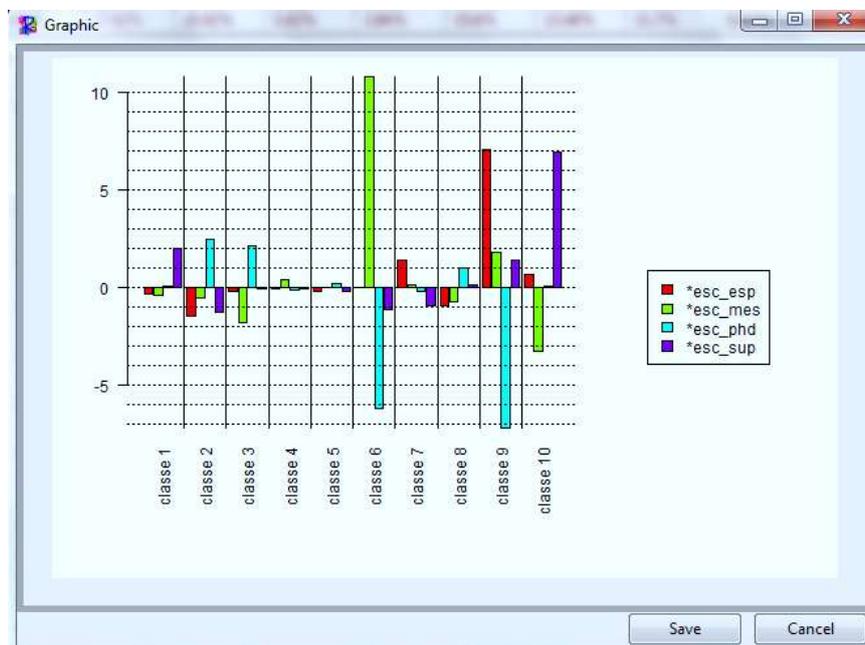


Fig. 37 – Método de Reinert: Perfis – Chi2 modalities of variable

- **Ferramentas do CNRTL:** busca na base de dados do Centre National Ressources Textuelles et Lexicales (<http://www.cnrtl.fr/>) à partir da forma/tema selecionado. Permite obter uma definição Lexicografica, dos sinônimos (Synonymie), dos Antônimos (Antonymie), a etimologia L'étymologie (Etymologie), a morfologia (Morphologie) ou um gráfico de proximidade da forma. Para executar esse procedimento é necessário acessar a Internet e funciona apenas para o francês. Os resultados são mostrados no navegador da internet existente no computador. Esta opção não faz sentido se o *corpus* não estiver em francês.

- **Graph of cluster:** independente da linha selecionada. Trata de uma análise de semelhança realizada sobre uma tabela de ausência/presença (0/1) que cruza as unidades escolhidas na linha (texto ou ST) e as formas ativas da classe na coluna. A matriz de semelhança é construída sobre as colunas (as formas ativas da classe), Figura 40.

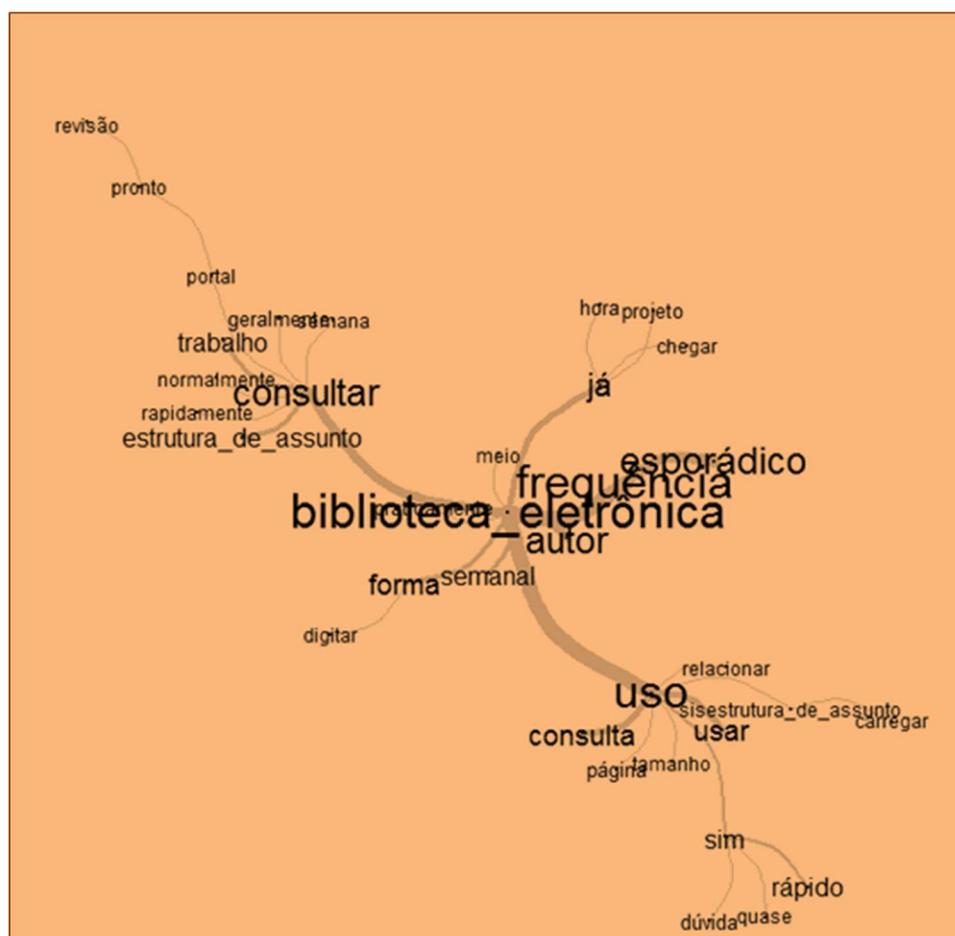


Fig. 40 – Método de Reinert – Graph of cluster para a forma Biblioteca eletrônica

- **Repeated segments:** independente da linha selecionada. Abre uma nova janela que mostra a frequência e o tamanho dos segmentos repetidos da classe, Figura 41.

Segmento	Frequência	Tamanho
da embrapa_cerrados	5	2
por exemplo	4	2
e eu	3	2
eu não	3	2
o artigo	3	2
os artigos_na_mídia	3	2
para a	3	2
site da	3	2

Fig. 41 – Método de Reinert – Repeated segments

- **Typical text segments:** independente da linha selecionada. Abre uma outra janela onde lista os ST e as características da classe, Figura 42. Dois valores são propostos:

. **absoluto:** os ST são classificados em função da soma dos χ^2 ligados à classe das formas ativas que os contém.

. **relativo:** os ST são classificados em função da média dos χ^2 ligados à classe das formas ativas que os contém.

Segmento	Score
**** *suj_06 *uni_cpac *car_pesq *sex_mas *ida_30-39 *esc_phd *anos_m2 *fora_2-5 *int_dia são artigos de pesquisadores que sairam na pab entendi ai eu vi que era uma coisas de anos 90 e poucos e não tinha mais pra cá ela é acesso livre o periódico tem direito	185.84
**** *suj_16 *uni_cpac *car_bol *sex_mas *ida_m30 *esc_mes *anos_m2 *fora_2-5 *int_dia sei quais os tipos_de_publicação que a biblioteca eletrônica possui eu gostaria de ter acesso pelo site da embrapa_cerrados aos artigos da pab quase todas as embrapas têm uma biblioteca_eletrônica mas elas não são iguais cada uma tem a sua	159.97
**** *suj_06 *uni_cpac *car_pesq *sex_mas *ida_30-39 *esc_phd *anos_m2 *fora_2-5 *int_dia porque às vezes o portal capes vai até um ano por exemplo até 95 e eu quero um artigo de 93 ai eu queria saber se a biblioteca tem esse periódico	143.08
**** *suj_06 *uni_cpac *car_pesq *sex_mas *ida_30-39 *esc_phd *anos_m2 *fora_2-5 *int_dia então foi uma dúvida que me surgiu também achei esse artigo aqui vi que você tem o resumo eu vi isso consolida informação também aí eu clicando aqui ele volta para a página da embrapa_cerrados e já vai direto	141.03
**** *suj_03 *uni_cpac *car_pesq *sex_mas *ida_30-39 *esc_phd *anos_6-10 *fora_m2 *int_dia o poster tem um pouco mais de informação que pode te ajudar a saber o que você quer exatamente se é aquilo mesmo eu sou favorável a incluir o poster tem os artigos pab	134.85
**** *suj_10 *uni_cpac *car_bol *sex_fem *ida_m30 *esc_sup *anos_m2 *fora_m2 *int_dia acesso online muito mais fácil porque é muito difícil às vezes você ir atrás do pesquisador e pedir o artigo ir na biblioteca pedir o artigo então é muito mais fácil você sentar no computador e procurar o artigo	134.71

Fig. 42 – Método de Reinert – Typical text segments

- **Word cloud of cluster:** Permite apresentar um gráfico contendo as formas existentes na classe escolhida, Figura 43.

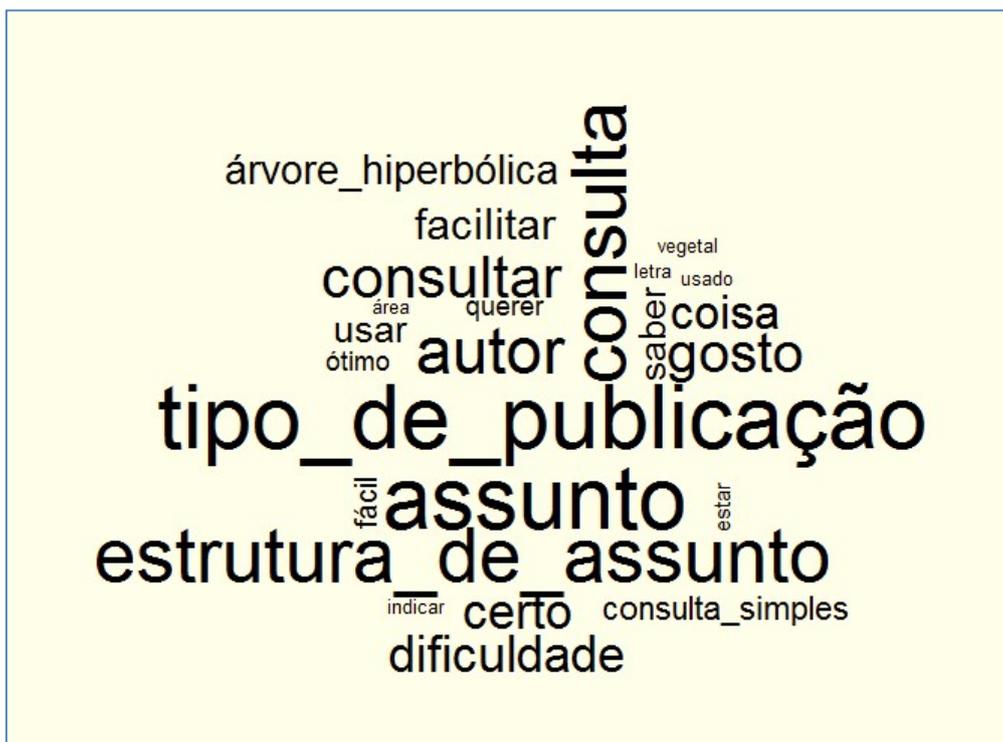


Fig. 43 – Método de Reinert – Wordcloud cluster

- **Export:** exporta todos os ST presentes na classe escolhida para um arquivo em formato txt que é gravado na pasta de relatórios. Esse arquivo poderá constituir um novo *corpus*. Poderá também ser utilizado para proceder à análise de conteúdo.

- **Export for Tropes:** exportação para o aplicativo Tropes Text Analysis.

- **Export for Owledge:** exportação para o aplicativo Owledge - Semantic-Knowledge's Built-To-Order Solution for Semantic Analysis.

b) Opções disponíveis pela classificação

É possível, também, obter outros relatórios e procedimentos no menu de Navegação, janela do lado esquerdo do Iramuteq. Clicando-se com o botão direito do mouse sobre o relatório do Dendrograma, obtém-se o menu apresentado na Figura 44.

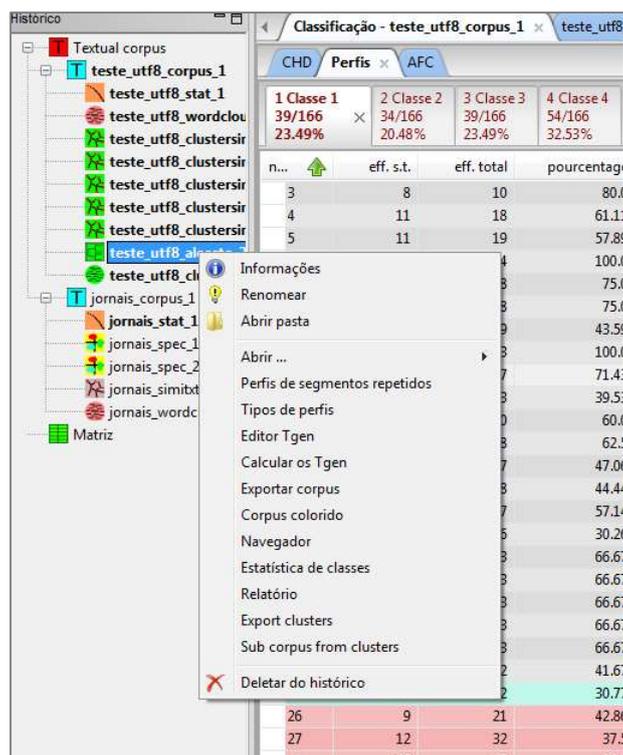


Fig. 44 – Método de Reinert – Outras funções da janela de navegação

- **Informações:** traz as características do *corpus*.
- **Abrir:** permite abrir os antiperfis (formas significativamente ausentes da classe), Figura 45.
- **Perfis de segmentos repetidos:** calcula o perfil dos segmentos repetidos. Diferentemente dos Perfis gerados por formas, esse procedimento é gerado por segmento.
- **Tipos de perfis:** Calcula os perfis dos tipos gramaticais de acordo com o Alceste ou com o Léxico, conforme escolha. Grava arquivo denominado `profil_type.csv` na pasta de resultados.
- **Exportar *corpus*:** esta opção permite exportar o *corpus* decomposto em segmentos de texto indexado pela variável estrela e pela classe que os contém. Os segmentos de texto que ficam à parte da classificação aparecem rotulados como `*classe_0`, Figura 46.

CHD Perfis AFC Perfis de segmentos repetidos Anti-perfis x								
classe 1 x	classe 2	classe 3	classe 4					
n...	↑	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0		2	38	5.26	-9.11		poster	0.00253
1		1	30	3.33	-8.28		informação	0.00400
2		0	19	0.0	-6.59		publicação	0.01026
3		0	18	0.0	-6.2		resumo	0.01277
4		0	13	0.0	-4.33		incluir	0.03741
5		0	12	0.0	-3.97		congresso	0.04625
6		0	11	0.0	-3.62		embrapa_cerrados	NS (0.05716)
7		1	18	5.56	-3.61		trabalho	NS (0.05728)
8		2	24	8.33	-3.59		colocar	NS (0.05820)
9		0	10	0.0	-3.27		artigo	NS (0.07065)
10		1	16	6.25	-2.93		artigos_na_midia	NS (0.08698)
11		0	9	0.0	-2.92		simples	NS (0.08736)
12		1	15	6.67	-2.6		dever	NS (0.10701)
13		0	8	0.0	-2.58		pegar	NS (0.10814)
14		0	8	0.0	-2.58		dificil	NS (0.10814)
15		0	8	0.0	-2.58		sair	NS (0.10814)
16		0	7	0.0	-2.24		posteres	NS (0.13411)
17		5	69	7.25	-17.34		*ida_30-39	< 0,0001
18		1	26	3.85	-6.62		*suj_05	0.01007
19		13	80	16.25	-4.51		*fora_m2	0.03373
20		1	20	5.0	-4.33		*suj_06	0.03751
21		3	28	10.71	-3.06		*anos_m2	NS (0.08022)
22		0	7	0.0	-2.24		*suj_11	NS (0.13411)

Fig. 45 – Método de Reinert – Antiperfis

```

**** *suj_02 *uni_cpac *car_pesq *sex_fem *ida_40-49 *esc_phd *anos_6-10 *fora_10-15 *int_dia *classe_0
a frequência de uso de biblioteca eletrônica ser semanal eu consultar por autor por titulo e tipo_de_publicação
árvore_hiperbólica nunca estrutura_de_assunto nunca consulta_simples que entrar com palavra chave também

**** *suj_02 *uni_cpac *car_pesq *sex_fem *ida_40-49 *esc_phd *anos_6-10 *fora_10-15 *int_dia *classe_1
o que que eu achar sobre rapidez facilidade ou dificuldade a vez eu entrar em tipo_de_publicação eu achar que eu não
saber se ser em outro época mas achar que o tipo_de_publicação poder dar um link de autor para a gente talvez

**** *suj_02 *uni_cpac *car_pesq *sex_fem *ida_40-49 *esc_phd *anos_6-10 *fora_10-15 *int_dia *classe_4
porque a vez eu ir buscar uma publicação minha eu saber que ela estar em boletim em hora que aparecer
tipo_de_publicação ou quando ir por autor e vice verso achar que uma coisa dever abrir para outro

**** *suj_02 *uni_cpac *car_pesq *sex_fem *ida_40-49 *esc_phd *anos_6-10 *fora_10-15 *int_dia *classe_1
se você procurar por autor poder ter alguma coisa que abrir para tipo_de_publicação ou em tipo_de_publicação alguma
coisa que te abrir para autor eu vir aqui por exemplo por tipo_de_publicação

**** *suj_02 *uni_cpac *car_pesq *sex_fem *ida_40-49 *esc_phd *anos_6-10 *fora_10-15 *int_dia *classe_1
ai eu querer mas ser onde eu ter a dificuldade a vez minha dificuldade ser a seguinte por exemplo ai eu vir aqui mas não
estar aqui ser onde eu estar bater a cabeça então

```

Fig. 46 – Método de Reinert – Exportação do corpus

- **Corpus colorido:** Esta função cria um arquivo *.html rescrevendo o *corpus* inteiro sendo que cada segmento de texto tem a cor da sua classe. Os ST não classificados são escritos em preto. Este arquivo com o nome *corpus_couleur.html* e fica disponível na pasta de relatórios, Figura 47.

podem ter acesso a essas informações de formas mais rápidas e interessante

e principalmente para você ter esse arquivo daquilo e também p

e também ver o que é que está se publicando com que qualidade esses artigos_na_mídia

porque às vezes não tem necessidade de você estar fazendo tem acho muito interessante

não eu acho que inclusive para cada trabalho que a gente leva p

fica perdido e se depender de você consultar anais de congresso

leva um poster nem que seja uma impressão resumida daquele p

inclusive eu estou atrás agora de um poster que saiu de uma tec

o povo está atrás disso e eu não consigo localizar nem poster ne biblioteca_eletrônica

eu posso consultar para saber se teve alguma coisa por ocasião o papel está armazenado aqui eletronicamente

Fig. 47 – Método de Reinert – Exportação em html colorido

- **Navegador:** abre uma nova janela contendo uma matriz que traz todas as formas com o valor qui-quadrado de ligação a cada classe, Figura 48.

formas	classe 1	classe 2	classe 3	classe 4
0 achar	-0.132	5,64	-8.816	0.941
1 estar	3.574	-4.617	-0.465	0.573
2 saber	8.308	-4.453	-0.212	-0.14
3 coisa	11.453	-3.273	-3.231	0.015
4 gente	0.978	-0.082	-1.371	0.17
5 autor	15.894	-5.2	-3.614	0.006
6 poster	-9.113	1.03	37.6	-13.627
7 biblioteca_eletrônica	-0.018	2.047	-3.614	0.372
8 consultar	14.126	-1.306	-2.007	-1.288
9 só	-1.15	-6.969	-2.379	21.534
1 vez	0.862	-1.149	0.205	-0.107
1 informação	-8.28	0.183	32.334	-8.469
1 exemplo	-0.728	-4.585	0.502	3.901
1 entrar	0.78	-0.29	-0.709	0.182
1 colocar	-3.588	-1.098	-1.887	14.896
1 tipo_de_publicação	22.788	-4.247	-2.598	-1.18
1 usar	7.89	-0.718	-0.072	-2.461
1 estrutura_de_assunto	18.903	-2.741	-0.072	-5.13
1 querer	4.93	-5.2	1.087	-0.977
1 ficar	0.201	-0.73	0.978	-0.319
2 procurar	1.467	-0.884	-1.45	0.645
2 uso	0.005	-1.056	1.602	-0.106
2 consulta	19.091	-1.172	-3.972	-1.483
2 fácil	5.851	-2.479	-5.816	1.821
2 dar	-0.185	-0.939	10.561	-2.949
2 bom	0.888	1.672	-5.064	0.005
2 publicação	-6.589	0.004	-0.709	9.17

Fig. 48 – Método de Reinert – Navegador

- **Estatística por classe:** esta função cria um arquivo csv na pasta de relatórios (stat_par_classe.csv), contendo estatísticas de cada uma das classes: o número total de ocorrências; o número de formas diferentes; o número de hápax; o número de segmentos de texto; e a taxa do número de hápax pelo número de formas, Figura 49.

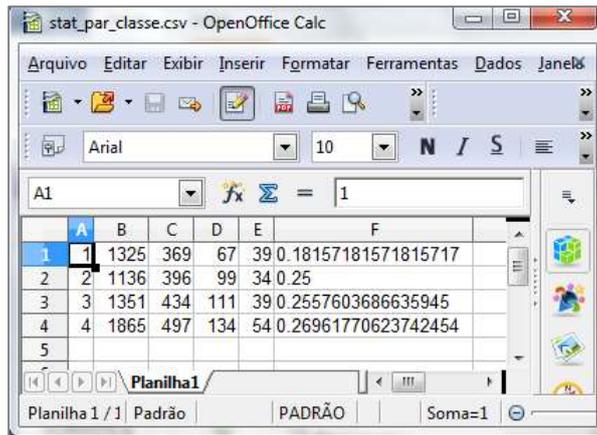


Fig. 49 – Método de Reinert – Estatística por classe

- **Relatório:** esta função cria um ficheiro RAPPORT.TXT, que contém as características gerais da classificação e os perfis, Figura 50.

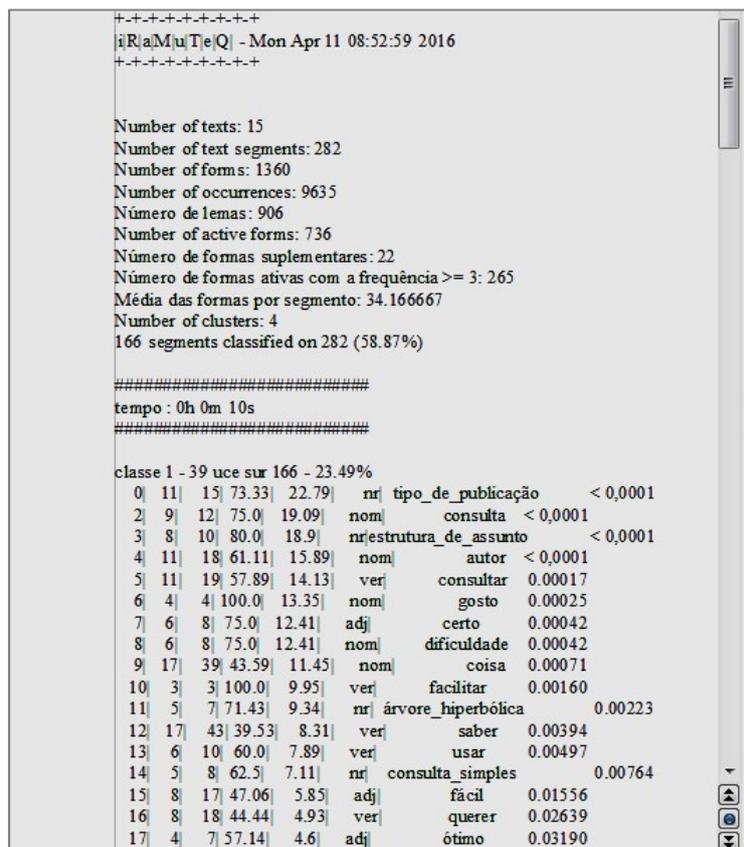


Fig. 50– Método de Reinert – Relatório

- **Deletar do histórico:** Elimina a análise selecionada do histórico do Iramuteq.

4. Análises de Classificação arquivos criados

Arquivos produzidos pela Análise Classificação são:

- **AFC2DCL.png**: Gráfico AFC : Classes - Coordenadas- fator ½.
- **AFC2DEL.png**: Gráfico AFC : Variáveis ilustrativas - Coordenadas - fator ½.
- **AFC2DL.png**: Gráfico AFC : Variáveis ativas - coordenadas - fatores ½.
- **AFC2DSL.png**: Gráfico AFC : variáveis suplementares - coordenadas- fatores ½.
- **afc_col.csv**: Resultados da AFC; Coordenadas, correlação, MASS, contribuição das classes.
- **afc_facteur.csv**: Resultados da AFC; Valores próprios, % de inércia e percentagem acumulada dos fatores.
- **afc_row.csv**: Resultados da AFC; coordenadas, correlação, MASS, contribuição das formas.
- **Analyse.ira**: Arquivo do Iramuteq, permite abrir uma análise.
- **antiprof_segments.csv ***: Antiperfis dos segmentos repetidos.
- **antiprof_type.csv ***: Antiperfis dos tipos gramaticais.
- **Antiprofile.csv**: Antiperfis das formas.
- **Arbre_1.png**: Dendrograma da primeira CHD.
- **Chistable.csv**: χ^2 da associação de cada forma às classes.
- **Classe_mod.csv**: Tabela de contingência formas ativas.
- **Corpus_couleur.htm***: *corpus* com a respetiva cor da classe a que pertencem.

- **Dendro1.png**: Dendrograma da classificação.

- **Export_corpus.txt***: Arquivo criado com a opção de exportação do *corpus*.

- **Info.txt**: Resumo da análise de classificação.

- **Liste_graph_afc.txt**: arquivo do Iramuteq.

- **Liste_graph_chd.txt**: arquivo do Iramuteq.

- **ListeUCE1.csv**: arquivo do Iramuteq.

- **Prof_segments.csv***: Perfis dos segmentos repetidos.

- **Profil_type.csv***: Perfis dos tipos gramaticais.

- **Profiles.csv**: Perfis das classes.

- **Ptable.csv**: Nível de significância da associação do χ^2 de cada uma das formas às classes.

- **RAPPORT.txt***: Arquivo criado pela opção “Rapport”.

- **sbyClasseOut.csv**: Arquivo do Iramuteq.

- **Segment_classe.csv***: Tabela de contingência dos segmentos repetidos/classes.

- **Stat_par_classe.csv***: Arquivo criado pela opção Estatística por classe.

- **Tablet.csv**: Tabela de contingência das variáveis ilustrativas/classes.

- **Tablesup.csv**: Tabela de contingência das formas suplementares/classes.

- **TableUC1.csv**: Matriz utilizada para análise no formato Matrix Market (repositório de dados de teste de matriz para utilização em estudos comparativos de algoritmos).

- **Type_cl.csv***: Tabela de contingência dos tipos gramaticais/classes.

- **Uce.csv**: Posicionamento dos segmentos de texto por classes.

X - Análises de similitude

A Análise de similitude é baseada na teoria dos grafos¹⁸ cujos resultados auxiliam no estudo das relações entre objetos de um modelo matemático.

No Iramuteq, a análise de similitude mostra um grafo que representa a ligação entre palavras do *corpus* textual. A partir desta análise é possível inferir a estrutura de construção do texto e os temas de relativa importância, a partir da coocorrência entre as palavras.

Ela auxilia o pesquisador na identificação da estrutura da base de dados (*corpus*), distinguindo as partes comuns e as especificidades, além de permitir verificá-las em função das variáveis descritivas existentes.

Esta análise produz gráficos a partir da biblioteca Igraph do R. A tabela de entrada é uma tabela de presença/ausência. A matriz de similitude é calculada a partir de um dos escores escolhidos. A maior parte dos escores propostos estão contidos na biblioteca (*proxy*) do R.

A escolha de gráfico dinâmico ao invés do estático possibilita sua edição direta para melhorar a apresentação sem necessidade de outro aplicativo. No entanto, para gerar um gráfico mais elaborado com cores e demarcações, é necessário exportá-lo para o formato graphml e, então editá-lo com o software Gephi (<https://gephi.org/>) que também opera com a interface R.

1. Parâmetros

Ao escolher a análise de similitude, surge o menu de Preferências cujas orientações estão no tópico **Tratamento comuns nas análises** do capítulo VI. Depois de se escolher as opções no menu de Preferências, aparece a janela de opções, Figura 51.

Este menu é composto de três partes: Lista das formas; Configurações Gráficas e Ajustes Gráficos.

A **Lista das formas** e as suas frequências é apresentada no quadro da esquerda. Por *default* estão todas selecionadas, mas o usuário pode selecionar as formas desejadas, segurando-se a tecla CTRL enquanto clica com o mouse.

¹⁸ Para significado de Grafo, veja o capítulo III – Nomenclatura utilizada neste documento.

Na aba de **Configurações gráficas**, pode-se editar a análise, trocar o índice de coocorrências por algum outro, escolher se será uma árvore máxima ou não, etc.

Na aba **Ajustes gráficos**, 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. A seguir, apresenta-se cada um dos parâmetros passíveis de configuração.

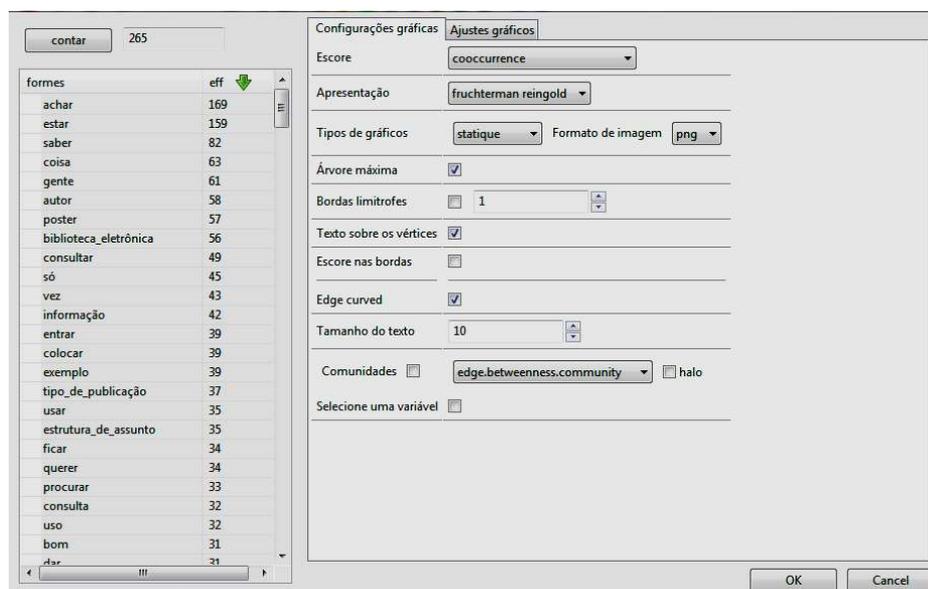


Fig. 51 – Análise de Similitude - Parâmetros de configurações gráficas

Configurações gráficas: Esta aba fica ativa logo desde a abertura do menu. Ela nos permite escolher as características do gráfico:

- **Escore:** permite escolher o tipo de índice utilizado (coocorrência por *default*)
- **Apresentação:** modo de apresentação dos dados. As opções são aleatórias random, cercle, Fruchterman Reingold, Kamaad-Kawai, graphot¹⁹.
- **Tipo de Gráfico:** Escolha do método de apresentação do gráfico. Estático: Produz uma imagem com o formato png ou svg. Dinâmico: utiliza a interface tk do gráfico. Esta interface permite editar o gráfico no próprio aplicativo, reorganizando os vértices, girando e

¹⁹ Veja a documentação da biblioteca igraph do R para mais detalhes.

centralizando, mudando cores dos vértices e etc. 3D: produz um gráfico a três dimensões que se abre numa janela “rgl”. Veja nota no Capítulo IX item 2 Resultados.

- **Formato de imagem:** permite salvar a imagem estática em dois tipos de formatos png ou svg que podem ser editados pelos programas de edição de imagem tais como Corel Photopaint ou Adobe Photoshop.

- **Árvore máxima:** calcula a árvore máxima.

- **Bordas limítrofes:** permite colocar ou eliminar as arestas para as quais o valor do índice é inferior ou igual ao limite escolhido.

- **Texto sobre os vértices:** afixa o texto sobre os vértices dos gráficos.

- **Texto sobre as arestas:** afixa o texto sobre as arestas do gráfico.

- **Edge curved:** cantos arredondados.

- **Tamanho do texto:** indicar o tamanho da fonte desejada.

- **Comunidades:** Permite o cálculo das comunidades²⁰.

- **Halo:** representa as comunidades por um halo de cor.

- **Selecione uma variável:** depois de validado abre um novo menu com as variáveis (ou várias modalidades), Figura 52. O grafo será restrito às variáveis escolhidas e as formas relativas às modalidades serão ressaltadas com cor.

²⁰ Veja a documentação da biblioteca igraph do R para mais detalhes.

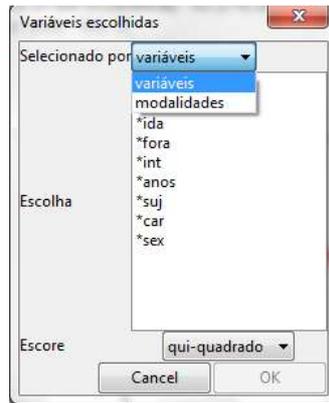


Fig. 52 – Análise de Similitude - Parâmetros de configurações gráficas

Ajustes gráficos, Figura 53, é a segunda aba do menu onde é possível escolher as características visuais do gráfico.

- **Tamanho do gráfico:** em pixels. Para apresentação com melhor visualização, deve-se configurá-lo levando-se em conta a resolução do monitor utilizado.
- **Tamanho do vértice proporcional à frequência:** permite ajustar o tamanho dos vértices proporcional à frequência da forma ou ao seu χ^2 .

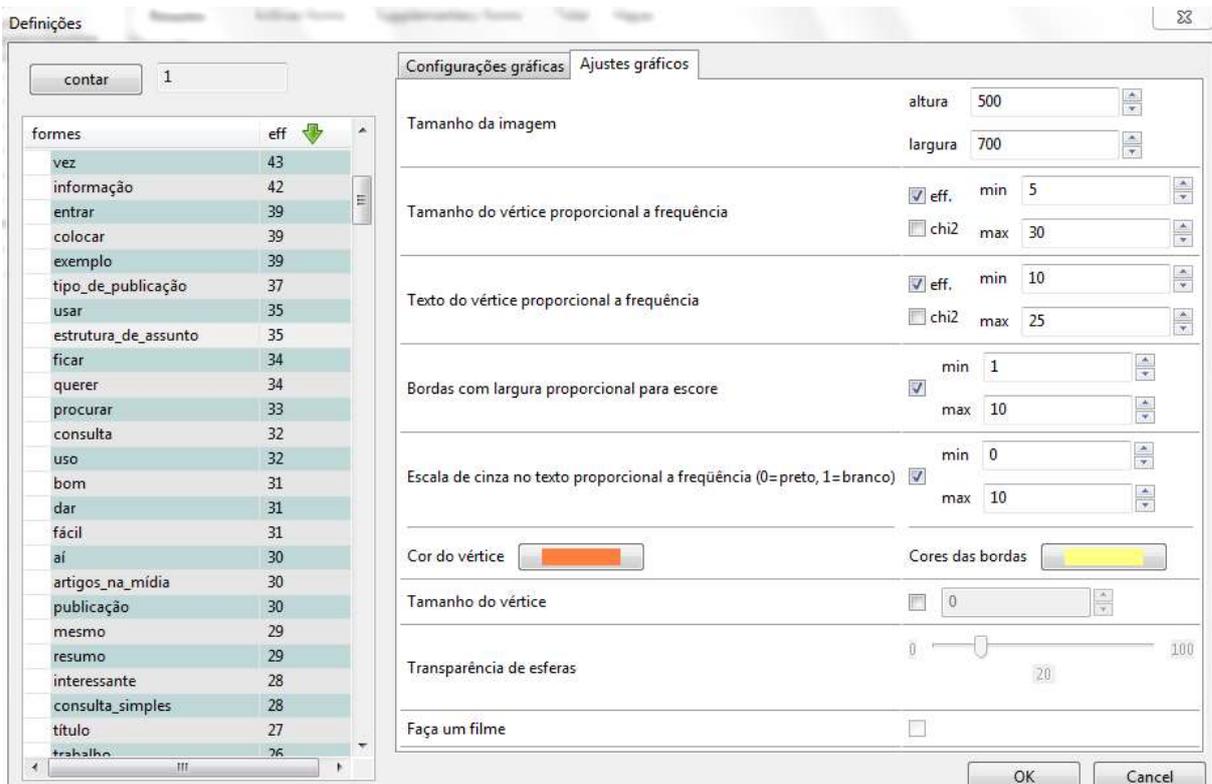


Fig. 53 – Análise de Similitude – Parâmetros de ajustes gráficos

- **Texto do vértice proporcional à frequência:** permite ajustar o texto dos vértices proporcional à frequência da forma ou ao seu χ^2 .
- **Bordas com largura proporcional para escore:** permite ajustar a largura das arestas proporcional ao valor da frequência ou ao seu χ^2 .
- **Escala de cinza no texto proporcional à frequência (0: preto; 1: branco):** permite atribuir a cor do texto dos vértices em função da frequência da forma.
- **Cor do vértice:** atribui a cor escolhida aos vértices e/ou arestas.
- **Tamanho do vértice:** usa um tamanho único para os vértices do gráfico.
- **Transparência de esferas:** regula a transparência de esferas para os gráficos em 3D.
- **Faça um filme:** permite fazer um filme representando ciclo completo de um gráfico em 3D. Observe que a opção tipo de gráfico deve estar em 3D. Veja nota no Capítulo IX item 2 Resultados.

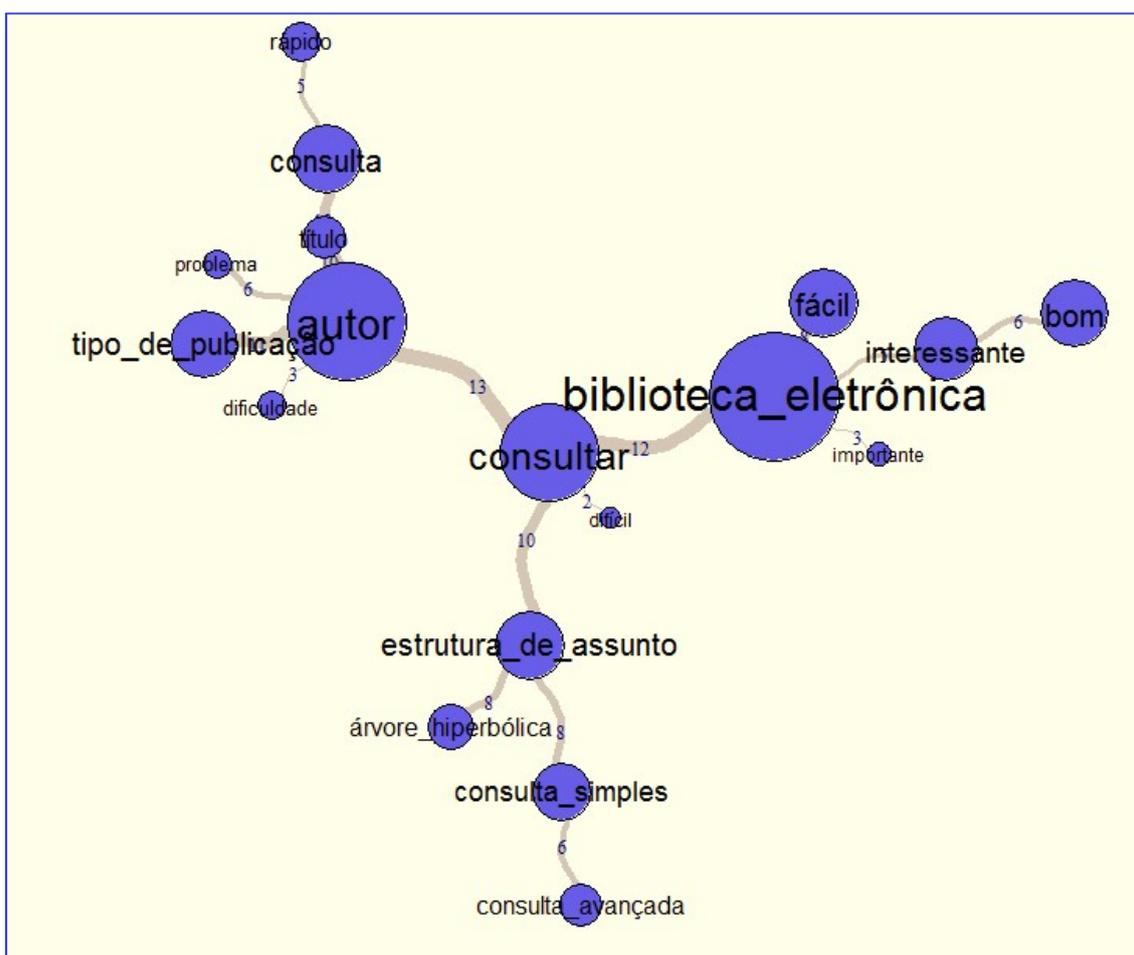
2. Resultados

O gráfico é apresentado em formato de árvore (hierárquico) na janela dos resultados (gráfico estático) e em nova janela gráficos dinâmicos e 3D.

No caso de gráficos estáticos, no canto superior esquerdo da janela, aparecem dois botões. O primeiro deles mostra pequeno gráfico com traços vermelhos e pontos pretos. Ele permite que se modifiquem os parâmetros da análise, abrindo novamente a janela para edição dos parâmetros. O segundo botão no qual está escrito EXPORT, exportará a imagem para arquivo que é armazenado na pasta de resultados em formato graphml, que posteriormente pode ser editado pelos softwares Gephi ou Visone.

Apresentam-se nas Figuras 54 a 60, a seguir, alguns gráficos de similitude escolhidos para que se possa visualizar as diferenças conforme parâmetros escolhidos.

Para gráficos em 3D veja nota no Capítulo IX item 2 Resultados.



**Fig. 54 – Análise de Similaridade – Apresentação Fruchterman Reingold
Escore Coocorrência, com escore nas bordas**

A árvore de coocorrência é composta por um núcleo central, a partir do qual, surgem algumas ramificações. O núcleo central é representado, neste exemplo, pela palavra Biblioteca eletrônica do qual surgem ramificações que descendem. Os ramos que apresentam maiores graus de conexão com o núcleo são: consultar, fácil, bom, rápido, interessante, importante. Outros dois núcleos são igualmente relevantes: Consultar e Autor. Em Consultar, encontramos as diversas formas de consulta: estrutura de assunto, árvore hiperbólica, consulta simples e consulta avançada. Em Autor, encontram-se os termos relevantes: problema e difícil.

Dica: É importante ressaltar que esses gráficos devem ser realizados para análise de parte do *corpus*, caso contrário as interações ficam muito confusas de difícil visualização. Para dividi-los

escolha as formas mais interrelacionadas na tabela de formas do lado esquerdo do menu por temas mais específicos.

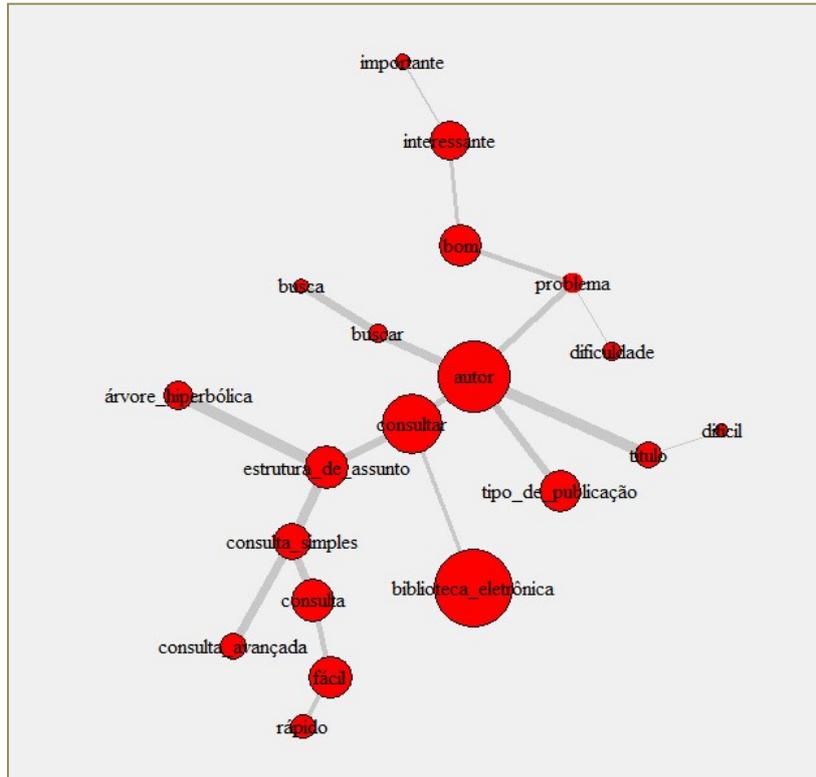


Fig. 55 – Análise de Similitude – Apresentação Kamada Kawai, Escore Pearson

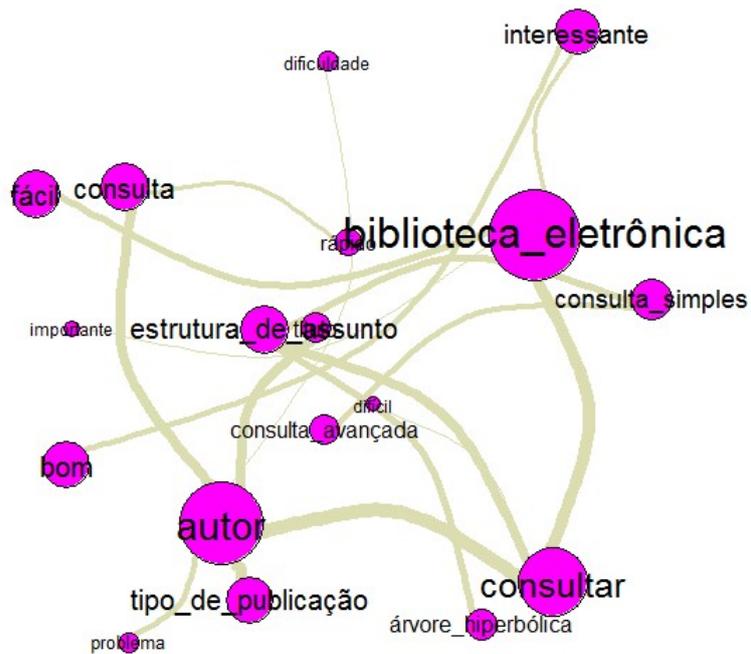


Fig. 56 – Análise de Similitude – Apresentação random, Escore Coocorrência

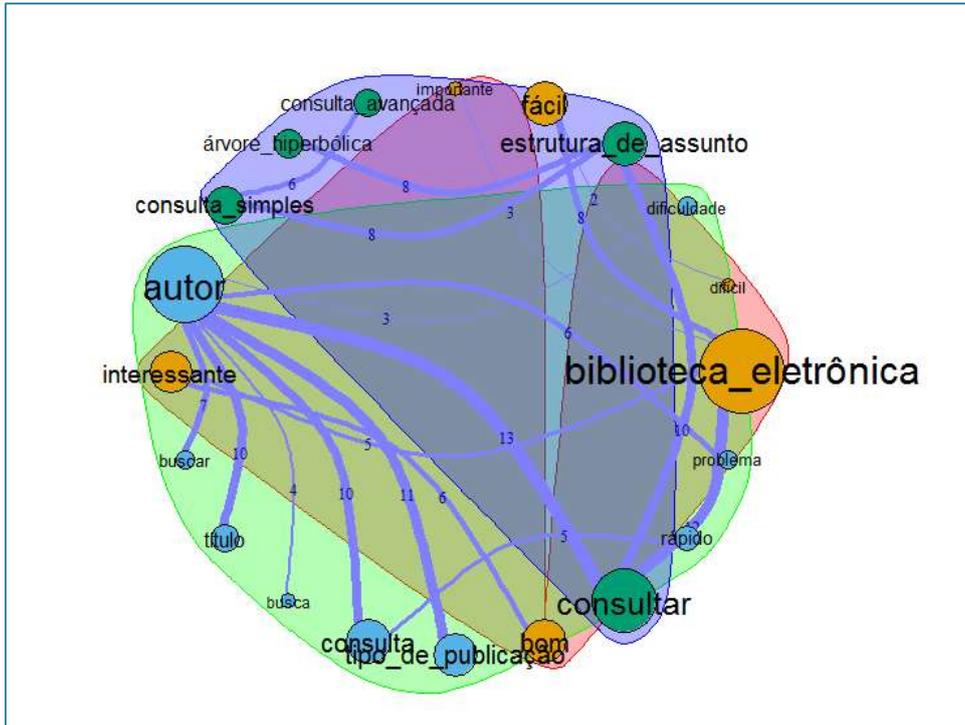


Fig. 57– Análise de Similitude – Apresentação cercle, Escore Cocorrência , com Comunidades, Halo, e texto nos vértices

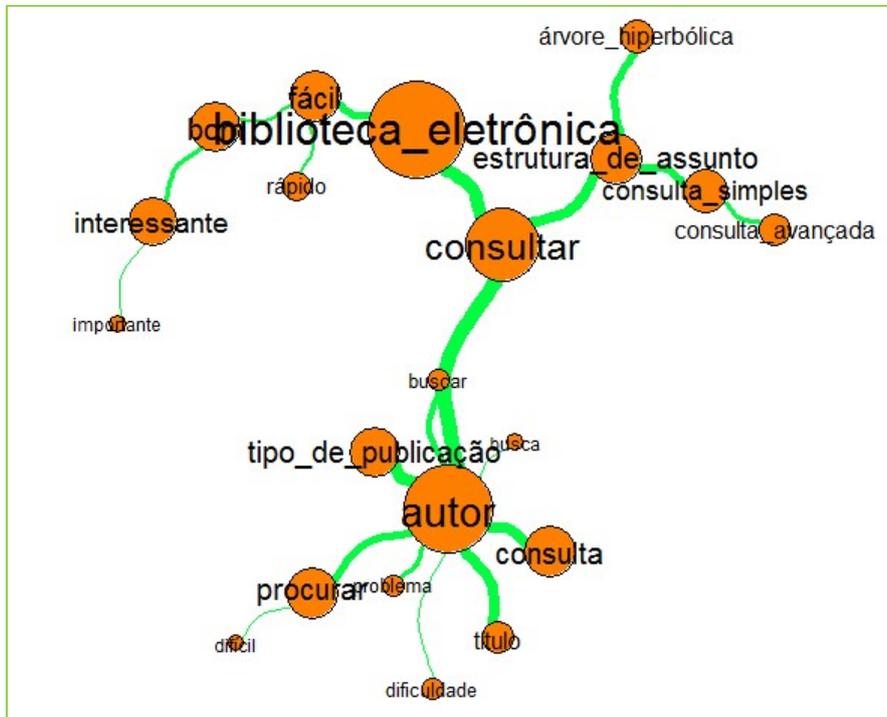


Fig. 58– Análise de Similitude – Apresentação Graphopt, Escore Cocorrência

Nota: Os gráficos de similitude diferem das árvores apresentadas na Análise do Método de Reinert (Word graph e Cluster graph) no sentido que não se pode escolher apenas uma forma e o sistema apresenta as suas relações com as demais. No caso presente, é sempre necessário escolher um conjunto de formas e então ver quais são as relações entre elas.

3. Análises de Similitude arquivos criados

- **Actives.csv:** lista das formas ativas.

- **Analyse.ira:** arquivo do Iramuteq. Permite abrir uma análise.

- **graph_simi_X.png:** Gráfico com formato png. A cada reedição do gráfico é criado um novo arquivo.

- **Liste_graph.txt:** arquivo do Iramuteq.

- **listeuce1.csv:** arquivo do Iramuteq.

- **mat01.csv:** Matriz utilizada para análise no formato MarketMatrix.

- **RData.RData:** Arquivos R.

- **Selected.csv:** Lista de identificação das formas selecionadas (é recarregado em cada edição).

XI - Nuvem de palavras

A análise por meio de nuvem de palavras mostra um conjunto de palavras agrupadas, organizadas e estruturadas em forma de nuvem. As palavras são apresentadas com tamanhos diferentes, ou seja, as palavras maiores são aquelas que detêm maior importância no *corpus* textual, a partir do indicador de frequência ou outro escore estatístico escolhido. É uma análise lexical mais simples, porém, bastante interessante, na medida em que possibilita rápida identificação das palavras-chaves de um *corpus*, isto é, a rápida visualização de seu conteúdo, pois as palavras mais importantes estão mais perto do centro e graficamente são escritas com fonte maiores.

1. Parâmetros

Depois de escolher as opções comuns a todas as análises no menu das Preferências, no qual deve-se seguir as orientações do tópico Tratamento comuns nas análises, capítulo IV, aparece a janela de configuração, Figura 61.

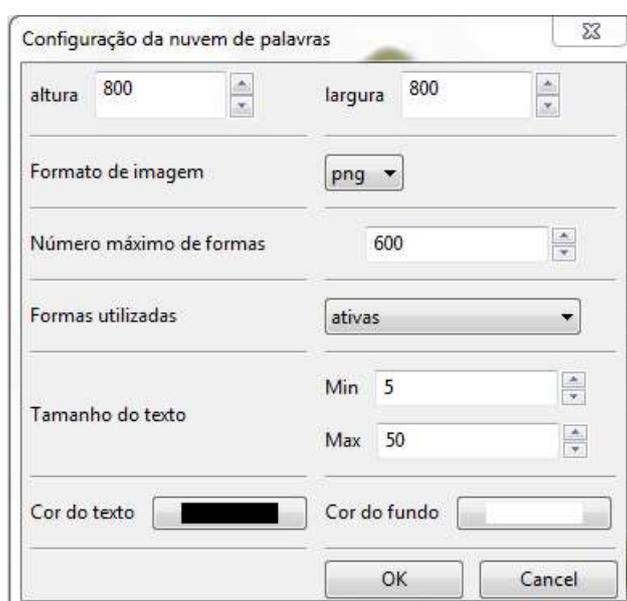


Fig. 61 – Nuvem de palavras – Configuração

- **Altura/Largura:** fixa a dimensão da nuvem em pixels. Para apresentação com melhor visualização, deve-se configurá-lo levando-se em conta a resolução do monitor utilizado.

Fig. 62 – Nuvem de palavras – Resultado

A Figura 62 traz um exemplo de Nuvem de palavras elaborada.

Dica: É importante ressaltar que esses gráficos devem ser realizados para análise de parte do *corpus*, caso contrário, as interações ficam muito confusas de difícil visualização. Para restringir utilize as opções de Número máximo de formas e/ou também, Formas utilizadas. Há possibilidade de limitar, também, escolhendo-se as formas mais interrelacionadas na tabela de formas do lado esquerdo do menu por temas mais específicos.

3. Análise de Nuvem de palavras arquivos criados

- **Arquivos:** Nome_do_corpus_wordcloud_x.

- **Actives_eff.csv:** Lista das formas ativas e sua frequência.

- **Analyse.ira:** Arquivo do Iramuteq. Permite abrir uma análise.

- **Nuage1.png:** Imagem da nuvem e formato png.

- **Select.csv:** Lista dos identificadores das formas selecionadas.

XII - Reedição dos Gráficos de Análise Textual

1. Reedição dos Dendrogramas

Esta caixa de diálogo é acessível pelo ícone  existente à esquerda dos dendrogramas, Figura 63.

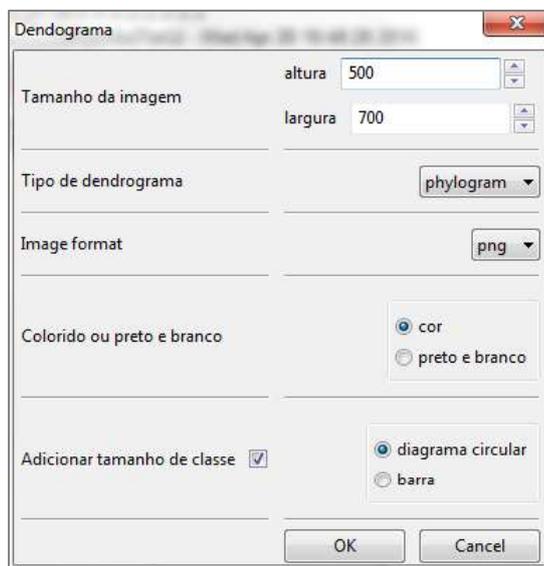


Fig. 63 – Reedição dos Dendrogramas: Configuração

- **Altura/Largura:** tamanho da imagem gerada em pixels.
- **Tipo de dendrograma:** permite escolher a forma de apresentação do dendrograma (phylograma, cladograma, fan, unrooted e radial). Veja as diferentes formas na Figura 64.
- **Image format:** permite escolher se formato png ou svg. A escolha entre um e outro formato deve levar em conta se o usuário pretende abri-lo com outro aplicativo de edição gráfica.
- **Colorido ou preto e branco:** permite escolher se colorido ou preto e branco.
- **Adicionar tamanho de classe:** permite sobrepor ao dendrograma o tamanho das classes (em barras ou diagrama circular).

As Figuras 64 a 68 mostram exemplos de formatos escolhidos.

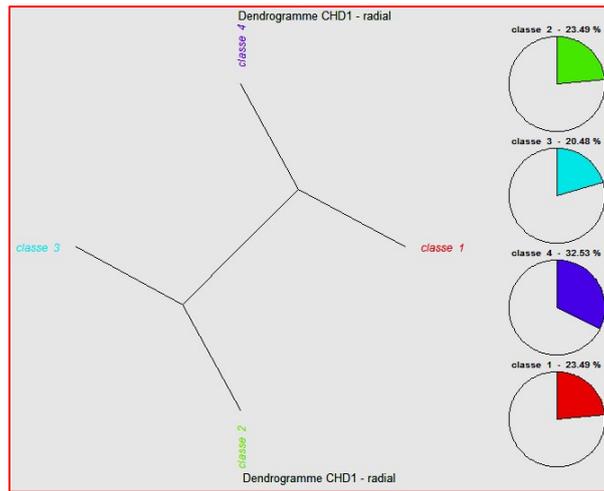


Fig. 64 – Dendrograma radial colorido com adição do tamanho da classe circular

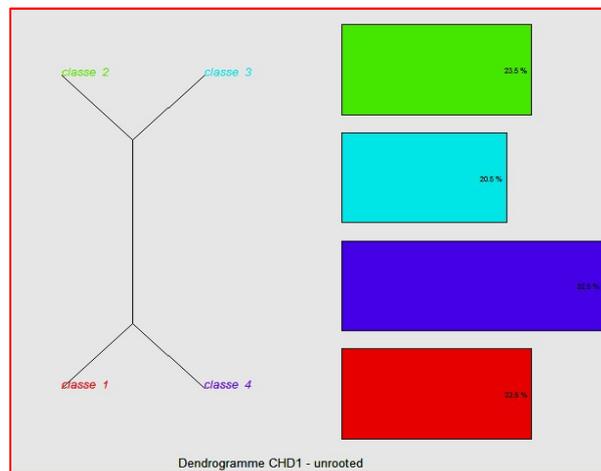


Fig. 65 – Dendrograma unrooted colorido com adição do tamanho da classe em barras

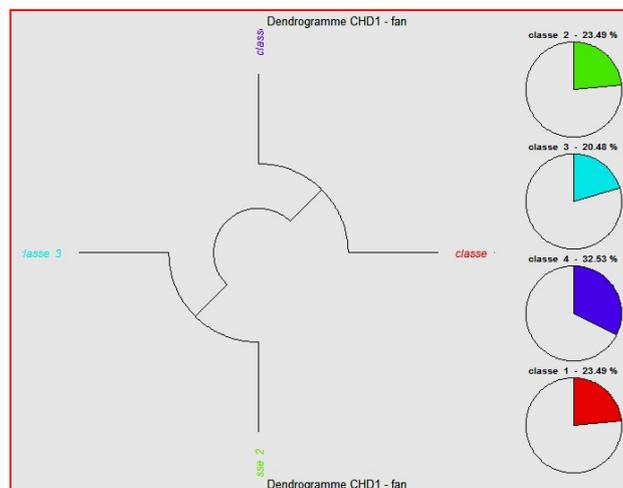


Fig. 66 – Dendrograma fan colorido com adição do tamanho da classe em círculos

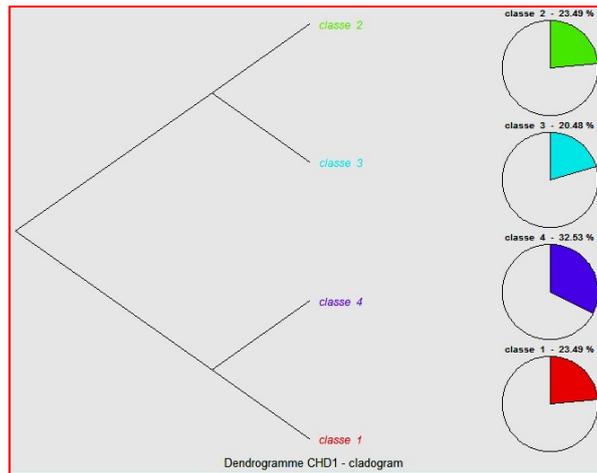


Fig. 67 – Dendrograma cladograma colorido com adição do tamanho da classe em círculos

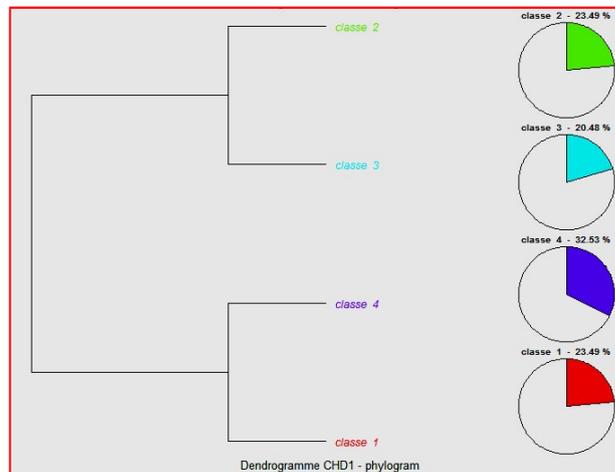


Fig. 68 – Dendrograma phylograma colorido com adição do tamanho da classe em círculos

2. Reedição da AFC – Análise Fatorial de Correspondência

A caixa de diálogo, Figura 69, é acessível pelo ícone à esquerda das AFC.

- **Tipos de gráficos:** permite escolher os gráficos em duas ou três dimensões. Para gráficos em 3D, veja nota no Capítulo IX item 2 Resultados.

- **Formato de imagem:** permite escolher o formato de gravação da imagem. As opções são png e svg. Esses formatos só importam se o usuário pretende abri-los com outra aplicação de edição de imagens.

- **Representação:** permite escolher entre representação por coordenadas ou por correlacionamentos da variável.

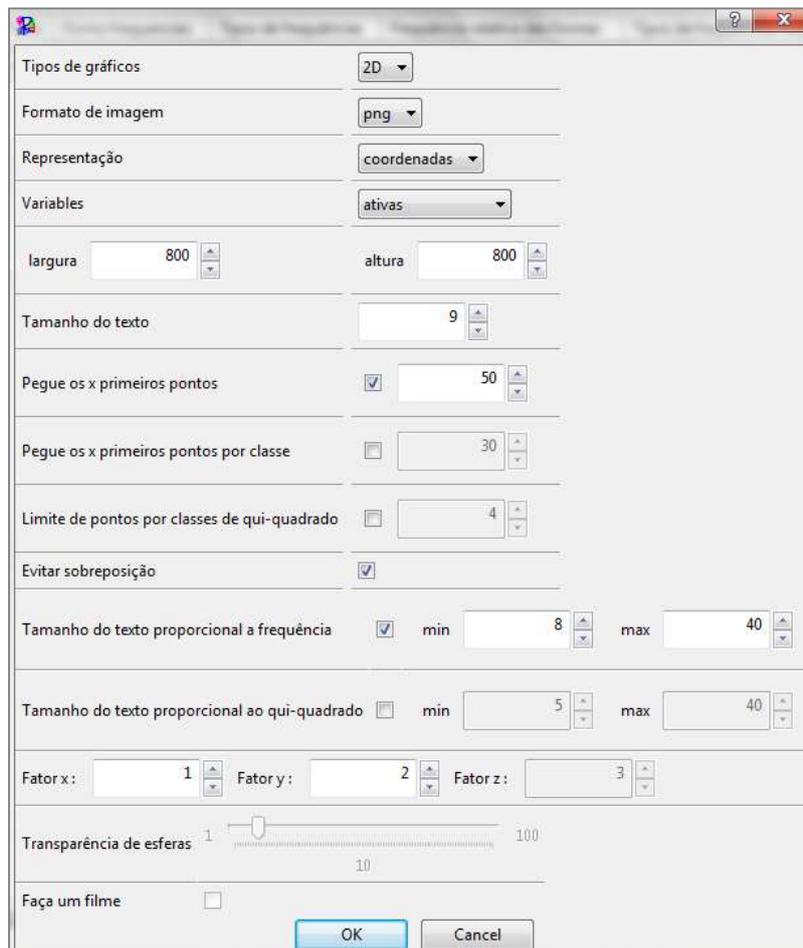


Fig. 69 – Reedição das AFC: Configuração

- **Variáveis:** permite escolher as formas que serão analisadas de variáveis: ativas, suplementares, classes ou estrelas.

- **Largura/Altura:** define as dimensões do gráfico em pixels.

- **Tamanho do texto:** define o tamanho do texto. Porém a opção abaixo “Tamanho do texto proporcional à frequência” é melhor porque estabelece o tamanho mínimo e máximo de acordo com o peso no *corpus*.

- **Pegue os x primeiros pontos:** o sistema toma em conta os dez primeiros pontos.

- **Pegue os x primeiros pontos por classe:** pega os primeiros pontos com qui-quadrado mais alto da classe.
- **Limite de pontos por classe de qui-quadrado:** estabelece o mínimo de ligação à classe para aparecer.
- **Evitar sobreposição:** esta opção evita sobreposição das etiquetas dos pontos, isto produz um viés, mas permite uma melhor legibilidade.
- **Tamanho do texto proporcional à frequência:** determina o tamanho do texto em função da frequência da forma dentro do *corpus*.
- **Tamanho do texto proporcional ao qui-quadrado:** determina o tamanho do texto em função do qui-quadrado de ligação à classe.
- **Fator x/y/z:** respetivamente fator em abcissas, fator em ordenadas e fator em cotas. Para gráficos em duas dimensões só podem ser preenchidos os fatores x e y, ou seja, somente para gráficos em 3D pode-se marcar o fator z.
- **Transparência de esferas:** permite ajustar a transparência das esferas nos gráficos em 3D.
- **Faça um filme:** disponível unicamente para os gráficos em 3D, produz um gif animado que representa uma rotação de 360° do gráfico. Veja nota no Capítulo IX item 2 Resultados.

3. Reedição de Gráficos Dinâmicos

Algumas análises possibilitam a execução de gráficos dinâmicos. Esses gráficos podem ser editados no próprio aplicativo. A Figura 70 mostra a janela de edição.

São possíveis as seguintes operações:

- **Visão do gráfico**

Pode-se escolher como o gráfico é mostrado na janela a fim de que fique com todos os vértices e arestas visíveis. Escolha View/Fit on screen para que o tamanho dele se ajuste à janela. Não esqueça, primeiramente de maximizar a janela, clicando no quadradinho indicado pela seta no topo da janela, Figura 70.

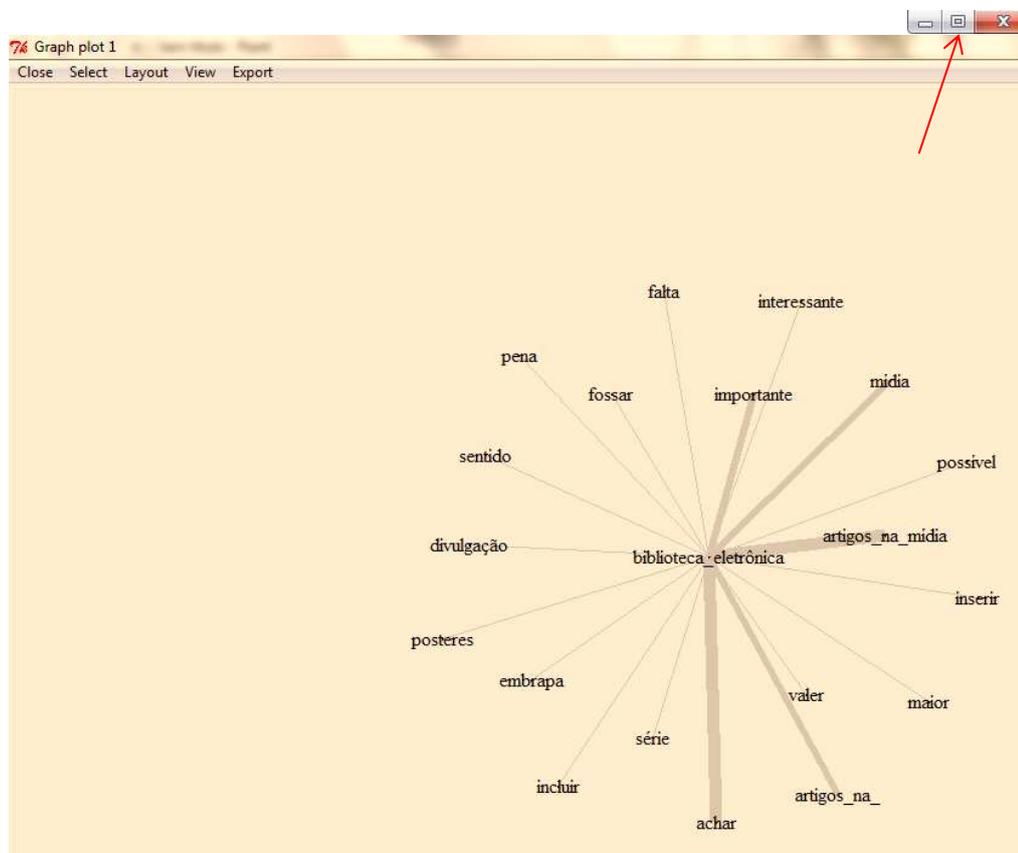


Fig. 70 – Janela de Reedição dos Gráficos Dinâmicos

Pode-se escolher, também, para que o gráfico fique centralizado na janela com a escolha da opção View/Center on screen. Nessa opção o gráfico não muda de tamanho, apenas é centralizado na janela.

A opção Labels selecionada indica que as etiquetas dos vértices estão mostradas. Se a desmarcamos, as etiquetas desaparecem. A utilidade de retirá-las seria para posterior digitação das etiquetas com outro aplicativo gráfico que possibilite a escolha da fonte, tamanho e cor.

A opção Rotate permite girar o gráfico na medida dos graus escolhidos e modificando a posição dos vértices.

- Seleção de vértice

Escolha a opção Select/Select all vértices para selecionar todos os vértices e aplicar um formato único para todos eles. Pode-se escolher ainda Select some vértices.

Nesse caso, o sistema solicita indicar os vértices desejados, Figura 71.

Basta clicar no vértice desejado que o sistema automaticamente o seleciona. Observe que o nó apresenta um ponto vermelho indicando que ele está selecionado. Pressione a tecla CTRL para selecionar os demais. No fim da seleção, clique OK e então escolha o formato desejado. Alternativamente é possível selecionar vértices individualmente, bastando clicar nele.

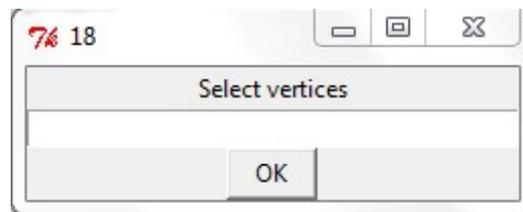


Fig. 71 – Janela de Seleção de Vértices

Após selecioná-lo é possível mudar a cor e tamanho, clicando-se com o botão direito do mouse e escolhendo-se: vertex color ou vertex size.

- Escolha de tamanho do vértice

Escolha primeiramente vertex size, Figura 72.

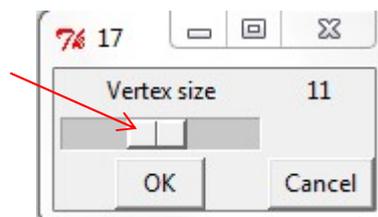


Fig. 72 – Janela de Escolha de Tamanho de Vértices

Clique e arraste os botões, indicados pela seta, para a esquerda para aumentar o tamanho e para direita para diminuir. Depois clique em OK e verifique como ficou o tamanho. Se desejar modificá-lo, basta selecionar e modificá-lo novamente.

- Escolha a cor do vértice

Para escolher a cor, selecione vertex color e o sistema abre uma janela com várias opções de cor, Figura 73.

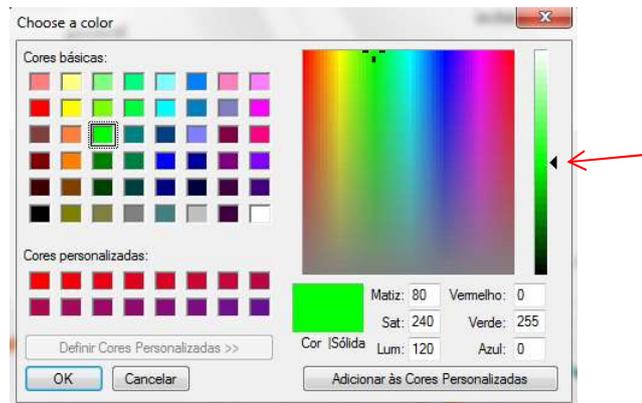


Fig. 73 – Janela de Escolha de Cor de Vértices

Do lado esquerdo, no alto da janela existe uma tabela de cores padronizadas (Cores básicas). Pode-se clicar em qualquer cor desejada, como, no exemplo, na cor verde. Em seguida, se desejado, pode-se clarear ou escurecer essa cor, mexendo-se na seta, do lado direito, arrastando-a para cima e para baixo. Essa nova cor é salientada na amostra (Cor Sólida). Se desejado, essa nova cor pode ser guardada para uso futuro, por exemplo, em outro vértice, bastando clicar em Adicionar às Cores Personalizadas, de forma que para aplicá-la em outro vértice, bastará clicar na cor customizada que foi acrescentada na tabela de Cores personalizadas (lado esquerdo e em baixo da Figura).

- Seleção de aresta

Escolha a opção Select/Select all edges para selecionar todas as arestas e aplicar um formato único para todas elas. Pode-se escolher ainda Select some edges. Nesse caso, o sistema solicita indicar as arestas desejadas, da mesma maneira que na escolha dos vértices.

Basta clicar na aresta desejada que o sistema automaticamente a seleciona. Observe que a aresta fica com a linha tracejada, indicando que ela está selecionada. Pressione a tecla CTRL para selecionar as demais. No fim da seleção, clique OK e então escolha o formato desejado. Alternativamente é possível selecionar arestas individualmente, bastando clicar nela.

- Escolha de tamanho da aresta

Da mesma maneira da escolha de tamanho do vértice, escolha o tamanho (grossura da aresta) clicando e arrastando os botões, indicados pela seta, para a esquerda para aumentar o tamanho e para direita para diminuir, Figura 74. Depois clique em OK e verifique como ficou o tamanho. Se desejar modificá-lo, basta selecionar e modificá-lo novamente.

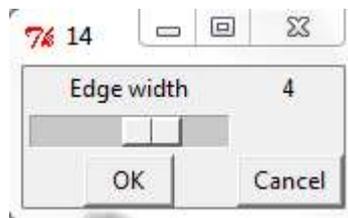


Fig. 74 – Janela de Escolha de Tamanho de Arestas

- Escolha a cor da aresta

Para escolher a cor, selecione edge color e o sistema abre uma janela com várias opções de cor, tal qual para seleção de cor de vértice. Escolha a cor obedecendo às instruções dadas no item Escolha a cor do vértice.

- Desmarcação de tudo

Depois de aplicar o formato desejado é necessário desmarcar a(s) aresta(s) ou vértice(s) selecionados a fim de não fazer mais modificações nele(s). Escolha Select/Deselect everything para esse procedimento. Tudo que estiver selecionado será, automaticamente, desmarcado.

- Layout

Nesta opção é possível alterar o formato do gráfico, escolhendo-se entre as opções apresentadas, Figura 75.



Fig. 75 – Janela de Escolha de Leiaute

- Correção de superposição de vértices e comprimento de arestas

Para mudar a posição de um vértice sobreposto, clique nele e o arraste para um local livre na janela.

Para diminuir ou aumentar o tamanho de uma aresta, clique nela e arraste aumentando ou diminuindo seu comprimento.

- Exportação do gráfico

As modificações aplicadas no gráfico por meio da janela de edição são gravadas automaticamente no arquivo eps (formato Postscript) gerado ao se clicar na opção de Exportação e escolher a pasta onde ele será armazenado. Esse formato poderá ser editado por outro programa de edição de imagens.

Após a exportação basta fechar a janela (opção Close do menu horizontal). Observe, no entanto, que após fechar a janela não é mais possível retornar à edição do gráfico dentro do Iramuteq.

Referências

BORTOLOSSI, H.J.J.; QUEIROZ, J.D.B.; SILVA, M.M. da. **A Lei de Zipf e Outras Leis de Potência em Dados Empíricos**. Rio de Janeiro: Sociedade Brasileira de Matemática (Projeto Klein de Matemática em Português), 2011.

CAMARGO, B. V. **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*. João Pessoa: Editora da UFPB, 2005. p.511-539.

CAMARGO, B.V.; JUSTO, A.M. IRAMUTEQ: um software gratuito para análise de dados textuais. **Temas em Psicologia**, v. 21, n. 2, p.513-518, 2013.

CAMARGO, B.V.; JUSTO, A.M. **Tutorial para uso do software de análise textual IRAMUTEQ**. [Florianópolis]: Universidade Federal de Santa Catarina, 2013b. 18p.

FREIRES, L.A. **Atitudes frente a parentalidades**: uma explicação da aceitação/oposição a partir de variáveis explícitas e implícitas. João Pessoa, PB: Universidade Federal da Paraíba. Centro de Ciências Humanas e Letras. Departamento de Psicologia, abril 2015. 236 p. (Tese de doutorado).

GHIGLIONE, R.; MATALON, B. **O inquérito**: teoria e prática. Oeiras: Celta, 1993.

GUEDES, V. L. da S.; BORSCHIVER, S.. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: VI CINFOM, 2005, Salvador. **Anais Eletrônico do VI CINFOM**. Salvador -BA: UFBA, 2005.

HAZBOUN, A.M. ALCHIERI, J.C. Justificativas e concepções de psicólogos que não utilizam avaliação psicológica. **Avaliação Psicológica**, v. 12, n. 3 p. 361-368, 2013.

LOUBÈRE, L.; RATINAUD, P. **Documentation IRaMuTeQ 0.6 alpha 3 version 0.1**. 2014. 37 p. <http://www.iramuteq.org/documentation/fichiers/documentation_iramuteq_21_12_2013.pdf> Acesso em 02/04/2016.

LOUBÈRE, L.; RATINAUD, P. **Manual Iramuteq versão 0.1**. Trad. de Baltazar Fernandes. 35 p. <http://www.academia.edu/9312034/Manual_Iramuteq> Acesso em 02/04/2016.

[MANUAL do iramuteq em francês]. 30p. <http://www.iramuteq.org/documentation/html> Consulta em 2/04/2016.

NASCIMENTO, A.R.A do; MENANDRO, P.R.M. Análise lexical e análise de conteúdo: uma proposta de utilização conjugada. **Estudos e Pesquisas em Psicologia**, Rio de Janeiro, ano 6, n. 2, 2006, p. 72-88.

OLIVEIRA, L.F.R. de. **Tutorial (básico) de utilização do Iramuteq**. [Goiânia]: Universidade Federal de Goiás, [2015]. 14 p.

RABUSKE, M. **Introdução à teoria dos grafos**. Florianópolis: UFSC, 1992.

RATINAUD, P. **Introduction à IraMuTeQ**. Toulouse : Université de Toulouse-Le Mirail, 2013. 7p.

SILVA, J.P. DA; BOUSFIELD, A.B. DA S.; CARDOSO, L.H. A hipertensão arterial na mídia impressa: análise da revista *Veja*. **Psicologia e Saber Social**, v.2, n.2, p.191-203, 2013.