

**FFLCH/DLCV-USP- 2º SEMESTRE DE 2015
'FILOLOGIA E LÍNGUA PORTUGUESA'- FLC0284
TURMAS: 2015201/ 2015214/ 2015212
PROFA. DRA. MÁRCIA S. D. DE OLIVEIRA**

**AULA 13 – AULA-CONFERÊNCIA
TEMA: LINGUÍSTICA DE CORPUS; LINGUÍSTICA COMPUTACIONAL
– O EXEMPLO DO 'E-DICTOR'**

por

**LUCAS RAFALDINI
lucas.rafaldini.silva@usp.br**

**MÁRCIA OLIVEIRA
marcia.oliveira@usp.br**

CORPUS/CORPORA

“A corpus is a remarkable thing,
not so much because it is a collection of language text,
but because of the properties that it acquires if it is well designed
and carefully-constructed.”

Sinclair (2005)

Concepção de corpus para a Linguística e para a *Linguística de Corpus* – Aluísio & Almeida (2006: 156 -158)

- A utilização de corpus sempre foi um recurso empregado em pesquisas linguísticas.
- O que muda, entre a *Linguística* e a atual *Linguística de Corpus*, é a concepção de corpus.
- Para Berber Sardinha (2004), a *Linguística de Corpus* é tida como uma:

“... abordagem que se ocupa da coleta e da exploração de corpora ou um conjunto de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador.” (o grifo é nosso)

- A *Linguística de Corpus* descarta livros e outros textos não computadorizados como *corpus*.
- A *Linguística de Corpus* descarta a *web* como *corpus*, ainda que os textos estejam disponíveis e em formato eletrônico. Isto pelo fato de suas dimensões serem desconhecidas, de estar em contínua mudança e ainda pelo fato do *corpus* não ter sido projetado a partir de uma perspectiva linguística.

- Há várias tipologias de *corpus* que indicam parâmetros importantes a serem considerados.
- Uma tipologia atual é a que se vê em Berber Sardinha (2004), que inclui 7 critérios importantes acerca do 'corpus'.
- O critério 'modalidade' pode ser considerado o mais importante: o texto falado; o escrito; ou ambos

e-filologia

- Para alguns pesquisadores, a união entre os estudos históricos da língua, o trabalho filológico e o trabalho computacional do texto traz desafios inéditos. Nasce a ‘e-filologia’.
- Para outros pesquisadores, no entanto, as rupturas são de tal forma profundas que chegam a determinar o nascimento de um campo inteiramente novo de investigação: as *Humanidades Digitais*.
- ver: Paixão de Souza (2013)

- As pesquisas baseadas em corpus têm tido, na última década, um amplo desenvolvimento no contexto brasileiro.
- A sua relevância e pertinência é atestada nos domínios da Linguística (como a ‘Linguística Aplicada’ e a ‘Linguística Computacional’) - ver Aluísio & Almeida (2006: 156).
- A relevância das pesquisas em corpus tem sido atestada ainda no desenvolvimento do que vem sendo chamado de e-filologia, como se atesta em projetos como o *Corpus Thyco Brahe* que busca diferentes combinações de procedimentos filológicos, linguísticos e computacionais envolvidos em sua construção.
- *O Corpus Anotado do Português Histórico Tycho Brahe*
<http://www.tycho.iel.unicamp.br/~tycho/corpus/>
- Para detalhes sobre o *Tycho Brahe* – ver Paixão de Souza (2014a)

- Os corpora eletrônicos anotados:
- Não se trata de uma mera coleção ‘de’ textos eletrônicos, mas sim de um banco de dados ‘sobre’ textos, englobando diferentes camadas de representação sobre sua linguagem e sobre sua materialidade.
- Sobre corpora anotados (primeiras codificações sobre o texto que serão “conhecidas e anotadas”) ver Paixão de Souza (2014: 82-86)
- e-Dictor – uma ferramenta para edição filológica eletrônica e análise linguística automática: uma ferramenta para as humanidades digitais – ver Paixão de Souza (2014b)

Editor



HD.br

e-dictor

O que é o e-dictor?

Como o e-dictor funciona?

Por que isso é importante?

e-dictor

O que é o e-dictor? ←

Como o e-dictor funciona?

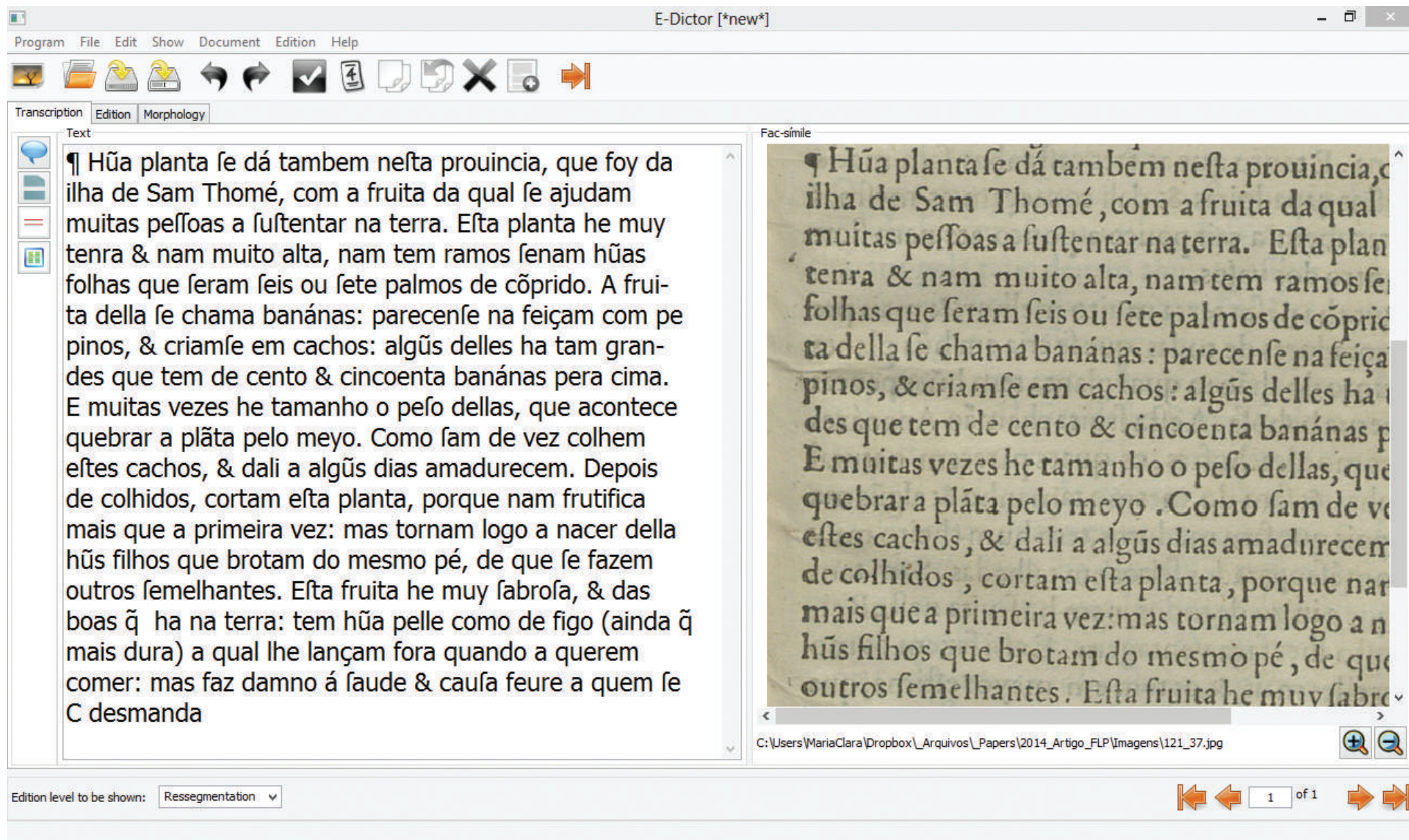
Por que isso é importante?

e dictator

"Uma ferramenta
para edição filológica eletrônica
e análise linguística automática"

e dictator

"Uma ferramenta
para **edição filológica eletrônica**
e análise linguística automática"



e-dictor

Com o Projeto Edições Filológicas na Brasileira Digital (2009-2013) criou-se uma nova possibilidade de digitalização com o e-dictor, a partir de um banco de dados de erros de reconhecimento automático (OCR) gerado no desenvolvimento de edições corrigidas e modernizadas para algumas obras do acervo.

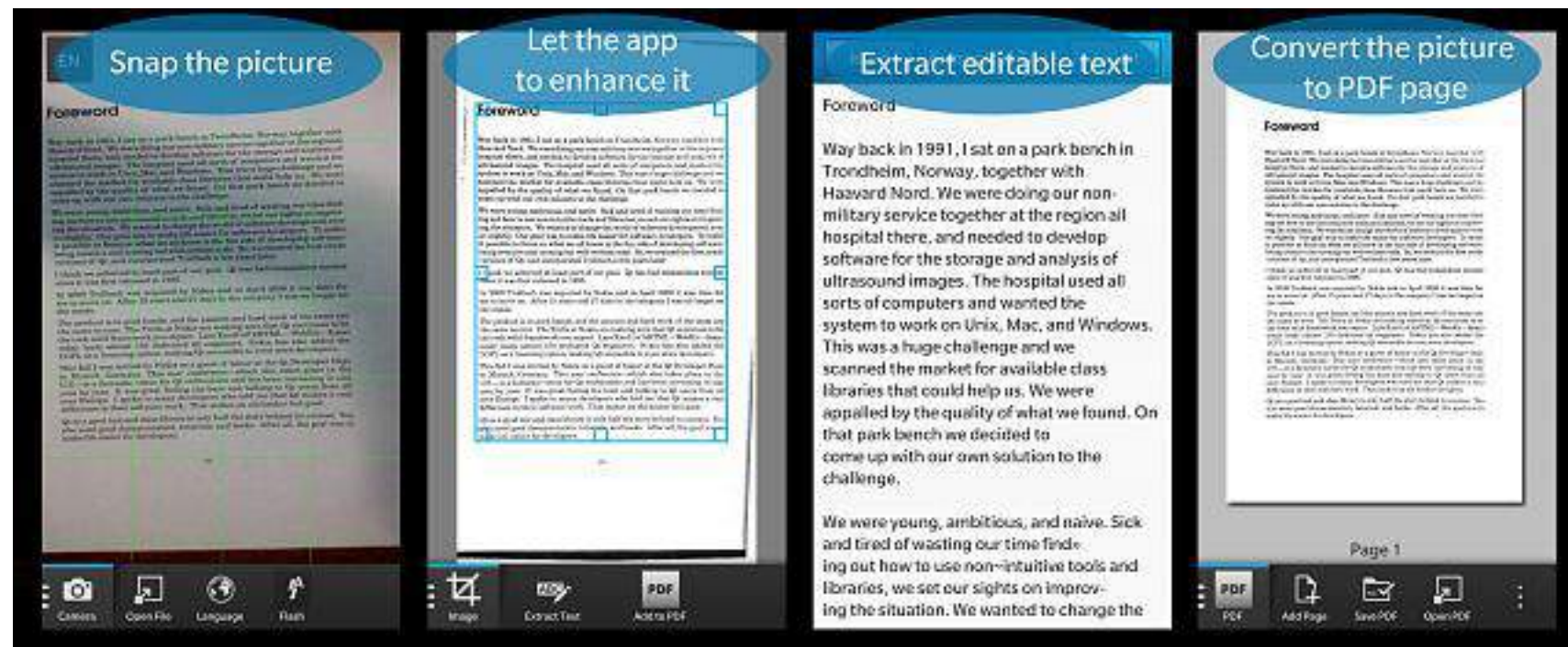
e-dictor

Com o Projeto Edições Filológicas na Brasileira Digital (2009-2013) criou-se uma nova possibilidade de digitalização com o e-dictor, a partir de um banco de dados de erros de reconhecimento automático



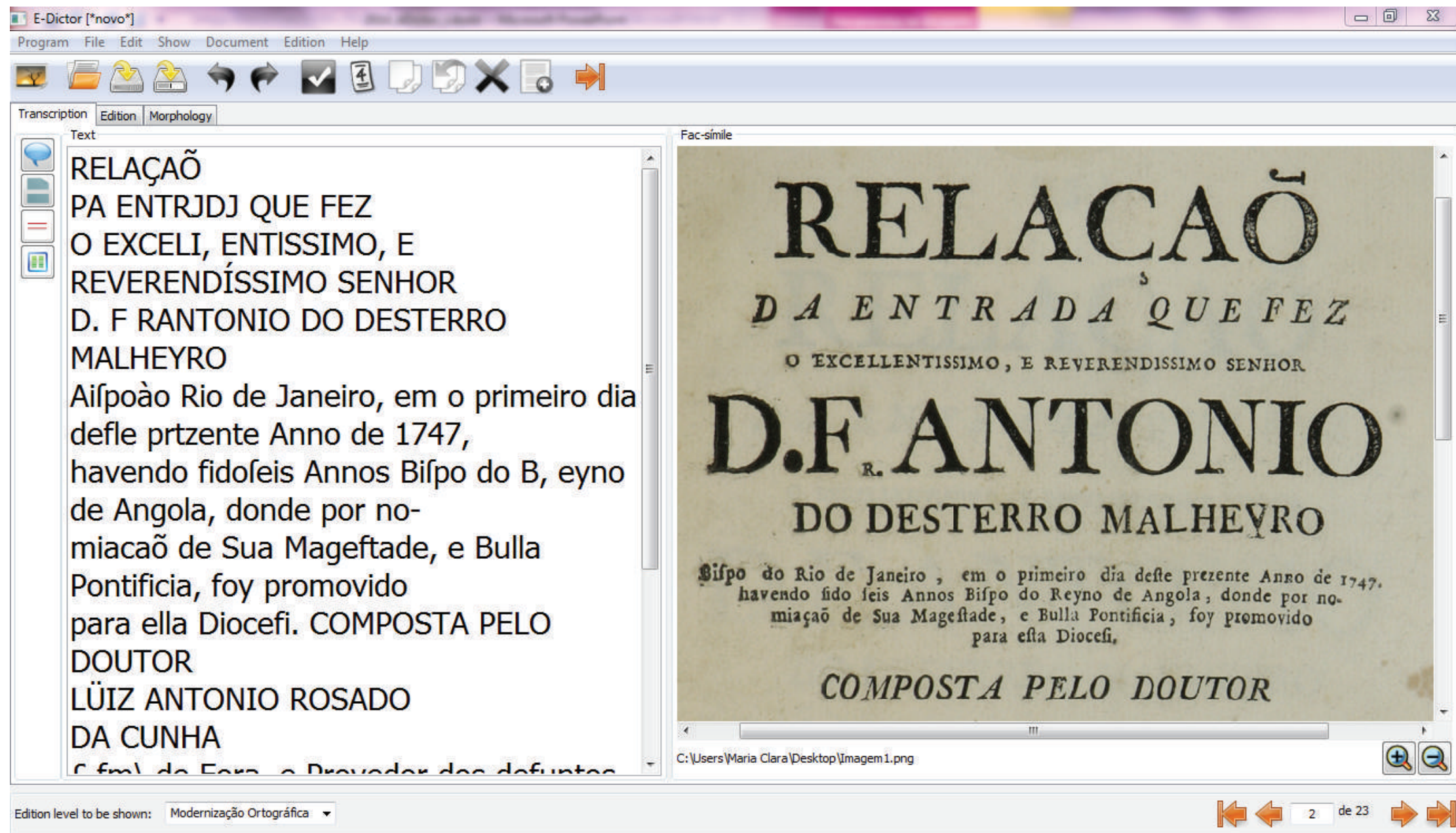
(OCR) gerado no desenvolvimento de edições corrigidas e modernizadas para algumas obras do acervo.

Optical Character Recognition



Optical Character Recognition





ANTES

RELAÇÃO
DA ENTRADA QUE FEZ
O EXCELLENTISSIMO, E REVERENDISSIMO SENHOR
D. F. R. ANTONIO
DO DESTERRO MALHEYRO

Bispo do Rio de Janeiro, em o primeiro dia deste presente Anno de 1747.
havendo sido seis Annos Bispo do Reyno de Angola, donde por no-
mição de Sua Magestade, e Bulla Pontificia, foy promovido
para esta Diocese.

COMPOSTA PELO DOUTOR
LUIZ ANTONIO ROSADO
DA CUNHA

*Juiz de Fóra, e Provedor dos defuntos, e au-
zentes, Capellas, e Refidos do Rio de Janeiro.*



RIO DE JANEIRO
Na Segunda Officina de ANTONIO ISIDORO DA FONCECA

Anno de M. CC. XLVII.
Com licenças do Senhor Bispo.

RELAÇÃO
DA ENTRADA QUE FEZ
O EXCELLENTISSIMO, E REVERENDISSIMO SENHOR
D.F. R. ANTONIO
DO DESTERRO MALHEYRO

Bispo do Rio de Janeiro, em o primeiro dia deste presente Anno de 1747.
havendo sido seis Annos Bispo do Reyno de Angola, donde por no-
mição de Sua Magestade, e Bulla Pontificia, foy promovido
para esta Diocese.

COMPOSTA PELO DOUTOR
LUIZ ANTONIO ROSADO
DA CUNHA

*Juiz de Fóra, e Provedor dos defuntos, e au-
zentes, Capellas, e Refidos do Rio de Janeiro.*



RIO DE JANEIRO
Na Segunda Officina de ANTONIO ISIDORO DA FONCECA.

Anno de M. CC. XLVII.
Com licenças do Senhor Bispo.

DEPOIS

e dictator

"Uma ferramenta
para edição filológica eletrônica
e análise linguística automática"

e-dictor

O e-dictor realiza uma conversão do texto comum para um texto **XML**, realizando, assim, a categorização sintática e morfológica de todos os elementos nele presentes.



XML

e**X**tensible **M**arkup **L**anguage

ou, simplesmente,
uma linguagem de rotulação simples

Editor

Por ora, tenhamos em mente que o **XML** somente é incorporado ao texto, tornando-o uma outra coisa.

e-dictor

O que é o e-dictor?

Como o e-dictor funciona?



Por que isso é importante?

Um exercício de abstração



1 hora



2 horas



3 horas

Um exercício de abstração



1 hora

**Computador
Celular**



2 horas

Computador



3 horas

Celular

Um exercício de abstração



1 hora

**Computador
Celular**

**Escrever
um conto**



2 horas

Computador

**Fazer
um trabalho**



3 horas

Celular

**Ler
"A Sonata a Kreutzer"**

Um exercício de abstração

Propriedades



1 hora
Computador
Celular

Escrever
um conto



2 horas
Computador

Fazer
um trabalho



3 horas
Celular

Ler
"A Sonata a Kreutzer"

Um exercício de abstração



1 hora

**Computador
Celular**

**Escrever
um conto**



2 horas

Computador

**Fazer
um trabalho**



3 horas

Celular

**Ler
"A Sonata a Kreutzer"**



**Atividades com
uma propriedade em comum (tempo).**

Um exercício de abstração

**Elemento
de foco
(atividades)**



1 hora

**Computador
Celular**

**Escrever
um conto**



2 horas

Computador

**Fazer
um trabalho**



3 horas

Celular

**Ler
"A Sonata a Kreutzer"**

No E-dictor...

Palavra / Morfema



**Característica
sintática / morfológica**



**Característica
sintática / morfológica**



**Outra palavra / morfema
com ao menos uma
característica em comum**



1 hora
Computador
Celular

Escrever
um conto



2 horas
Computador

Fazer
um trabalho



3 horas
Celular

Ler
"A Sonata a Kreutzer"

¶ Hũa planta se dá também nesta prouincia, que foy da ilha de Sam Thomé, com a fruita da qual se ajudam muitas pessoas a sustentar na terra. Esta planta he muy tenra & nam muito alta, nam tem ramos senam hũas folhas que seram seis ou sete palmos de cõprido. A fruita della se chama banáanas: parecen-se na feiçã com pepinos, & criam-se em cachos: algũs delles ha tam grandes que tem de cento & cincoenta banáanas pera cima.

Anotação de edição da sentença

<s id="s_209"><w id="01"><o>**A**</o></w><w id="02"><o>**frui-**<bk id="bk_5"/>**ta**</o><e
t="jun">frui-ta</e><e t="mod">fru-ta</e><e t="hif">fruta</e></w><w id="03"><o>-
della</o><e t="mod">dela</e></w><w id="04"><o>**fe**</o><e t="gra">se</e></w><w
id="05"><o>**chama**</o></w><w id="06"><o>**banáanas**</o><e t="mod">bananas</e></
w><w id="07"><o>:</o></w><w id="08"><o>**parecenfe**</o><e t="gra">parecense</
e><e t="mod">parecem-se</e></w><w id="09"><o>**na**</o></w><w id="10"><o>**feijam**</
o><e t="mod">feijão</e></w><w id="11"><o>**com**</o></w><w id="12"><o>**pe**<bk
id="bk_6"/>**pinos**</o><e t="jun">pepinos</e></w><w id="13"><o>,</o></w><w
id="14"><o>**&**</o><e t="gra">e</e></w><w id="15"><o>**criamfe**</o><e t="gra">criamse</
e><e t="mod">criam-se</e></w><w id="16"><o>**em**</o></w><w id="17"><o>**cachos**</
o></w><w id="18"><o>:</o></w><w id="19"><o>**algûs**</o><e t="mod">alguns</
e></w><w id="20"><o>**delles**</o><e t="mod">deles</e></w><w id="21"><o>**ha**</o><e
t="mod">há</e></w><w id="22"><o>**tam**</o><e t="mod">tão</e></w><w id="23"><o>-
gran-<bk id="bk_7"/>**des**</o><e t="jun">gran-des</e><e t="hif">grandes</e></w><w
id="24"><o>**que**</o></w><w id="25"><o>**tem**</o></w><w id="26"><o>**de**</o></w><w
id="27"><o>**cento**</o></w><w id="28"><o>**&**</o><e t="gra">e</e></w><w id="29"><o>-
cincoenta</o><e t="mod">cinquenta</e></w><w id="30"><o>**banáanas**</o><e t="mod">
bananas</e></w><w id="31"><o>**pera**</o><e t="mod">para</e></w><w id="32"><o>-
cima</o></w><w id="33"><o>.<bk id="bk_8"/></o></w></s>

Anotação de edição da sentença

<s id="s_209"><w id="01"><o>A</o></w><w id="02"><o>frui-<bk id="bk_5"/>ta</o><e
t="jun">frui-ta</e><e t="mod">fru-ta</e><e t="hif">fruta</e></w><w id="03"><o>-
della</o><e t="mod">dela</e></w><w id="04"><o>fe</o><e t="gra">se</e></w><w
id="05"><o>chama</o></w><w id="06"><o>banáñas</o><e t="mod">bananas</e></
w><w id="07"><o>:</o></w><w id="08"><o>parecenfe</o><e t="gra">parecense</
e><e t="mod">parecem-se</e></w><w id="09"><o>na</o></w><w id="10"><o>feijam</
o><e t="mod">feijão</e></w><w id="11"><o>com</o></w><w id="12"><o>pe<bk
id="bk_6"/>pinos</o><e t="jun">pepinos</e></w><w id="13"><o>,</o></w><w
id="14"><o>&</o><e t="gra">e</e></w><w id="15"><o>criamfe</o><e t="gra">criamse</
e><e t="mod">criam-se</e></w><w id="16"><o>em</o></w><w id="17"><o>cachos</
o></w><w id="18"><o>:</o></w><w id="19"><o>algûs</o><e t="mod">alguns</
e></w><w id="20"><o>delles</o><e t="mod">deles</e></w><w id="21"><o>ha</o><e
t="mod">há</e></w><w id="22"><o>tam</o><e t="mod">tão</e></w><w id="23"><o>-
gran-<bk id="bk_7"/>des</o><e t="jun">gran-des</e><e t="hif">grandes</e></w><w
id="24"><o>que</o></w><w id="25"><o>tem</o></w><w id="26"><o>de</o></w><w
id="27"><o>cento</o></w><w id="28"><o>&</o><e t="gra">e</e></w><w id="29"><o>-
cincoenta</o><e t="mod">cinquenta</e></w><w id="30"><o>banáñas</o><e t="mod">
bananas</e></w><w id="31"><o>pera</o><e t="mod">para</e></w><w id="32"><o>-
cima</o></w><w id="33"><o>.<bk id="bk_8"/></o></w></s>



Anotação morfosintática

**A/D-F fruta/N dela/P+PRO se/CL chama/VB-P bananas/N-P :/.
parecem-se/VB-P+CL na/P+D-F feijão/N com/P pepinos/N-P,/
e/CONJ criam-se/VB-P+CL em/P cachos/N-P :/. alguns/Q-P
deles/P+PRO há/HV-P tão/ADV grandes/ADJ-G-P que/C tem/TR-P
de/P cento/NUM e/CONJ cinquenta/NUM bananas/N-P para/P cima/N ./.**



Anotação morfosintática

A/D-F fruta/N dela/P+PRO se/CL chama/VB-P bananas/N-P :/.
parecem-se/VB-P+CL na/P+D-F feição/N com/P pepinos/N-P,/,
e/CONJ criam-se/VB-P+CL em/P cachos/N-P :/. alguns/Q-P
deles/P+PRO há/HV-P tão/ADV grandes/ADJ-G-P que/C tem/TR-P
de/P cento/NUM e/CONJ cinquenta/NUM bananas/N-P para/P cima/N ./.

Características / Propriedades dos elementos

e-dictor

O que é o e-dictor?

Como o e-dictor funciona?

Por que isso é importante? ←

e-dictor

Com o e-dictor, os *corpora* podem ser rapidamente analisados para, por exemplo, encontrar padrões recorrentes de diferenças estruturais entre o português contemporâneo e o português arcaico.

e-dictor

"A principal finalidade do e-dictor é oferecer uma interface amigável aliada a um alto nível de controle e flexibilidade na codificação de textos eletrônicos com finalidade de pesquisa linguística."

REFERÊNCIAS BIBLIOGRÁFICAS

Aluísio, S. M. & G. M. de Barcellos Almeida. 2006. O que é e como se constrói um *corpus*? Lições apreendidas na compilação de vários corpora para pesquisa linguística. *Calesidoscópio* 4(3), 156-178.

Berber Sardinha, T. 2004. *Linguística de corpus*. São Paulo: Manoele.

Paixão de Sousa, M. C. 2013. Humanidades Digitais: O digital e as novas formas de construção do conhecimento. Comunicação ao Seminário Internacional Sistemas de Informação e Acervos Digitais de Cultura. São Paulo, 12 de março de 2013.

Disponível em Vídeo – Canal da Secretaria de Políticas Culturais do Ministério da Cultura:

<http://youtu.be/m0s-iAfZPD>

Acesso em 15 de agosto de 2015.

_____. 2014a. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa* (16), 53-93. Disponível em:

<http://www.revistas.usp.br/flp/article/view/88404>.

Acesso em: 15/08/201

_____. 2014b. E-Dictor. *Uma ferramenta para as humanidades digitais*. Conferência na “Semana do Projeto Libolo”. Slides.

REFERÊNCIAS BIBLIOGRÁFICAS

Sinclair, J. 2005. Corpus and text – basic principles. In M. Wyne (ed.). *Developing linguistics corpora: a guide to good practice*, p. 1-16. Oxford, Oxbow Books. Disponível em:

<http://ahds.ac.uk/linguistic.corpora/>

Acesso em 25/06/2014

Obrigado!

e dictator

AHDig
Associaç[ã]o das Humanidades Digitais

HD.br