9 anos de desenvolvimento

Ana Frankenberg-Garcia

Linguateca - Fundação para a Computação Científica Nacional (FCCN)

Instituto Superior de Línguas e Administração (ISLA)

O que é o COMPARA?

Um corpus paralelo bidirecional de português e inglês

Um dos diversos recursos criados de raiz no âmbito da Linguateca

http://www.linguateca.pt/COMPARA/

Como surgiu o COMPARA?

- Diana 1996: doutoramento baseado num pequeno corpus paralelo EN-PT e PT-EN
- Ana 1999: utilização no ensino de tradução de um corpus paralelo EN-PT extraído do PE
- Diana 1999: Projecto PCP (primórdios da Linguateca)
- Ana 1999-2000: subsídio da FCT e ano sabático em Oxford para iniciar a construção de um corpus a sério.
- Ana e Diana 2000: início da parceria COMPARA

Preocupações iniciais

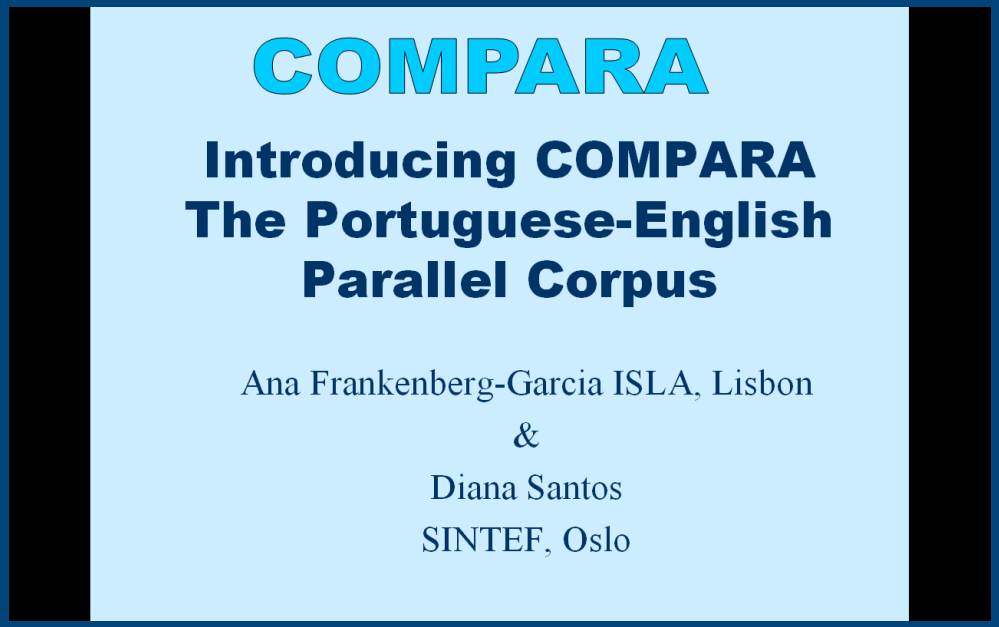
- Determinar estrutura do corpus
- Encontrar número suficiente de bitextos PT-EN
- Conseguir autorizações
- Criar regras de digitalização e alinhamento
- Escolher ferramentas de corpora apropriadas
- Criar interface pública PT e EN para conhecedores e leigos

Características básicas

- Estrutura bidirecional, só traduções diretas PT-EN e EN-PT e inicialmente só literatura publicada
- Disponibilização imediata
- Notas de tradução preservadas e anotação de palavras estrangeiras, títulos, ênfase, etc.
- Alinhamento direcional por frase do texto original com anotação detalhada
- Ferramentas IMS CWB
- Interface pública DISPARA

Lançamento do COMPARA

- Maio 2000: Primeiros testes em www.portugues.mct.pt/COMPARA/
- Novembro 2000: Primeira apresentação pública na CULT 2K, Itália



...e anúncio na Corpora

http://gandalf.aksis.uib.no/corpora/2001-1/0017.html

Corpora: COMPARA: the Portuguese-English parallel corpus

From: Diana Maria de Sousa Marques Pinto dos Santos (*Diana.Santos@informatics.sintef.no*)
Date: Mon Jan 08 2001 - 16:21:01 MET

- **Next message:** Mohamed Noamany: "Corpora: Arabic-English Parallel Corpora needed."
- **Previous message:** Diana Maria de Sousa Marques Pinto dos Santos: "Re: Corpora: Morfological ambiguity"
- **Messages sorted by:** [date] [thread] [subject] [author]

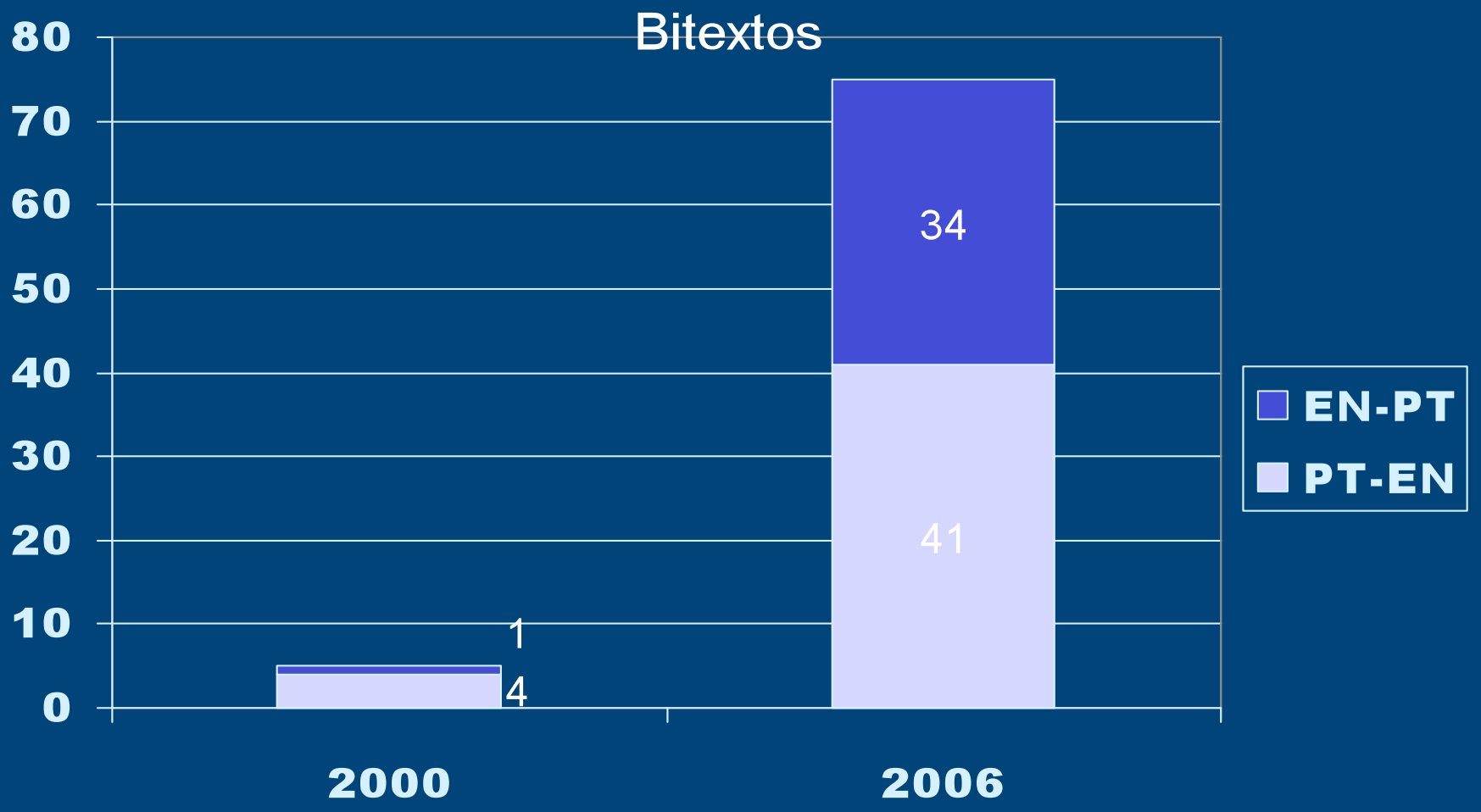
We are pleased to announce that COMPARA, the Portuguese-English parallel corpus, is now available at http://www.portugues.mct.pt/COMPARA/Welcome.html

COMPARA is open-ended, freely available on the Web, and made for people who have never used corpora before as well as for experienced corpus users. COMPARA's criteria for text alignment allow corpus users to investigate translational discourse changes such as when and where translators have chosen to join, separate, delete, add and reorder sentences. Users can also inspect translators' notes, and the corpus admits more than one translation per source text.

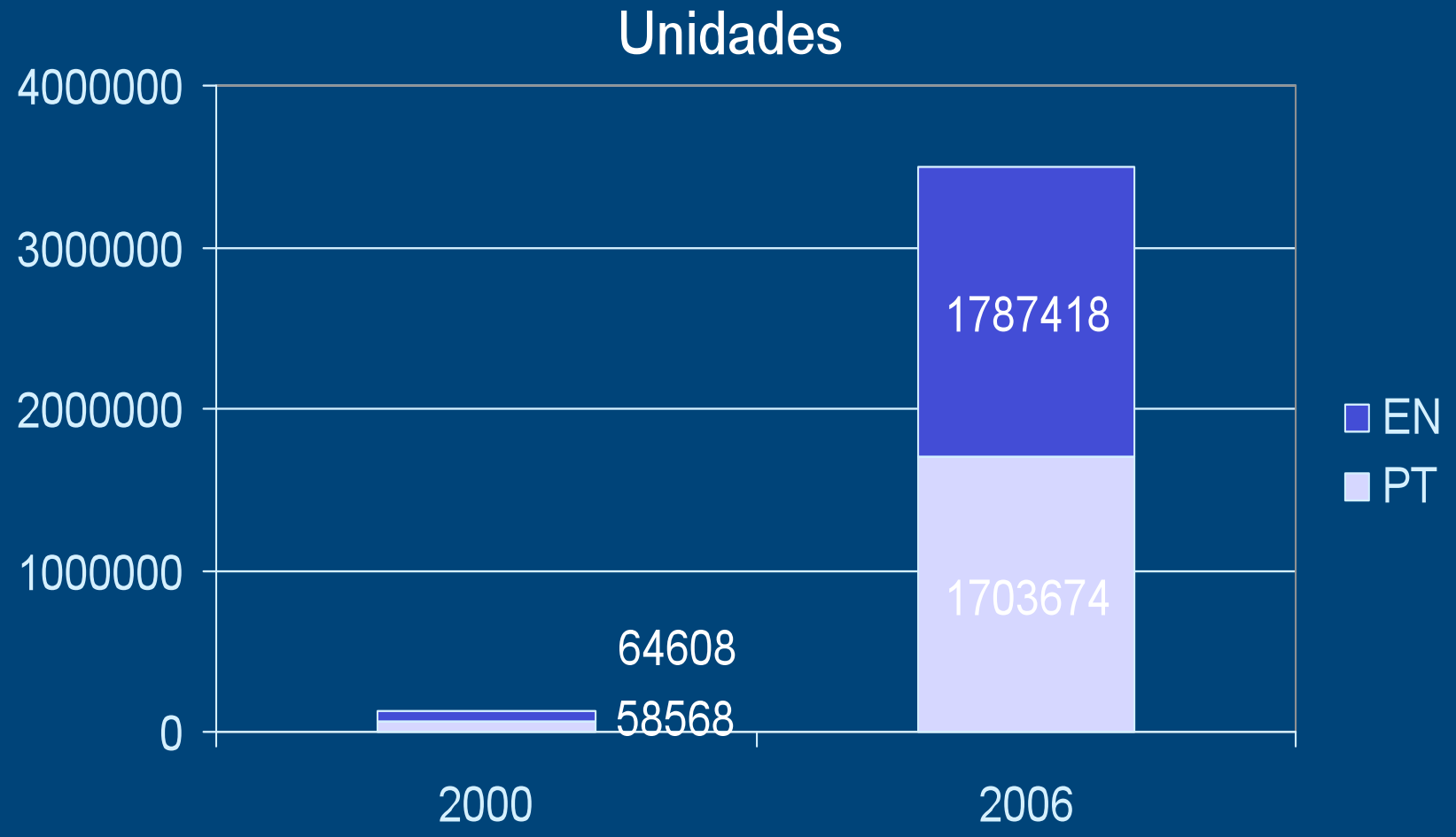
Only six parallel fiction texts have been fully processed so far, but permission has been obtained to incorporate many more.

Find: compara Match case Reached end of page, continued from top

Expansão do corpus 2000-2006



Expansão do corpus 2000-2006



Necessidade de mais mão de obra

- **Estagiários voluntários**
  - Sofia Sommer Ribeiro (ISLA, Lisboa), 2001 a 2002
  - Vera Almeida (ISLA, Lisboa), 2001 a 2002
  - Rosário Silva (ISLA, Lisboa), 2002 a 2003
  - Anne Silveira (Universidade de Melbourne), 2002 a 2003
  - Elisabete Ferreira (Universidade do Porto), seis meses em 2003
  - Susana Inácio (Universidade de Lisboa), seis meses em 2004
  - Rosa Pires (ISLA, Lisboa), 2002 a 2005
  - Cláudia Gomes (ISLA, Lisboa), 2004 a 2005
- **Criados tutoriais de processamento de textos paralelos**

Assistentes de investigação vinculados à Linguateca

- 2003: Rosário Silva
- 2004: Susana Inácio
- 2007: Pedro Sousa

- Formou-se uma equipa estável de 5 pessoas

Além da expansão, outras melhorias

- Interface em constante desenvolvimento tendo em vista novas funcionalidades e usabilidade (desde sempre)
- Ações de divulgação em palestras e workshops (desde sempre)
- Ajuda para pesquisar (desde 2003)
- Documentação detalhada sobre a construção do corpus disponibilizada (desde 2003)
- Tutorias de utilização do corpus (desde 2004)

COMPARA

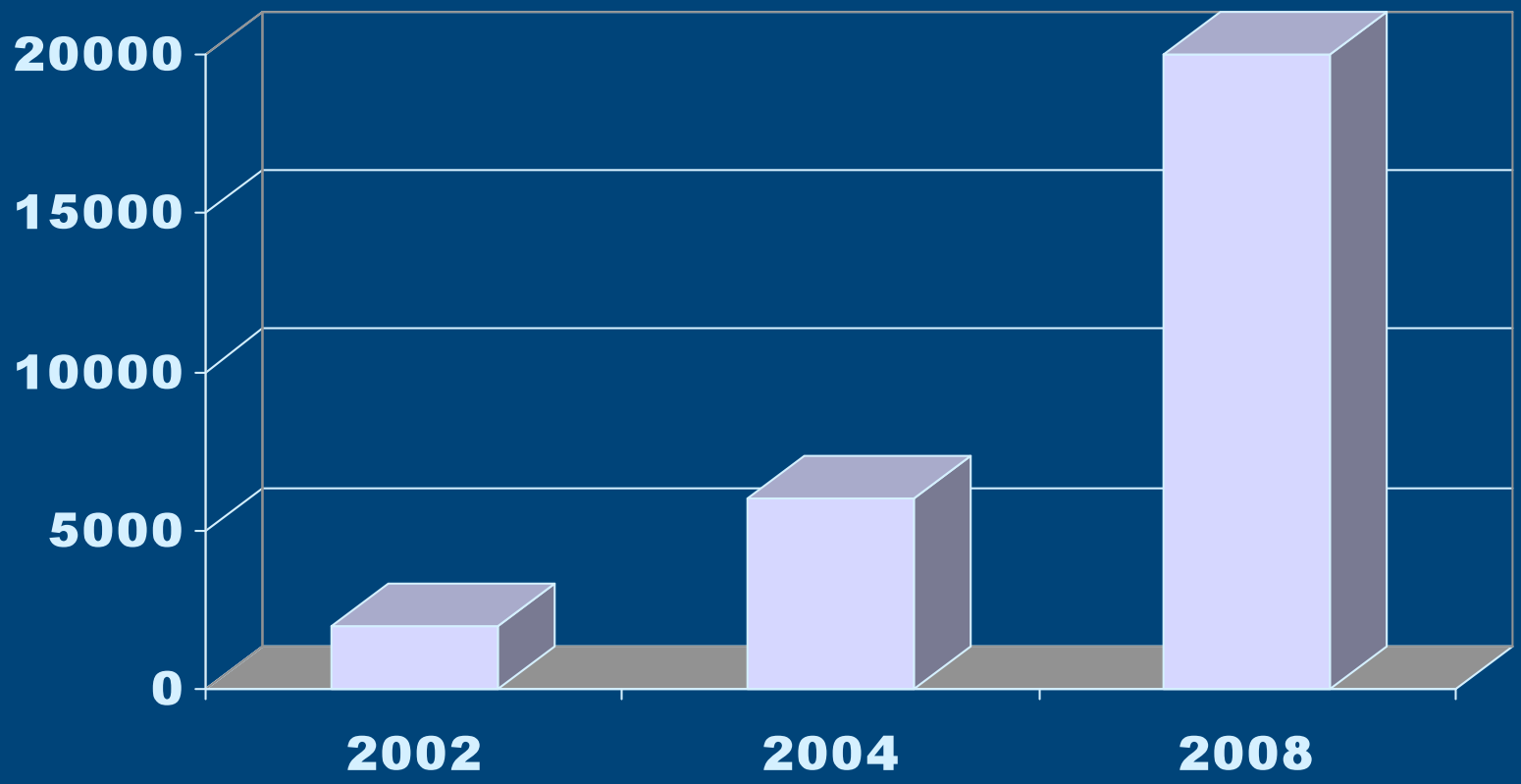
Melhorias continuam...

- Revisão completa da etiquetagem textual inicial (2005)
- Prospecção de bitextos não literários (2005)
- Anotação gramatical PT com o PALAVRAS (2004)
- Revisão manual da anotação PT (desde 2004)
- Re-introdução de marcas de parágrafo (2007)
- Revisão parcial da digitalização dos textos obtidos por download (2007)

E ainda...

- Anotação semântica para cor (2007)
- Interface totalmente remodelada (2008)
- Anotação gramatical EN com o CLAWS (2008)
- Revisão manual da anotação EN (desde 2008)
- Contatos para liberar textos protegidos (em curso)

Acessos mensais (picos)



Balanço

- Nenhum corpus é perfeito, mas o COMPARA é comparativamente um corpus muito cuidado (digitalização, alinhamento, anotação, interface, ajuda, documentação...)
- Em nove anos avançamos bastante mas tivemos falhas e ainda há muito por fazer...
- Mas o futuro é incerto

Obrigada

Coordenadoras do COMPARA
Ana Frankenberg-Garcia & Diana Santos

Assistentes de pesquisa
Pedro Sousa, Rosário Silva & Susana Inácio

Linguateca (Pólos SINTEF e FCCN)
www.linguateca.pt

Financiamento
Governo português e União Europeia (FEDER & FSE)
ref. POSC/339/1.3/C/NAC