Análise de Dados Textuais em Pesquisas de Mobilidade Urbana e Transporte com IRaMuteQ

Kevin Masinda Mahema
kevinmasinda16@gmail.com

Grupo de pesquisa comportamento em transportes e novas tecnologias - GCTNT
Programa de Pós Graduaçaõ em Transportes - PPGT/UnB
Departamento de Eng. Civil e Ambientais
Universidade de Brasília - UnB

10 de janeiro de 2022

Apresentação

Software IRaMuTeQ é um software livre sob licença de GNU GPL (v2). É um software que usa a interface do R (www.r-project.org) e pode ser expandido a partir da linguagem de programação Python (www.python.org). Permitindo assim, uma análise textual multidimensional e estatística. A análise textual ou análise sobre os corpus de textos, é feito a partir do método de classificação hierárquica descendente descrito por Reinert; além de outras análise lexicais tais como a análise de similitude de um segmento de texto, a estatística clássica textual e uma busca específica a partir de um segmento de texto bem definido, a fim de auxiliar na explicação/interpretação de textos.

Objetivos

Este manual (em formato de minicurso) tem como objetivo de apresentar a ferramenta IRaMuTeQ como uma ferramenta de análise de conteúdo textual que ajuda na engenharia comportamental para tomada de decisão, na analise de dados textuais ou lexical. O software IRaMuTeQ será apresentada desde a sua instalação, passando pela preparação de corpus, análise do corpus, manipulação de software e por fim a obtenção e exportação dos resultados finais.

Ao final deste manual, o aprendiz deverá ser capaz de:

Objetivos (cont.)

- Analisar dados textuais ou lexical;
- Compreender a dicotomia clássica entre quantitativo e qualitativo na análise de dados;
- Quantificar e empregar cálculos estatísticos sobre variáveis essencialmente qualitativas - os textos;
- Tornar possível, a partir da análise textual, a descrição material produzido por um determinado produtor;
- Utilizar a análise textual com a finalidade comparativa, relacional, comparando produções diferentes em função de variáveis específicas que descrevem quem produziu o texto;
- Interpretar os dados resultantes da análise textual.

Pré-requisito

O presente manual (minicurso) é destinado aos

- Alunos tanto da graduação como da pós graduação;
- Professores;
- Empresas que trabalham com análise qualitativa de textos tais como:
    1. Entrevista;
    2. Livro;
    3. Pesquisa de opinião;
    4. Artigo científicos;
    5. Jornais;
    6. Etc.

Pré-requisito (cont.)

- Pesquisadores voltados não somente nas áreas de ciências de saúdes, humanas, sociais, mas também na área de ciências exatas e das engenharias.

Não é necessário ter conhecimento em linguagem de programação Python e/ou R

Instalação

IRaMuteQ é uma interface do software R. **É indispensável a instalação do software R**.

- É recomendavél baixar uma das versões do R a partir da versão 3.6
- IRaMuteQ não é compatível com as **versões 4.0 ou posterior de R.**
- IRamuTeQ pode ser baixado a partir da página do projeto no sourceforge
- Há versões disponíveis para Windows, Mac OS X et GNU/Linux.
- Código fonte do projeto no GitHub

Instalação - Windows

1. Baixar e instalar R a partir www.r-projet.com
2. Baixar e instalar IRaMuTeQ

Instalação - GNU/Linux

1. Linux Mint: é preciso instalar o package **r-base-dev** (sudo apt-get install r-base-dev) a partir de um terminal.
2. Não é mais necessário instalar as bibliotecas R (rlg, ape, gee, igraph, proxy, rgl) como nas primeiras versões de IRaMuteQ
3. Na versão atual de IRaMuTeQ, pode acontecer que o pacote rgl não seja instalado. PS: não preocupa-se, o pacote serve somente para visualização 3D.
4. Para verificar se a instalação for bem sucedida, abre IRamuTeQ, vá em edição - preferência - verifique instalação.

Instalação - MacOS

1. Primeiro baixar e instalar o software R
2. Baixae e Instalar o **xquartz**
3. Baixar e Instalar IRaMuTeQ
4. Baixar o arquivo Rgraph e em seguida ir em aplicação - IRaMuTeQ (clique direto) - mostrar o conteudo do pacote - Rscripts e por fim, substituir o arquivo Rgraph.
5. PS: sobre MacOS X pode acontecer que IRaMuTeQ não seja aberto por ser reconhecido como aplicação vindo de um fonte desconhecido. Será preciso abrir IRaMuTeQ pelo terminal. Com isso será preciso executar os seguintes comandos:
   - cd */Aplications/iramuteq.app*
   - ls
   - ./iramuteq

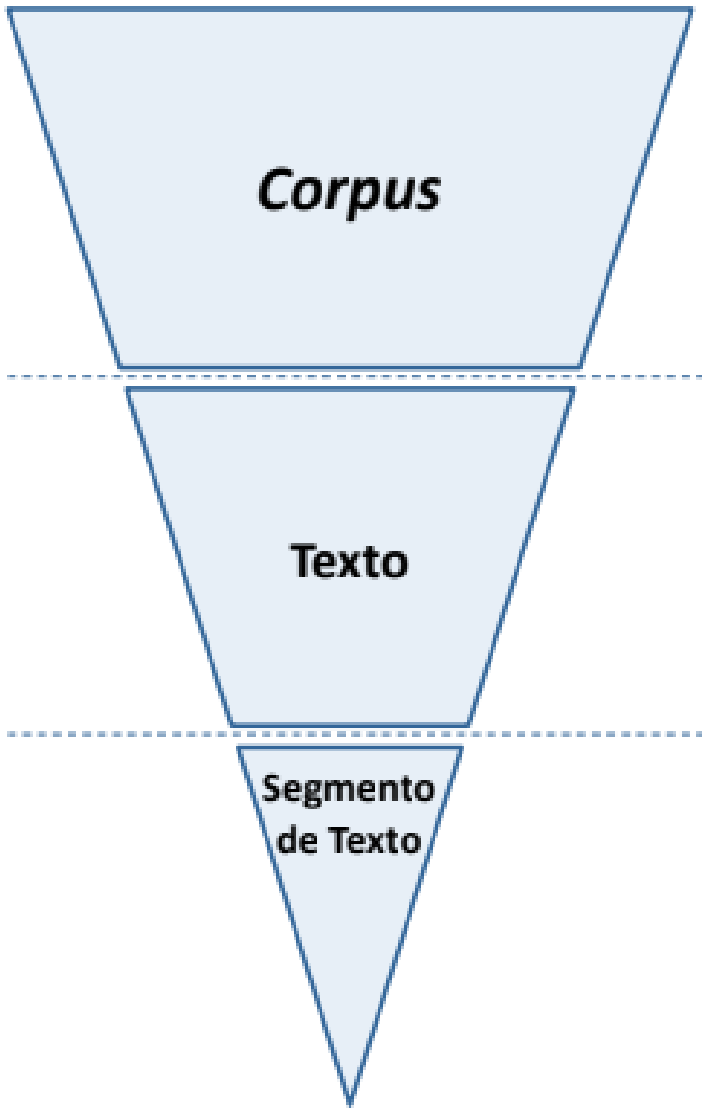
   Por fim, quando o programa abrir, é necessario aceitar a instalação de todos os pacotes R (caso IRaMuTeQ pedir)

Introdução: Conceitos básicos

- **Segmentos de Texto - ST:** são geralmente os conjuntos de palavras que tem um tamanho de aproximadamente de três linhas de textos. Na maioria das vezes, é delimitado automaticamente pelo software em função do tamanho do corpus ou pode ser montado pelo pesquisador. É considerado como a unidade do corpus
- **Textos:** são os conjuntos de segmentos de texto geralmente definidos pelo pesquisador dependendo da natureza da sua pesquisa. O texto deve obrigatoriamente começar sempre por quatro asterisco: ******** e conter pelo menos uma variável. A separação entre dois textos é feito por uma **linha em branco**
- **Corpus:** É o conjunto de textos. De forma mais simples, corpus, é o nosso arquivo de análise. Só existe e tem um único corpus.

Um corpus contém vários textos, e esse último contém vários segmentos de texto

Introdução: Conceitos básicos (cont.)



Introdução: Conceitos básicos (cont.)

- **Classe:** pode ser definido como um conjunto de segmento de texto ou vocabulários de mesma homogeneidade ou natureza.
- **Lematização:** é o processo de trazer os substantivos nas suas formas masculino - singular, verbos nas suas formas no infinitivo e adjetivos nas suas formas masculino - singular também. O software realiza a lematização de forma automática (precisa ser ativado durante a análise pelo pesquisador)
- **Hapax:** são palavras que ocorrem/aparecem somente uma única vez em todo o corpus.

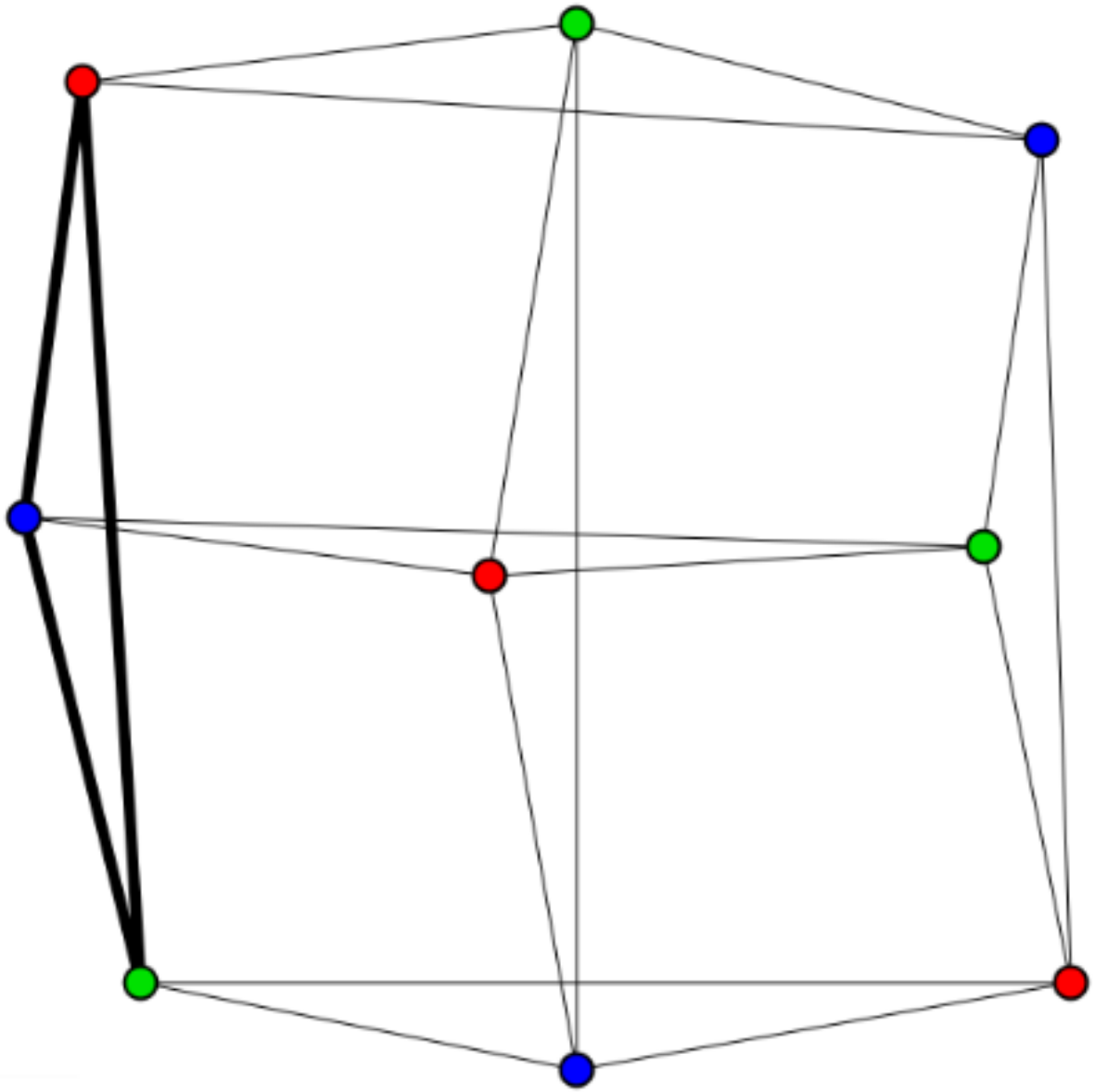
Noção de grafos

- **Grafos:** é um ramo da matemática que estuda as relações entre objetos de um mesmo conjunto. Por definição, um grafo denominado por **G(V,E)** é um conjunto não vazio de vértices (nós) - **V** interligados por um conjunto de arestas - **E**. [1]

Dependendo da aplicação, arestas podem ou não ter direção, pode ser permitido ou não arestas ligarem um vértice a ele próprio e vértices e/ou arestas podem ter um peso (numérico) associado.

- **subgrafos:** um subgrafo de um grafo G é um pedaço de G. O conjunto de vértices e o conjunto de arestas do pedaço de G devem ser coerentes. Assim, é melhor formular o conceito como uma relação entre dois grafos: Um grafo H é subgrafo de um grafo G se todos os vértices e todas as arestas de de H são vértices e arestas de G também. [1]

Noção de grafos (cont.)



Codificação de caráteres

- IRaMuTeq é uma interface de R que permite a inserção/anexação de um arquivo .txt ou em outras palavras, a inserção de umas cadeias de caracteres em um buffer de sistema de arquivos.
- Quando o arquivo é salvo, ele usa uma codificação de texto para decidir quais bytes cada caractere se tornará.

Um caráter na computação é o nome que se dá a cada um dos símbolos que se podem usar para produzir um programa de computador, bem como os textos e imagens apresentados na tela quando se executa um programa em modo texto. Já, uma **codificação de caracteres** é um padrão de relacionamento entre um conjunto de caracteres. [2]

- **Bit:** é a unidade básica que os computadores e sistemas digitais utilizam. Pode assumir somente dois valores, 0 ou 1. [3]
- **Bytes:** é uma unidade de informação digital equivalente a oito bits [3]

Codificação de caráteres (cont.)

Existem vários tipos de codificação de caráteres que o IRaMuTeQ aceita. entre eles, pode-se citar:

1. **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. [4]
2. **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. [5]
3. etc.

Noções: informática, matemática e estatística

- **Pixels:** a palavra pixel é uma junção de termos "**pic**ture" e "**el**ement". Ou seja, "elemento de imagem". É a menor unidade de uma imagem digital, sendo que, um conjunto de pixels com diferentes cores formam a imagem inteira. [6]
- **Default:** é um termo de origem inglês, que em português, significa padrão. A palavra default é um termo técnico em computação que quer dizer, configuração padrão, configuração pré-definida.[7]
- **permilagem ou por mil (‰):** proporção relativamente a mil; número em relação a mil; fração cujo denominador é 1000.
- **Distribuição de frequência:** uma distribuição de frequência é um agrupamento de dados em classes de modo a fornecer a quantidade e/ou a percentagem. Com isso, podemos agrupar e visualizar um conjunto de dados sem precisar levar em conta os valores individuais de dados em cada classe.

já uma **distribuição de frequência (absoluta ou relativa)** pode ser apresentada em tabelas ou gráficos. Uma distribuição de frequência agrupa os dados por classes de ocorrência. [8]

- **Distribuição hipergeométrica:** uma variável aleatória X tem uma distribuição hipergeométrica com parâmetros N, r e n quando esta representa o número de sucessos dentro de uma amostra de tamanho n extraída, sem reposição, de uma população de tamanho N formada por r sucessos e N-r fracas. [9]
- **Distribuição Gama:** uma variável aleatória contínua X tem distribuição gama com parâmetros $\alpha > 0$ e $\lambda > 0$ se sua função densidade é dada por:

$$f_x(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \lambda^{\alpha} e^{-\alpha x}, & x > 0 \\ 0 & \text{caso contrário} \end{cases}$$

Note que, quando $\alpha = 1$, resulta a densidade exponencial com parâmetro λ, ou seja, a densidade exponencial é um caso particular da densidade gama. [9]

- **Distribuição Qui-Quadrado ou X^2:** é um caso particular da distribuição Gama, onde, o parâmetro de forma α é igual a $\frac{n}{2}$, com n inteiro positivo, e o parâmetro λ é $\frac{1}{2}$ com n graus de liberdade. [9]

É uma das distribuições mais utilizadas em estatística inferencial, principalmente para realizar testes de X^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. [7]

Noções: informática, matemática e estatística (cont.)

- **Lei de Zipf:** trata-se de uma lei de potências sobre a distribuição de valores de acordo com o número de ordem numa lista. Na linguística, ela conta a frequência que cada palavra aparece em um determinado texto e segue uma distribuição que pode-se aproximar pela fórmula:

$$P_n \sim 1 \cdot n^{-\alpha}$$

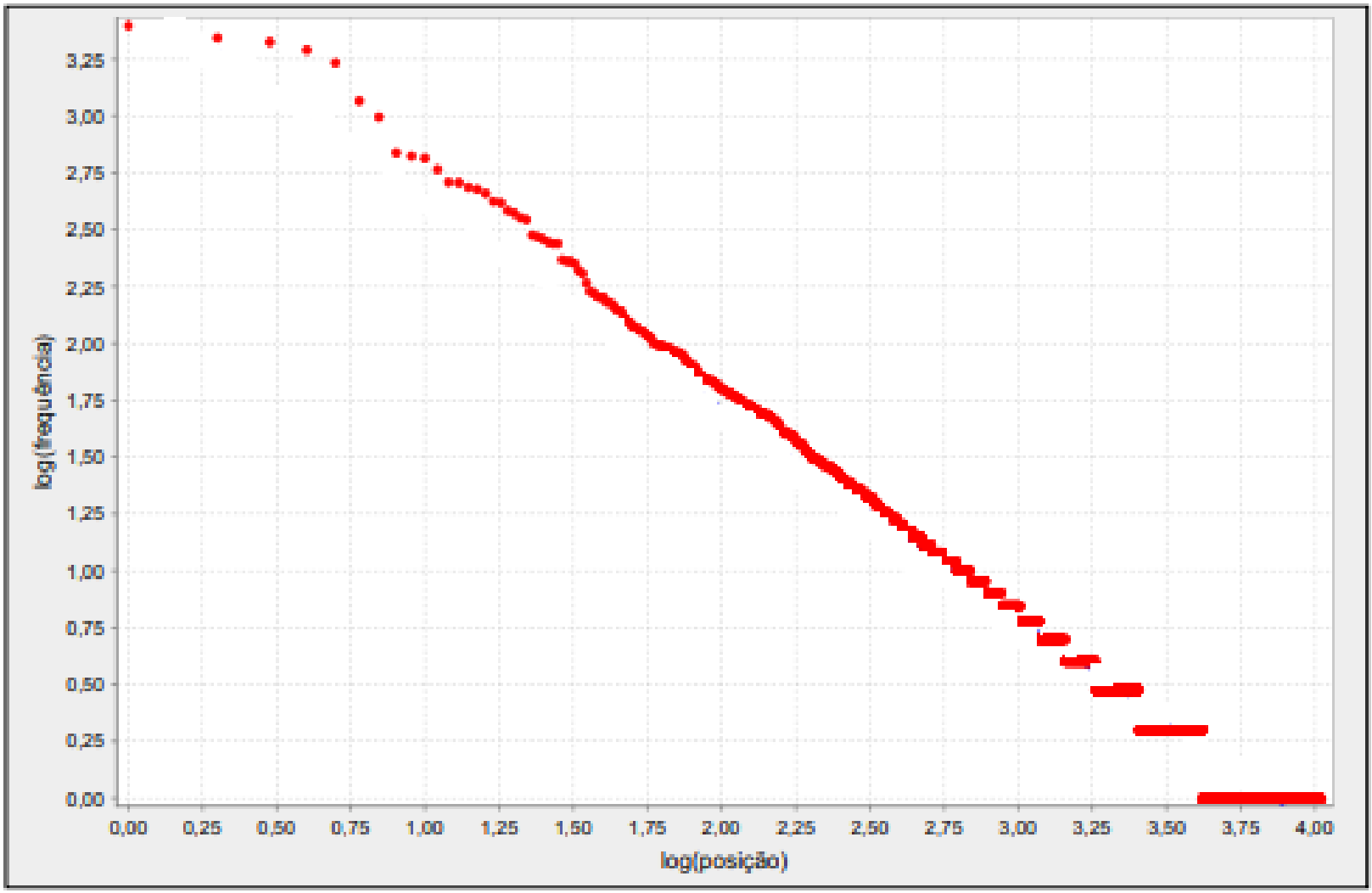
Exemplo: suponha que após a contagem de ocorrência de cada palavra dentro de um texto e classificada em ordem decrescente pela sua frequência, representada na tabela abaixo: [10]:

Noções: informática, matemática e estatística (cont.)

Posição(x)	Frequência(y)	Palavras	$\widehat{\mathbf{x}} = \log x$	$\widehat{\mathbf{y}} = \log y$	**palavra**
1	2483	a	**0,000...**	**3,396...**	**a**
2	2203	o	**0,301...**	**3,343...**	**o**
3	2112	que	**0,477...**	**3,324...**	**que**
4	1949	bom	**0,602...**	**3,2898...**	**bom**
⋮	⋮	⋮	⋮	⋮	⋮
178	37	mestrado	**2,250...**	**1,568...**	
⋮	⋮	⋮	⋮	⋮	⋮
10447	1	unb	**4,018...**	**0,000...**	**unb**
10448	1	congo	**4,019...**	**0,000...**	**congo**
10449	1	brasil	**4,019...**	**0,000...**	**brasil**

Nas leis de potencias, ao invés de analisar a posição(x) e a frequência(y), são analisados $\widehat{x} = \log x$ e $\widehat{y} = \log y$ e os valores são apresentados em uma figura.

Noções: informática, matemática e estatística (cont.)



Criação e preparação de corpus

A criação de um corpus é simples e segue os seguinte passos:

- Abrir um editor de texto da sua escolha. word, libreOffice, etc. (recomendo o sublime 3)
- Colar o texto a ser analisado no seu editor de texto, salvar o arquivo com extensão de texto (.txt)

Qualquer arquivo .txt pode ser considerado um corpus. Mas para que o seu corpus seja reconhecido e analisado no software IRaMuTeq, é preciso se atentar nas seguinte regras de formatagem:

1. Os textos são introduzidos por quatro estrelas **(****)** seguido de uma variável estrelada.
2. Um texto é separado de um outro texto, pela uma quebra de linha **(uma única linha em branco)**.
3. Cada texto deve conter pelo menos uma variável.
4. Uma variável é introduzido por uma estrela (*).

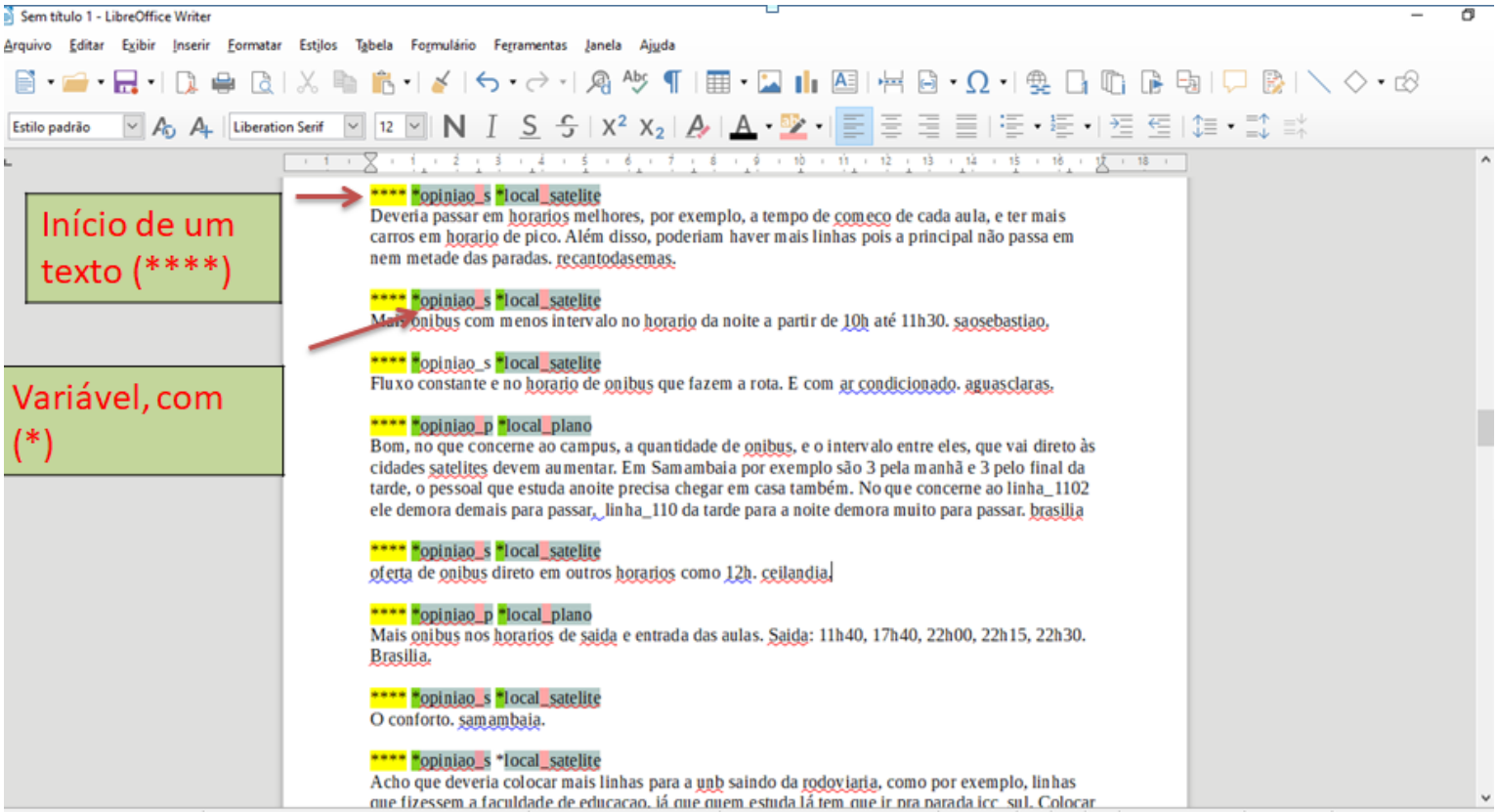
Criação e preparação de corpus (cont.)

5. É possível colocar uma variável estrelada dentro de um texto, se é somente se, essa variável começa por um traço seguido de estrela **(-*)**. Nesse caso, fala-se então de um **corpus com temático ou corpus temático**.
6. No corpus temático, todos os textos devem pertencer a uma temática.
7. Não usar caráter tal como traço (;). Palavras com traço, trocar por sublinha durante a montagem do corpus. ex: Jean-Pierre por Jean_Pierre ou JeanPierre.

Cuidado: IRaMuTeQ é case sensative, ou seja, Asa_Sul e Asa_sul são formas diferentes. (a palavra sul é uma modalidade de asa)

Criação e preparação de corpus (cont.)

A figura abaixo, ilustra o exemplo de um corpus sem temática já pronto. Percebe-se que, cada texto começa por **** + espaço + * + nome_variável. (cada variável, começa com uma estrela colada na frente dela)

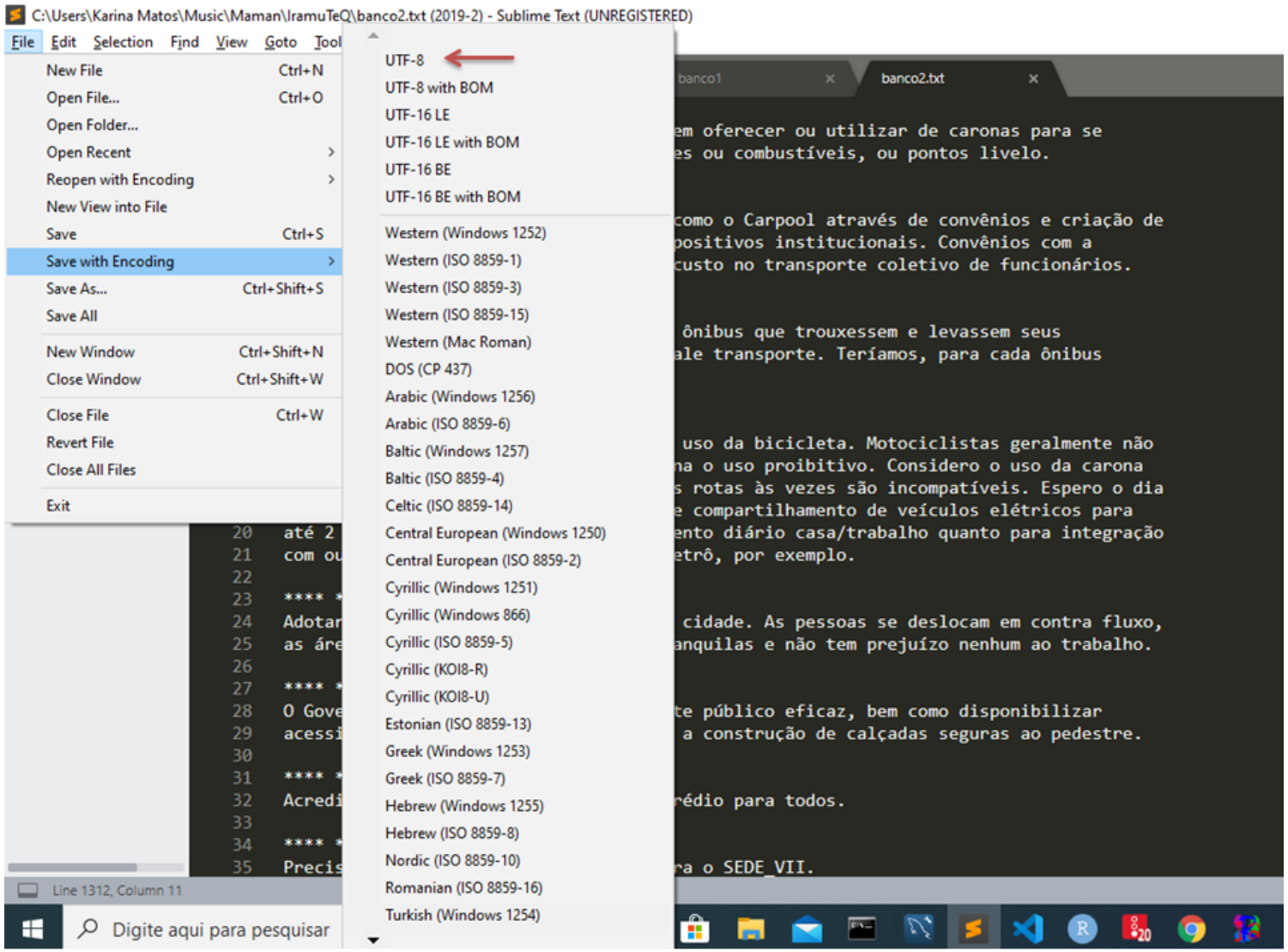


Criação e preparação de corpus (cont.)

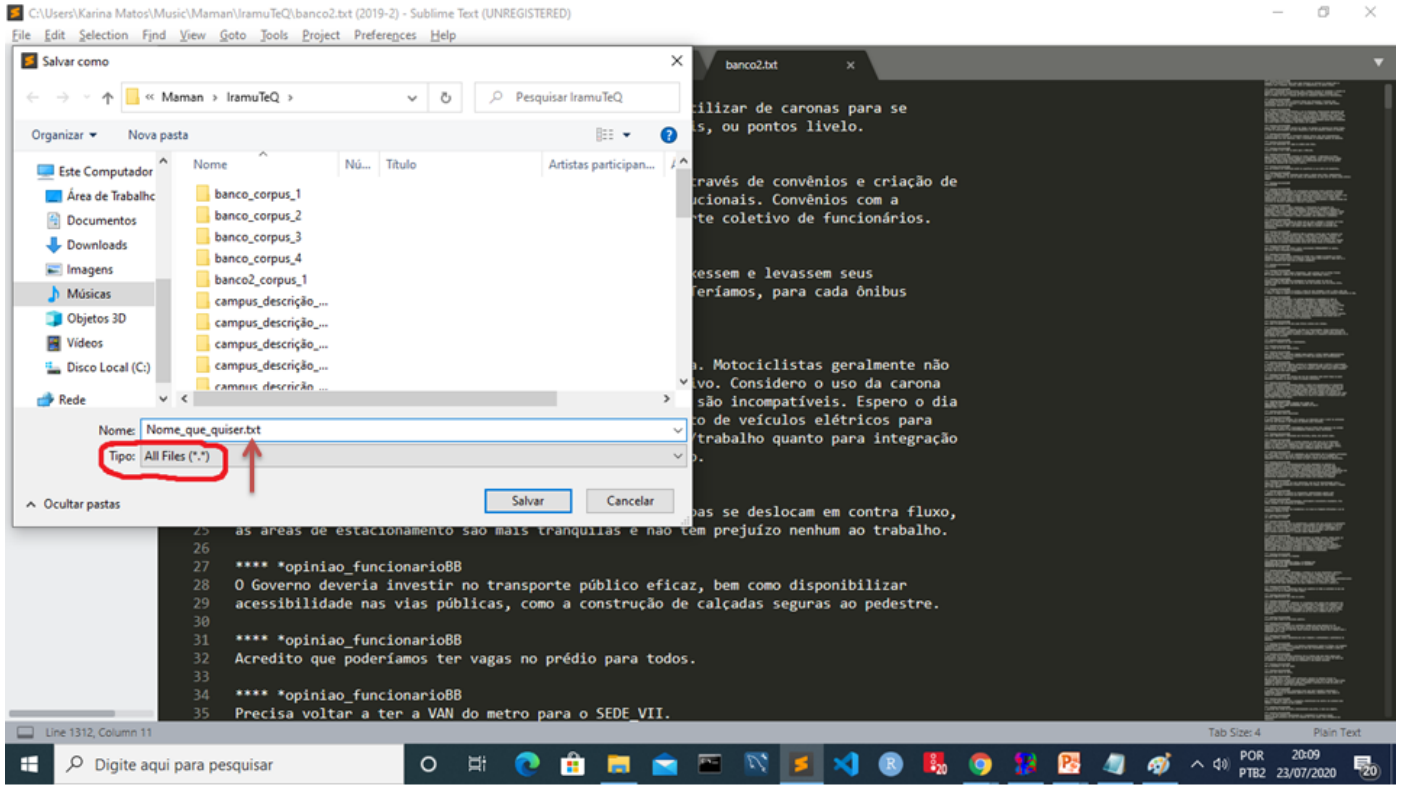
Para evitar incompatibilidade de codificação de caráteres entre aquela que está salva no arquivo e a escolhida no IRaMuTeQ (o que acontece muito), as figuras abaixo, mostram como deve-se escolher a codificação da sua escolha dentro das aceitas pelo IRaMuTeQ com ajuda do Sublime Text 3 e por fim, a maneira certa de salvar o corpus.

Criação e preparação de corpus (cont.)

Criação e preparação de corpus (cont.)



Criação e preparação de corpus (cont.)



Erro comum na criação de um corpus

Erro: **** sexo_m
Texto texto

certo: **** *sexo_m, ou seja (**** + espaço + * + variável)

Erro: **** Texto texto

certo: **** *var_qualquer

Erro: **** *var_qualquer
ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok
-*thematique
texto texto

Erro comum na criação de um corpus (cont.)

certo: **** *var_qualquer
-*thematique
ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok ok

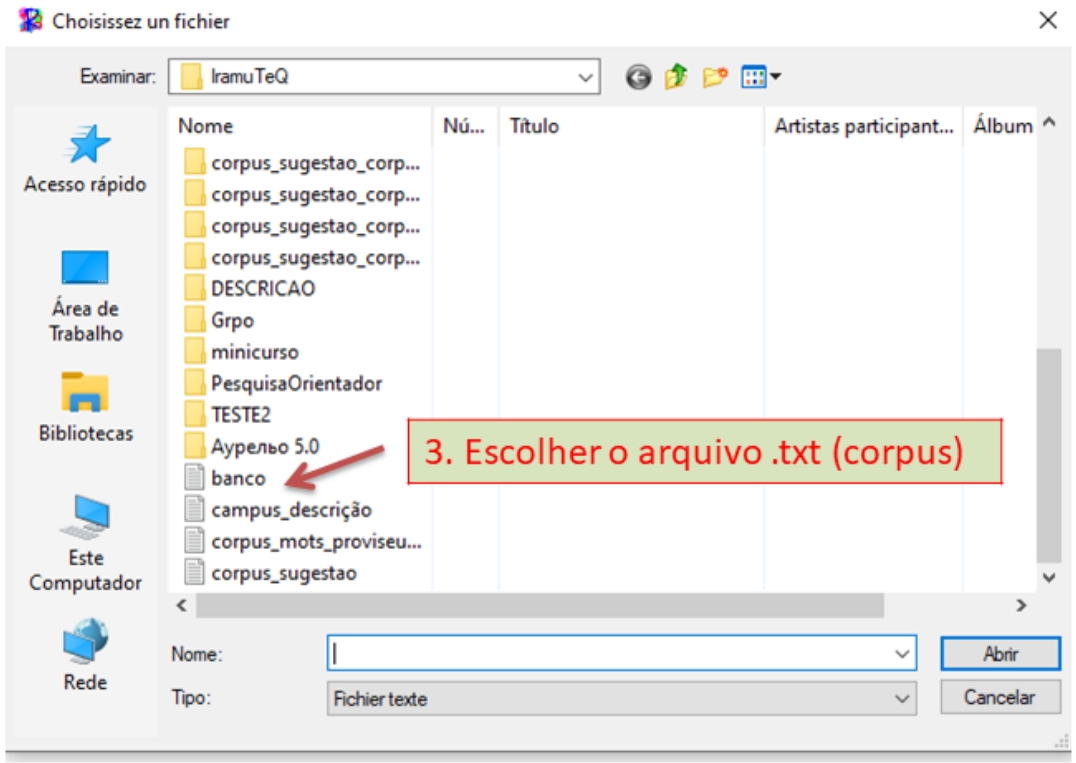
texto texto.

Ou seja,entre a inicialização do texto e a declaração da tema, não pode haver segmento de texto

A declaração da tema deve começar com (-) seguido de (*).

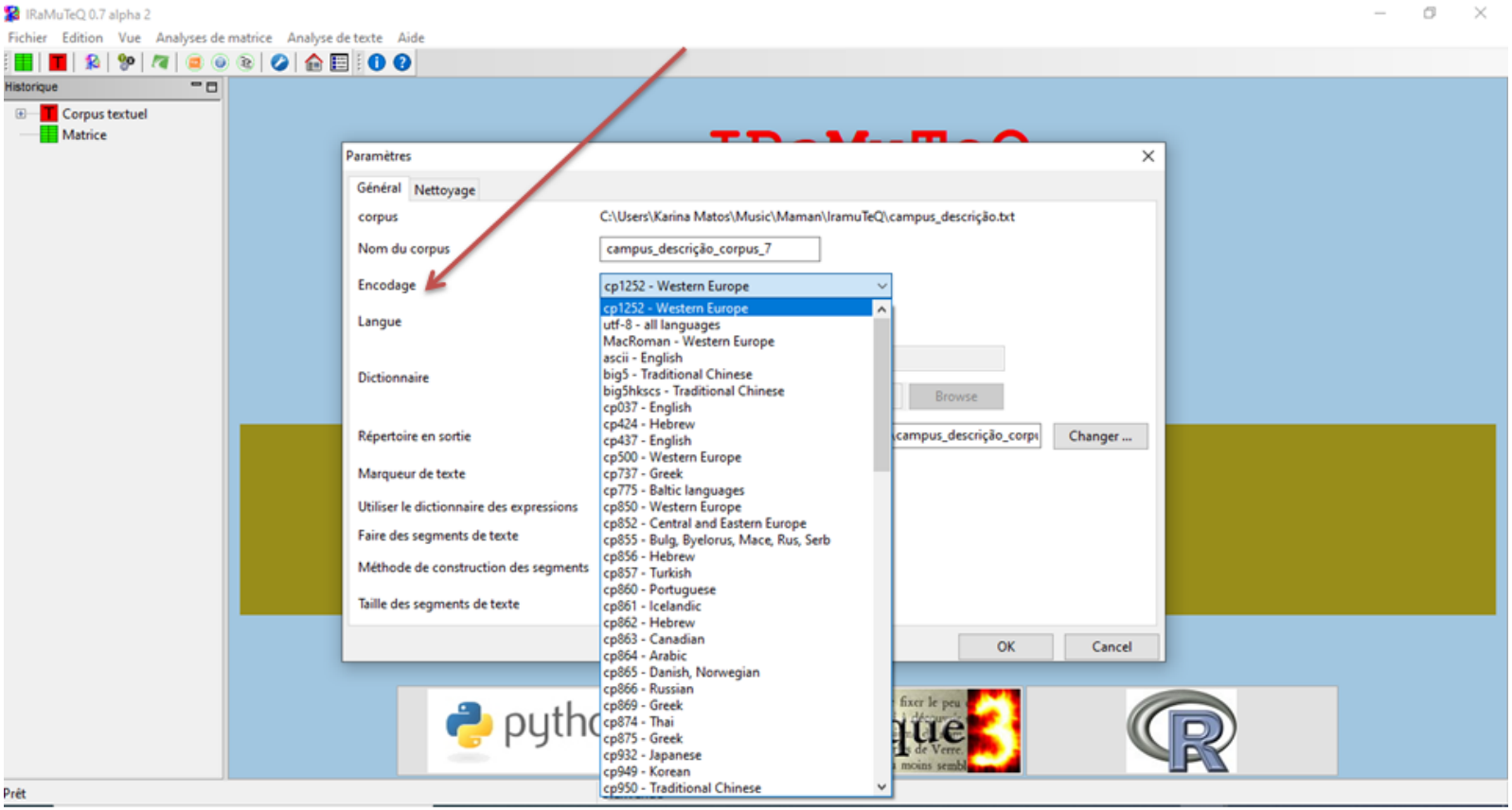
Carregando um corpus

Carregando um corpus (cont.)



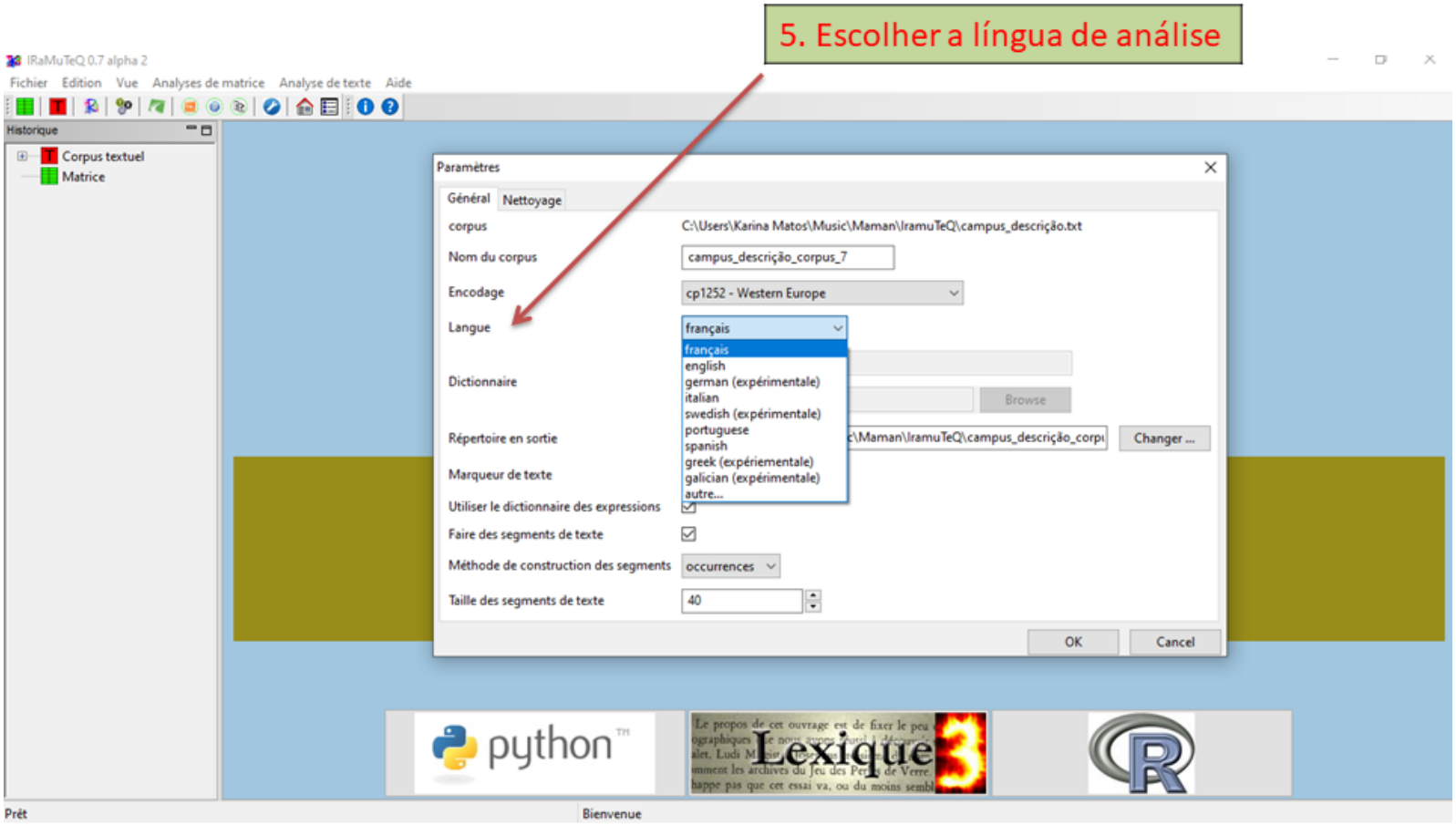
Configurações: Aba Geral - IRaMuTeQ

4. Escolher a codificação adequada

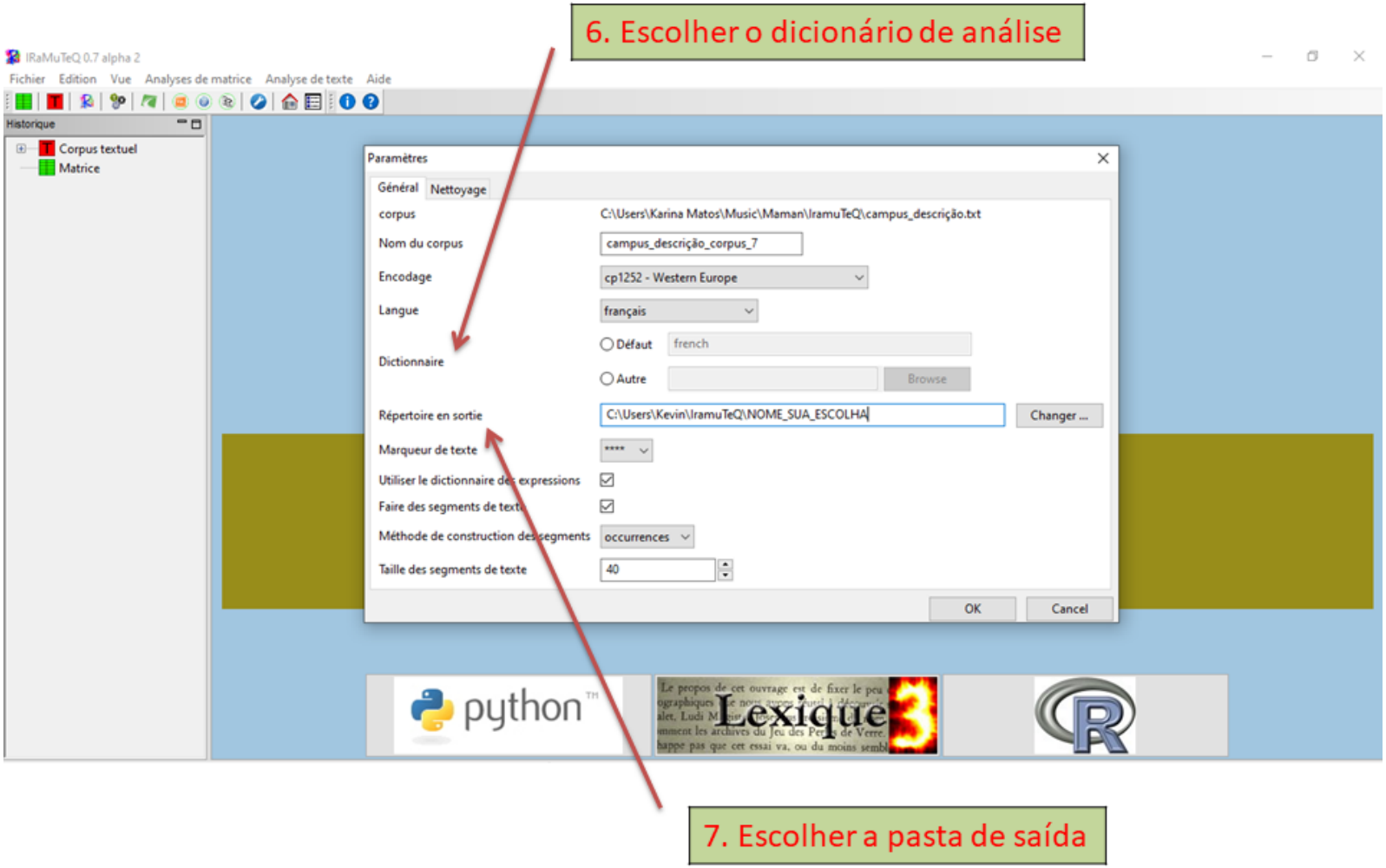


Recomenda-se escolher: cp1252(windows) ou utf-8 all languages

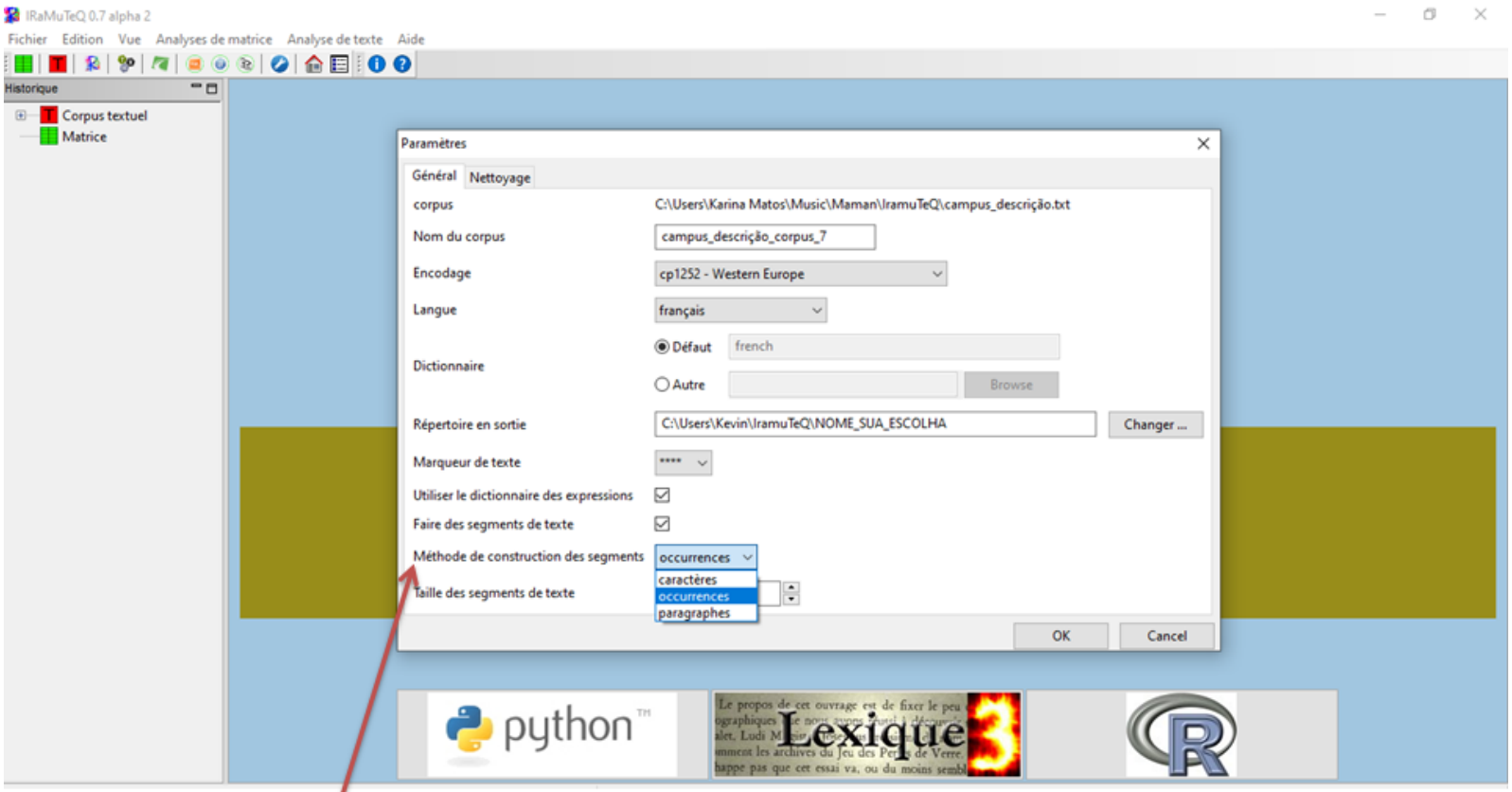
Configurações: Aba Geral - IRaMuTeQ (cont.)



Configurações: Aba Geral - IRaMuTeQ (cont.)

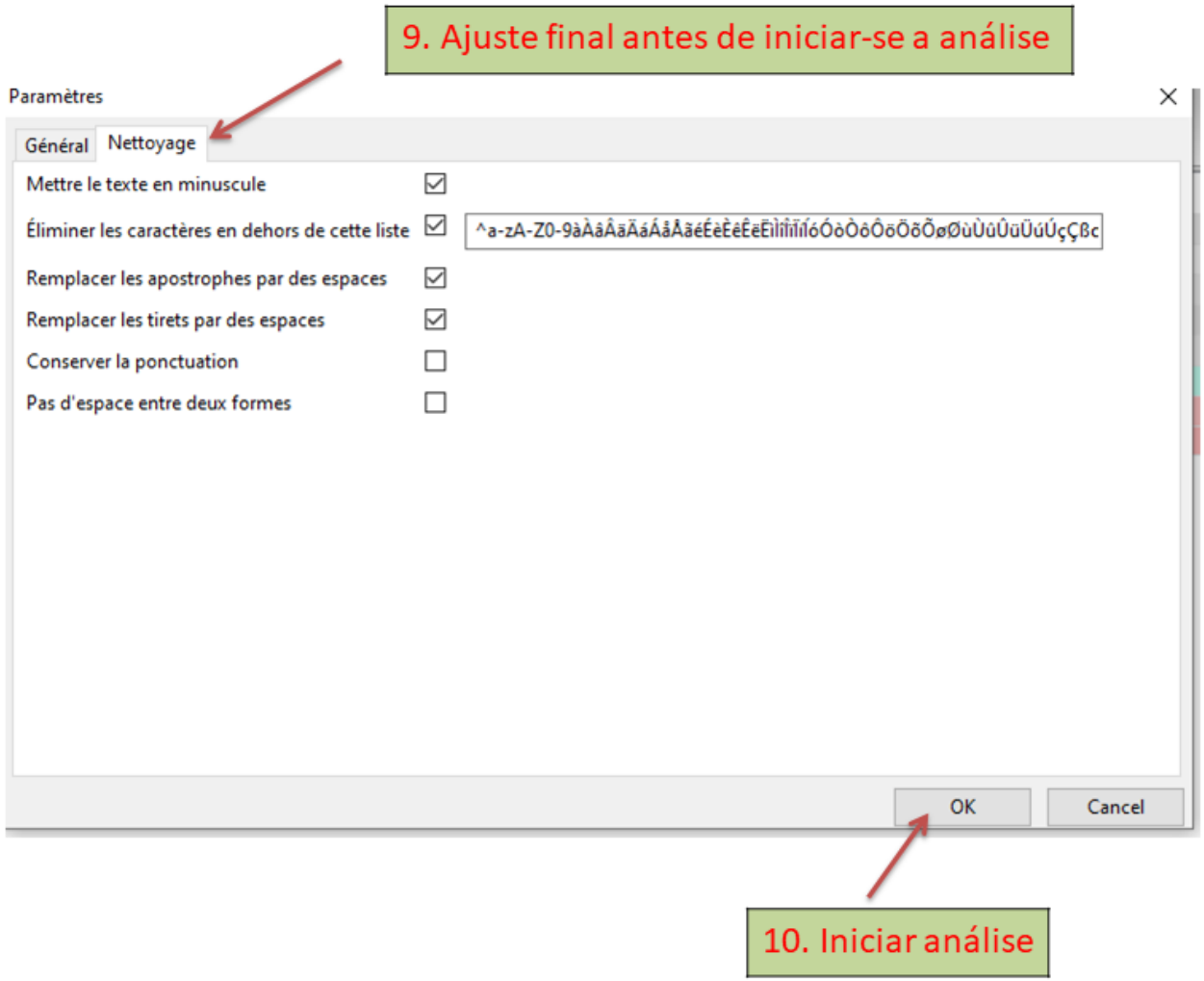


Configurações: Aba Geral - IRaMuTeQ (cont.)



8. Escolher o método de construção de segmento de texto(ST)

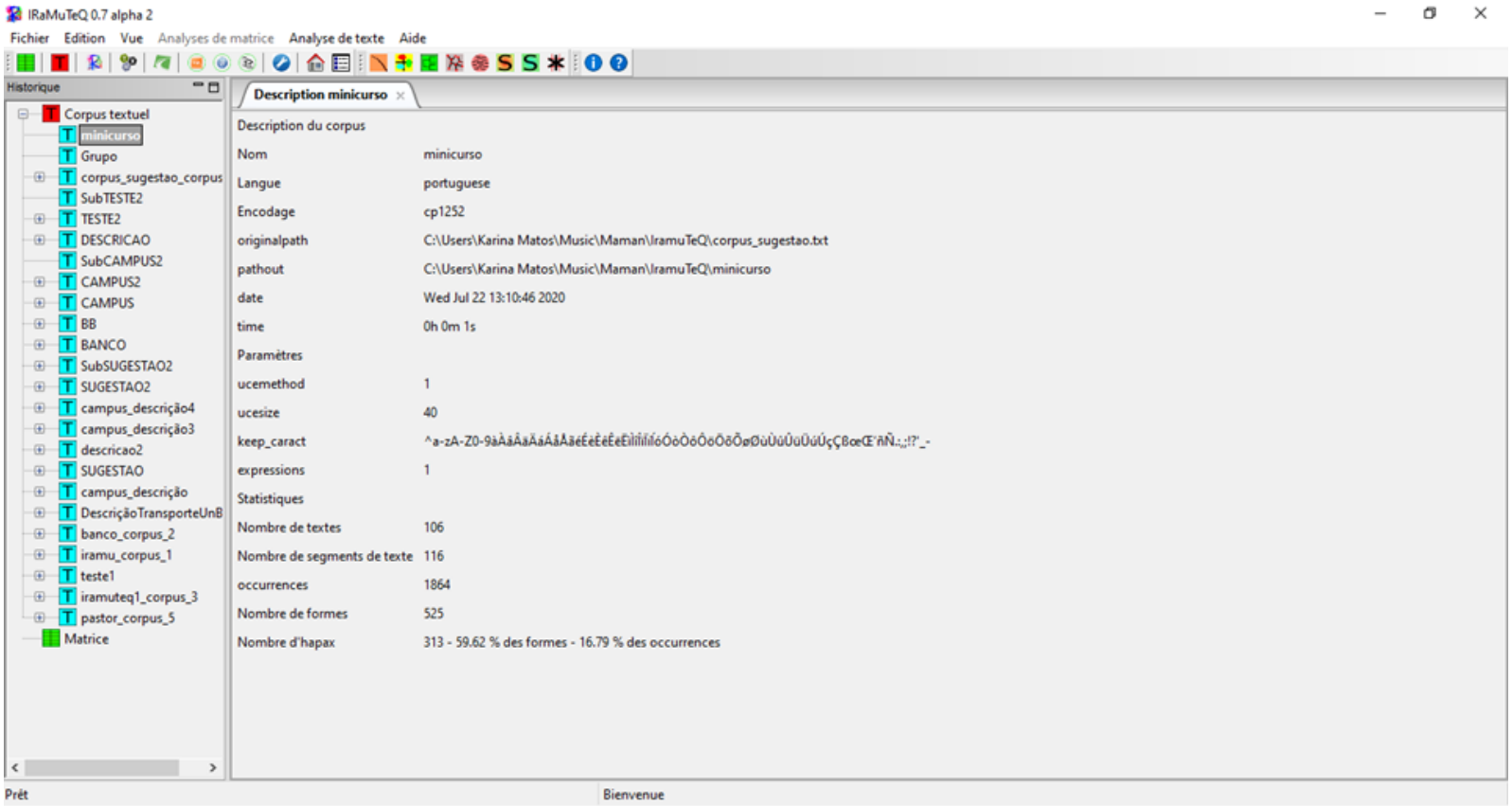
Aba Limpeza: IRaMuTeQ



Aba Limpeza: IRaMuTeQ (cont.)

- **Colocar o texto em minúsculo:** IRaMuTeQ converte de forma automática, todas as formas (palavras) em maiúsculas para minúsculas. Se desativar essa opção, IRaMuTeq vai considerar por exemplo, **UnB** e **unb**, como duas formas diferentes.
- **Eliminar as caráteres fora dessa lista:** Por padrão, somente as palavras alfanuméricas e acentuadas são levadas em contas.
- **Trocar apóstrofos por espaço:** Substituí apóstrofos por espaço.
- **Trocar o traço por espaço:** Substituí os - pelos espaços.
- **Manter a pontuação:** Se preferir manter a pontuação, deve-se evitar de usar (;) dentro do corpus.
- **Sem espaço entre duas formas:** Se essa opção for ativado, IRaMuTeQ não irá usar o espaço como delimitador entre formas.

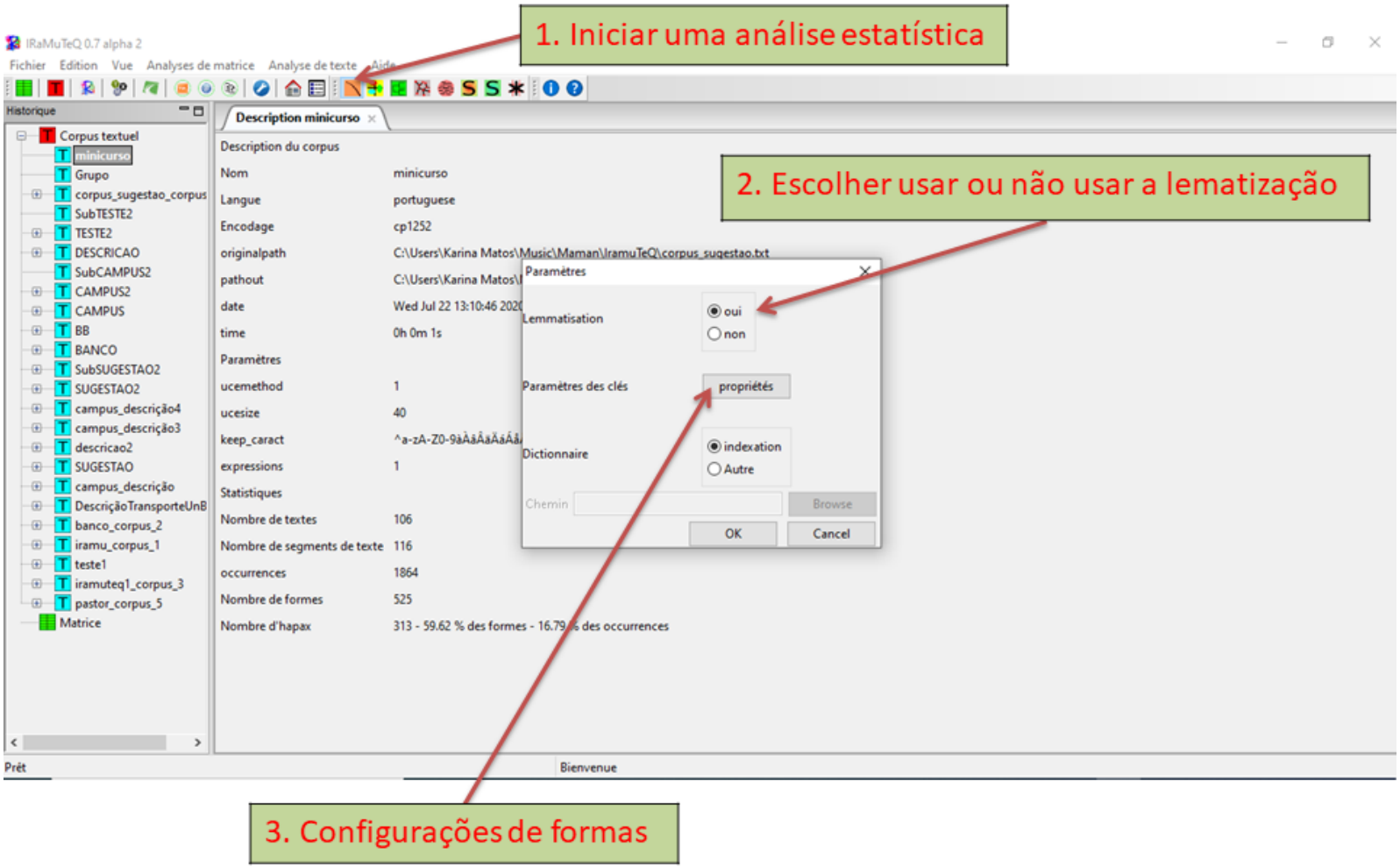
Descrição do corpus



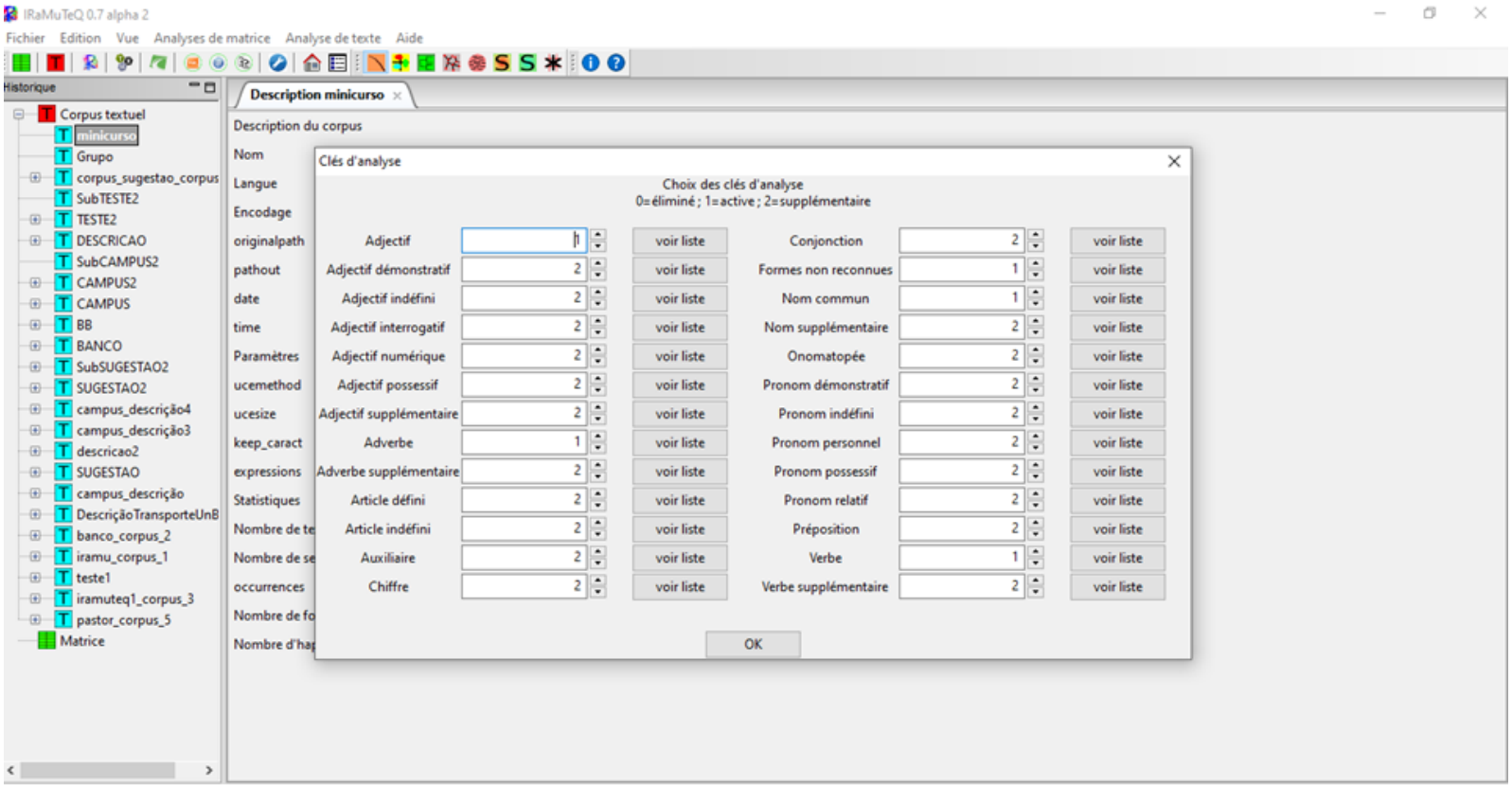
Análise: Estatística

A análise estatística ou análise lexicográfica, apresenta uma análise simples sobre o corpus de texto. Identifica e reformata as unidades de texto, transformando textos em ST; detecta as formas (palavras) com suas respectivas frequências, calcula a frequência média das formas e o total dos hapax no corpus; faz uma redução de palavras (forma reduzida) com base nos vocabulários que têm a mesma raiz ou os lematizam; monta um dicionário de formas reduzidas e mostra todas as formas ativas e suplementares.

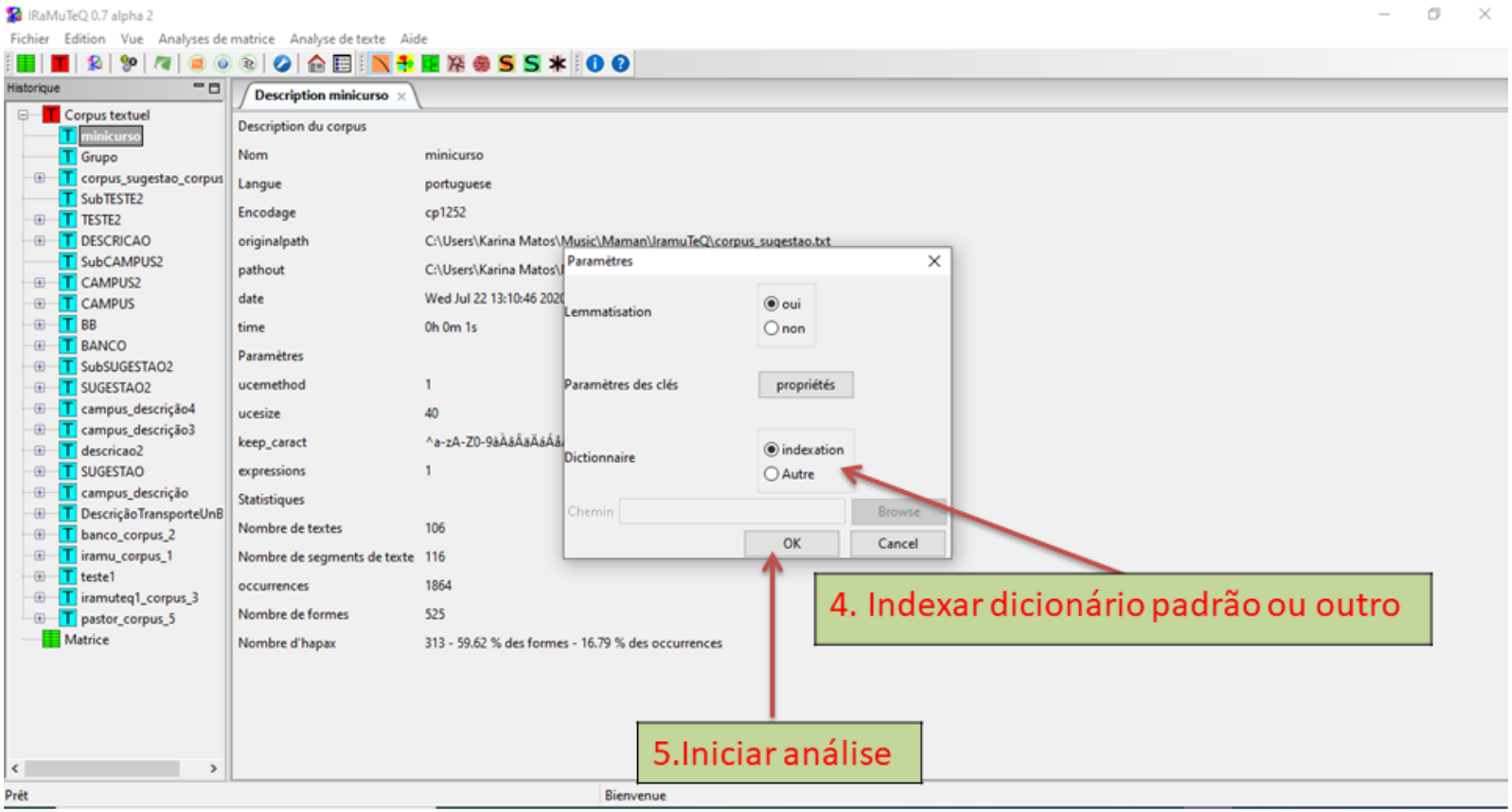
Análise: Estatística (cont.)



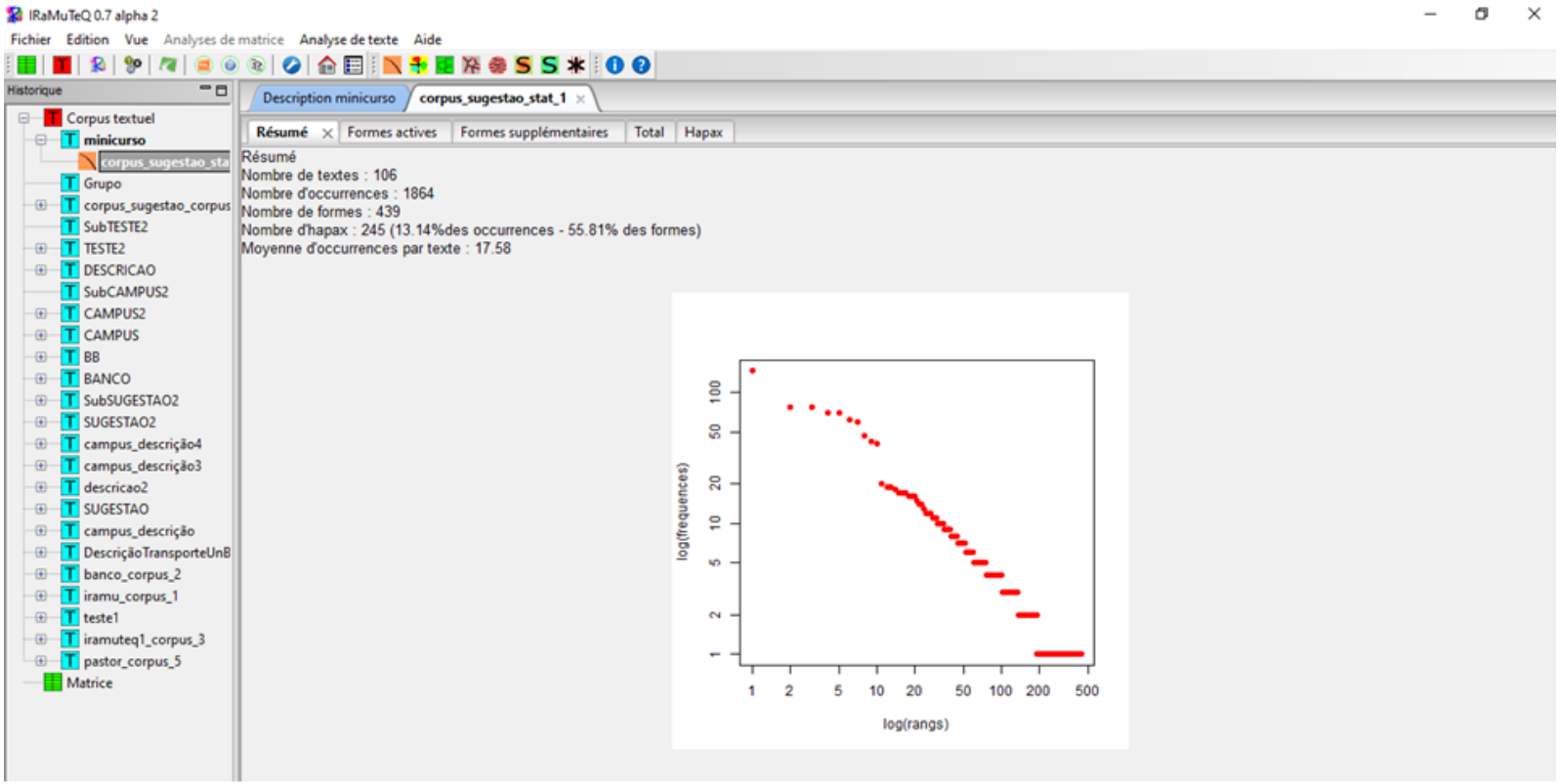
Análise: Estatística (cont.)



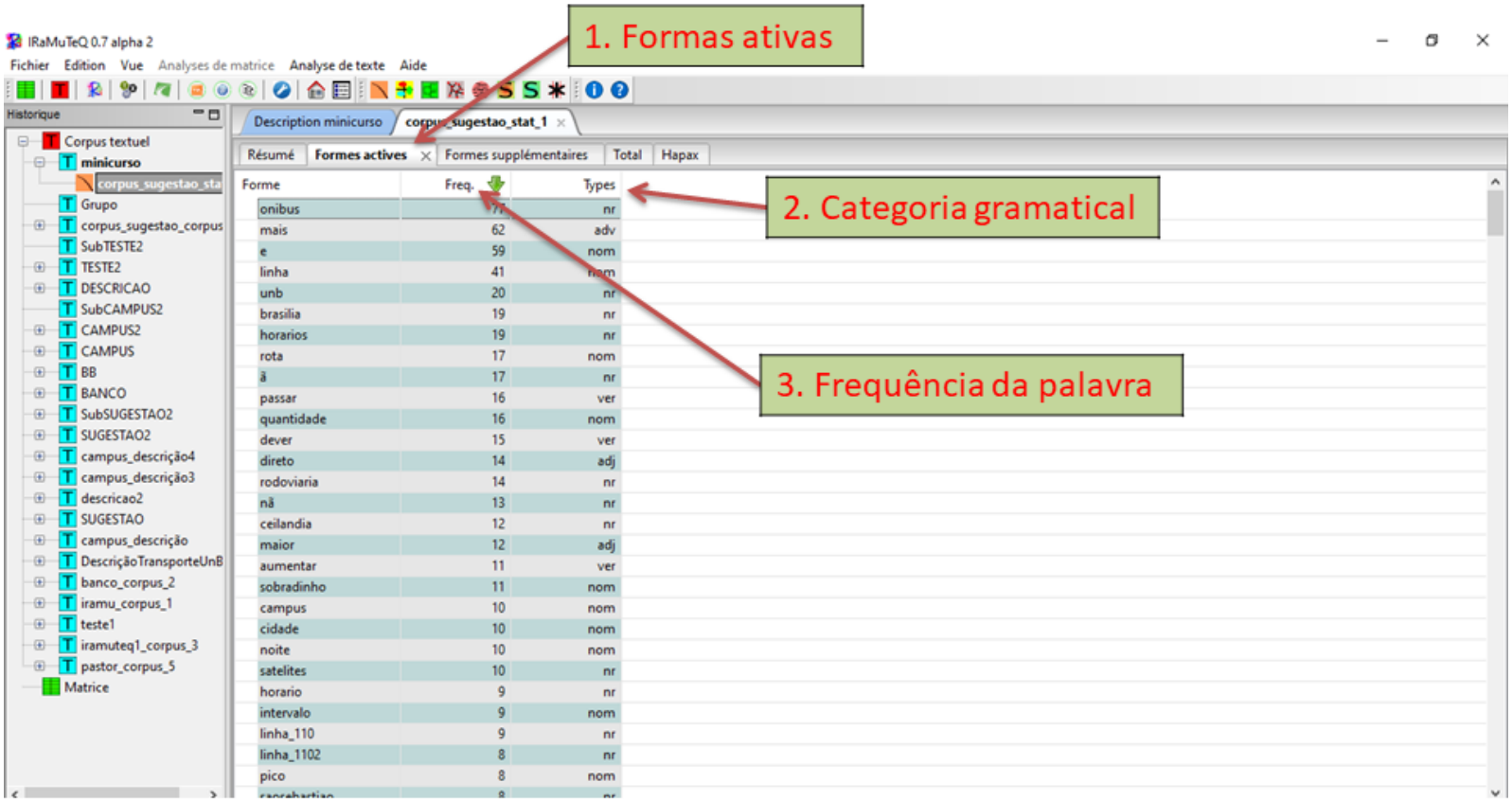
Análise: Estatística (cont.)



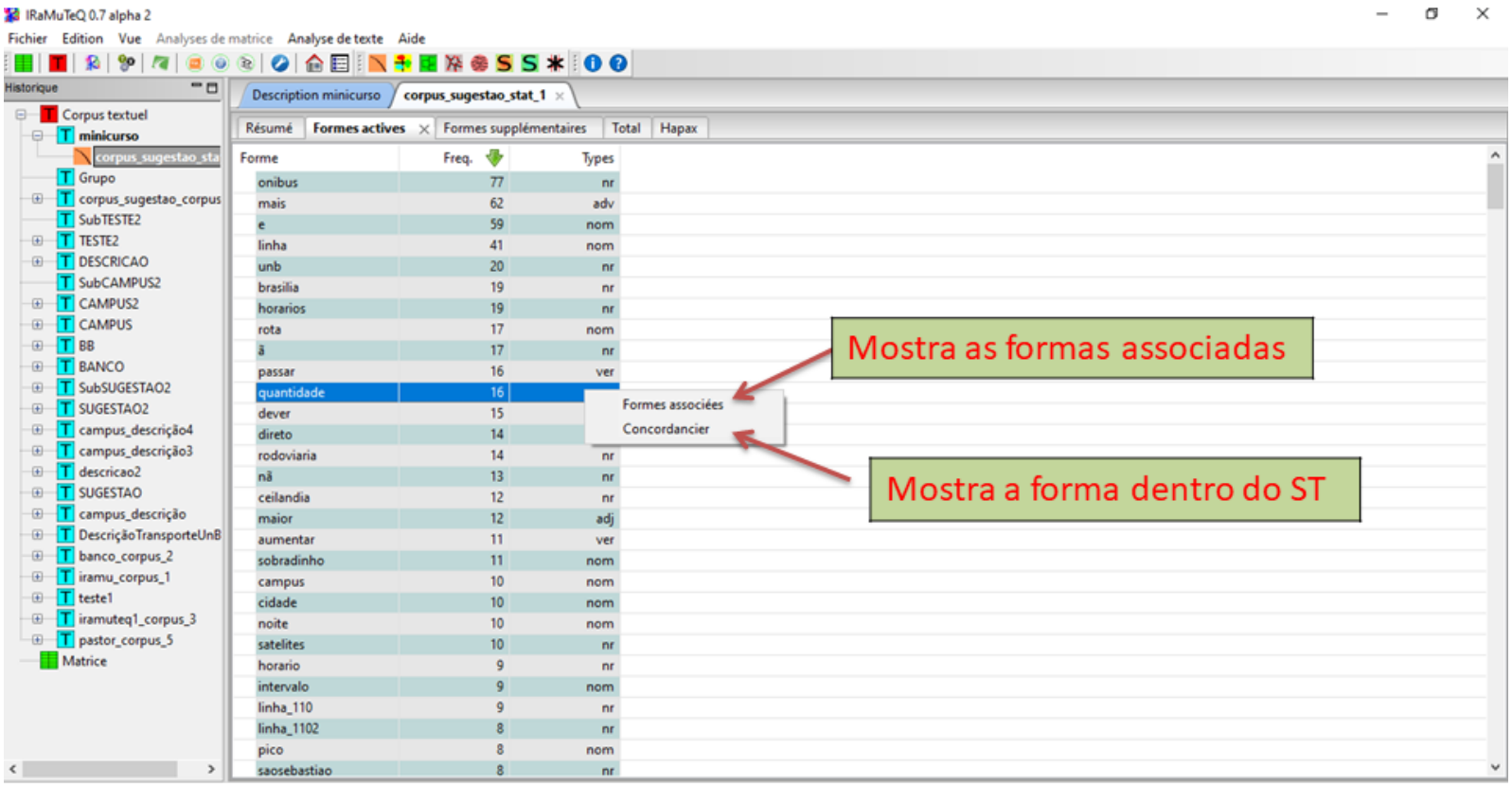
Resultado: análise estatística



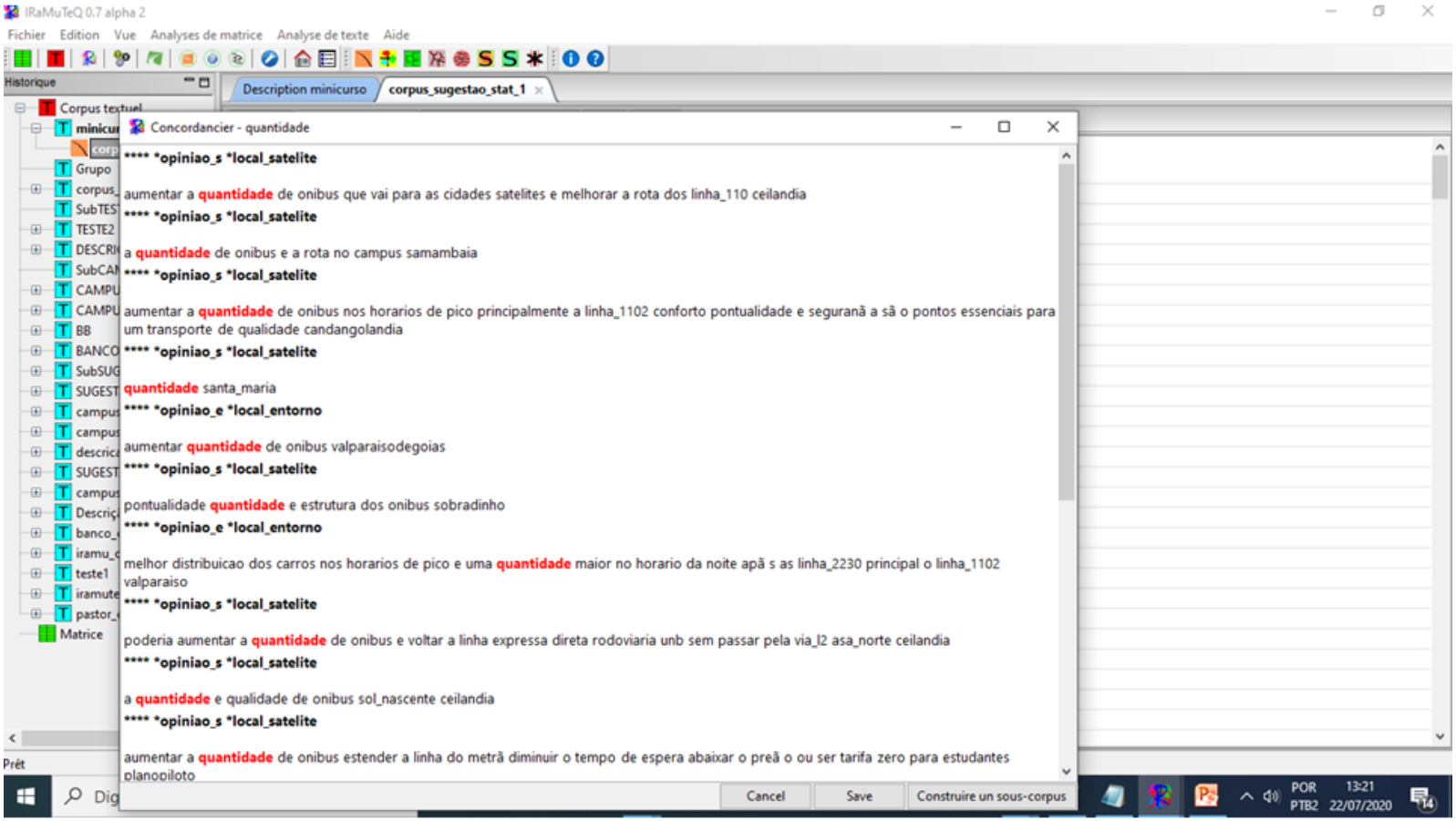
Resultado: análise estatística (cont.)



Resultado: análise estatística (cont.)



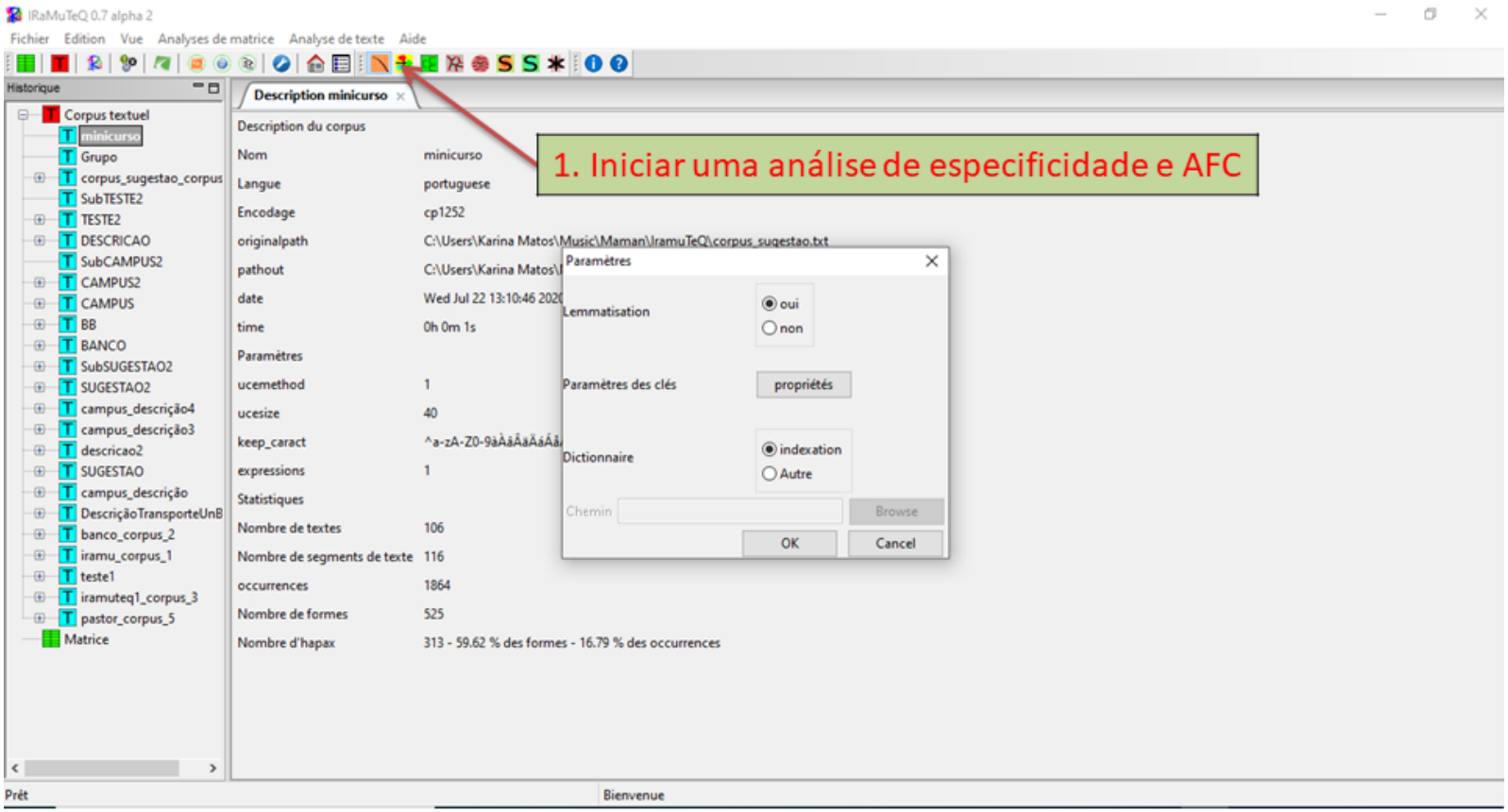
Resultado: análise estatística (cont.)



Análise de especificidade e AFC

Produz uma análise fatorial de correspondência em uma tabela de continência, cruzando as formas ativas com as variáveis. ou seja, possibilita a análise da produção textual em função das variáveis de caracterização com no mínimo duas modalidades.

Análise de especificidade e AFC (cont.)



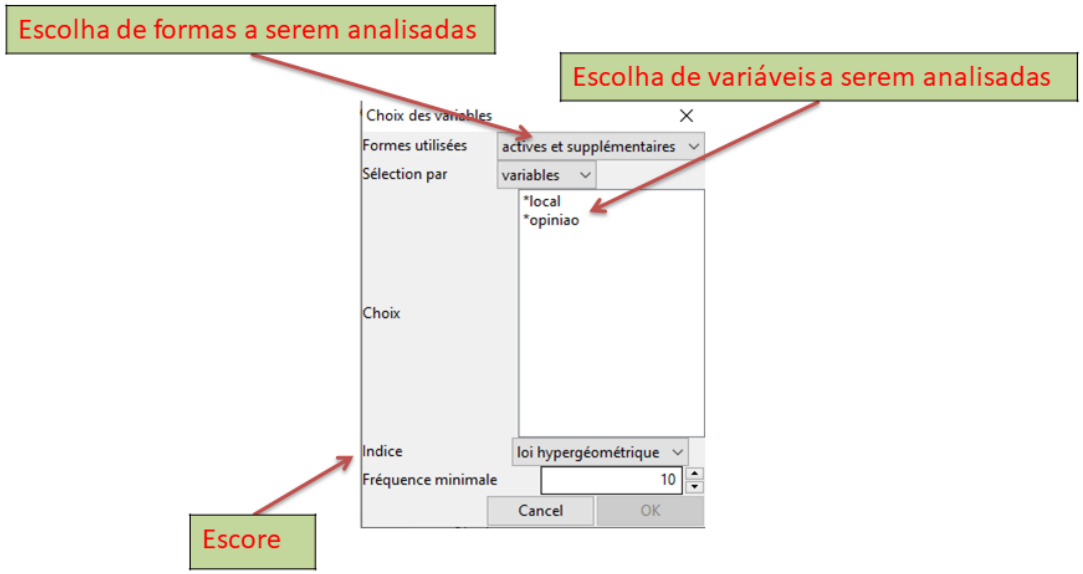
Análise de especificidade e AFC (cont.)

A seleção de uma variável é obrigatória. A variável selecionada, deve ter no mínimo duas modalidade. Caso selecionar duas ou mais variáveis, somente uma será levada em conta para análise.

A Análise Fatorial de Correspondência (AFC:) é uma representação gráfica que auxilia na visualização
dos dados e apresenta quão as classes ou formas se aproximam/distanciam.

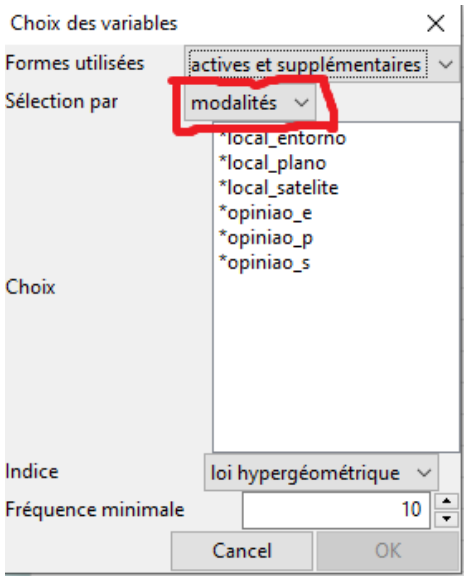
Existem duas possibilidade de cálculo nessa análise, a hipergeomêtrica e a X^2 (qui-quadrado)

Análise de especificidade e AFC (cont.)



Análise de especificidade e AFC (cont.)

Se escolhermos a seleção por modalidade, todas as modalidades devem pertencer a uma variável, caso contrário, a análise não terá sentido.



Análise de especificidade e AFC (cont.)

Formes | Formes banales | Types | Fréquences des formes | Fréquences des typ

formes	*opiniao_e	*opiniao_p	*opiniao_s
noite	2.8988	1.3872	-4.3355
quantidade	0.6527	-0.5261	0.2532
horarios	0.5421	0.4159	-0.641
mais	0.3329	-0.5488	0.4462
aumentar	0.3257	1.2032	-1.332
maior	0.2987	0.308	-0.3842
direto	0.2532	-0.4046	0.346
rodoviaria	0.2532	-0.7936	0.6508
passar	0.2163	0.9888	-0.9672
unb	-0.1602	-1.3147	1.2258
linha	-0.2263	-1.0445	1.0441
onibus	-0.2481	-0.7408	0.751
campus	-0.2522	-0.1968	0.3319
satelites	-0.2522	-0.4787	0.7009
cidade	-0.2522	-0.4787	0.7009
sobradinho	-0.2777	-1.1655	1.5229
ceilandia	-0.3033	-1.2731	1.6636
nao	-0.3547	1.2539	-0.8357
dever	-0.3806	-0.4641	0.7398
rota	-0.4324	0.5402	-0.2986
brasilia	-0.4844	8.4547	-6.8024

Spécificités - MiniCurso_spec_1

Formes | Formes banales | **Types** | Fréquences des formes | Fréquences des typ

formes	*opiniao_e	*opiniao_p	*opiniao_s
nr	1.3089	0.5499	-1.1535
pre	0.9157	-0.748	0.3517
pro_per	0.3378	0.9467	-0.9772
adj	0.3312	0.3804	-0.4262
conj	0.2469	-0.5803	0.5439
adv_sup	-0.0238	0.7347	-0.6241
adj_sup	-0.0238	-0.0884	0.1178
ono	-0.0238	-0.0884	0.1178
pro_int	-0.0477	-0.1769	0.2357
adj_num	-0.0477	-0.1769	0.2357
num	-0.1194	0.6372	-0.4672
pro_pos	-0.1194	-0.4426	0.5899
nom_sup	-0.263	-0.9756	1.3003
nom	-0.3026	-0.9831	0.9651
art_def	-0.3304	-0.3564	0.4489
ver	-0.333	0.8223	-0.6631
pro_rel	-0.4782	0.6335	-0.3842
adv	-0.6707	0.6522	-0.3642
pro_ind	-0.9158	-0.5654	1.0897
ver_sup	-1.48	0.3453	0.5587

Spécificités - MiniCurso_spec_1

Formes | Formes banales | Types | **Fréquences des formes** | Fréquences des typ

formes	*opiniao_e	*opiniao_p	*opiniao_s
onibus	4	13	60
mais	4	11	47
noite	4	5	1
horarios	2	5	12
quantidade	2	2	12
linha	2	5	34
direto	1	2	11
unb	1	1	18
aumentar	1	5	5
maior	1	3	8
passar	1	6	9
rodoviaria	1	1	12
nao	0	6	8
campus	0	2	8
rota	0	5	12
brasilia	0	16	3
ceilandia	0	0	12
satelites	0	1	9
dever	0	2	13
cidade	0	1	9
sobradinho	0	0	11

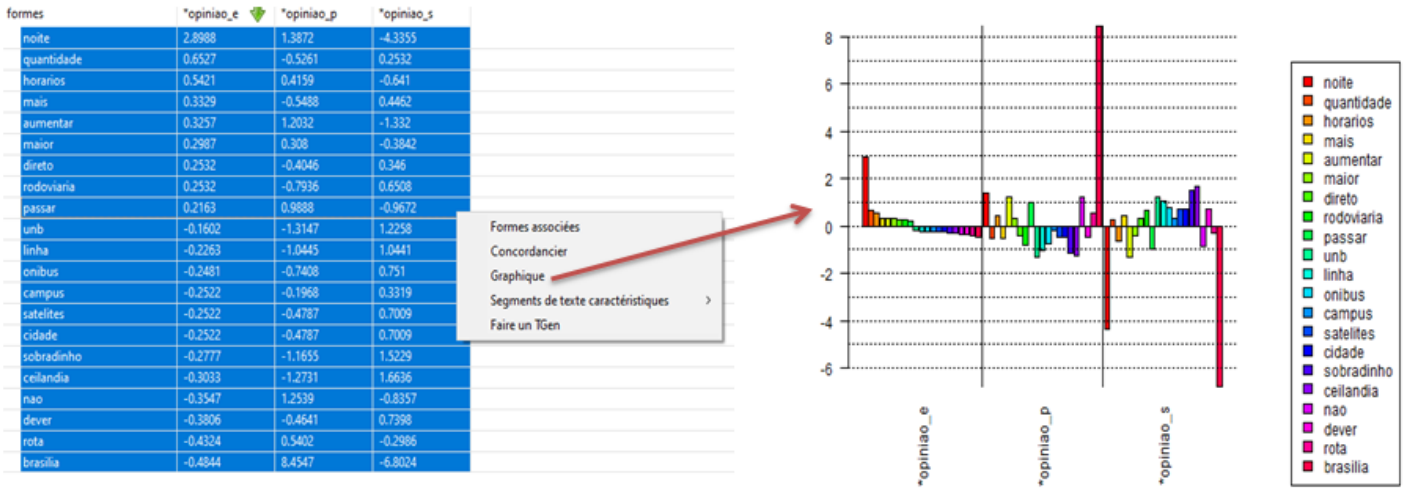
Spécificités - MiniCurso_spec_1

Formes | Formes banales | Types | Fréquences des formes | Fréquences des types

formes	*opiniao_e	*opiniao_p	*opiniao_s
onibus	166.67	141.3	191.08
mais	166.67	119.57	149.68
noite	166.67	54.35	3.18
horarios	83.33	54.35	38.22
quantidade	83.33	21.74	38.22
linha	83.33	54.35	108.28
direto	41.67	21.74	35.03
unb	41.67	10.87	57.32
aumentar	41.67	54.35	15.92
maior	41.67	32.61	25.48
passar	41.67	65.22	28.66
rodoviaria	41.67	10.87	38.22
nao	0	65.22	25.48
campus	0	21.74	25.48
rota	0	54.35	38.22
brasilia	0	173.91	9.55
ceilandia	0	0	38.22
satelites	0	10.87	28.66
dever	0	21.74	41.4
cidade	0	10.87	28.66
sobradinho	0	0	35.03

Figura 2: Resultado da análise de especificidade

Análise de especificidade e AFC (cont.)



formes	*opiniao_e	*opiniao_p	*opiniao_s
noite	2.8988	1.3872	-4.3355
quantidade	0.6527	-0.5261	0.2532
horarios	0.5421	0.4159	-0.641
mais	0.3329	-0.5488	0.4462
aumentar	0.3257	1.2032	-1.332
maior	0.2987	0.308	-0.3842
direto	0.2532	-0.4046	0.346
rodoviaria	0.2532	-0.7936	0.6508
passar	0.2163	0.9888	-0.9672
unb	-0.1602	-1.3147	1.2258
linha	-0.2263	-1.0445	1.0441
onibus	-0.2481	-0.7408	0.751
campus	-0.2522	-0.1968	0.3319
satelites	-0.2522	-0.4787	0.7009
cidade	-0.2522	-0.4787	0.7009
sobradinho	-0.2777	-1.1655	1.5229
ceilandia	-0.3033	-1.2731	1.6636
nao	-0.3547	1.2539	-0.8357
dever	-0.3806	-0.4641	0.7398
rota	-0.4324	0.5402	-0.2986
brasilia	-0.4844	8.4547	-6.8024

Figura 3: Gráfico de escore para cada forma

Análise de especificidade e AFC (cont.)

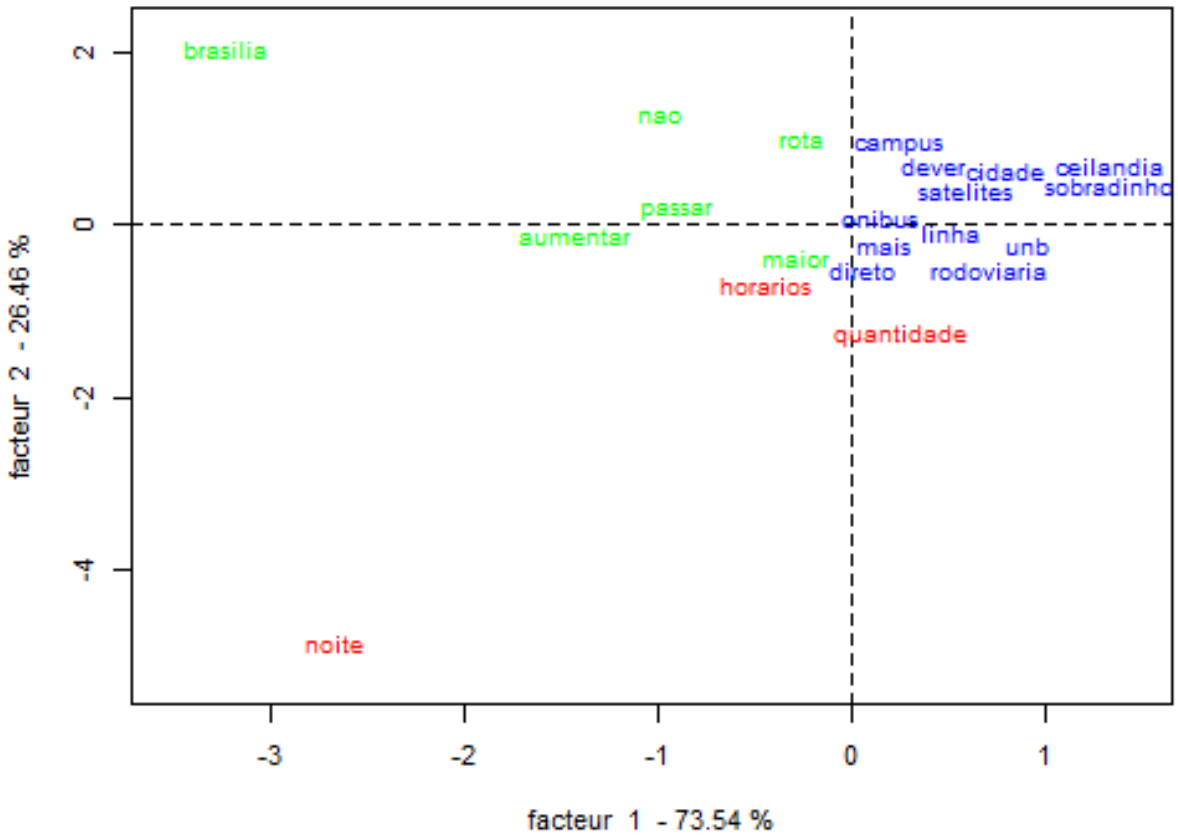


Figura 4: Gráfico AFC

Análise por classificação hierárquica descendente - CHD

Essa análise está baseado em princípios de teoria de grafos baseado na teoria descrito por Reinert, permitindo detectar as coocorrências entre formas, auxiliando na identificação da estrutura de conteúdo de um corpus. [11]

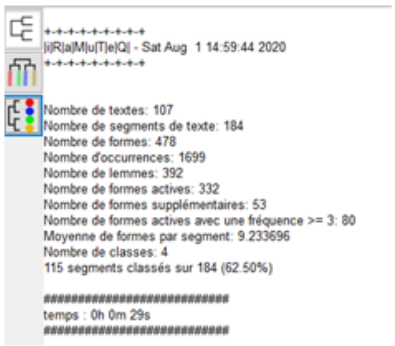
A classificação hierárquica descedente descrito por Reibert, visa obter classes de ST que apresentam vocabulários próximos ou distantes lexicalmente e com a ideia de que as palavras usadas em contexto similar, sejam associadas em um mesmo mundo léxico. [7]

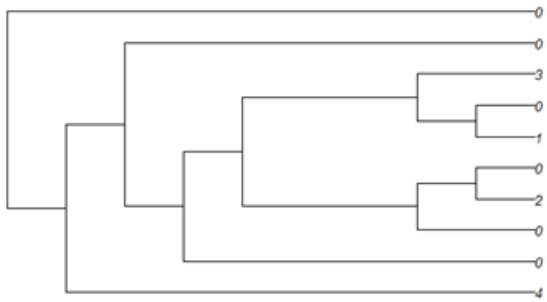
Esta classificação pode ser realizada sob três formas de modalidades diferentes: **classificação simples sobre o texto**, **classificação simples sobre o segmento do texto** e por fim a **classificação dupla sobre o segmento de texto**.

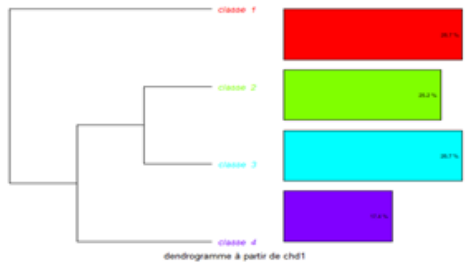
Análise por classificação hierárquica descendente - CHD (cont.)

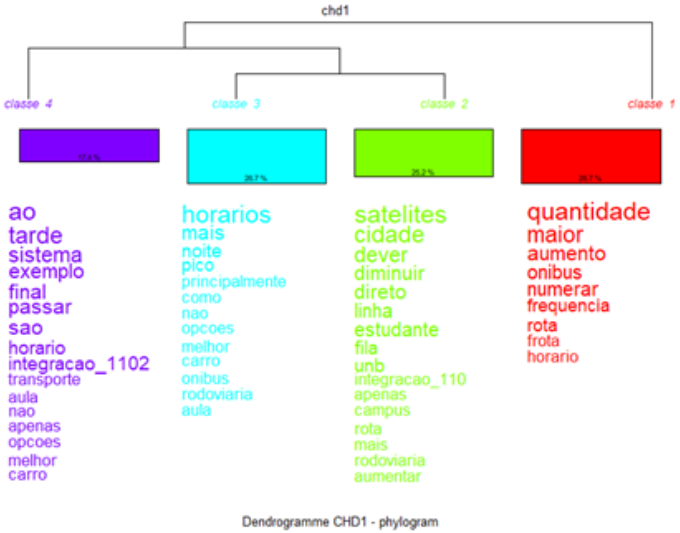
- Classificação simples sobre o texto: os textos mantêm a sua integridade, a classificação permite o reagrupamento dos textos mais próximos
- Classificação simples sobre o segmento do texto: a classificação incide sobre os segmentos de texto
- Classificação dupla sobre o segmento de texto: A classificação 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. [11]

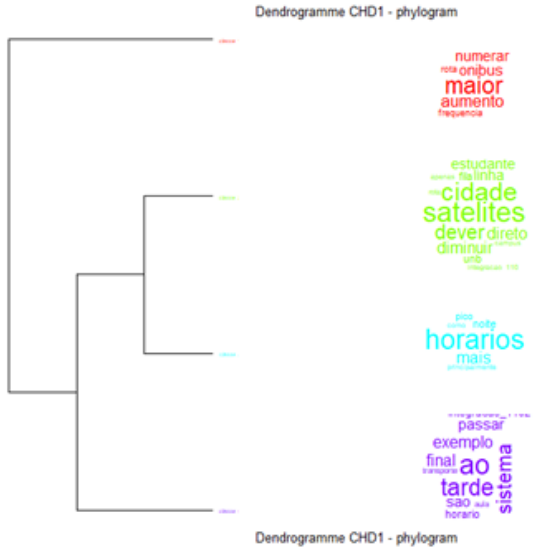
Análise por classificação hierárquica descendente - CHD (cont.)



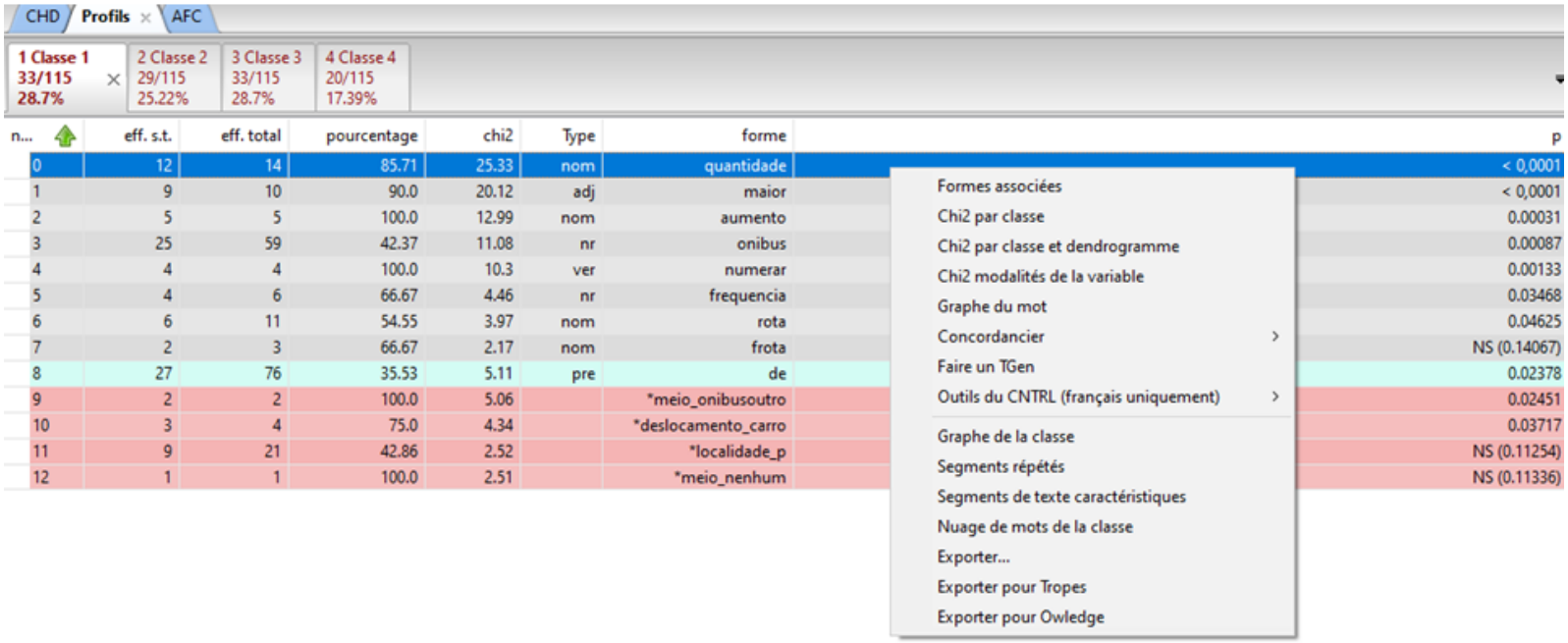








Análise por classificação hierárquica descendente - CHD (cont.)



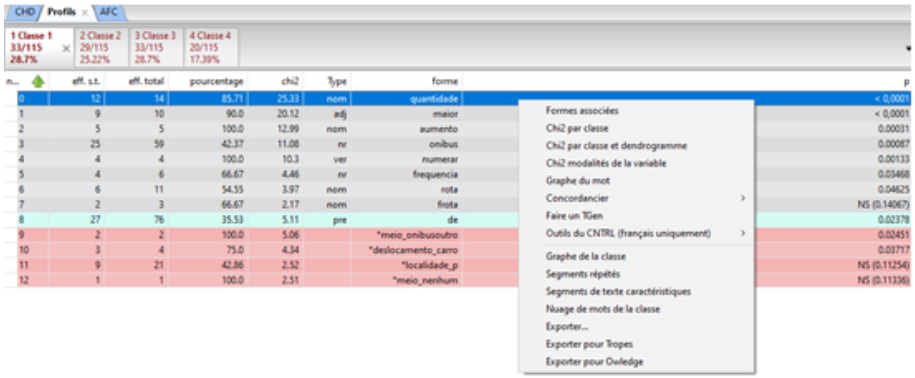
CHD | Profils | AFC

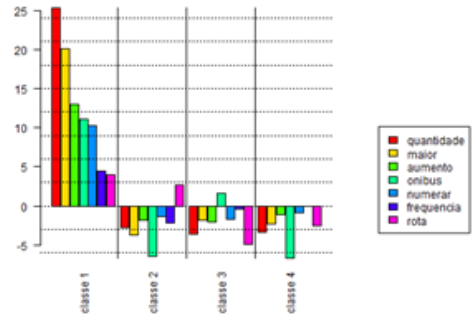
1 Classe 1 33/115 28.7%	2 Classe 2 29/115 25.22%	3 Classe 3 33/115 28.7%	4 Classe 4 20/115 17.39%

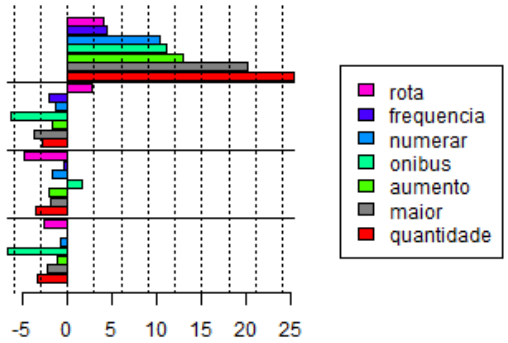
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	12	14	85.71	25.33	nom	quantidade	< 0,0001
1	9	10	90.0	20.12	adj	maior	< 0,0001
2	5	5	100.0	12.99	nom	aumento	0.00031
3	25	59	42.37	11.08	nr	onibus	0.00087
4	4	4	100.0	10.3	ver	numerar	0.00133
5	4	6	66.67	4.46	nr	frequencia	0.03468
6	6	11	54.55	3.97	nom	rota	0.04625
7	2	3	66.67	2.17	nom	frota	NS (0.14067)
8	27	76	35.53	5.11	pre	de	0.02378
9	2	2	100.0	5.06		*meio_onibusoutro	0.02451
10	3	4	75.0	4.34		*deslocamento_carro	0.03717
11	9	21	42.86	2.52		*localidade_p	NS (0.11254)
12	1	1	100.0	2.51		*meio_nenhum	NS (0.11336)

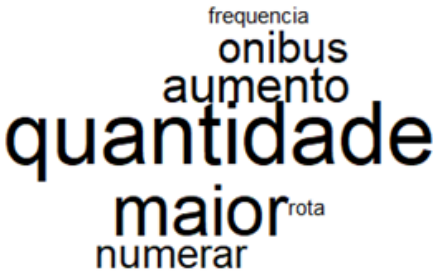
- Formes associées
- Chi2 par classe
- Chi2 par classe et dendrogramme
- Chi2 modalités de la variable
- Graphe du mot
- Concordancier ›
- Faire un TGen
- Outils du CNTRL (français uniquement) ›
- Graphe de la classe
- Segments répétés
- Segments de texte caractéristiques
- Nuage de mots de la classe
- Exporter...
- Exporter pour Tropes
- Exporter pour Owledge

Análise por classificação hierárquica descendente - CHD (cont.)

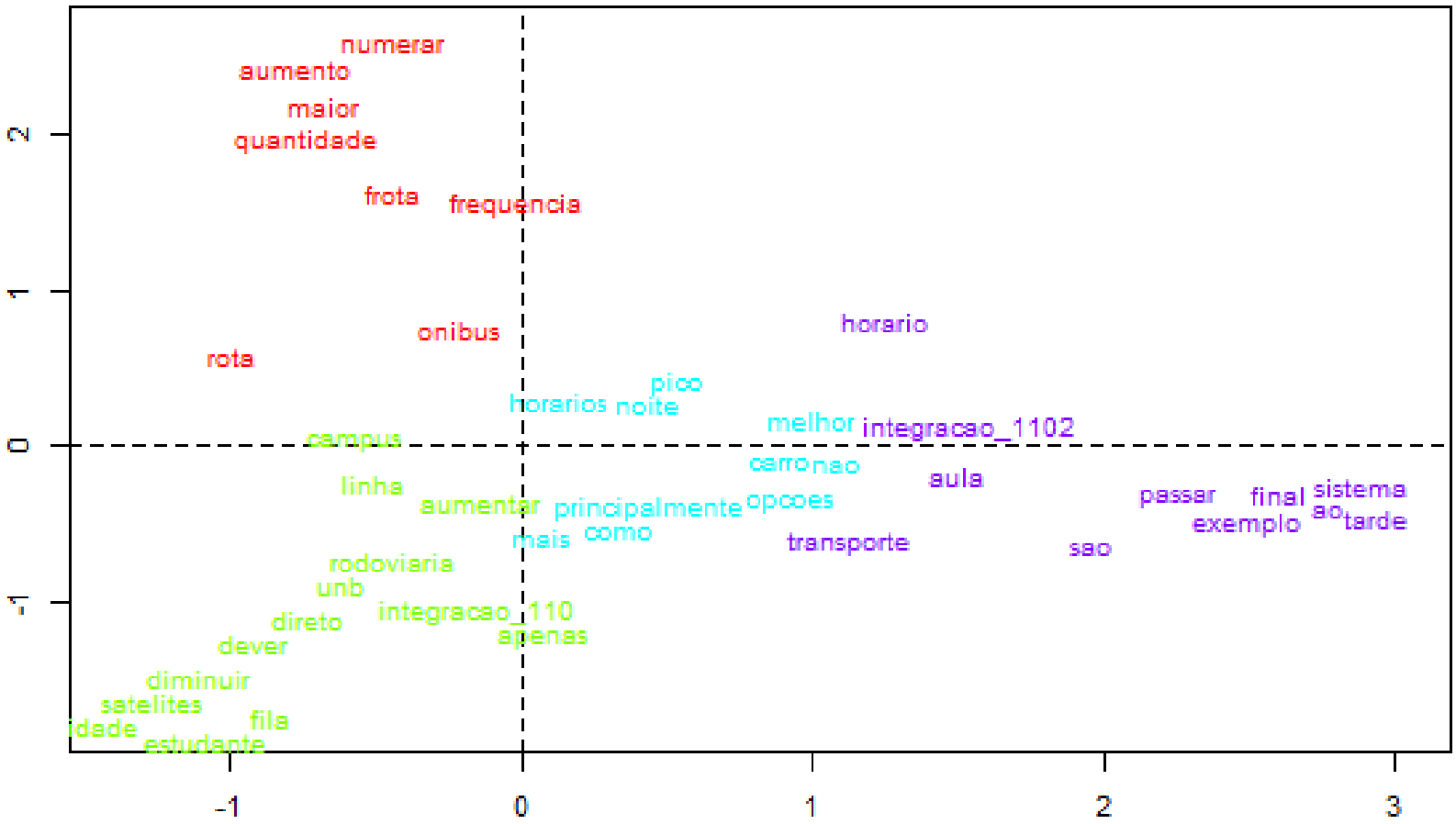








Análise por classificação hierárquica descendente - CHD (cont.)



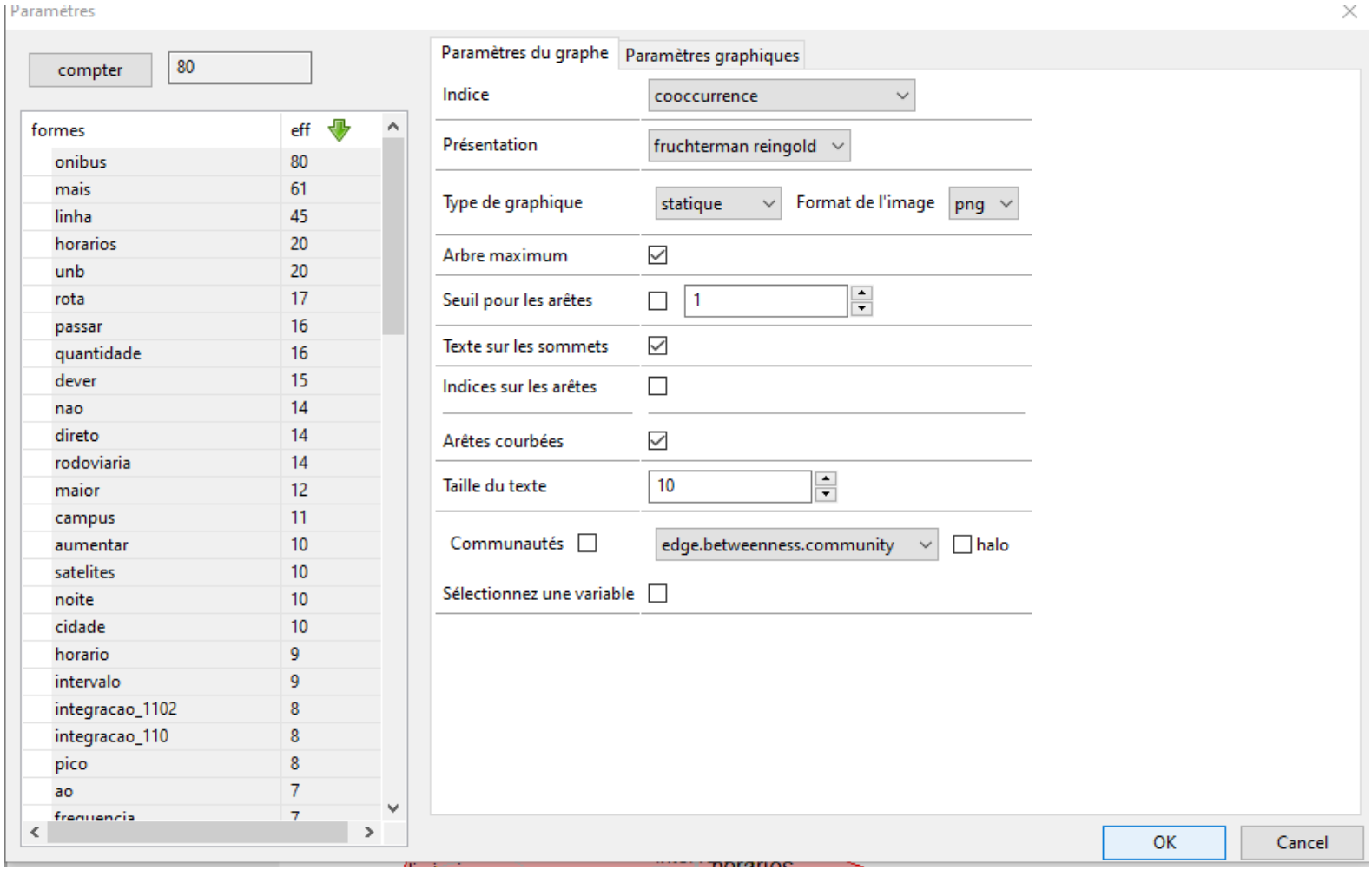
Análise de similitude

A Análise de similitude é baseada na teoria dos grafos cujos resultados auxiliam no estudo das relações entre objetos de um modelo matemático. [7]

A análise de similitude no software IRaMuTeQ apresenta um grafo que mostra a ligação entre formas de um corpus. 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. [11]

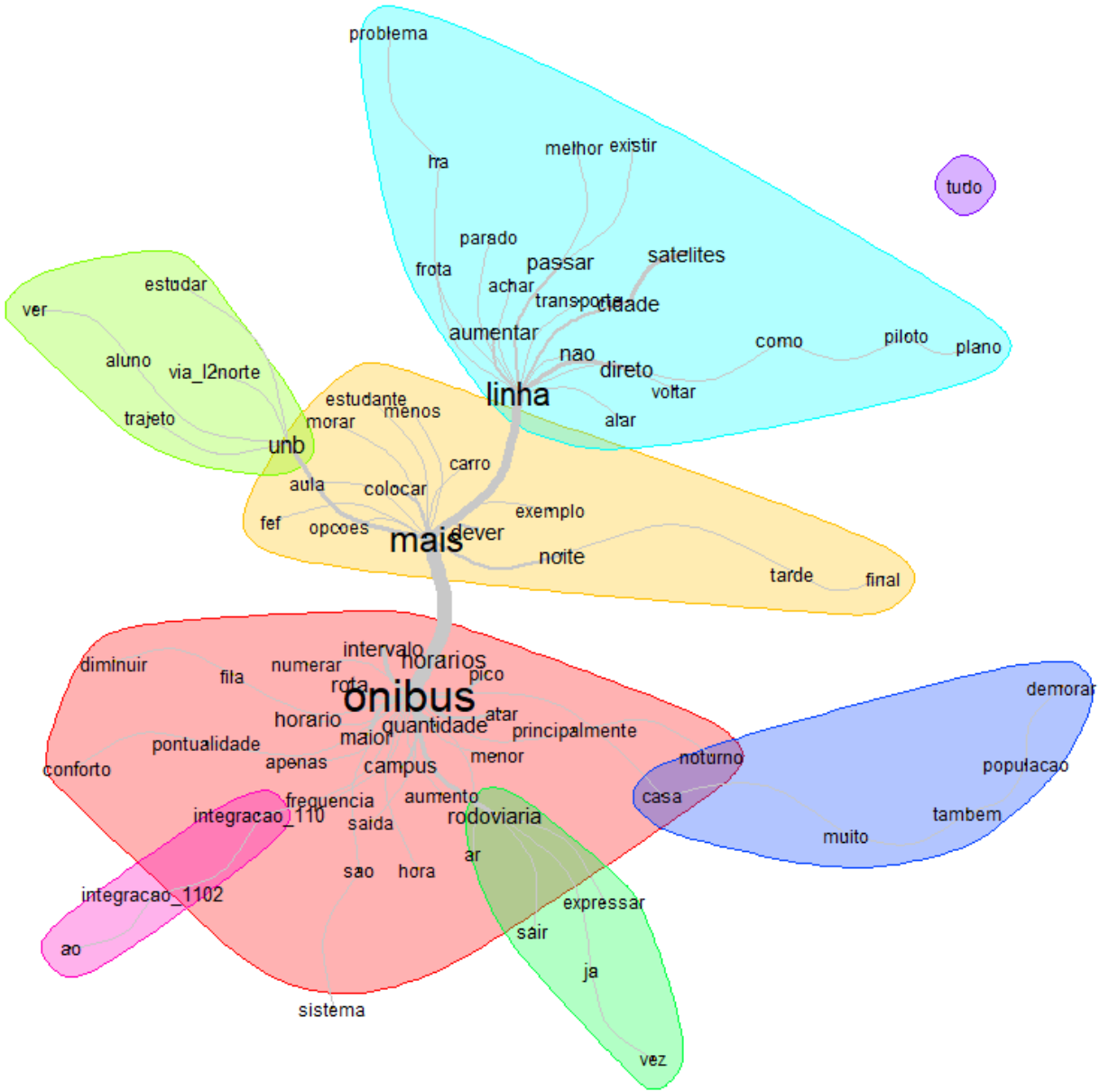
Ela auxilia o pesquisador na identificação da estrutura da base de dados (corpus), distinguindo as partes comuns e as especificidades, além de permitir verificá-las em função das variáveis descritivas existentes.

Análise de similitude (cont.)

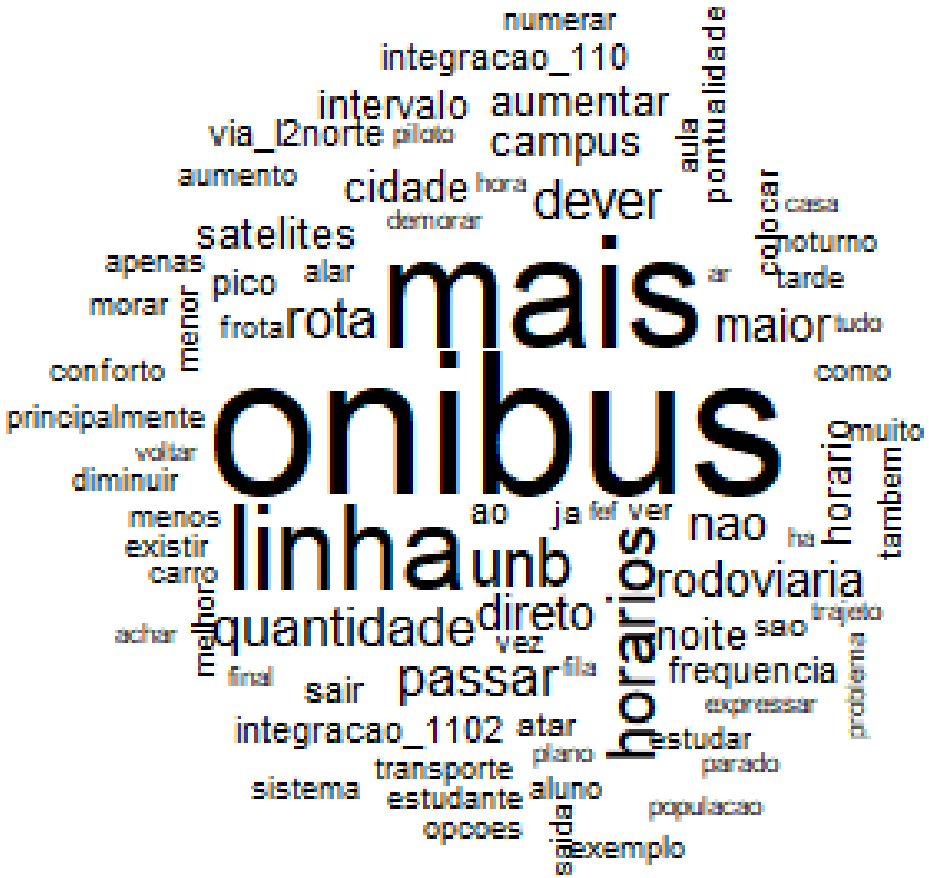


Análise de similitude (cont.)

Análise de similitude (cont.)



Nuvem de palavras



Bibliografia

[1] P. F. .-. IME-USP. (2014), Teoria de grafos, URL: https://www.ime.usp.br/~pf/algoritmos_para_grafos/aulas/graphs.html.

[2] Microsoft. (2020), Codificação de caráteres, URL: https://docs.microsoft.com/pt-br/powershell/scripting/dev-cross-plat/vscode/understanding-file-encoding?view=powershell-7.

[3] significados. (2020), bit - Significado de Byte, URL: https://www.significados.com.br/.

[4] M. E. Salviati. (2017), Manual do Aplicativo Iramuteq, URL: http://www.iramuteq.org/documentation/fichiers/manual-do-aplicativo-iramuteq-par-maria-elisabeth-salviati.

[5] P. Feofiloff. (2018), Unicode e UTF-8, URL: https://www.ime.usp.br/~pf/algoritmos/.

[6] wikipedia. (2015), Pixels, URL: https://pt.wikipedia.org/wiki/Pixel.

[7] M. E. SALVIATI, "Manual do aplicativo Iramuteq", *UNB planaltina. Recuperado em novembro*, vol. 20, p. 2018, 2017.

Bibliografia (cont.)

[8] eecis. (2016), distribuição de frequencia, URL: https://www.eecis.udel.edu/~portnoi/classroom/prob_estatistica/2006_1/lecture_slides/aula04.pdf.

[9] E. Reis, P. Melo, R. Andrade e T. Calapez, "Estatıstica aplicada", *Lisboa: Edições Sılabo*, 1999.

[10] H. J. Bortolossi, J. J. D. B. Queiroz e M. M. da Silva, "A Lei de Zipf e Outras Leis de Potência em Dados Empıricos",, 2012.

[11] P. Marchand e P. Ratinaud, "L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011)", *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT*, vol. 2012, pp. 687–699, 2012.