DESENVOLVIMENTO DE UM PARSER DE CONECTORES TEXTUAIS E SUA APLICAÇÃO PARA ANÁLISE DE GÊNEROS TEXTUAIS

Leonardo Zilio (Letras/UFRGS)

Rodrigo Wilkens (PPG-Comp/UFRGS)

IX ELC

Porto Alegre – 08 de Outubro de 2010

- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

Introdução

- Dois trabalhos
  - Desenvolvimento de uma ferramenta
  - Aplicação da mesma

- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

Ferramenta - Objetivos

- Identificação e classificação automática de conectores
- Emprego em corpora

Ferramenta - Referencial

- Computacional:
  - DiZer 2.0
  - RST
- Linguístico:
  - Neves (2000) – Gramática de Usos do Português

Ferramenta - Método

- Compilação de listas de conectores de Neves (2000)
- Corpus DiZer
  - 40 extratos de textos de diferentes áreas
  - 4.105 tokens
- Anotação manual do corpus

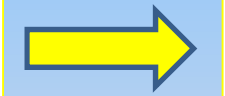
Ferramenta - Método

- Extração manual de regras e de conectores a partir do corpus
- Teste do funcionamento das regras em relação ao corpus anotado
- Aperfeiçoamento da ferramenta com base no teste

Ferramenta - Formatação

- Pré-processamento e PALAVRAS (Bick, 2000)
- Segmentação da saída em orações
- Classificação

BLA BLA BLA.
BLA BLA BLA.
BLA BLA BLA.



```
BLA     <artd> DET F S @>N  #1->2
BLA     <cjt-head> <act> N F S @SUBJ>  #2->12
BLA     <vH> <fmc> <mv> V PR 3P IND VFIN @FS-STA  #12->0
```

BLA / BLA BLA.
BLA BLA / BLA.
BLA / BLA BLA.

BLA – Sem conector
BLA BLA. – Adversativa

BLA BLA – Final
BLA. – Sem conector

BLA – Causal
BLA BLA. – Comparativa

Ferramenta – Segmentação

- Regras de segmentação
  - V FIN + V FIN
  - para + INF
  - além=de + INF
  - etc.

13 regras de segmentação

Regras de segmentação

```
para + INF - para - 3
além=de + INF - além=de - 3
depois + INF - depois - 3
antes + INF - antes - 3
apesar=de + INF - apesar=de - 3
a=fim=de + INF - a=fim=de - 3
sem + INF - sem - 3
o=que + INF - INF - 3
como + INF - INF - 3
semantes + INF - INF - 3
de=maneira=a + INF - de=maneira=a - 3
do=que + INF - do=que - 2
```

Ferramenta – Classificação

- Dicionário de conectores não previstos pelo PALAVRAS (20 expressões compostas acrescentadas)

- Regras de classificação
  - Listas de conectores
  - Regras especiais
    - apesar=de + INF → concessiva
    - seja + seja → alternativa
    - etc.

27 regras especiais de classificação

Regras especiais de classificação

```
para + INF - final - para - 3
depois + INF - temporal - depois - 3
antes + INF - temporal - antes - 3
além=de + INF - aditiva - além=de - 3
apesar=de + INF - concessiva - apesar=de - 3
a=fim=de + INF - final - a=fim=de - 3
sem + INF - condicional - sem - 3
o=que + INF - complementadora - INF - 3
como + INF - modal - INF - 3
seja + seja - alternativa - seja - 3
istoé + VFIN - parafrástica - VFIN - 3
ou=seja + VFIN - parafrástica - VFIN - 3
desde=que + IND - temporal - IND - 3
desde=que + SUBJ - condicional - SUBJ - 3
consoante/ADV - conformativa - consoante - 3
conforme/ADV - conformativa - conforme - 3
segundo/ADV - conformativa - segundo - 3
consoante/PRP - conformativa - consoante - 3
conforme/PRP - conformativa - conforme - 3
segundo/PRP - conformativa - segundo - 3
que/SPEC - relativa - que - 1
que/KS - complementadora - que - 1
se/KS - condicional/hipotética - se - 1
caso/KS - condicional - caso - 1
como/KS - causal - como - 1
como/ADV - comparativa - como - 1
nem/KC - aditiva - nem - 1
```

Ferramenta – Listas de Conectores

- 14 Listas
  - **Aditivas** (e, assim como, também, além disso)
  - **Adversativas** (mas, porém, entretanto etc.)
  - **Causais** (porque, por isso, pois etc.)
  - **Comparativas** (como se, do mesmo modo que, tal qual etc.)
  - **Concessivas** (embora, conquanto, ainda que)
  - **Condicionais** (contanto que, a não ser que, a menos que etc.)

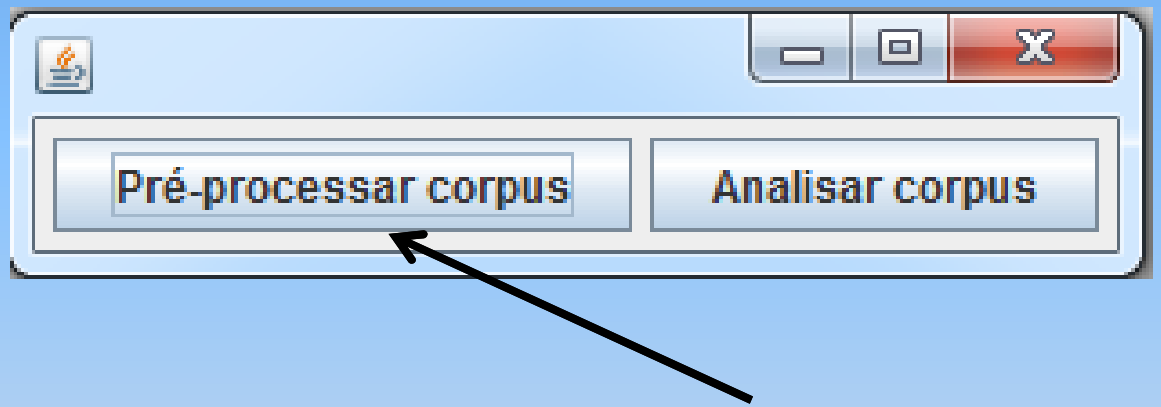
Ferramenta – Listas de Conectores 2

- **Conformativas** (à medida que, de acordo)
- **Consecutivas** (de tal maneira que, a tal ponto que, de tal modo que etc.)
- **Disjuntivas** (ou)
- **Finais** (para que, a fim de, de modo que)
- **Modais** (sem que)
- **Parafrásticas** (de outro modo, em outras palavras, de outra maneira etc.)
- **Temporais** (quando, antes que, depois que etc.)
- **Relativas** (o que, onde, tudo o que etc.)

Ferramenta – Listas de Conectores 3

porque
pois
pois=que
porquanto
já=que
uma=vez=que
dado=que
visto=que
visto como
pois=que
tanto=mais=que
porcausaque
porcausadeque
por=isso=que
portanto
porisso
por=conseqüência
porconsequência
assim

Ferramenta - Funcionamento



Primeiro passo – Pré-processar o corpus

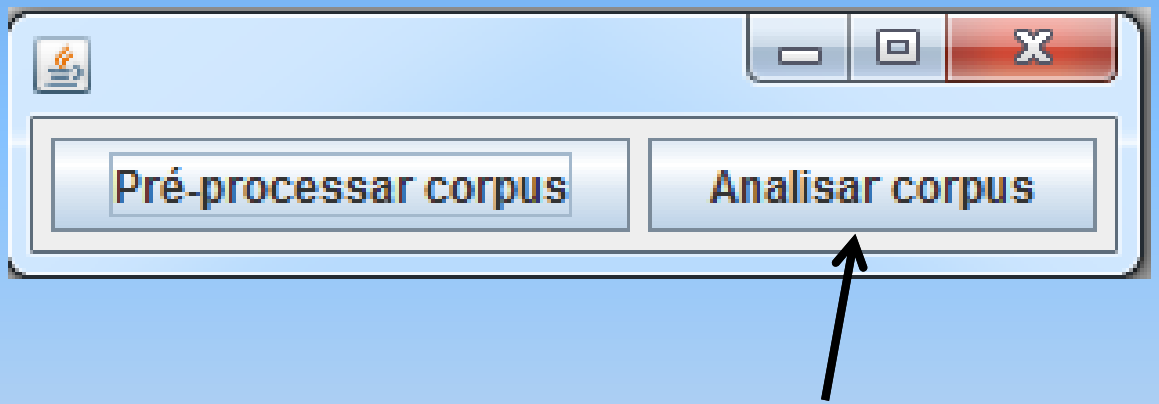
Essa etapa faz alguns ajustes ao corpus antes de passar pelo PALAVRAS.

Ferramenta - Funcionamento

?

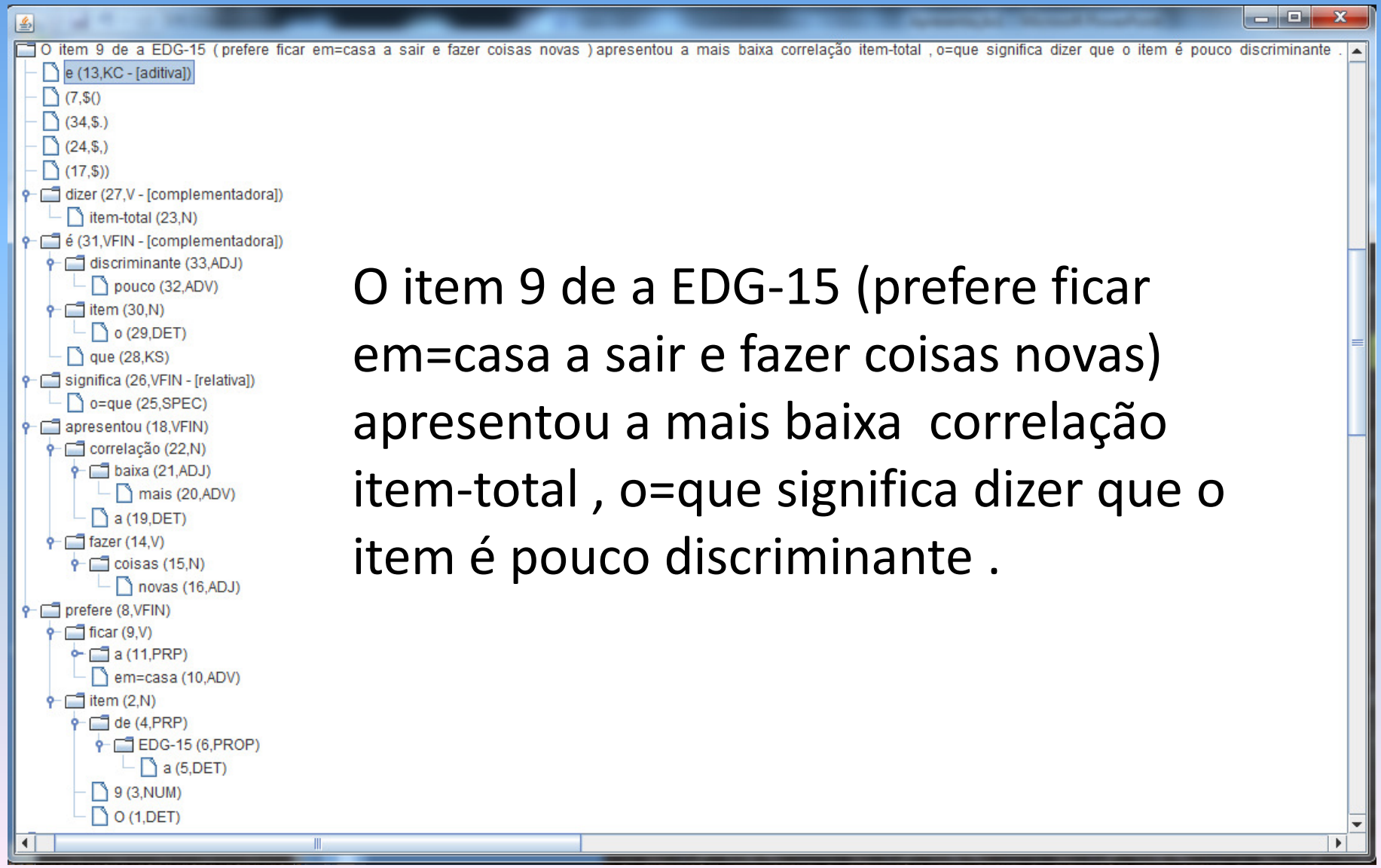
Segundo passo – PALAVRAS

Ferramenta - Funcionamento



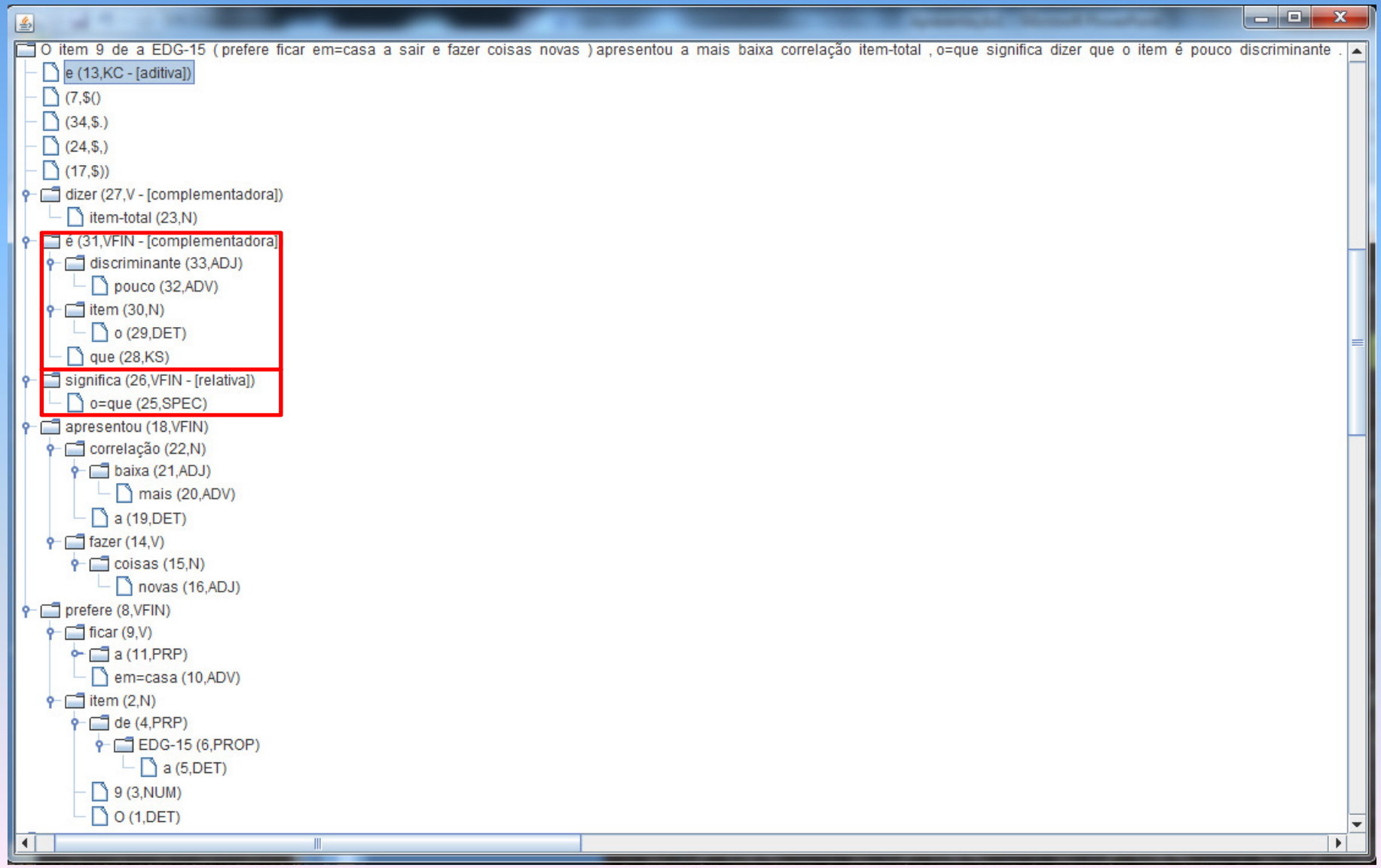
Terceiro passo – Análise

Ferramenta – Na prática

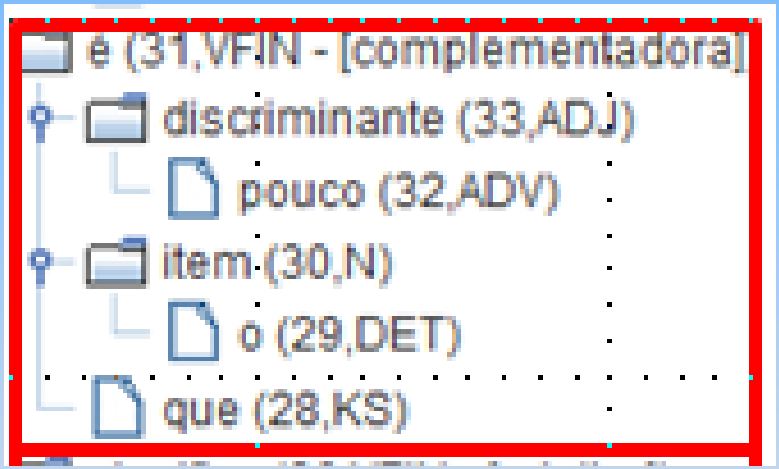


O item 9 de a EDG-15 (prefere ficar em=casa a sair e fazer coisas novas) apresentou a mais baixa correlação item-total , o=que significa dizer que o item é pouco discriminante .

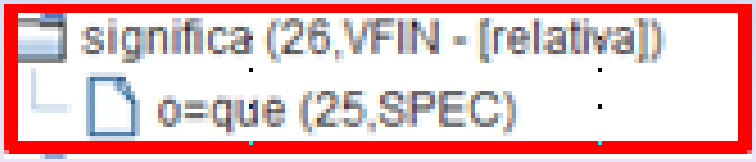
Ferramenta - Acertos



O item 9 de a EDG-15 (prefere ficar em=casa a sair e fazer coisas novas) apresentou a mais baixa correlação item-total , o=que significa dizer que o item é pouco discriminante .

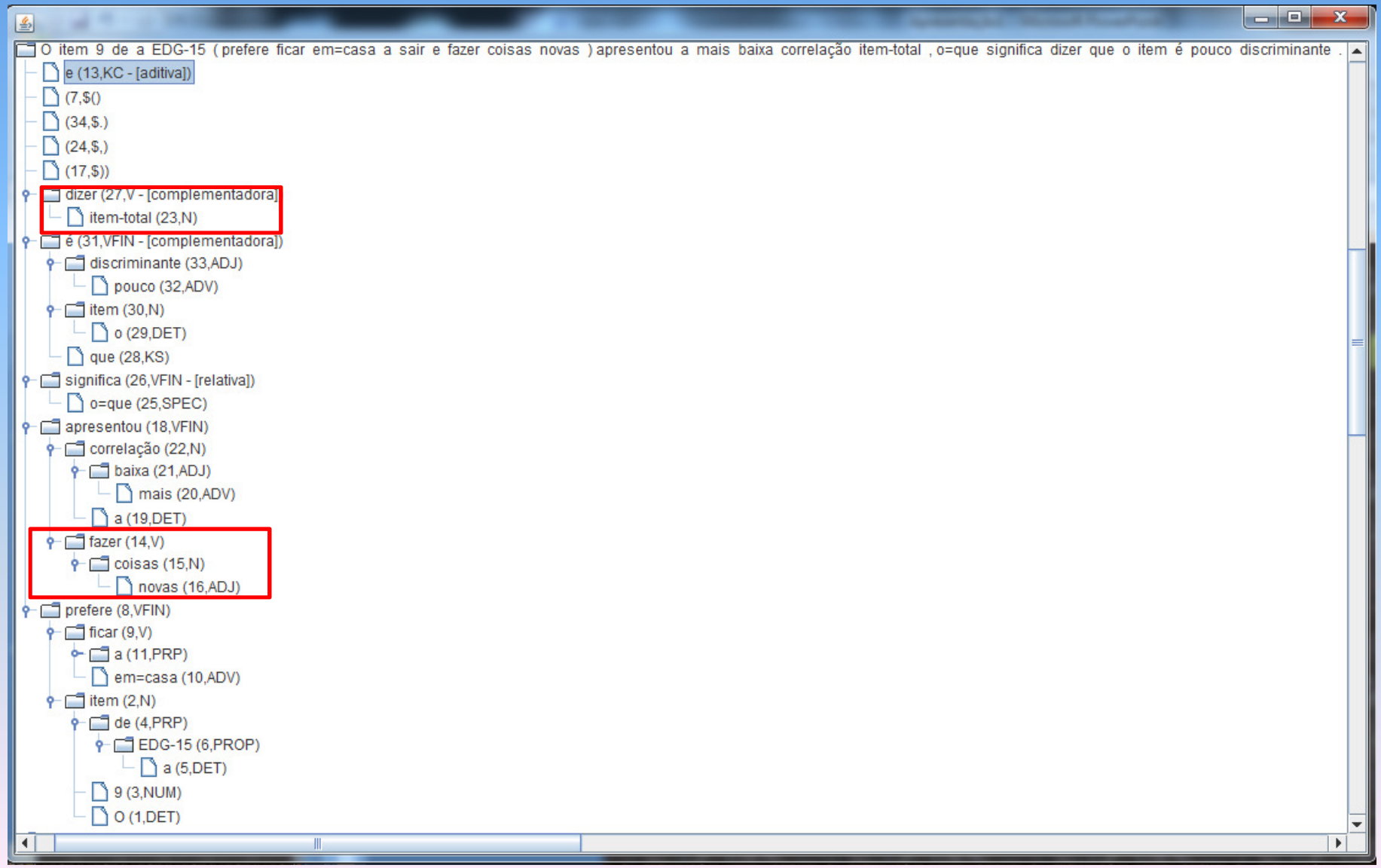


...dizer que o item é pouco ... – complementadora

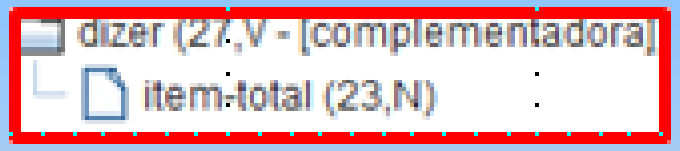


..., o que significa dizer... – relativa

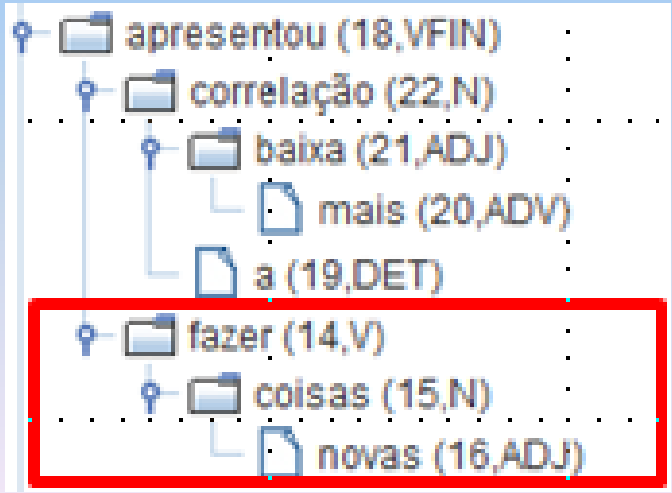
Ferramenta – Erros PALAVRAS



O item 9 de a EDG-15 (prefere ficar em=casa a sair e fazer coisas novas) apresentou a mais baixa correlação item-total , o=que significa dizer que o item é pouco discriminante .

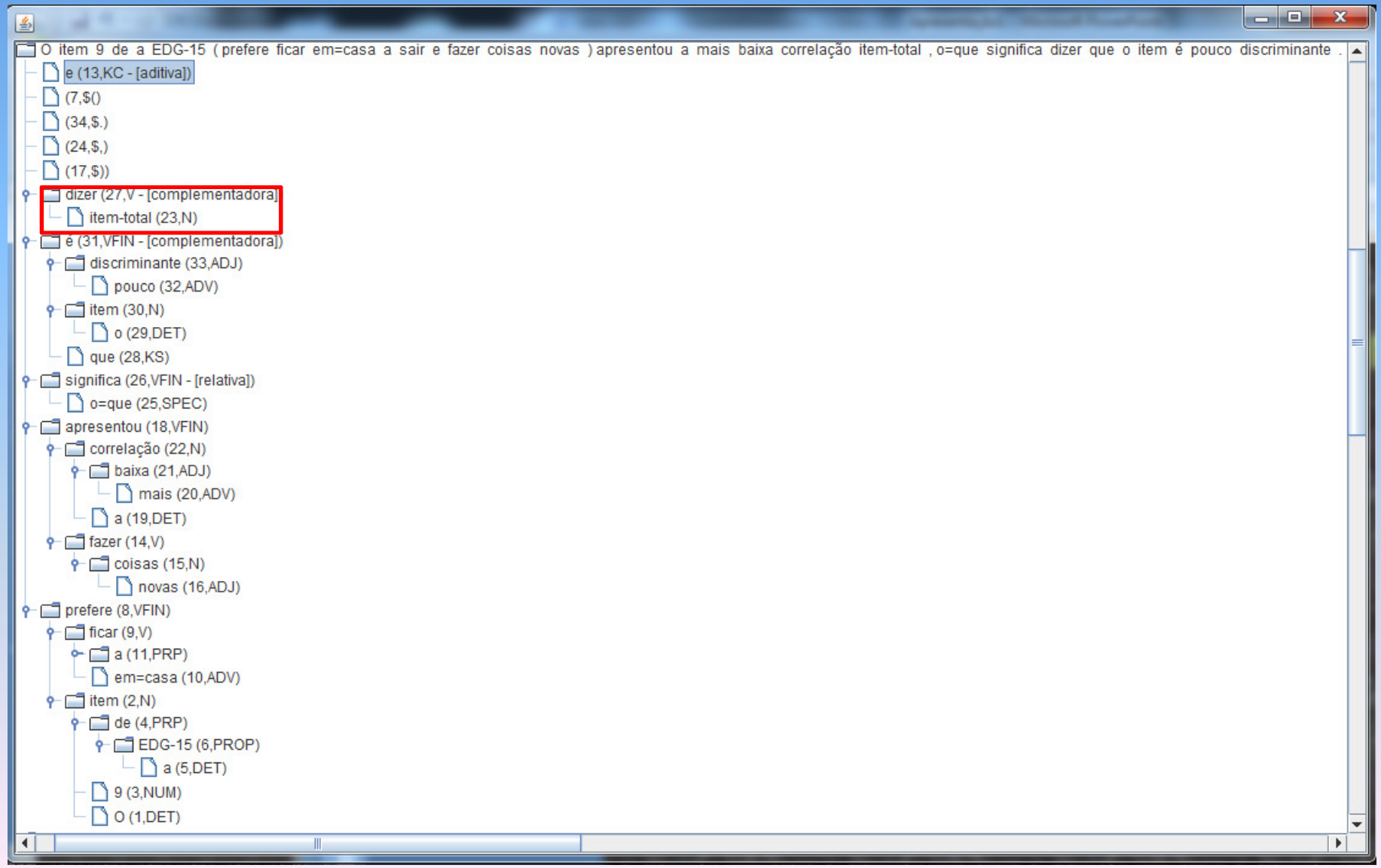


Má organização das dependências

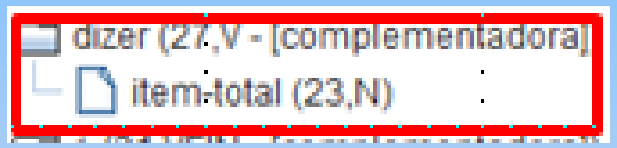


Má organização das dependências

Ferramenta – Erros do nosso parser

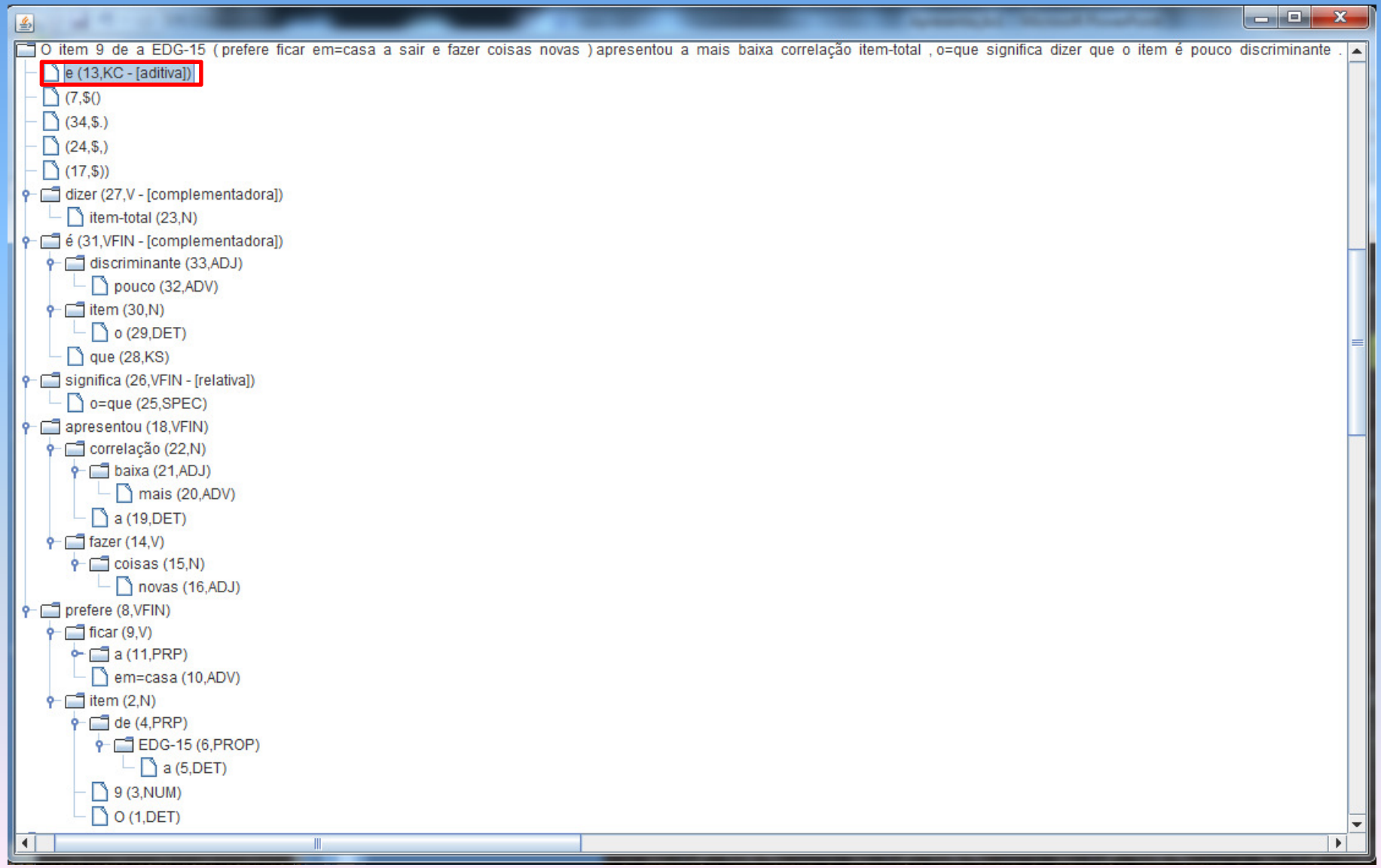


O item 9 de a EDG-15 (prefere ficar em=casa a sair e fazer coisas novas) apresentou a mais baixa correlação item-total , o=que significa dizer que o item é pouco discriminante .



..., o que significa dizer que... – complementadora

Ferramenta - Problemas



O item 9 de a EDG-15 (prefere ficar em=casa a sair e fazer coisas novas) apresentou a mais baixa correlação item-total , o=que significa dizer que o item é pouco discriminante .

e – aditiva – separado do resto da sentença

Algumas conjunções coordenadas (e, mas, ou) são separadas da sentença e colocadas junto à raiz.

Ainda resta tratar os casos em que isso acontece.

- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

Ferramenta - Resultados

- 86,22% de acerto em relação ao corpus anotado manualmente
- 93,63% de acerto se não forem levadas em consideração as coordenadas que o PALAVRAS separa

Ferramenta - Resultados

- Dos 13,78% de erro
  - 56,45% foi da separação das coordenadas pelo PALAVRAS
  - 17,02% casos em que mais de um conector igual ocorria na mesma sentença e o parser só reconhecia o primeiro (já corrigido)
  - 12,78% erros diversos
  - 12,77% erro nas POS-tags do PALAVRAS
  - 6,38% erro nas dependências do PALAVRAS

- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

Aplicação - Objetivos

- Mostrar um dos possíveis empregos da ferramenta desenvolvida
- Observar as semelhanças no uso de conectores em textos de áreas diferentes
- Observar essas semelhanças em partes diferentes dos textos
- Não ter caráter conclusivo

Aplicação - Método

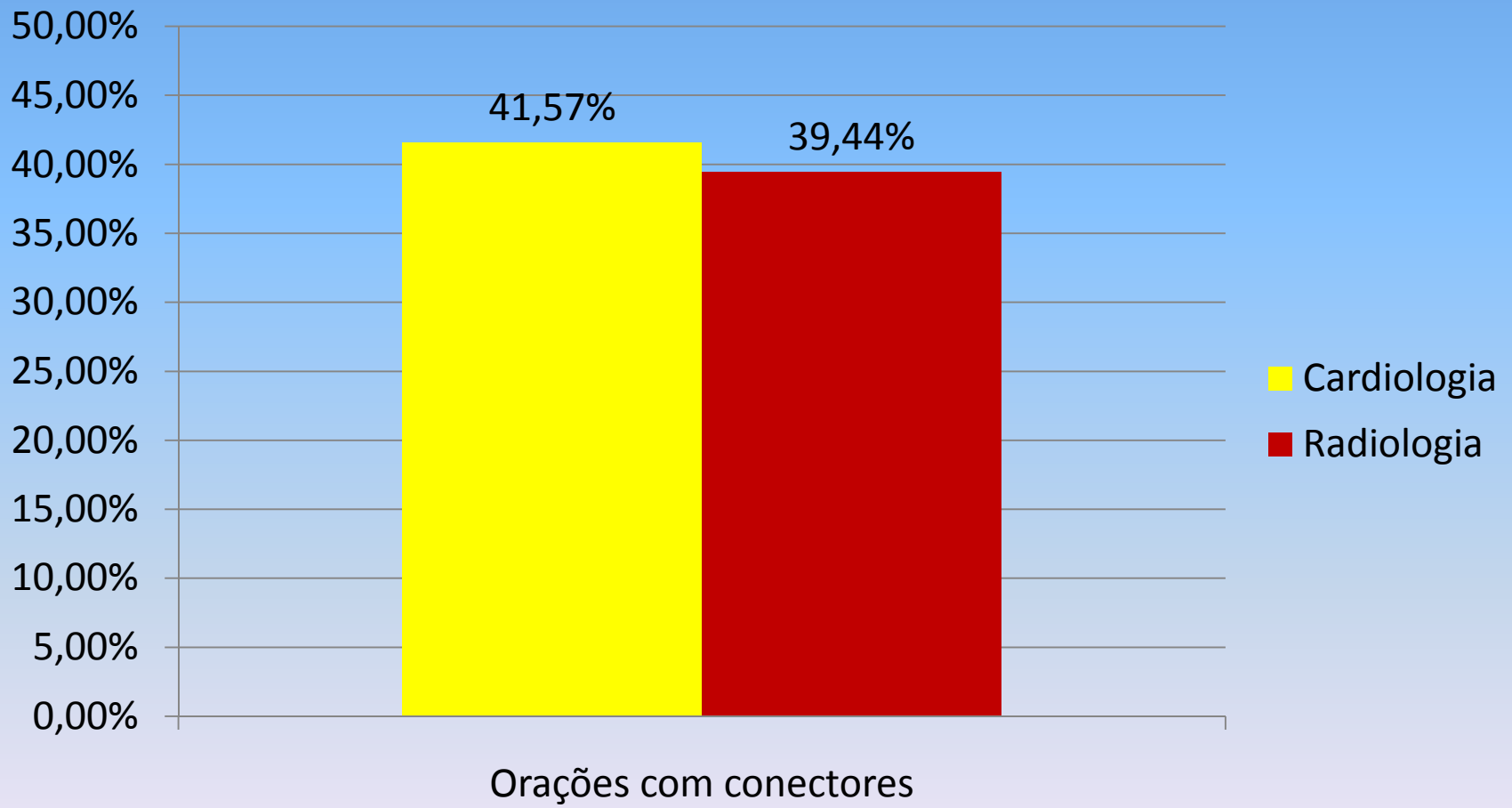
- Corpus
  - 10 artigos dos Arquivos Brasileiros de Cardiologia
    - 10 introduções (3.383 tokens)
    - 10 discussões (11.341 tokens)
  - 10 artigos da Radiologia Brasileira
    - 10 introduções (4.760 tokens)
    - 10 discussões (8.129 tokens)
  - Total = 27.613 tokens

Aplicação - Método

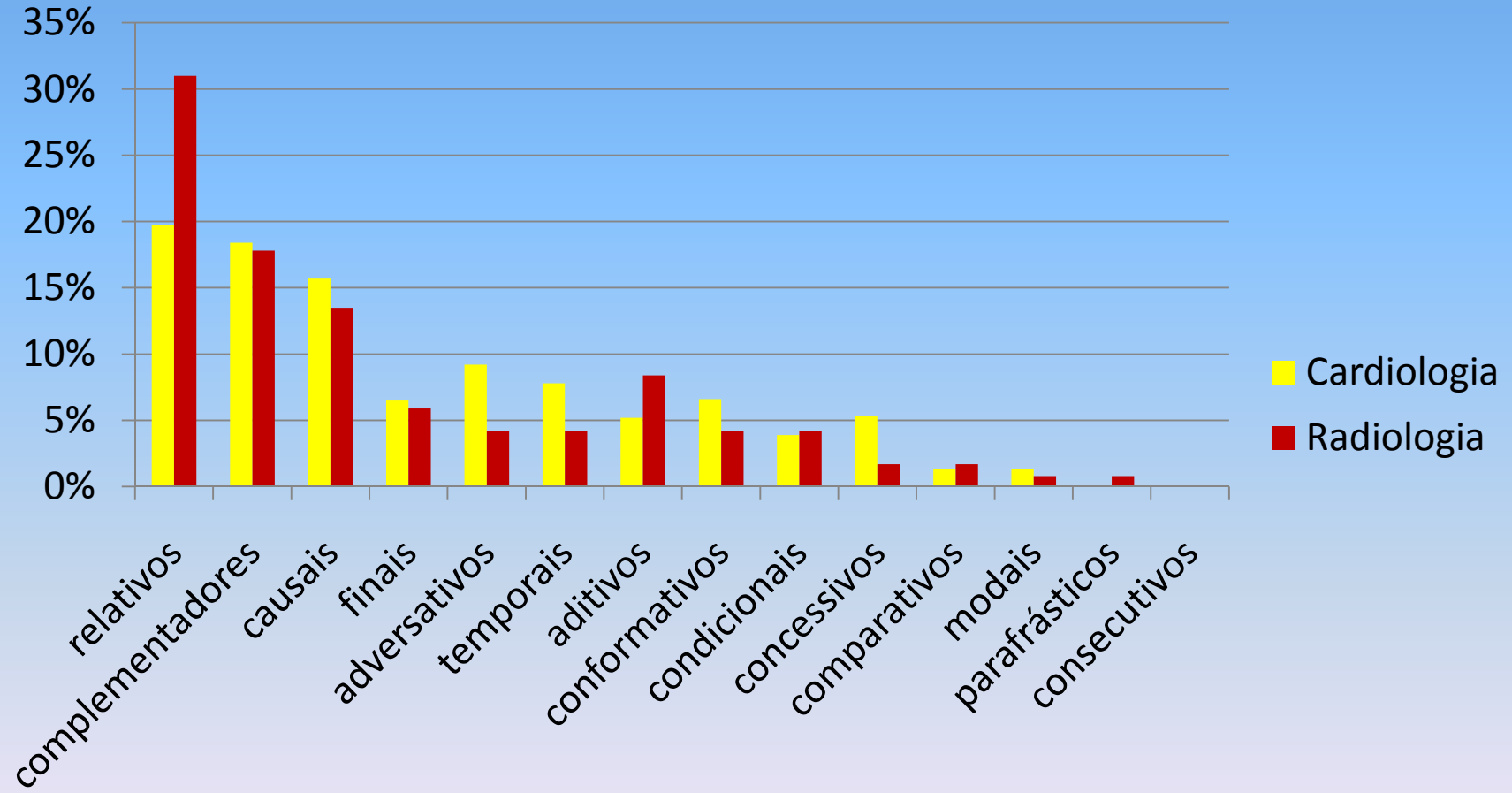
- Utilização do nosso parser em cada uma das seções dos artigos
- Listagem dos tipos e das quantidades de conectores empregados
- Contraste entre as seções e as áreas
- Statistica7 – usado para realizar o teste t

- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

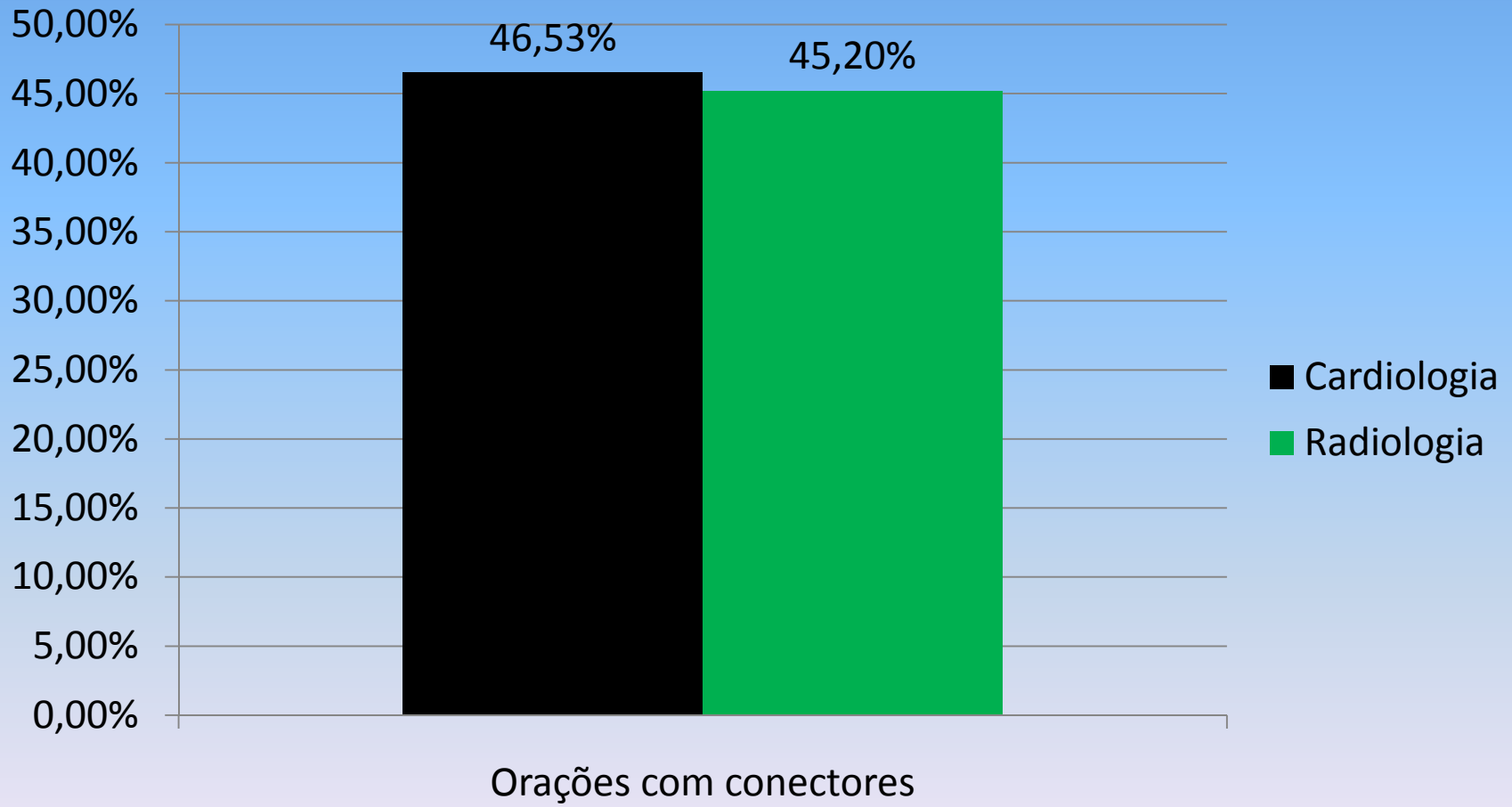
Introduções



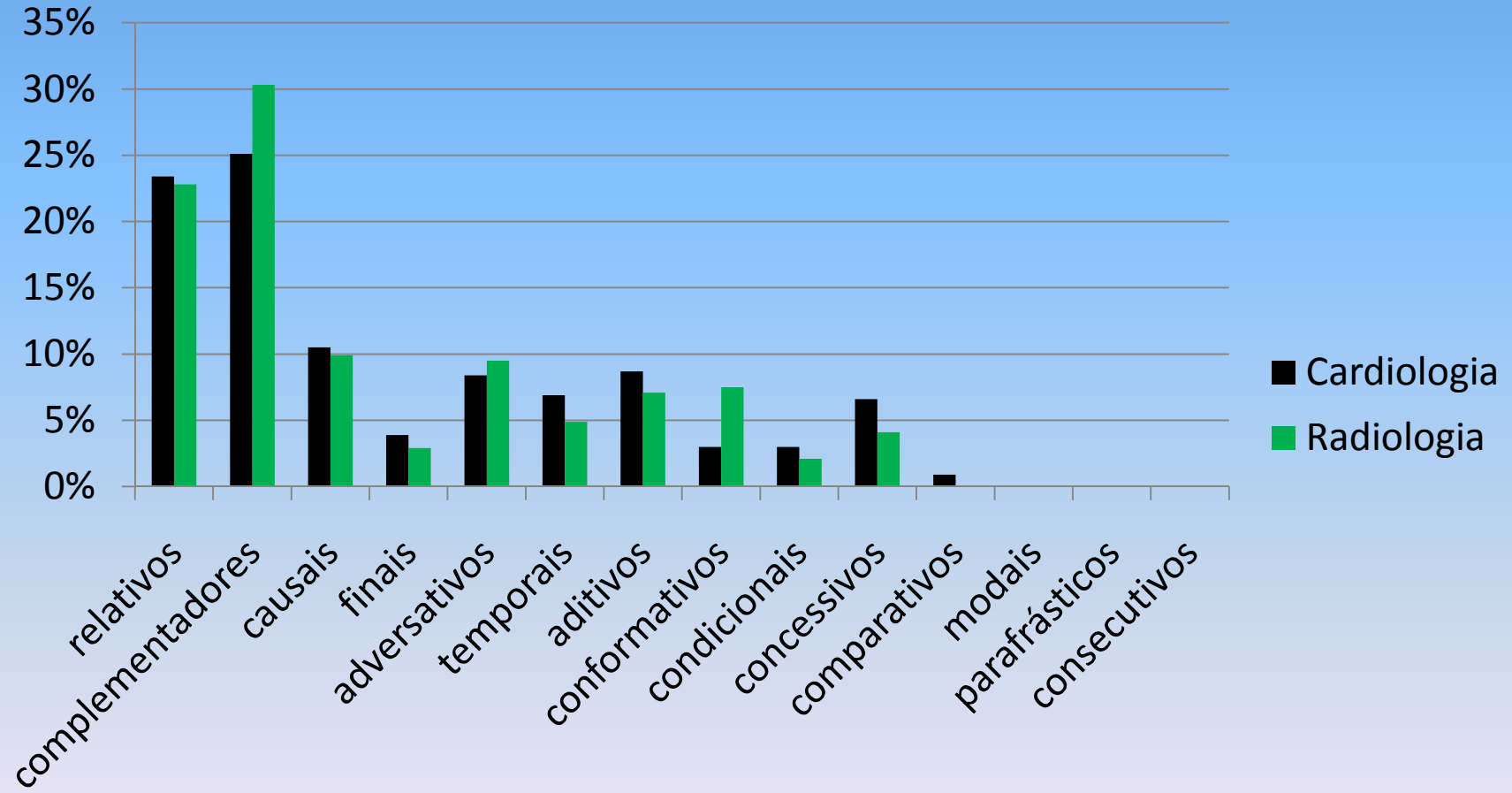
Distribuição de conectores
Introduções



Discussões


50,00%
45,00%
40,00%
35,00%
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
46,53%
45,20%
Cardiologia
Radiologia
Orações com conectores

Distribuição de conectores Discussões



- Introdução
- Ferramenta
- Ferramenta - Resultados
- Exemplo de aplicação
- Aplicação – Resultados
- Comentários

Comentários - Ferramenta

- Positivos
  - Se mostrou bastante eficiente
  - É flexível e simples de ajustar
  - Atingiu os objetivos
- Nem tão positivos
  - Precisa do PALAVRAS
  - Ainda não trata conjunções fundamentais como e, mas e ou, pois elas estão separadas das orações

Comentários - Aplicação

- Positivos
  - Serviu para testar a ferramenta em um corpus um pouco maior
  - Permitiu observar os estilos de conexão de textos de duas áreas diferentes
  - Indicou que a Cardiologia e a Radiologia são semelhantes em questões de conectores
- Nem tão positivos
  - O corpus estudado foi muito pequeno para se fazerem asserções mais conclusivas

Referências Citadas

- BICK, Eckhardt. (2000) *The Parsing System* PALAVRAS*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press. Disponível em: <http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf>

- DiZer (Versão 2.0) - http://www.nilc.icmc.usp.br/dizer2/

- NEVES, Maria Helena de Moura. (2000) *Gramática de usos do português*. São Paulo: Editora UNESP.

Muito obrigados!

Leonardo Zilio

leonardozilio@yahoo.de