The COMET Project: Comparable and Parallel Corpora for the English-Portuguese Pair

Stella E. O. Tagnin
University of São Paulo
UCCTS – Ormskirk
27-29 July 2010

Brief history

- 1998 – Projeto CoMET is conceived:
  - Technical Corpus – CorTec
  - Translation Corpus – CorTrad
  - Learner Corpus – CoMAprend
- Originally: 5 languages
  - English, French, German, Italian, and Spanish

CorTec

- **2001**– Technical Translation subject at Specialization in Translation Course
  - 11 different glossaries: http://www.fflch.usp.br/citrat/citrat.htm
  - 11 bilingual comparable corpora
- **Subsequent years**: more corpora and more glossaries (not published)
- **Plus**: corpora from graduate students

2005 – 1st launching

- CorTec (Technical Corpus)
  - http://www.fflch.usp.br/dlm/comet/consulta_cortec.html
- CoMAprend (Learner Corpus)
  - http://www.fflch.usp.br/dlm/comet/comaprend.html

CorTec 2005

- 5 comparable corpora:
  - Cooking recipes
  - Ecotourism - environment
  - Computer Science
  - Cardiology – Hipertension
  - Law – agreements
- English – Portuguese original texts
- approximately 200,000 words each

CorTec 2005

- Online Tools
  - Frequency List (also alphabetical)
  - Concordancer
    - equal to (exact word)
    - starting with (prefixes)
    - finishing with (suffixes)
    - containing (root of word)
  - n-grams

CoMAprend - 2005

- Writings by students
  - undergraduate courses
  - extracurricular courses
- Languages
  - English, French, German, Italian, and Spanish
- Only corpora for download
- **2008**: inclusion of investigation tools

CorTec 2008 – 2nd launching

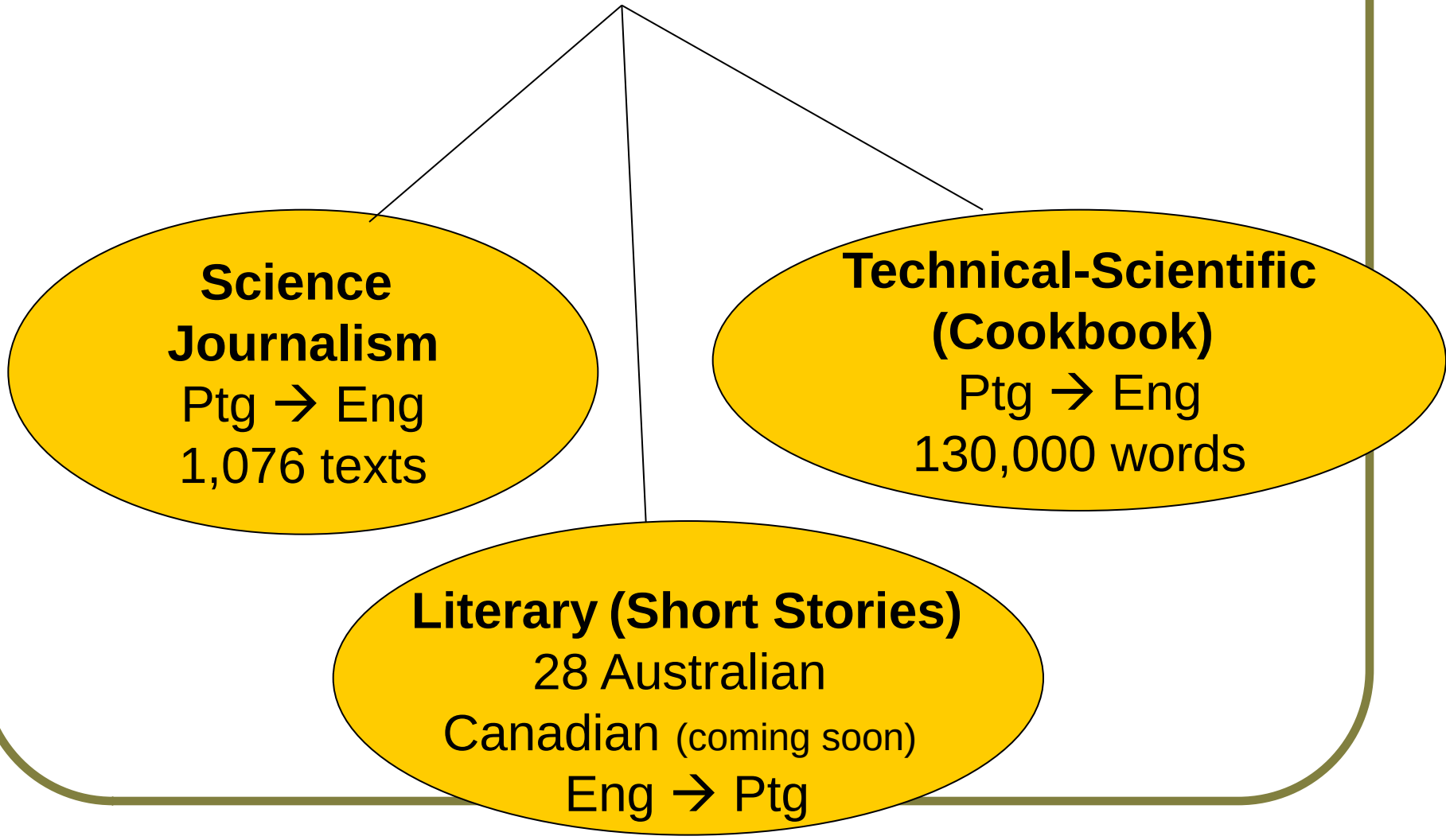
14 corpora

- Ecotourism
- Hipertension
- Legal agreements
- Astronomy
- Renal failure
- Linguistics
- Magnetic flowmeters
- Nutritional Supplements
- Computer Science
- Football
- Coffee
- Cultural Tourism
- Cooking recipes 1 & 2

CorTrad 2009 – new!

- Cooperation began May 2008:
  - **CoMET**: collection and preparation of texts
  - **Linguateca**: computational implementation - DISPARA (Santos, 2002); alignment, POS tagging and semantic annotation
- **Parallel Corpus**
  - English →Portuguese
  - Portuguese → English
- Interface: only in Portuguese (being translated into English)
- http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html

CorTrad – 3 parallel subcorpora



CorTrad 2009

- **Population**: availability
- **Special features**:
  - ✓ **Multiversion** – comparison of various stages of translation
  - ✓ **Elaborate search queries** – specific for each subcorpus

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

ovas publicações no link

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

Início | Jornalístico | Literário | Técnico-científico

- Agradecimentos
- Projeto conjunto
- Como citar
- Perguntas já respondidas
- Fale conosco

Bem-vin rad

O **CorTrad** é o c ção (português-inglês) do **COMET**. Além das possibilidades de pesquisa normal rpora paralelos, o CorTrad dispõe de pelo menos duas funcionalidades i lidade de se comparar diferentes versões de um mesmo texto (original, versões ublicada); (ii) mecanismos de busca diferenciados para cada gênero pesquisado - per esquisar seções específicas dos diferentes tipos textuais.

O CorTrad é um corpus aberto e conta atualmente com três subcorpora:

- CorTrad **jornalístico** (por ora, divulgação científica)
- CorTrad **literário** (por ora, contos)
- CorTrad **técnico-científico** (por ora, culinária)

A lista das obras incluídas, assim como os agradecimentos devidos, encontram-se em Agradecimentos.

A disponibilização do CorTrad na rede é um projeto conjunto COMET/NILC/Linguateca, usando o sistema DISPARA.

Comentários ou questões para a equipe do CorTrad

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

Responsáveis: **Stella E. O. Tagnin** e **Elisa D. Teixeira**
Desenho/concepção gráfica do site do CorTrad: **Patricia Tagnin**
Coleta e preparação de textos, revisão de alinhamento: **Sabrina Matuda** e **Soraia Manzela**

Linguateca: aplicação do sistema DISPARA ao CorTrad; criação dos corpora em formato IMS (Open) CWB; desenvolvimento de novas funcionalidades para este projeto específico, em particular a procura em várias traduções e a adaptação da interface para textos técnicos; documentação do processo a seguir para aumento posterior do corpus.
Responsável: **Diana Santos**

NILC: alojamento do CorTrad em seu servidor; manutenção e aperfeiçoamento da parte computacional do serviço a partir da data de uma primeira versão estável.
Responsável: **Sandra Aluísio**

Copyright - Disclaimer

Comentários ou questões para a equipe do CorTrad

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

ublicações no link Artigos

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

Perguntas já respondidas

FAQ

Como faço para ver os adjetivos mais frequentes na parte cumulativa?

Procure por [pos="ADJ.*"] e peça **Distribuição dos lemas**

Como faço para procurar advérbios no corpus de divulgação científica?

Procure [pos="ADV"]

O que significa principal?

Principal é o corpus a partir do qual será feita a pesquisa. Por exemplo: se você quer saber quantas ocorrências de "vermelho" há no texto traduzido finalmente publicado, deverá marcá-lo como principal. Isto significa que, se alguma ocorrência de "vermelho" da primeira versão da tradução foi eliminada na revisão, ela não aparecerá no resultado.

A sua pergunta ainda não foi respondida?
Contate-nos que tentaremos responder muito em breve.

Science Journalism: Revista Fapesp

Original

Brazilian Portuguese)

Published translation

(online publication)

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

Início | Jornalístico | Literário | Técnico-científico

Textos incluídos no corpus jornalístico

A **parte jornalística do CorTrad** conta atualmente com textos da edições de 2001, 2002 e 2003 da Revista Pesquisa FAPESP, totalizando 20 números. As seções incluídas foram: Humanidades, Ciência, Tecnologia, Estratégias, Laboratório, Linha de Produção e Política de C&T.

Lista de revistas FAPESP incluídas:

- 2001: edições 62 (mar-01), 63 (apr-01), 64 (mai-01), 65 (jun-01), 67 (ago-01), 68 (set-01), 69 (out-01) e 70 (nov-01)
- 2002: edições 76 (jun-02), 77 (jul-02), 78 (ago-02), 79 (set-02), 80 (out-02), 81 (nov-02) e 82 (dez-02)
- 2003: edições 83 (jan-03), 84 (fev-03), 85 (mar-03), 86 (abr-03) e 87 (mai-03)

Atenção, esta página ainda é preliminar, em particular a classificação quanto ao gênero e assunto, segundo a tabela proposta pelo Lácio-Web, está sendo revista neste momento.

ID	Gênero	Assunto
ca-abr01v01	Carta	Direito
ca-abr01v02	Carta	Cultura
ca-abr01v03	Carta	Cultura

PROJETO COMET
Corpus Multilíngue para Ensino e Tradução

USP

bem vindo ao site do Pr

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

Examples of Search Queries

Pesquisar

Exemplos de pesquisas

	Procurar:	Resultado:
a palavra *violino*	`violino`	concordância
palavras começando por *açúc*	`"açúc.*"`	concordância
palavras antecedidas por dois pontos	`":" @[]`	distribuição das formas
formas do verbo *reunir* em contexto	`reunir.*`	concordância
palavras modificadas por *muito*	`"muito" @[]`	distribuição de formas

Veja dados quantitativos para mais informações.

Comentários ou questões para a equipe do CorTrad

Última atualização da página: 06 de dezembro de 2009
Última atualização do corpus: 13 de outubro de 2009

Internet 100%

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

são português / inglês

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

Início | Jornalístico | Literário | Técnico-científico

CorTrad jornalístico divulgação científica

O CorTrad é um corpus aberto, sujeito a alterações. Clique em atualizações para saber mais.

Pesquisar no corpus

Help

Original	principal	pesquisa	ver
Tradução publicada	principal		ver

Ignorar maiúsculas/minúsculas

Pesquisar

Resultado

- Concordância
- Distribuição das formas
- Distribuição dos lemas
- Distribuição da categoria gramatical (PoS)

Internet 100%

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

ções no link Artigos

Ajuda

Se tem dúvidas sobre a sintaxe da pesquisa, pode ver por ora as seguintes páginas feitas para projetos relacionados:

- Exemplos de pesquisa no AC/DC
- Ajuda à pesquisa no COMPARA

Em breve mais informação específica de ajuda ao próprio CorTrad será colocada aqui.

Comentários ou questões para a equipe do CorTrad

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

a as novas publicações n

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

COMPARA

- Início
- Pesquisa
- Ajuda

  Pesquisa

  Aula prática
- Textos do COMPARA
- Informações gerais
- Documentação específica
- Linguateca

Notícias [19 de Set. de 2008]

Ajuda para pesquisar

Tópicos de ajuda

Imprimir

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

:S

O que quer pesquisar	Exemplo do que deve digitar	Possíveis resultados
palavra isolada	"assim"	assim
sequência de mais de uma palavra	"a" "pensar"	a pensar
duas palavras intercaladas por uma palavra qualquer	"por" ".*" "tempo"	por um tempo, por muito tempo, po tempo, por pouco tempo, etc.
duas palavras intercaladas por zero, uma, duas ou três outras quaisquer	"bebeu" []* "vinho" within 3	bebeu vinho, bebeu um vinho, bebe vinho, bebeu muito vinho...
palavra com grafias diferentes	"a(c)?ção"	acção, ação
	"lo[iu]ça"	loiça, louça
palavras alternativas	"(enfim|finalmente)"	enfim, finalmente
palavras começadas com "des"	"des.*"	desrespeito, desfavorecer, desânim desta, destaque...
palavras terminadas em "ção"	".*ção"	canção, lição, relação, Conceição,

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

Seja bem vindo

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

Original	principal [lema="pesquisar"]	ver
Tradução		ver

Search possibilities for Science Journalism

Pesquisar

Resultado

Concordância	Distribuição das formas
Distribuição dos lemas	Distribuição da categoria gramatical (PoS)
Distribuição do tempo verbal e/ou do caso pronominal	Distribuição de pessoa e/ou número
Distribuição do gênero morfológico	Distribuição da função sintática
Distribuição por documento	Distribuição por data de publicação
Distribuição por gênero de texto	Distribuição por tema
Distribuição por campo semântico	Distribuição por grupo (de cor, de vestuário, etc.)

Opções

Resultados por ordem alfabética (só distribuições)

Resultados numerados (só concordância)

Internet 100%

Comparing results

- How are verbs “acreditar” (= believe) and “achar” (= think) used in different text types in the journalistic corpus?

- [lema=“acreditar”] + Distribuição por gênero de texto
- [lema=“achar”] + Distribuição por gênero de texto

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

CorTrad jornalístico divulgação científica

O CorTrad é um corpus aberto, sujeito a alterações. Clique em atualizações para saber mais.

Pesquisar no corpus

Original	◉ principal	[lema="acreditar"]	☑ ver
Tradução publicada	○ principal		☑ ver

☐ Ignorar maiúsculas/minúsculas

Pesquisar

Resultado

○ Concordância	○ Distribuição das formas
○ Distribuição dos lemas	○ Distribuição da categoria gramatical (PoS)
○ Distribuição do tempo verbal e/ou do caso pronominal	○ Distribuição de pessoa e/ou número
○ Distribuição do gênero morfológico	○ Distribuição da função sintática
○ Distribuição por documento	○ Distribuição por data de publicação
◉ Distribuição por gênero de texto	○ Distribuição por tema
○ Distribuição por campo semântico	○ Distribuição por grupo (de cor, de vestuário, etc.)

Opções

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

CorTrad jornalístico divulgação científica

196 casos.

Expressão de busca: **[lema="acreditar"]**
Resultado escolhido: **distribuição de gênero**
Corpus pesquisado: **originais**

= believe

Voltar
Nova pesquisa

Houve 11 valores diferentes de genero.

Reportagem	92
Notícia	40
article	23
Artigo	21
Entrevista	11
Notícia-OU-section	3
section	2
opinion	1
article-OU-section	1
notícia-OU-section	1
Notícia -OU-Notícia	1

Para informação sobre os códigos internos do atributo **genero**, consulte a página do subcorpus respectivo.

8-janeiro-2010:*Esperamos que o CorTrad jornalístico divulgação científica lhe tenha sido útil!*

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

CorTrad jornalístico divulgação científica

160 casos.

Expressão de busca: **[lema="achar"]**
Resultado escolhido: **distribuição de genero**
Corpus pesquisado: **originais**

= think

Vo
Nova pesq

Houve 11 valores diferentes de genero.

Entrevista	57
Reportagem	50
article	24
Notícia	14
Artigo	8
Carta	2
-OU-Editorial	1
editorial	1
Editorial	1
Notícia-OU-section	1
section	1

Para informação sobre os códigos internos do atributo **genero**, consulte a página do subcorpus respectivo.

8-janeiro-2010:*Esperamos que o CorTrad jornalístico divulgação científica lhe tenha sido útil!*

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

CorTrad jornalístico divulgação científica

196 casos.

Expressão de busca: **[lema="acreditar"]**
Resultado escolhido: **distribuição de gênero**
Corpus pesquisado: **originais**

= believe

Voltar
Nova pesquisa

Houve 11 valores diferentes de genero.

Reportagem	92
Notícia	40
article	23
Artigo	21
Entrevista	11
Notícia-OU-section	3
section	2
opinion	1
article-OU-section	1
notícia-OU-section	1
Notícia -OU-Notícia	1

Para informação sobre os códigos internos do atributo **genero**, consulte a página do subcorpus respectivo.

8-janeiro-2010:*Esperamos que o CorTrad jornalístico divulgação científica lhe tenha sido útil!*

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

CorTrad jornalístico divulgação científica

160 casos.

Expressão de busca: **[lema="achar"]**
Resultado escolhido: **distribuição de genero**
Corpus pesquisado: **originais**

= think

Vo
Nova pesq

Houve 11 valores diferentes de genero.

Entrevista	57
Reportagem	50
article	24
Notícia	14
Artigo	8
Carta	2
-OU-Editorial	1
editorial	1
Editorial	1
Notícia-OU-section	1
section	1

Para informação sobre os códigos internos do atributo **genero**, consulte a página do subcorpus respectivo.

8-janeiro-2010:*Esperamos que o CorTrad jornalístico divulgação científica lhe tenha sido útil!*

Technical-Scientific: Cookbook

Original (Brazilian Portuguese)

Translators' first version (English)

Revised text (by native speaker)

Published translation (not yet in the corpus)

How are adverbs distributed among the 3 parts of the Cooking corpus:

filling – introduction - conclusion?

[pos="ADV"]

distribuição por parte da obra
→ distribution by part of file

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

/ersão português / inglês

CorTrad técnico-científico culinária

7207 casos.

Expressão de busca: **"[pos=ADV.*]"**
Resultado escolhido: **distribuição de idparte**
Corpus pesquisado: **originais**

Voltar

Nova pesquisa

Houve 3 valores diferentes de idparte.

recheio 6089
introd 857
concl 261

15-dezembro-2009:*Esperamos que o CorTrad técnico-científico culinária lhe tenha sido útil!*

Comentários ou questões para a equipe do CorTrad

When is
"natural" in Portuguese
NOT translated as
natural in English?

natural vs !natural

Resultado: “natural” ≠ “*natural*”

PROJETO COMET
Corpus Multilíngue para Ensino e Tradução

USP

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

Início | Jornalístico | Literário | Técnico-científico

CorTrad técnico-científico culinária

Expressão de busca: **"natural" :CORTRAD_CULI_TRAD1 !"natural" :CORTRAD_CULI_TRAD2 !"natural"**
Resultado escolhido: **concordância em contexto**
Corpus pesquisado: **originais**

[ajuda]

14 ocorrências.

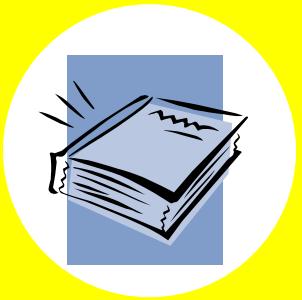
[voltar]
[nova pesquisa]

Original	Primeira tradução	Tradução revisada
O mais **natural** é imaginar um almoço lá fora, mas por que não um jantar, com o frescor da noite?	When we think of an outdoor meal, it is usually lunch that comes to one's mind but why not dinner, enjoying the freshness of the night?	When we think of an outdoor meal, it is usually lunch that comes to one's mind but why not dinner, enjoying the freshness of the night?
1/2 xícara de iogurte **natural**	½ cup whole milk plain yogurt	½ cup plain whole milk yogurt
3/4 de xícara de iogurte **natural**	1/3 cup sugar	? cup sugar ½ cup white wine vinegar
1 1/2 xícara de iogurte **natural**	1 ½ cups whole milk plain yogurt	1½ cups plain whole milk yogurt
2 1/3 de xícaras de iogurte **natural**	2 1/3 cups whole milk plain yogurt	2? cups plain whole milk yogurt
1 xícara de iogurte **natural**	1 cup whole milk plain yogurt	1 cup plain whole milk yogurt
600 ml de iogurte **natural** (3 copinhos)	600 ml (21 fl oz) whole milk plain yogurt	600 ml (21 fl oz) plain whole milk yogurt
100 ml de iogurte **natural**	100 ml (3.5 fl oz) whole milk plain yogurt	100 ml (3½ fl oz) plain whole milk yogurt
600 ml de iogurte **natural** (3 copinhos)	600 ml (21 fl oz) whole milk plain yogurt	600 ml (21 fl oz) plain whole milk yogurt
200 ml de iogurte **natural** (1 copinho)	200 ml (7 fl oz) whole milk plain yogurt	200 ml (7 fl oz) plain whole milk yogurt
200 ml de iogurte **natural** (1 copinho)	200 ml (7 fl oz) whole milk plain yogurt	200 ml (7 fl oz) plain whole milk yogurt
1 litro de suco de laranja **natural**	1 L (1.1 qt) freshly squeezed orange juice	1 L (1 qt) freshly squeezed orange juice
1 litro de suco de laranja **natural**	1 L (1.1 qt) freshly squeezed orange juice	1 L (1 qt) freshly squeezed orange juice

Internet | Protected Mode: Off 100%

Literary: Australian short stories

Original (Australian English)



Student's translation (Brazilian Portuguese)

Revised draft (after teacher's suggestions)

Published translation

Search word: house

PROJETO COMET
Corpus Multilíngue para Ensino e Tradução

USP

Se

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
Artigos, etc.
Links
Informativo
Contato
Site FFLCH
Site USP

Início | Jornalístico | Literário | Técnico-científico

CorTrad literário contos

Expressão de busca: **"house" %c**
Resultado escolhido: **concordância em contexto**
Corpus pesquisado: **originais**

[ajuda]

89 ocorrências.

[voltar]
[nova pesquisa]

Original	Primeira tradução	Tradução revisada	Tradução publicada
nothing has ever been secular in this **house**.	Penso com irritação: nada jamais foi secular nesta casa.	Penso com irritação: nada jamais foi secular nesta casa.	Penso com irritação: nada jamais foi secular nesta casa.
But I also move out of the shaft of light that falls from the **house**, knowing, with a rush of annoyance, that if they see me with weeping they will discern the Holy Spirit who hovers always with his bright demanding wings.	Mas eu também me escondo da claridade que sai da casa, sabendo, com súbita irritação, que se eles me virem chorando eles irão enxergar o Espírito Santo que sempre paira sobre nós com suas asas brilhantes e exigentes.	Mas eu também me escondo do facho de luz que vem da casa, sabendo, com súbita irritação que, se eles me virem chorando, vão enxergar o Espírito Santo que sempre paira sobre nós com suas asas brilhantes e exigentes.	Mas eu também me escondo da claridade que sai da casa, sabendo, com súbita irritação que, se eles me virem chorando, vão enxergar o Espírito Santo que sempre paira sobre nós com suas asas brilhantes e exigentes.
The **house** was full of the whispering of women, and all of us felt melancholy.	SEM TRADUÇÃO	A casa estava cheia de sussurros de mulher, e todos nós nos sentíamos melancólicos.	A casa estava cheia de sussurros de mulher, e todos nós nos sentíamos melancólicos.
Soon the **house** began to live again.	SEM TRADUÇÃO	Logo a casa voltou à vida outra vez.	Logo a casa voltou à vida outra vez.
There was a coming and going, and music, in the old **house** at Schutz.	SEM TRADUÇÃO	Havia um entra-e-sai, e música também, na velha casa em Schutz.	Havia um entra-e-sai, e música também, na velha casa de Schutz.
That year my eldest sister Phrosso thought old **house** at Schutz.	SEM TRADUÇÃO	Naquele ano, Phrosso, minha irmã mais velha, achou que estava apaixonada por um atleta italiano, e meu irmão Aleko havia decidido tornar-se estrela de cinema.	Naquele ano, Phrosso, minha irmã mais velha, achou que estava apaixonada por um atleta italiano, e meu irmão Aleko decidiu tornar-se estrela de cinema.
Sometimes I felt this bitterly, but I could not	Às vezes eu encarava isso de uma forma um	Às vezes eu encarava isso de uma forma um	Às vezes eu encarava isso de uma forma um

Done | Internet | Protected Mode: Off | 100%

Semantic Tagging

Clothes

- Color

Principal
O Projeto
Equipe
CorTec
oMAprend
CorTrad
rtigos, etc.
Links
nformativo
Contato
ite FFLCH
Site USP

Início | Jornalístico | Literário | Técnico-científico

CorTrad jornalístico divulgação científica

93 casos.

Expressão de busca: **[sema="roupa.*"]**
Resultado escolhido: **distribuição de lema**
Corpus pesquisado: **originais**

Voltar
Nova pesquisa

Houve 26 valores diferentes de lema.

roupa	17
coroa	12
sapato	9
manta	7
malha	7
carteira	5
manto	5
calçado	4
camiseta	4
gabão	3
acessório	2
bota	2
meia	2
vestido	2
camisa	1
[illegible]	1

Clothes

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

nultiversão português / inglês

CorTrad técnico-científico culinária

1139 casos.

Expressão de busca: **[sema="cor.*"]**
Resultado escolhido: **distribuição de lema**
Corpus pesquisado: **originais**

Voltar

Nova pesquisa

Houve 43 valores diferentes de lema.

dourar	341
branco	144
vermelho	122
verde	118
louro	79
cru	45
preto	39
cor	39
laranja	32
dourado	27
amarelo	15
roxo	14
rosado	12
creme	12
castanho	10
castanho do pará	9

Color

PROJETO COMET
Corpus Multilíngue para Ensino e Tradução

Trad - nosso novo corpus pa

CorTrad técnico-científico culinária

1139 casos.

Expressão de busca: **[sema="cor.*"]**
Resultado escolhido: **distribuição de func**
Corpus pesquisado: **originais**

Voltar
Nova pesquisa

Houve 40 valores diferentes de func.

N<	475
FMV	98
P<	92
>N	84
IMV_#ICL-P<	82
<SC	72
<ACC	42
IMV_#ICL-AUX<	31
IMV_#ICL-<SC	30
IMV_#ICL-<ACC	24
<PRED	19
NPHR	13
IMV	9
N<PRED	9
IMV_#ICL-N<	6
<SUBJ	5

Syntactic Function

PROJETO COMET

Corpus Multilíngue para Ensino e Tradução

USP

ça o CorTrad - nosso no

- Principal
- O Projeto
- Equipe
- CorTec
- CoMAprend
- CorTrad
- Artigos, etc.
- Links
- Informativo
- Contato
- Site FFLCH
- Site USP

Expressão de busca: **"environment"**
Resultado escolhido: **distribuição de docid**
Corpus pesquisado: **tradução final**

Voltar

Nova pesquisa

Houve 152 valores diferentes de docid.

ci-mai01v12	13
ci-out02v12	7
ci-fev03v14	7
ci-mar01v07	6
hu-abr01v08	5
tec-mar03v07	4
su-set01v24	4
po-ago01v13	4
ci-mai03v17	4
ci-jul02v13	4
tec-out02v03	3
tec-abr01v08	3
hu-mai01v09	3
ci-set01v16	3

Journalistic by document

Concluído

Internet

100%

CorTrad técnico-científico culinária

1139 casos.

Expressão de busca: **[sema="cor.*"]**
Resultado escolhido: **distribuição de pos**
Corpus pesquisado: **originais**

pos = *part-of-speech*

Voltar
Nova pesquisa

Houve 6 valores diferentes de pos.

ADJ	619
V	270
N	139
V_fmc	93
ADJ_n	14
V_n	4

Para informação sobre os códigos internos do atributo **pos**, consulte por ora a página do projeto AC/DC sobre anotação.

5-janeiro-2010:*Esperamos que o CorTrad técnico-científico culinária lhe tenha sido útil!*

Comentários ou questões para a equipe do CorTrad

Principal
O Projeto
Equipe
CorTec
CoMAprend
CorTrad
rtigos, etc.
Links
nformativo
Contato
Site FFLCH
Site USP

Início | Jornalístico | Literário | Técnico-científico

CorTrad técnico-científico culinária

1139 casos.

Expressão de busca: **[sema="cor.*"]**
Resultado escolhido: **distribuição de sema**
Corpus pesquisado: **originais**

Semantic field: color

Voltar
Nova pesquisa

Houve 3 valores diferentes de sema.

cor	1041
cor:humana	91
cor:ausência	7

5-janeiro-2010:*Esperamos que o CorTrad técnico-científico culinária lhe tenha sido útil!*

Comentários ou questões para a equipe do CorTrad

CorTrad - specifics

Improvements over other English-Portuguese parallel corpora

- **Multiversion** – comparison of various stages of translation process
- **Elaborate Search queries: specific for each corpus**

Computational background

- **DISPARA** (Santos, 2002) – system to make parallel corpora available online
  - ✓ **Corpus processing**
    - → IMS-CWB (Christ *et al*., 1999), now **Open CWB**
  - ✓ **PoS tagging**
    - → Portuguese: **PALAVRAS** (Bick, 2000) http://visl.hum.sdu.dk/visl/pt/
    - → English: **CLAWS** (Rayson & Garside, 1998) http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/
  - ✓ **Interface**
    - → conceived by team and implemented by Patricia Tagnin

Thanks to

- **Eckhard Bick** and **Paul Rayson** for permission to use PALAVRAS and CLAWS, respectively, for the CorTrad.
- **Sandra Aluísio** and **Arnaldo Candido Júnior** from NILC for hosting the CoMET Project
- **Diana Santos - Linguateca**, co-financed by the Portuguese Government, by EU (FEDER e FSE), under agreement POSC/339/1.3/C/NAC, by UMIC and by FCCN.
- **CNPq**, for grants to develop COMET (2005) and COMET (2008).

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

POS_Conhecimento
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

UMIC
Agência para a Sociedade do Conhecimento

FCCN
Fundação para a Computação Científica Nacional

Thank you

Stella

(seotagni@usp.br)