ELC - 2011 X Encontro de Linguística de Corpus

EBRALC - 2011 - V Escola Brasileira de Linguística Computacional

Aquisição de linguagem escrita e conhecimento fonológico: o corpus do projeto e-Labore

Autores:
Gustavo Mendonça
Thaïs Cristófaro Silva
Leonardo Almeida
Eduardo Gonçalves

Novembro 2011

Organização Geral

- Sobre o projeto e-Labore
  - coleta
  - cadastro
  - disponibilização
- Sobre a classificação dos desvios ortográficos
- Sobre a estrutura do banco de dados
- Sobre a verificação de consistência dos dados
- Sobre o algoritmo de determinação dos desvios
- Sobre as pesquisas possíveis
- Exemplos de pesquisas
- Considerações finais

Objetivos

1. Apresentar a organização dos dados do Projeto e-Labore,

1. Realizar um levantamento inicial do corpus avaliando, sobretudo, desvios ortográficos,

1. Elaborar uma interface gráfica para a web de modo a facilitar o acesso da comunidade científica ao banco de dados.

Sobre o projeto e-Labore

O **e-Labore** (Laboratório Eletrônico de Oralidade e Escrita) consiste em um projeto coordenado por Thaïs Cristófaro-Silva, Daniela Guimarães, Leonardo Almeida e Raquel Fontes-Martins que tem por propósito **coletar**, **cadastrar** e **disponibilizar** para a comunidade científica um banco de dados de material escrito por crianças de 6 a 12 anos.

Sobre o projeto e-Labore

“O corpus do projeto e-Labore permitirá o mapeamento do vocabulário infantil do português brasileiro contemporâneo. A partir dos dados do projeto e-Labore será possível formular um mapeamento do vocabulário infantil que pode oferecer contribuições para a investigação de teorias de aquisição da linguagem em geral. Particularmente, o projeto procura contribuir com os debates a respeito da interação entre a linguagem adulta e infantil em um contexto de mudança lingüística e evolução da linguagem.” [4]

Sobre a coleta das redações

A coleta das redações do projeto e-Labore se deu de acordo com os seguintes critérios:

- Somente escolas de Belo Horizonte foram selecionadas;

- As escolas foram dividas uniformemente entre as 9 regionais de Belo Horizonte (Barreiro, Centro-sul, Leste, Nordeste, Noroeste, Norte, Oeste, Pampulha, Venda-Nova);

- Em cada regional, 4 escolas foram escolhidas para participar do projeto, sendo 2 públicas e 2 particulares.

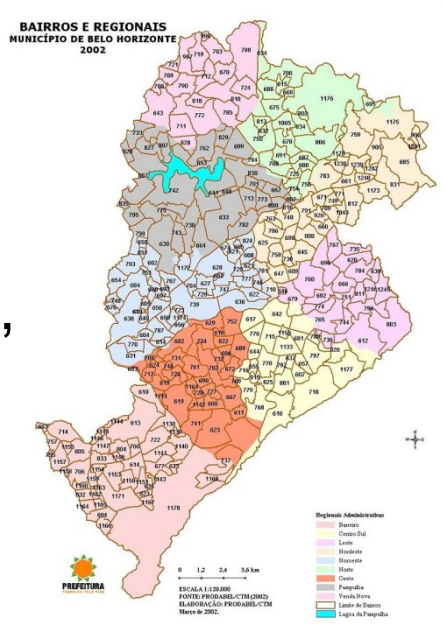
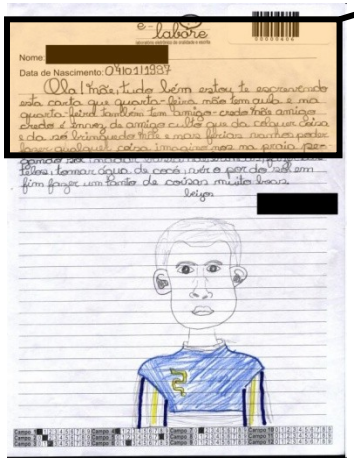


Figura 1: as 9 regionais de Belo Horizonte

Sobre o cadastro das redações

"Cada uma das crianças que participa do projeto recebe uma folha pautada que é utilizada na produção do texto. Nessa folha as crianças podem escrever e desenhar livremente."

Todas as redações produzidas pelos alunos que participaram do projeto e-Labore foram **digitadas** e **digitalizadas**.



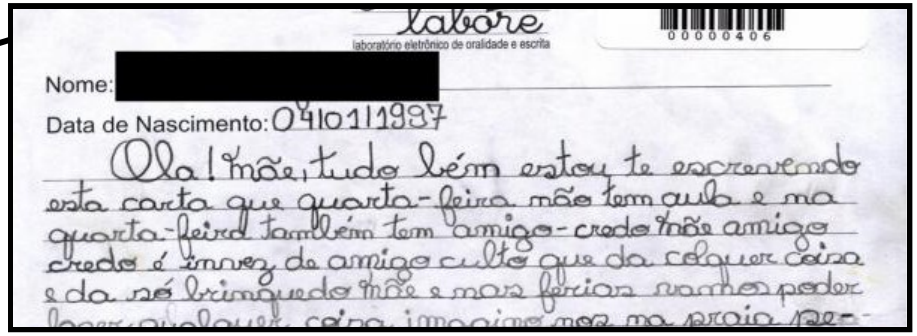


Figura 2: exemplo de redação escaneada

${Ola}[Olá]! Mãe, tudo {bém}[bem] estou te escrevendo esta carta que quarta-feira não tem aula e na quarta-feira também tem amigo-credo {Mãe}[mãe] amigo credo é {invez}[em vez] de {amigo culto}[amigo-oculto] que {da}[dá] {colquer}[qualquer] coisa e {da}[dá] só brinquedo {Mãe}[mãe] e nas {ferias}[férias] vamos poder

Sobre o cadastro das redações

A digitação foi feita por um dos colaboradores do projeto e-Labore, seguindo-se **7 regras**:

1. Organização do texto: quebra de linha <ENTER>
2. Organização do texto: paragrafação <ENTER> <ENTER>
3. Marcação de Erros {erro} [versão corrigida]
4. Dificuldade de Leitura *
5. Ausência de palavra +[palavra]
6. Início e fim de texto contínuo $...$
7. Hifenização _

Colaboradores:

Alba da Silva, Alessandra Deusdete, Amana Greco, Ariana Siqueira, Carla Vieira, Carolina Diniz, Cassandra Lima, Denise Veridiano, Frederico Fraga, Gisele Oliveira, Ignês Lara, Janaína Rabelo, Janayna Carvalho, Juliana Silva, Lucas Paiva, Luciana Cangussu, Marcelo Negri, Mariana Moreira, Rogério Brito | Ana Luisa Terto, Angélica Campos, Erick Leite, Estefânia Souza, Flávia Silveira, Flávia Carvalho, Jaqueline Castro, Joana Arzberger, Kelly Naves, Natália Oliveira, Michel Pires, Thiago Fraga.

Sobre o cadastro das redações

A digitalização foi feita a fim de manter-se toda a produção (textos, desenhos, palavras isoladas, acrósticos, etc) realizada pelos alunos participantes do projeto e-Labore. A frente e o verso de cada redação foram escaneados em alta resolução (3507 x 2480 pixels, 24 bits por pixel):

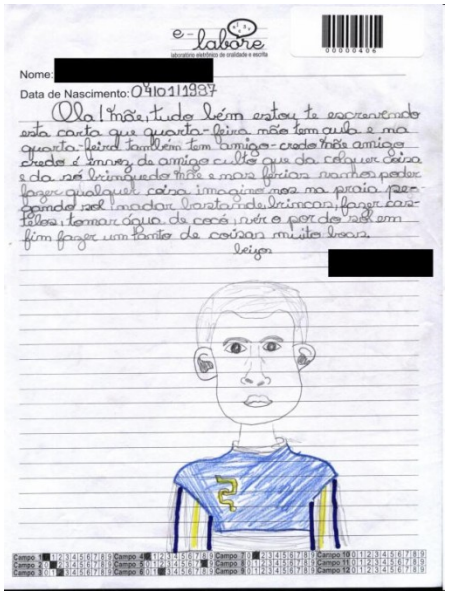


Figura 3: exemplo de redação escaneada

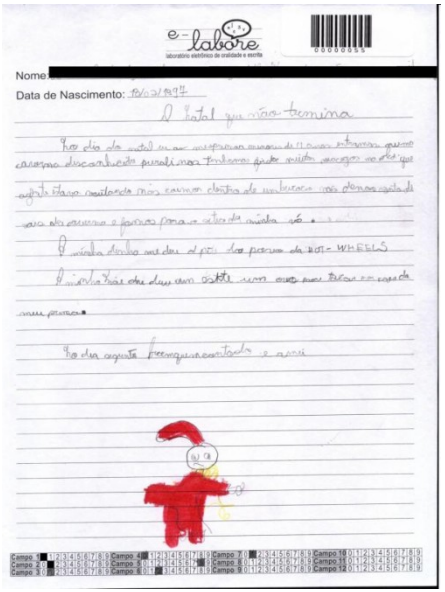


Figura 4: exemplo de redação escaneada

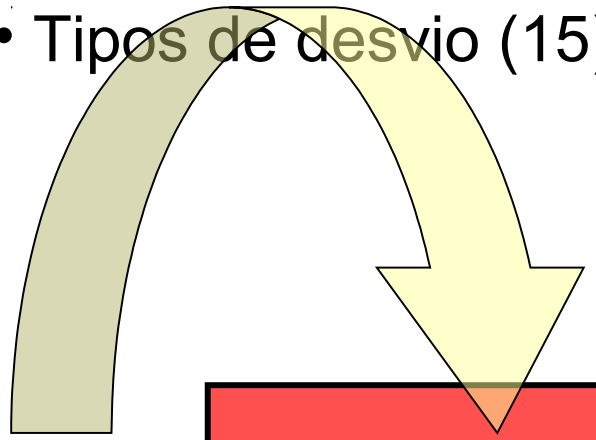
Sobre a organização dos dados

Como resultado da metodologia adotada no projeto e-Labore, podemos ter acesso às seguintes informações acerca das redações:

- Número da redação;
- Texto digitado;
- ~~Imagem digital da redação;~~
- Nome do aluno;
- Série;
- Sexo;
- Idade;
- Nome da escola;
- Tipo de escola: particular ou pública;
- Número e data da coleta.

Palavra por palavra:

- Forma desviante;
- Forma padrão;
- Tipos de desvio (15);



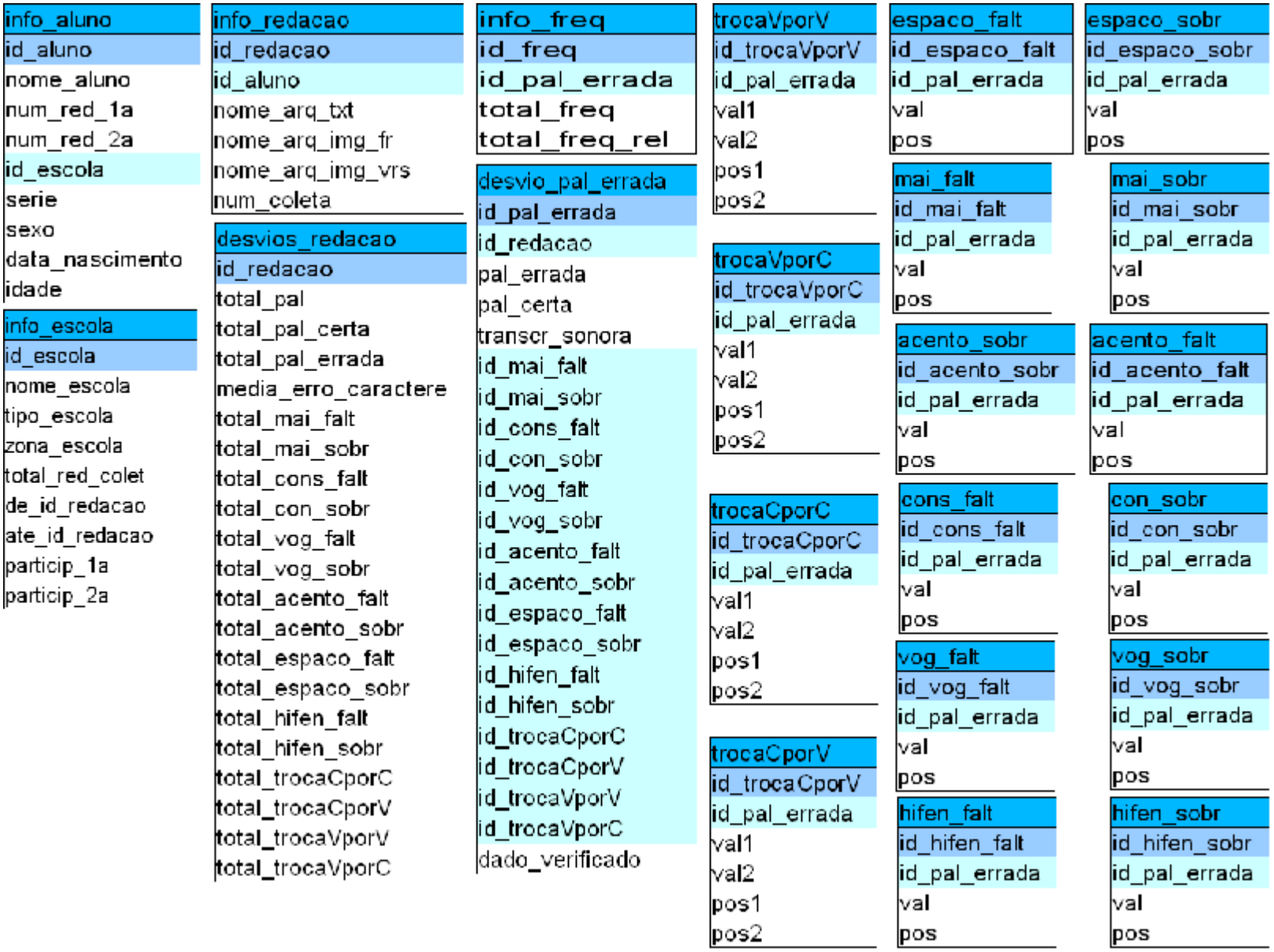
Banco de dados em SQL

Sobre a classificação dos desvios ortográficos

A classificação dos desvios foi feita tendo-se por base análises como as propostas por Scliar-Cabral (2003), Faraco (1997), Cagliari (1989) e Mollica (2003). Procurou-se atingir uma classificação geral dos desvios, de modo que fossem indicados os seguintes aspectos:

- troca, inserção ou apagamento de símbolos gráficos;
- troca, inserção ou apagamento de acento gráfico;
- troca entre letras maiúsculas e minúsculas;
- junção ou separação de palavras.

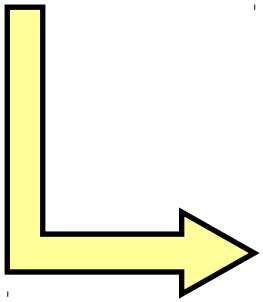
Sobre a estrutura do banco de dados



Sobre a verificação de consistência dos dados

A verificação dos dados das colunas formaDesviante e formaPadrao consistiu em:

- Apagamento de caracteres especiais.
- Apagamento de espaços em branco nos cantos das colunas.
- Verificação de diferença entre as colunas.



dado_verificado = 9;

1,1% do banco de dados
(998/85659 palavras)

Sobre o algoritmo de determinação dos desvios

De modo a facilitar a determinação dos desvios ortográficos, um algoritmo computacional implementado em PHP foi utilizado. Tal algoritmo visou automatizar o processo de determinação de erros, acelerando a indicação de desvios, bem como buscando minimizar o erro no processo.

O algoritmo, basicamente, compara, um a um, os caracteres presentes nas colunas formaDesviante e formaPadrao, marcando se há diferenças e qual a natureza dessas diferenças: se são inserções, trocas ou apagamentos.

Sobre o algoritmo de determinação dos desvios

id: 44191 - as palavras sao: formaPadrao(assassinaram!) e formaDesviante(Assasimara!)
0 - TROCA: carFP(a) foi trocado por carFD(A) (i: 0, j: 0, k: 1, l: 1)
a: o caractere e do tipo v: vogal minus
A: o caractere e do tipo V: vogal maius
1 - OK: carFP (s) e igual a carFD (s) (i: 1, j: 1, k: 2, l: 2)
2 - OK: carFP (s) e igual a carFD (s) (i: 2, j: 2, k: 3, l: 3)
3 - OK: carFP (a) e igual a carFD (a) (i: 3, j: 3, k: 4, l: 4)
4 - OK: carFP (s) e igual a carFD (s) (i: 4, j: 4, k: 5, l: 5)
5 - APAGAMENTO: carFP(s) foi apagado (i: 5, j: 5, k: 6, l: 6)
s: o caractere e do tipo c: consoante minus
6 - OK: carFP (i) e igual a carFD (i) (i: 6, j: 5, k: 7, l: 6)
7 - TROCA: carFP(n) foi trocado por carFD(m) (i: 7, j: 6, k: 8, l: 7)
n: o caractere e do tipo c: consoante minus
m: o caractere e do tipo c: consoante minus
8 - OK: carFP (a) e igual a carFD (a) (i: 8, j: 7, k: 9, l: 8)
9 - OK: carFP (r) e igual a carFD (r) (i: 9, j: 8, k: 10, l: 9)
10 - OK: carFP (a) e igual a carFD (a) (i: 10, j: 9, k: 11, l: 10)
11 - APAGAMENTO: carFP(m) foi apagado (i: 11, j: 10, k: 12, l: 11)
m: o caractere e do tipo c: consoante minus
12 - OK: carFP (!) e igual a carFD (!) (i: 12, j: 10, k: 13, l: 11, i(geral): 44191)
12 - FIM: Nao ha mais caracteres
desvioReducao: , MaiFalt: 0, MaiSobr: 1, ConsFalt: 2, ConsSobr: 0, VogFalt: 0, VogSobr: 0,
TrocaCons: 0, TrocaVog: 0,
AcentFalt: 0, AcentSobr: 0, EspacoFalt: 0, EspacoSobr: 0, DivisaoSil: 0, HifenFalt: 0,
HifenSobr: 0, Estrang: 0, dadoVerificado: 0
TrocaCporC: 1, TrocaCporV: 0, TrocaVporV: 0, TrocaVporC: 0
id: 44191

Sobre o êxito do algoritmo de desvios

O algoritmo utilizado obteve uma taxa de sucesso de 90,7%, reconhecendo os desvios existentes em 76839 das 84661 palavras do banco de dados.

As 7822 palavras nas quais o algoritmo falhou possuíam 3 ou mais desvios ortográficos em sequência, a exemplo de *prissizou* (precisou), *pulblica* (pública) ou BAT-MAN (Batman).

Tais palavras tiveram seus erros ortográficos preenchidos manualmente.

Sobre as pesquisas possíveis

A estruturação do banco de dados em SQL permite, através do cruzamento de informações, responder a diversas questões de cunho linguístico ou para-linguístico.

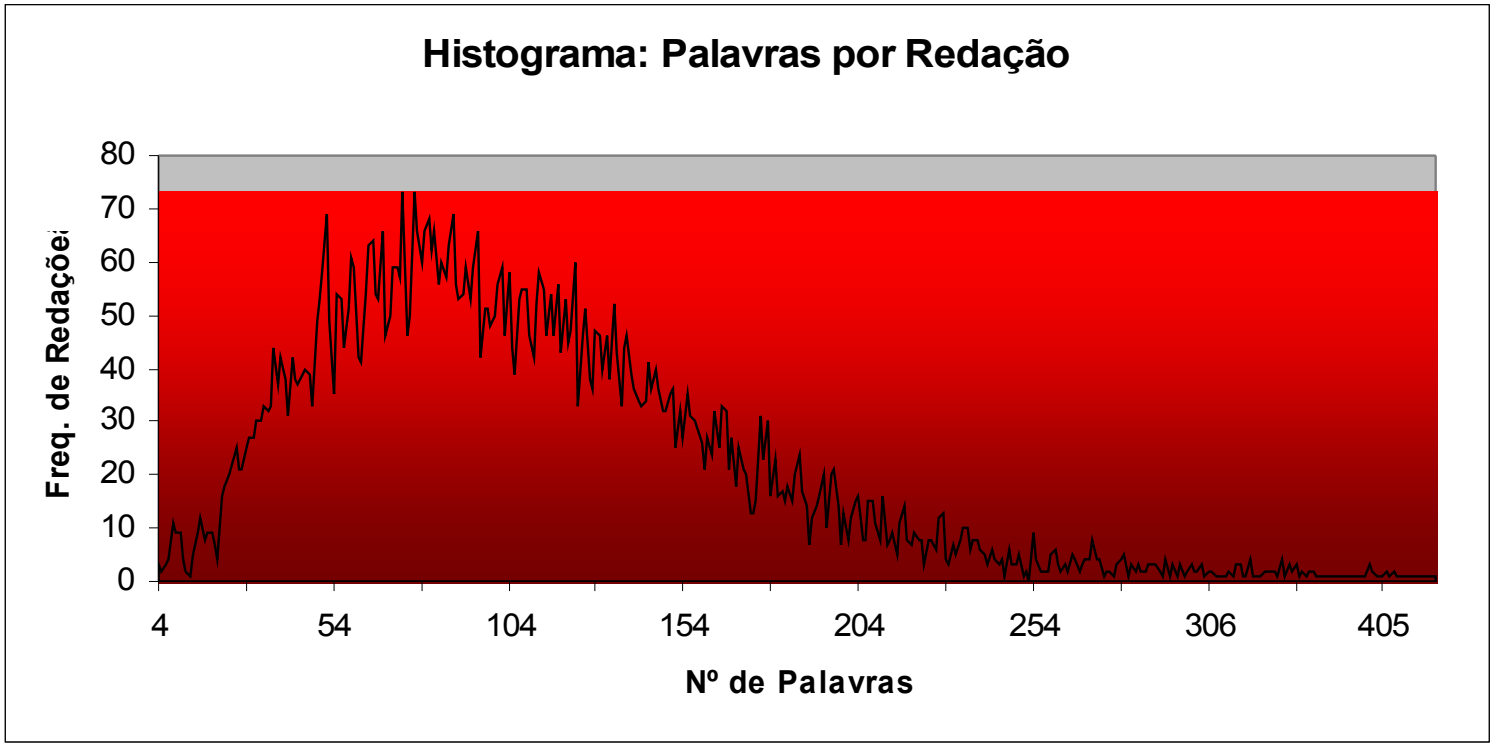
Exemplos de pesquisas possíveis

Observando-se a coluna ***formaDesviante***, por exemplo, pode-se observar **quais tipos de desvios ortográficos** as crianças cometem.

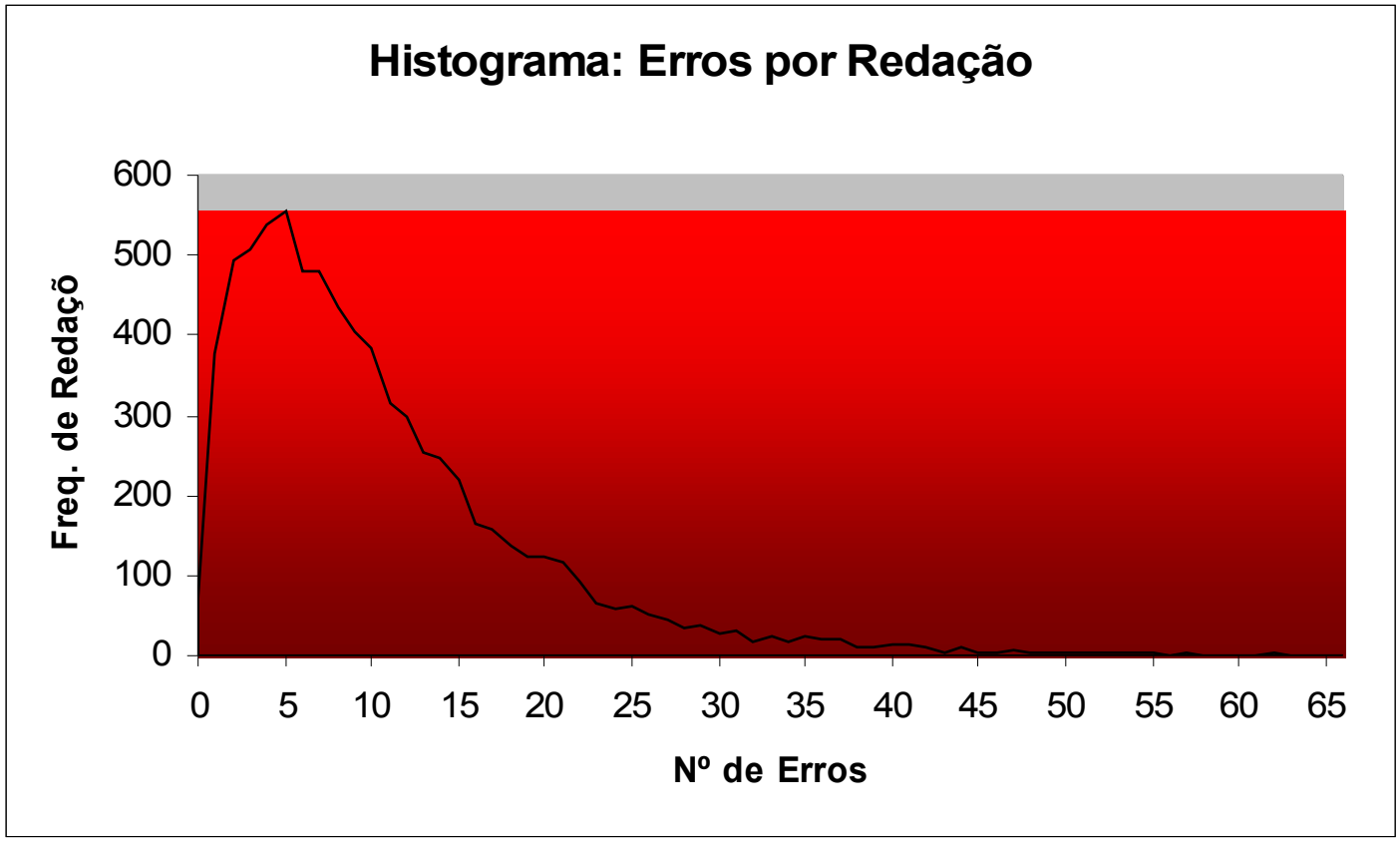
Cruzando-se os dados da coluna ***formaDesviante*** com os da ***formaSonora***, por exemplo, é possível verificar quais os desvios ortográficos **têm algum tipo de condicionamento fonológico** ou não.

Pode-se obter também respostas a perguntas de cunho para-linguístico: fazendo-se um cruzamento dos dados de **desvio** e a coluna ***tipoEscola*** é possível checar se **há diferenças entre os desvios encontrados entre escolar públicas e particulares**.

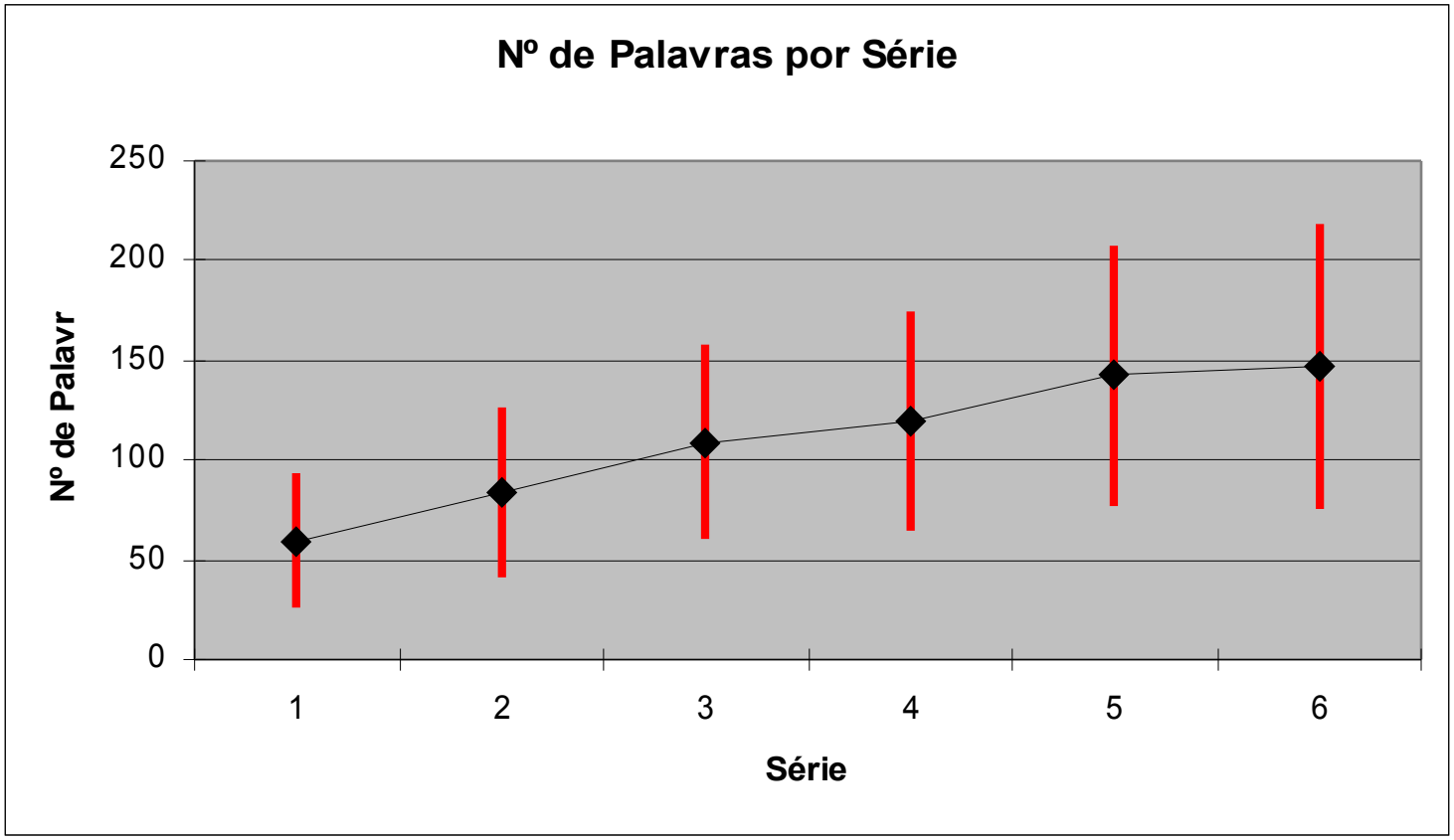
Exemplos de pesquisas possíveis



Exemplos de pesquisas possíveis



Exemplos de pesquisas possíveis



Exemplos de pesquisas possíveis

	formaPadrao	frequencia	serie	MaiFalt	MaiSobr	ConsFalt	ConsSobr	VogFalt	VogSobr	AcentoFalt	AcentoSobr	EspacoFalt	EspacoSobr	HifenFalt	HifenSobr	TrocaCporC	TrocaCporV	TrocaVporV	TrocaVporC	Tot. Desvios
1	é	431	1	79	3	0	0	0	0	349	1	0	0	0	0	0	0	32	0	464
2	Copa	131	1	129	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	133
3	também	130	1	0	4	33	0	0	0	104	1	0	1	0	0	30	1	0	0	174
4	natal	108	1	88	16	4	0	0	1	0	0	0	0	0	0	1	3	0	0	113
5	família	90	1	0	1	0	2	1	0	83	7	0	0	0	0	6	0	3	11	114
6	escola	85	1	2	72	2	0	2	1	0	0	0	0	0	0	0	1	8	0	88
7	aí	84	1	3	0	0	0	2	1	79	0	0	0	0	0	0	0	0	0	85
8	O	83	1	37	37	0	1	0	0	0	1	0	0	0	0	0	0	7	0	83
9	Brasil	80	1	56	0	4	0	1	0	0	0	0	0	0	0	22	5	1	0	89
10	Chapeuzinho	79	1	29	0	8	0	6	0	0	6	0	6	0	0	19	0	1	20	95
11	nós	78	1	1	2	0	0	0	0	64	9	0	0	0	0	0	0	3	0	79
12	Eu	76	1	29	39	0	1	1	2	0	1	0	1	0	0	0	0	2	2	78
13	lá	73	1	3	1	0	0	0	0	68	0	0	0	0	0	0	0	2	0	74
14	mas	72	1	5	6	3	0	0	51	0	2	0	0	0	0	10	1	0	0	78
15	vermelho	64	1	62	0	3	0	0	1	0	0	0	0	0	0	3	0	0	0	69
16	minha	64	1	23	19	13	0	1	0	0	2	0	1	0	0	8	0	2	0	69
17	Esta	59	1	0	0	2	0	1	0	52	5	0	0	0	0	2	0	0	0	62
18	a	59	1	17	21	0	0	0	0	6	16	0	0	0	0	0	0	0	0	60
19	mundo	57	1	41	1	10	2	0	0	0	0	0	0	0	0	3	0	2	0	59
20	começou	54	1	0	0	0	0	5	0	0	0	0	0	0	0	55	0	1	6	67
21	só	49	1	0	2	0	0	0	0	45	0	0	0	0	0	0	0	2	0	49
22	Papai	48	1	43	4	0	0	2	0	0	0	0	2	0	0	0	0	1	0	52
23	até	44	1	0	0	0	0	0	0	42	0	0	2	0	0	0	0	1	0	45
24	vez	41	1	0	2	0	1	0	0	0	0	0	0	0	0	39	0	1	0	43
25	vovó	41	1	0	0	0	0	0	0	37	18	0	0	0	0	0	0	0	0	55

Sobre a elaboração da interface web

Estudiosos da aquisição de linguagem escrita, de maneira geral, não detêm conhecimento sobre como manusear dados em um banco em SQL. Sendo assim, pretende-se elaborar uma interface gráfica na web, de modo a facilitar o acesso da comunidade científica ao corpus do Projeto e-Labore.

Considerações finais

O corpus do Projeto e-Labore mostra-se como uma ferramenta de relevância para os estudos que abordem a aquisição da linguagem escrita, bem como sua relação com a fonologia. A organização do corpus em um banco de dados SQL permite a realização de uma gama de opções de buscas, sendo possível e fácil o cruzamento das informações dentro do banco.

Bibliografia

[1] JOHNSON, K. “Speech perception without speaker normalisation.” In: JOHNSON, K; MULLENIX, J. W. (Ed.). *Talker variability without in speech perception*. San Diego: Academic Press, 1997. p. 145-165.

[2] PIERREHUMBERT, J. “Exemplar dynamics: word frequency, lenition and contrast.” In: BYBEE, J.; HOPPER, P. J. (Ed.). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 2001. p.137-158.

[3] BYBEE, J. *Phonology and Language Use*. Cambridge: CUP, 2001.

[4] E-LABORE. Laboratório Eletrônico de Oralidade e Escrita. Disponível em: <http://www.projetoaspa.org/elabore/index.php>. Acesso em: 26 de março de 2011.)

[5] E-LABORE. Laboratório Eletrônico de Oralidade e Escrita. Disponível em: <http://www.projetoaspa.org/elabore/metodologia/coleta.php>. Acesso em: 26 de março de 2011.)

Obrigado! =]