VIII Encontro de Linguística de Corpus – UERJ,13-14 nov. 2009

CorTrad:
um novo corpus paralelo multiversão para o par de línguas português-inglês

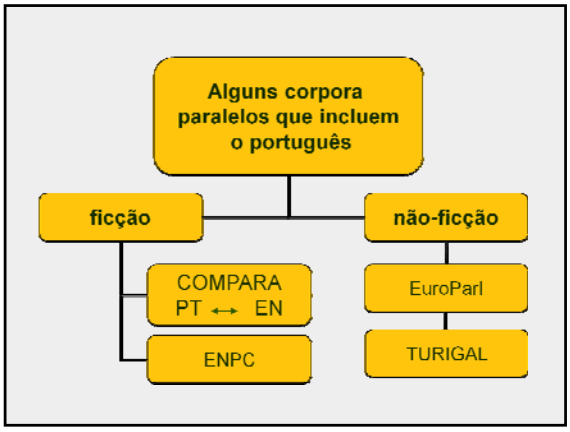
Elisa D. Teixeira (Projeto CoMET)
Diana Santos (Linguateca/SINTEF)
Stella E. O. Tagnin (DLM-USP)

Esta apresentação

- Corpora paralelos: usos e potencial
- Proposta do CorTrad
- Caracterização do corpus
- Caracterização do sistema de pesquisa
- Exemplos de pesquisas possíveis
- Etapas futuras

Corpora paralelos e tradução

- **Corpora paralelos: relevantes para o estudo, o ensino e a prática da tradução**
  - ✓ Isabelle (1992), McEnery & Wilson (1993) e Somers (1993) – potencial de corpora paralelos alinhados para a **tradução automática**
  - ✓ Johansson & Hofland (1994), Santos (1995) e Tagnin (2004) – potencial de corpora paralelos alinhados para **estudos contrastivos**
  - ✓ Malmkjaer (1998), Frankenberg-Garcia (1999) e Peters *et al.* (2000) – utilidade de corpora paralelos alinhados como **fonte de equivalentes na prática e no ensino da tradução**



CorTrad – projeto conjunto

- **Projeto CoMET** – DLM/USP, São Paulo (http://www.fflch.usp.br/dlm/comet/)
  - ✓ idealização
  - ✓ coleta e preparo dos textos
  - ✓ disponibilização do corpus para a comunidade
- **Linguateca** – SINTEF, Oslo (http://www.linguateca.pt/)
  - ✓ idealização
  - ✓ desenvolvimento e implementação da parte computacional
- **NILC** – USP, São Carlos (http://www.nilc.icmc.usp.br/)
  - ✓ hospedagem
  - ✓ futuramente: manutenção e melhoria da parte computacional

Características do CorTrad

- Projeto em andamento, iniciado em maio de 2008
- Conta atualmente com 3 subcorpora:
  - ✓ **Literário** (por ora, 27 contos australianos)
  - ✓ **Técnico-científico** (por ora, 1 livro de culinária de 350 páginas – aprox. 130.000 palavras)
  - ✓ **Jornalístico** (por ora, 1.076 textos provenientes de vários números da Revista Pesquisa Fapesp (2001 a 2003)
- Critério de coleta: disponibilidade
- **Diferenciais**:
  - ✓ **Multiversão** – permite comparar várias etapas da tradução / revisão
  - ✓ **Sistema de busca refinado** – permite buscas específicas para cada subcorpus

Corpus literário:
Contos australianos
(*corpus de aprendizes)

Corpus técnico-científico:
Culinária

Corpus jornalístico:
Divulgação científica

Sistema de pesquisa

- **DISPARA** (Santos, 2002) - sistema parametrizável de disponibilização de corpora paralelos na rede
  - ✓ **Sistema de processamento de corpus**
    → IMS-CWB (Christ *et al*., 1999), agora **Open CWB**
  - ✓ **Etiquetagem morfossintática**
    → Português: **PALAVRAS** (Bick, 2000)
    http://visl.hum.sdu.dk/visl/pt/
    → Inglês: **CLAWS** (Rayson & Garside, 1998)
    http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/
  - ✓ **Interface**
    → desenhada pela equipe e programada por Patricia Tagnin

Anotação

- Associada ao conteúdo
  - ✓ **Literário**: autor, tradutor, nome do conto
  - ✓ **Culinária**: estrutura textual (partes da obra; receita: título, lista de ingredientes, modo de fazer, tempo de preparo, etc.)
  - ✓ **Jornalístico**: título do artigo, assunto, gênero
- Global
  - ✓ Lema, categoria gramatical, tempo verbal, gênero, pessoa, número
  - ✓ Cor e grupo de cor, roupa e grupo de roupa (só em português)

Como usar o corpus

http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html

Tela inicial do CorTrad

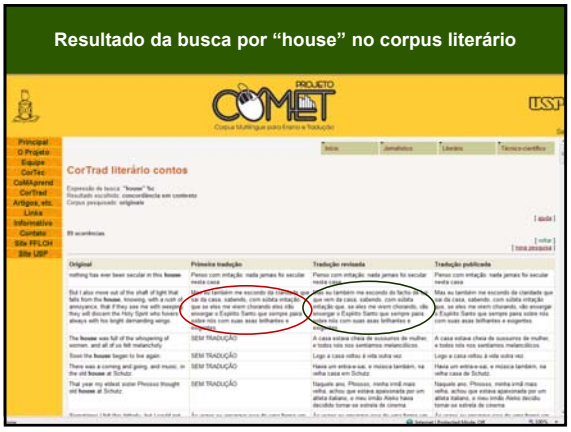
Tela de busca do corpus jornalístico

Resultado da busca por “research” nos textos traduzidos

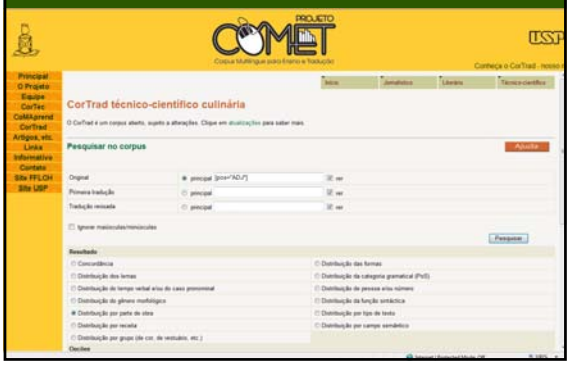
Restrição de busca por categoria gramatical

Resultado (jornalístico): “esquerda” = “left” (substantivo)

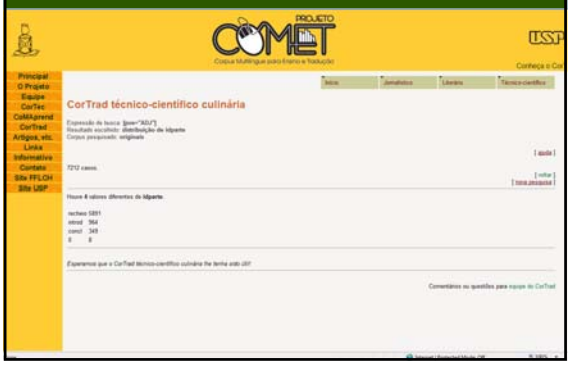
Resultado da busca por “house” no corpus literário



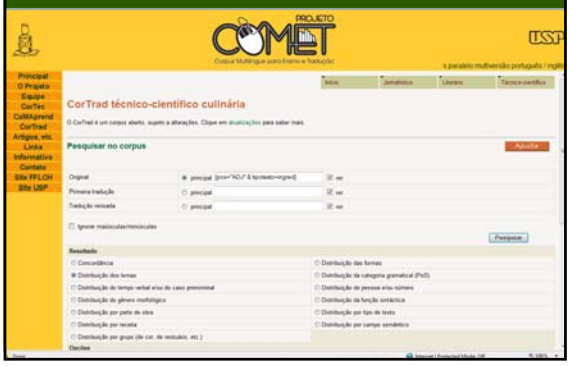
Tela de busca do corpus técnico-científico



Resultado da distribuição de adjetivos nas seções do livro



Distribuição de adjetivos em diferentes partes das receitas



Adjetivos em receitas (ingred) e reportagens

CorTrad técnico-científico culinária
Expressão de busca: [pos="ADJ" & tipotexto=ingred]
Resultado escolhido: distribuição de lema
Corpus pesquisado: originais
1056 casos.

Houve **167** valores diferentes de lema.

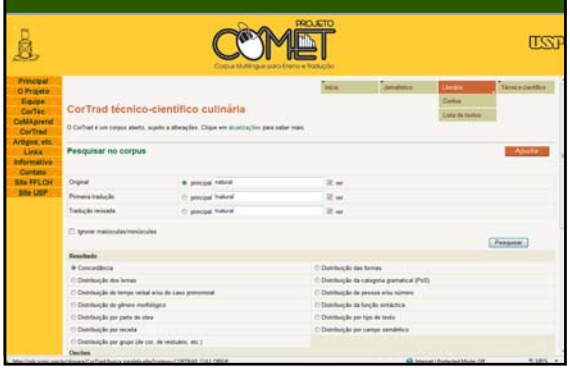
grande	103
fresco	74
vermelho	69
inteiro	54
médio	46
verde	45
branco	45
pequeno	45
maduro	25
limpo	24

CorTrad jornalístico divulgação científica
Expressão de busca: [pos="ADJ" & genero="Reportagem"]
Resultado escolhido: distribuição de lema
Corpus pesquisado: originais
23057 casos.

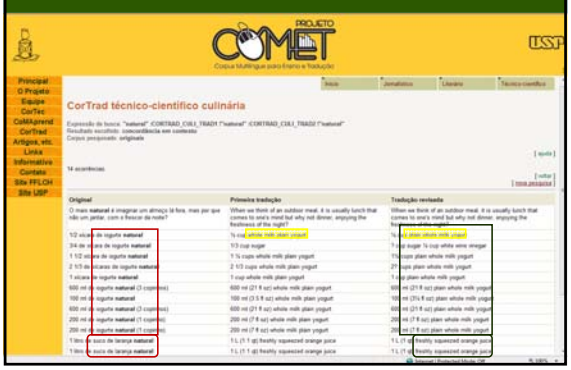
Houve **2842** valores diferentes de lema.

grande	843
novo	566
pequeno	328
brasileiro	285
bom	255
humano	243
alto	241
importante	208
científico	207
possível	207

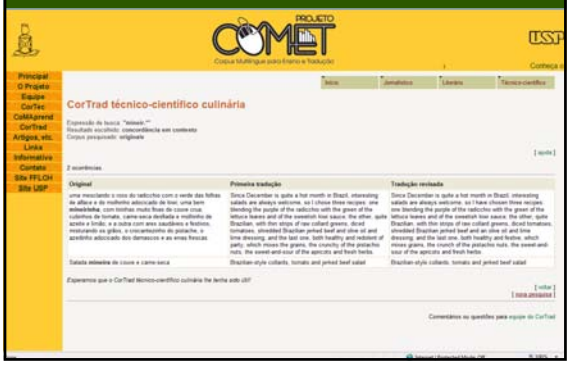
Casos em que "natural" não é vertido como "natural"



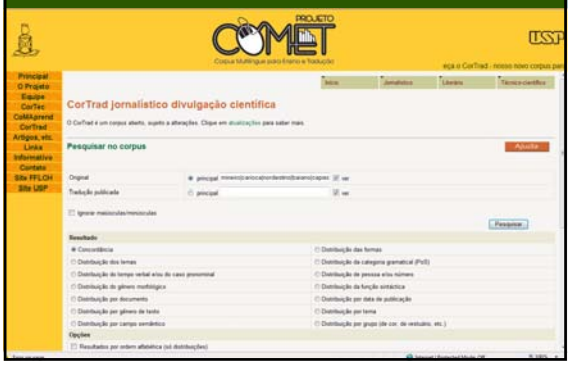
Resultado: "natural" ≠ "*natural*"



Traduções de “mineir.*” no corpus de culinária



Gentílicos brasileiros nas traduções do corpus jornalístico



Tradução de gentílicos brasileiros para o inglês

Próximos passos

- Revisão final do alinhamento
- Revisão da análise sintática
- Revisão da análise semântica
- Inserção de mais textos...

➔Em discussão:

– Anotação das diferenças entre as versões

Trabalho nos bastidores

- Recodificação das aspas simples e duplas no subcorpus literário
- Revisão e melhoria dos campos *assunto* e *gênero* no subcorpus jornalístico
- Criação automática das listas de conteúdo
- Finalização do conteúdo do site: documentação, ajuda, informação, etc.
- Observação dos diários (*logs*) e resposta aos usuários

Agradecimentos

- A Eckhard Bick e a Paul Rayson, pela autorização e pelo apoio prestado no uso do PALAVRAS e do CLAWS, respectivamente.
- A Sandra Aluísio e Arnaldo Candido Júnior do NILC pelo alojamento e apoio prestado
- Este trabalho foi parcialmente desenvolvido no âmbito da Linguateca, co-financiada pelo Governo Português, pela União Européia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

Obrigada!

Elisa (elisadut@usp.br)
Diana (Diana.Santos@sintef.no)
Stella (seotagni@usp.br)

➔ Temos cópias da lista de referência para quem se interessar