# COMPARA, um corpus paralelo de português e inglês na Web

Ana Frankenberg-Garcia (ISLA, Lisboa) & Diana Santos (SINTEF, Oslo)

*COMPARA is an open-ended corpus of Portuguese and English language texts aligned with their respective English and Portuguese translations. The corpus was designed to be useful to translators, language students and teachers as well as to lexicographers, linguists and researchers working with natural language processing. COMPARA's alignment and encoding criteria allow users to analyse not only how words and expressions have been translated, but also translators' notes and differences in source and translation sentence structure. In addition to this, COMPARA accommodates more than one translation per source text. The corpus is encoded according to the IMS Corpus Workbench system and is searchable on the Web via the DISPARA interface. Access to COMPARA is free.*

*O corpus COMPARA é uma coleção aberta de textos eletrônicos de língua portuguesa e inglesa alinhados com as suas respectivas traduções para inglês e português. O corpus foi preparado de forma a que possa ser útil tanto para tradutores, alunos e professores de línguas, como para investigadores, lingüistas e engenheiros da linguagem. Os critérios de alinhamento e codificação do COMPARA permitem inspecionar não só a tradução de palavras e expressões, como também as notas de tradução e diferenças entre a estrutura frásica do original e da tradução. Além disso, o COMPARA permite comparar múltiplas traduções de um mesmo original. O corpus funciona através do sistema de processamento de corpora IMS Corpus Workbench e encontra-se disponível na Web através da interface DISPARA. O acesso ao COMPARA é gratuito.*

## Introdução

O COMPARA é um corpus eletrônico paralelo[1] cuja estrutura foi inspirada no ENPC (English-Norwegian Parallel Corpus, Johansson et al. 1999). Os textos que constituem o corpus são originais em língua inglesa e portuguesa e as suas traduções para português e inglês. O corpus é extensível (podendo conter um número ilimitado de textos), é de acesso gratuito através da Web, e foi desenvolvido para ser útil tanto para pessoas com pouca ou nenhuma experiência prévia na utilização de corpora, como para utilizadores experimentados.

Determinou-se que o corpus seria extensível por duas razões de ordem prática. Primeiro, porque isso permitiria torná-lo operacional logo de início e ainda com poucos textos, uma vez que não existia (e ainda não parece existir) nenhum outro corpus paralelo de livre acesso que inclua o português. Segundo, porque assim poderiam ser os próprios utilizadores do corpus a indicar o melhor caminho para a sua expansão. Esta escolha também acabou por facilitar a correção de alguns problemas inicias e a incorporação de novas funcionalidades, já que o corpus ainda era pequeno e havia menos alterações a fazer.

Quisemos que o público-alvo do COMPARA abrangesse todos os que estudam, investigam ou trabalham com o português e o inglês, entre os quais encontram-se falantes nativos de português a aprender inglês, falantes nativos de inglês a aprender português, professores e autores de materiais didáticos de português e inglês língua estrangeira, tradutores, professores e estudantes de tradução, lexicógrafos, engenheiros da linguagem, lingüistas interessados no estudo da tradução e investigadores na área da literatura comparada. Todos estes grupos são utilizadores potenciais do COMPARA. Tivemos, por isso, especial preocupação em assegurar que uma gama vasta de utilizadores pudesse facilmente servir-se do COMPARA, e que o corpus não se limitasse a ser útil apenas às pessoas já habituadas a trabalhar com corpora. O nosso objectivo foi desenvolver um sistema que não afastasse – ou assustasse – quem nunca tivesse lidado com um corpus informatizado.

O acesso ao COMPARA é gratuito e efetua-se, sem necessidade de registro prévio, através do endereço http://www.portugues.mct.pt/COMPARA/BemVindo.html. Este *site* é da responsabilidade do Projecto Processamento Computacional do Português[2].

As consultas ao COMPARA também podem ser feitas por pessoas que conhecem pouco o português, pois toda a informação necessária à utilização do corpus encontra-se em português e em inglês.

Seleção de textos

A escolha dos textos que constituem o corpus não obedeceu a nenhum critério de seleção em termos de variante lingüística. Considerou-se interessante incluir o português de todos os países de expressão portuguesa e o inglês de todos os países de expressão inglesa. Com isto, o corpus permite comparar não só originais com traduções, mas também traduções para variantes diferentes. Neste momento, por exemplo, é possível utilizar o COMPARA para realizar um estudo contrastivo da tradução portuguesa e da tradução brasileira do romance "Therapy", do autor inglês David Lodge. Além disso, se o utilizador assim o desejar, poderá também delimitar o corpus de modo a compor um sub-corpus contendo unicamente as variantes que lhe interessam.

O COMPARA admite tanto textos contemporâneos como textos antigos. Não houve qualquer restrição a nível de data de publicação. Isto possibilita comparar traduções do mesmo original publicadas em épocas diferentes. É o caso do romance "Iracema", de José de Alencar (1865), que no COMPARA se encontra alinhado com uma tradução americana publicada em 2000 e com uma tradução inglesa de 1886, permitindo assim um estudo diacrônico. Mais uma vez, se o utilizador assim o entender, poderá restringir o corpus de maneira a construir um sub-corpus que exclua os textos anteriores ou posteriores a um ano de publicação qualquer.

Optou-se por começar o corpus a partir de uma coleção de textos de ficção, embora, numa fase posterior, esteja prevista a introdução de outros gêneros. A decisão baseou-se numa série de considerações, entre as quais inclui-se o fato de estas traduções (e originais) terem sido publicadas, o que por si só garante uma certa qualidade lingüística dos textos. Além disso, existe um número razoável de obras de ficção em língua portuguesa com traduções publicadas para inglês, realidade que não se aplica a gêneros como, por exemplo, o jornalístico e o científico e acadêmico.

Direitos de autor

Para se utilizar textos num corpus é antes de mais nada necessário obter licenças de utilização junto dos detentores dos direitos de autor das obras pretendidas. Para a parte inicial do COMPARA, de textos de ficção, procurou-se obter principalmente autorização para a utilização de publicações em língua original portuguesa e suas respectivas traduções para inglês, visto que a quantidade e variedade destas obras é muito menor do que a de originais em língua inglesa com tradução para português.

A resposta global dos autores, tradutores e editores contactados foi bastante encorajadora, especialmente se considerarmos que deram autorização para os seus textos serem pesquisados gratuitamente através da Web.

Neste momento, o COMPARA dispõe de licenças para incluir extratos (tipicamente da ordem de 30% do tamanho total da obra[3]) de sessenta e um pares de textos, de autores e tradutores provenientes da África do Sul, Angola, Brasil, Estados Unidos da América, Moçambique, Portugal e Reino Unido, nos quais está representado o trabalho de 34 autores e 32 tradutores[4].

A constituição do corpus em Novembro de 2001

O projecto de criação do COMPARA iniciou-se em meados de Outubro de 1999. O número de textos totalmente processados ainda é reduzido, mas já é possível utilizar o COMPARA com resultados práticos quer em estudos contrastivos gramaticais, quer na obtenção semi-automática de material de apoio ao ensino de línguas e da tradução (cf. Frankenberg-Garcia, no prelo). Para análises lexicais, o COMPARA, neste momento, é mais limitado, pois inclui apenas a linguagem ficcional. Não é muito provável, portanto, encontrar no corpus léxico que não seja comum em textos de ficção, como termos técnicos, por exemplo[5]. A Tabela 1 resume o conteúdo do corpus à data da escrita do presente artigo. Mais detalhes acerca dos textos que o compõem encontram-se disponíveis em http://portugues.mct.pt/COMPARA/Conteudo.html.

Tabela 1: Conteúdo do COMPARA em Novembro de 2001

| COMPARA Novembro 2001 | Língua portuguesa | Língua Inglesa | Total |
|---|---|---|---|
| Originais | 7 | 3 | 10 |
| Traduções | 4 | 8 | 12 |
| Palavras | 187 093 | 193 548 | 380 641 |

Opções de codificação

O principal objectivo da codificação de texto adoptada pelo COMPARA é dar acesso à tradução de frases de português para inglês e vice-versa. Isto implica que o COMPARA permite investigar particularidades a nível da frase ou dos constituintes desta (palavras, orações, sintagmas, etc.). Não é possível utilizar o COMPARA para inspecionar, por exemplo, parágrafos ou capítulos (não porque nos parecesse desinteressante, mas devido ao que tal significaria em termos de direitos de autor). Estas divisões, portanto, não foram explicitamente marcadas no corpus. Também não se guardou informação sobre disposição gráfica, figuras, diagramas, numeração de páginas, etc., omitindo-a na digitalização. Em suma, não se tentou preservar os textos de uma maneira que permitisse recuperar a sua forma original, visto que o COMPARA não tem licença de redistribuição dos originais e traduções que o compõem. Também por essa razão, não foi necessário seguir as normas do TEI (Text Encoding Initiative) (Sperberg-McQueen & Burnard 1994) ou de qualquer outro padrão de codificação de corpora concebido para o intercâmbio de textos, embora algumas soluções do TEI tenham servido de inspiração para resolver certos problemas de codificação.

Alinhamento

A unidade básica de alinhamento adotada no COMPARA é definida pela frase do texto original. Quando a correspondência frásica entre o original e a tradução não é direta, optou-se por manter intacta a separação por frases do original e introduzir ajustes ao alinhamento apenas na tradução.

Assim, cada frase do texto de partida encontra-se alinhada com o texto correspondente na tradução, seja ele uma, mais do que uma ou apenas parte de uma frase. As frases não traduzidas encontram-se alinhadas com entidades vazias. Inversamente, as frases introduzidas pelo tradutor sem texto correspondente no original são incluídas na unidade de alinhamento imediatamente precedente e marcadas de maneira a se poder identificar que são frases adicionais. Por outro lado, se tiver havido um reordenamento de frases na tradução, o alinhamento segue as regras anteriores, desde que seja possível identificar as correspondências. A mudança na ordem é codificada separadamente. Por exemplo, caso a ordem das frases ABC no original tenha sido alterada para ACB na tradução, a frase A da tradução será alinhada com a frase A do original e a ordem das frases C e B da tradução será invertida de modo a que possam ser alinhadas com as frases B e C do original. A Tabela 2 resume esses critérios de alinhamento.

Tabela 2: Alinhamento por frase (F) do texto original

| ORIGINAL | | TRADUÇÃO |
|---|---|---|
| F | → | F |
| F | → | F,F |
| F | → | ½ F |
| F | → | ø |
| F | → | F(+F) |
| Fa | | Fa |
| Fb | → | Fc |
| Fc | | Fb |

Todos os tipos de alinhamento identificados acima estão codificados de modo a que o utilizador possa, se assim o escolher, recuperar esta informação sempre que fizer uma busca. Além disso, é possível pesquisar automaticamente todos os casos em que houve junção, separação, omissão, adição e reordenamento de frase na tradução. No entanto, refira-se que a marcação do tipo de alinhamento se restringe apenas ao nível da frase. Não é possível, dada a complexidade acrescida em termos de preparação de texto e de programação, inspecionar automaticamente a adição ou a omissão de constituintes da frase, tal como palavras ou orações.

Cabe realçar que os critérios de alinhamento adotados, baseados sempre na divisão frásica do texto original, simplifica o alinhamento de um mesmo original com várias traduções, e permite, indiretamente, a comparação entre dois (ou mais) textos traduzidos, usando como denominador comum o original de que ambos derivam.

Preparação do corpus: do texto impresso ao hipertexto alinhado

O procedimento para se preparar um texto para o seu funcionamento no COMPARA é o seguinte:

1. Os textos que não conseguimos obter em versão eletrônica são digitalizados através de um programa de reconhecimento óptico de caracteres (OCR).
2. A leitura óptica é revista, todo o material não textual é eliminado, e são introduzidas marcas de título, palavras ou expressões estrangeiras, e ênfase. As notas de tradução são, além disso, introduzidas no ponto onde a sua chamada ocorre.
3. Faz-se um alinhamento manual por parágrafos, do texto original e da tradução.

4. Um conjunto de programas desenvolvidos no âmbito do projecto AC/DC (Santos & Bick, 2000) identifica as unidades básicas (*tokens*) e faz a separação de frases de cada texto (original e tradução).
5. O original e a tradução são alinhados automaticamente pelo programa EasyAlign, que integra o IMS Corpus Workbench - o ambiente para processamento de corpora utilizado no COMPARA[6].
6. O alinhamento obtido através do EasyAlign passa por uma revisão manual de modo a só aceitar alinhamentos do tipo *uma frase do original para x frases na tradução*. Durante essa revisão inclui-se a marcação manual de todos os casos de adição, junção e reordenamento de frases na tradução, assim como os casos complexos de 1+1/x (por exemplo, uma frase do original alinhada com uma frase e meia da tradução).
7. Faz-se a marcação automática dos casos de omissão e separação de frases e uma primeira versão do par original-tradução é posta em funcionamento.
8. Utiliza-se a própria interface DISPARA para recuperar as unidades de alinhamento que contenham separação de frases na tradução, uma vez que a separação automática de frases, apesar de ser de grande utilidade, não é 100% fiável. A partir dos resultados obtidos, faz-se uma nova revisão manual do alinhamento de modo a discriminar os casos em que houve realmente separação de frases dos casos em que não houve de fato separação[7].
9. Uma versão revista do par de textos é posta em funcionamento.

## Buscas no COMPARA: a interface DISPARA

O COMPARA é acedido através da interface DISPARA[8], desenvolvida para fazer de ponte entre o IMS Corpus Workbench e as especificidades próprias do COMPARA. Embora esta interface tenha sido originalmente criada para o COMPARA, é também facilmente adaptável a outros corpora paralelos codificados no sistema do IMS Corpus Workbench.

A interface DISPARA dá acesso a duas opções de busca no COMPARA. A BuscaSimples, feita para pessoas com pouca ou nenhuma experiência na utilização de corpora, permite procurar palavras ou seqüências de palavras em português (ou em inglês) em todos os textos do corpus e ver como estas palavras ou seqüências foram traduzidas para inglês (ou português). As instruções para se usar a BuscaSimples são elementares: os utilizadores precisam apenas escrever uma palavra ou expressão em inglês (ou português) e acionar o botão de procura (cf. http://www.portugues.mct.pt/COMPARA/BuscaSimples.html).

A BuscaComplexa destina-se a procuras mais sofisticadas. Ainda assim, procurou-se criar uma interface de fácil utilização, de modo a que mesmo quem nunca tenha usado um corpus paralelo se sinta capaz de explorar as suas potencialidades. De momento, a BuscaComplexa compreende quatro passos (cf. http://www.portugues.mct.pt/COMPARA/BuscaComplexa.html):

1º Passo
Neste passo os utilizadores devem escolher a direção de procura. Além de poderem escolher fazer uma busca de português para inglês ou de inglês para português, como na BuscaSimples, podem escolher procurar apenas de original para tradução ou só de tradução para original. Tal restrição é obviamente relevante nos casos em que a direcionalidade da tradução interessa.

2º Passo
Enquanto na BuscaSimples os resultados de uma pesquisa baseiam-se sempre na totalidade dos textos do corpus, este passo da BuscaComplexa serve justamente para se fazer uma pré-seleção dos textos que se pretende utilizar. Esta opção é obviamente importante visto que o COMPARA é um corpus aberto, e pode conter textos que não interessam a todos os utilizadores. Há três maneiras de se pré-selecionar textos:

1. Pode-se escolher automaticamente a variante ou combinação de variantes do português e do inglês que se pretende utilizar, excluindo qualquer variante indesejada. Por exemplo, é possível utilizar apenas português do Brasil e inglês britânico, ou apenas textos em inglês sul-africano e todas as variantes do português[9].

2. É também possível selecionar os textos por ano de publicação do original e/ou da tradução. Os utilizadores interessados apenas na linguagem de textos recentes podem omitir automaticamente os originais e as traduções anteriores a uma determinada data. Os utilizadores interessados apenas em textos mais antigos poderão igualmente excluir as obras mais recentes, posteriores a um ano de publicação qualquer.

3. A terceira opção de escolha é a mais fina, permitindo ao utilizador escolher qualquer combinação de pares de textos. Assim, é possível selecionar um sub-corpus apenas com textos de um mesmo autor, ou mesmo tradutor, ou com mais de uma tradução, etc.

Quando o corpus contiver outros gêneros além da ficção, será também possível fazer uma pré-seleção automática por gênero lingüístico.

3º Passo
Enquanto na BuscaSimples os resultados são sempre apresentados na forma de concordâncias, neste passo da BuscaComplexa o utilizador pode indicar que tipo de resultado pretende (ou combinação destes). Além de concordâncias, é possível obter a distribuição das formas presentes no corpus (por exemplo, para uma busca que inclua *for instance* e *for example*, saber quantas vezes cada uma das expressões aparece), a distribuição das fontes (em que textos é que foram encontradas) e, no caso de a pergunta ter sido formulada com um lado no texto original e outro na tradução, um resumo quantitativo (a distribuição dos resultados nas duas línguas).

4º Passo
Finalmente, a expressão de procura é digitada. Pode-se digitar uma simples palavra ou uma seqüência de palavras delimitadas individualmente por aspas, ou, através da sintaxe do IMS Corpus Workbench, pode-se efetuar buscas mais sofisticadas. É possível, por exemplo, procurar numa só expressão formas ortográficas diferentes (ex. *acto* e *ato*), diversidade morfológica de uma palavra (ex. as formas do verbo *imaginar*), todas as palavras começadas por uma seqüência de letras qualquer (ex. as letras que compõem o prefixo de negação inglês *un*), duas palavras com uma ou várias palavras indeterminadas entre elas (ex. *um* {*bom? mau? difícil?...*} *começo*), etc.[10].

Uma potencialidade que torna o sistema ainda mais interessante, na nossa opinião, é a hipótese de restringir a procura no corpus também pelo lado da tradução, ou seja, é possível procurar em paralelo nas duas línguas, e obter apenas os casos em que, por exemplo, *even* é traduzido por *até*, ou então somente os casos em que *even* não é traduzido por *até*.

Enquanto as facilidades acima provêm diretamente do IMS Corpus Workbench, há outras que foram implementadas de raiz no DISPARA, que permite observar o tipo de alinhamento, inspecionar notas de tradução e recuperar títulos, marcas de ênfase e palavras estrangeiras associados a cada expressão de busca. Através do DISPARA também é possível examinar estes fatores independentemente de qualquer expressão de busca, bastando, para tal, deixar a janela da expressão de busca vazia. Por fim, devido à manutenção da informação sobre o tipo de alinhamento, é possível recuperar todas as frases do original que tenham sido divididas, ou unidas, ou omitidas, ou reordenadas na tradução, assim como todas as frases da tradução que não constem do original.

## Apresentação dos resultados

As licenças de utilização dos textos que compõem o COMPARA permitem o uso dos resultados das pesquisas para efeitos de investigação e de ensino.

No entanto, para salvaguardar os direitos de autor, o número máximo de concordâncias mostradas cada vez que se faz uma busca é 500. Se o utilizador optar por não utilizar a totalidade do corpus, mas apenas uma parte reduzida dele, o número máximo de concordâncias que se pode apresentar passa a ser 200, de modo a que nenhum texto apareça na íntegra. Quando os resultados excedem estes limiares, mostra-se uma amostra aleatória de 500 (ou 200) concordâncias, embora se indique sempre o número total de instâncias encontradas. O utilizador receberá, nesses casos, uma mensagem informando que, para a proteção dos direitos de autor, apenas 500 (ou 200) casos serão facultados dentre os x>500 (ou x>200) encontrados.

As concordâncias são apresentadas em duas colunas verticais, com o texto português (ou inglês) procurado pelo utilizador a negrito do lado esquerdo, e o texto correspondente em inglês (ou português) do lado direito. Em vez de a concordância ser definida em termos de um número fixo de caracteres para a esquerda e para a direita, o utilizador vê sempre uma frase completa do texto original, alinhada com o texto correspondente na tradução (cf. apêndice).

Associado a cada concordância é apresentado um identificador que aponta para a descrição do par de textos e o número da unidade de alinhamento em questão. Seguindo o atalho, o utilizador tem acesso à referência bibliográfica completa dos textos em causa e a informação sobre direitos de autor, variante lingüística e dados quantitativos sobre o tamanho do extrato, em palavras e em unidades de alinhamento.

É possível navegar pelo resultado de forma a ver todas as concordâncias, assim como gravar os resultados em HTML, texto ou mesmo simplesmente cortar e colar para qualquer processador de texto, caso se queira aproveitar os resultados para fins didáticos (cf. Frankenberg-Garcia, 2000, Frankenberg-Garcia *no prelo*) ou para efeitos de investigação.

## Conclusão

O corpus COMPARA, apesar de se encontrar ainda numa fase inicial, tem a sua estrutura básica definida e, graças à interface DISPARA, já pode ser utilizado com finalidades práticas diversas. Ainda existe uma lista de obras para as quais já se obteve autorização de inclusão no corpus, mas que estão à espera de ser processadas. Há igualmente planos para expandir o corpus para outros gêneros de texto, bem como para aprimorar a interface DISPARA, de modo a melhorar a sua funcionalidade ou simplesmente torná-la mais fácil de usar. Estamos

convencidas que este é apenas o princípio, e que o COMPARA poderá vir a beneficiar uma vasta comunidade de pessoas interessadas na tradução do par português-inglês. Acreditamos que algumas das opções tomadas na concepção do corpus COMPARA e da interface DISPARA foram inovadoras, e esperamos que venham a contribuir para o avanço do campo mais largo do processamento e estudo de corpora paralelos como sub-disciplina da lingüística de corpora.

---

Notas

[1]Por *corpus paralelo* entende-se aqui uma coleção bilíngüe de textos alinhados com as suas traduções, chamado *corpus de traduções* na tradição da lingüística contrastiva. Johansson (1998) sugeriu que o progresso da área levaria à resolução deste conflito terminológico, o que parece não ter ainda acontecido – veja-se, também, a esse propósito, a introdução de Véronis (2000) para um breve enfoque histórico do conceito.

[2] Este projecto (http://www.portugues.mct.pt/), financiado pelo Ministério da Ciência e da Tecnologia de Portugal, tem como principais actividades a catalogação, a criação e disponibilização de recursos de língua portuguesa na Web, e a avaliação do processamento computacional da língua portuguesa. Veja-se Santos (2000) e Veiga & Santos (2001) para uma panorâmica das várias actividades do projeto.

[3] Ao contrário do ENPC (English-Norwegian Parallel Corpus), os extractos do COMPARA não são todos do mesmo tamanho nem são sistematicamente retirados do início da obra, devido à suspeita, referida em Santos e Oskefjell (1999), que essa opção restrinja de forma arbitrária o tipo de texto.

[4] A lista de autorizações do COMPARA é constantemente atualizada em http://www.portugues.mct.pt/COMPARA/Conteudo.html

[5] Cabe aqui lembrar que os estudos lexicais exigem corpora bem maiores que os estudos gramaticais (Biber et al, 1998).

[6] Consideramos o IMS Corpus Workbench (Christ, 1994; Christ et al. 1999) o sistema de corpora que melhor se adapta às necessidades do projecto Processamento Computacional do Português, no contexto do qual a interface DISPARA foi desenvolvida. Mais detalhes sobre esta motivação encontram-se descritos em Santos (1998), e Santos & Ranchhod (1999).

[7] As falhas ocorrem especialmente nos casos de discurso direto. Isso porque o programa de separação automática de frases interpreta o ponto de exclamação ou de interrogação seguido de uma palavra iniciada com letra maiúscula como sendo uma fronteira entre duas frases. Assim, enquanto o programa separa corretamente seqüências como *What a surprise! I love you!* ou *Quem chegou? Foi a Maria?,* acaba também por separar indevidamente seqüências como *"What a surprise!" I said.* e -*Quem chegou? Maria perguntou.* São precisamente estes os casos que requerem uma revisão manual.

[8] DISPARA é um sistema genérico de DIStribuição de corpora PARAlelos na Web.

[9] Qualquer combinação é possível, embora nem todas existam no corpus (por exemplo, não existe no corpus nenhum texto em inglês americano traduzido para português de Angola).

[10] Existe, no próprio formulário da BuscaComplexa, uma ligação direta para o manual do IMS Corpus Workbench, que fornece uma descrição detalhada das opções existentes. A sintaxe do IMS Corpus Workbench tem um poder expressivo muito elevado, mas a sua utilização requer um certo treino. Para facilitar a tarefa, está prevista a criação de um manual do utilizador do COMPARA, com exemplos relevantes para o par inglês-português.

Referências

Biber, Douglas, S. Conrad. & R. Reppen (1998) Corpus Linguistics: investigating language structure and use. Cambridge: Cambridge University Press.

Christ, Oliver, B. Schulze, A. Hofmann & E. Koenig (1999) "The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual", Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2) http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

Frankenberg-Garcia, Ana (2000) "Using a Translation Corpus to Teach English to Native Speakers of Portuguese". Op. Cit. - A Journal of Anglo-American Studies, 3, 65-78.

Frankenberg-Garcia, Ana (no prelo) "COMPARA, language learning and translation training" Actas da conferência Training the Language Service Provider for the New Millennium, Faculdade de Letras, Universidade do Porto, 25-26 Maio 2001.

Johansson, Stig (1998) "On the role of corpora in cross-lingüistic research" in Stig Johansson & Signe Oksefjell (eds) Corpora and crosslingüistic research: theory, method and case studies, Amsterdam: Rodopi, pp 3-24.

Johansson, Stig , J. Ebeling & S. Oksefjell (1999) English-Norwegian Parallel Corpus: Manual http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html [acedido 7/7/2000]

Santos, Diana (1998) "Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts", in A.Rubio, N.Gallardo, R.Castro & A.Tejada (eds.) Proceedings of The First International Conference on Language Resources and Evaluation, 1, 475-481.

Santos, Diana & S. Oksefjell (1999) "Using a Parallel Corpus to Validate Independent Claims", Languages in Contrast, 2:1, 117-132.

Santos, Diana & E. Ranchhod (1999) "Ambientes de processamento de corpora em português: Comparação entre dois sistemas", in Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR (Évora, 20-21 Setembro de 1999), 257-268.

Santos, Diana & E. Bick (2000) "Providing Internet access to Portuguese corpora: the AC/DC project", in Gavriladou M., G. Carayannis, S. Markantonatou, S.Piperidis & G. Stainhaouer (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000, 205-210.

Santos, Diana (2000) "O projecto Processamento Computacional do Português: Balanço e perspectivas", in M. Graça Nunes (ed) Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), 105-113.

Sperberg-McQueen, C. & Burnard, L. (eds.) (1994) "Guidelines for Electronic Text Encoding and Interchange" TEI P3. Association for Computers and Humanities/ Association for Computational Lingüistics/ Association for Literary and Lingüistic Computing. Chicago & Oxford.

Veiga, Pedro & Santos, D. (2001) "Contributo para o processamento computacional do português: o CRdLP", in Maria Helena Mira Mateus (ed.), Mais Línguas, Mais Europa: celebrar a diversidade linguística e cultural da Europa (Actas do colóquio de 25 a 26 de Janeiro de 2001), Lisboa: Edições Colibri, 103-109.

Véronis, Jean (ed.) (2000) Parallel Text Processing, Dordrecht: Kluwer Academic Publishers.

# Apêndice

## Resultados da pesquisa

Os resultados das buscas efectuadas no COMPARA podem ser usados para fins educacionais e investigação, desde que se mencione a fonte. Para citar textos específicos do corpus, seleccione o código azul ao lado de cada concordância de modo a obter a sua referência completa. Para se referir ao corpus como um todo, cite: **COMPARA http://www.portugues.mct.pt/COMPARA/** [25-Novembro-2001]

Procura: **"saudade(s)?"**.
Pedido de : concordância em contexto.
Corpus: COMPARA_PORT

11 ocorrências.

| | | |
|---|---|---|
| PBJA1T1(20): | Enquanto vogas assim à discrição do vento, airoso barco, volva às brancas areias a **saudade**, que te acompanha, mas não se parte da terra onde revoa. | While thou sailest thus at the mercy of the wind, graceful craft, let longing, which accompanies thee but does not depart from the land, return to the white sands where it soars. |
| PBJA1T1(185): | Foi a lembrança da pátria que trouxe a **saudade** ao coração pressago. | It was the memory of my homeland that brought longing to my foreboding heart.» |
| PBJA1T2(20): | Enquanto vogas assim à discrição do vento, airoso barco, volva às brancas areias a **saudade**, que te acompanha, mas não se parte da terra onde revoa. | But whilst thou sailest thus at the mercy of the winds, graceful barque, waft back to that white beach some of the yearning that accompanies thee, but which may not leave the land to which it returns. |
| PBJA1T2(185): | Foi a lembrança da pátria que trouxe a **saudade** ao coração pressago. | It was the memory of my native land that brought a saudade to my anxious soul.» |
| PBMA2(121): | Nenhuma água de Juventa igualaria ali a simples **saudade**. | No water from Iuventus could match simple nostalgia in that. |
| PBMA2(426): | Você sabe que eu morrerei também... que digo?... morro todos os dias, de paixão, de **saudades**... | You know that I would die, too... What am I saying?... I die every day, from passion, from longing...» |
| PBMA3(536): | Pádua começou s falar da administração interina, não somente sem as **saudades** dos honorários, nem o vexame da perda, mas até com desvanecimento e orgulho. | Pádua began to talk about the temporary directorship, not only with no regrets for the lost honoraria, no shame at having lost the job, but even with a certain conceit and pride. |
| PPEQ1(292): | Às vezes vinha-me como uma **saudade** dos meus tempos ocupados da repartição. | Sometimes I felt almost nostalgic for the days when I was busy at the office. |
| PPSC1(25): | A uma criatura como aquela não se podia ter afecto, embora no fundo ele fosse um excelente rapaz: mas ainda hoje evoco com **saudade** as nossas palestras, as nossas noites de café – e chego a convencer-me que, sim, realmente, o destino de Gervásio Vila-Nova foi o mais belo: e ele um grande, um genial artista. | It was impossible to feel affection for someone like that (although deep down he was an excellent fellow) , and yet even today I recall with nostalgia the talks we had, the nights spent in cafés and I can even convince myself that, yes, the fate of Gervásio Vila-Nova really was the most beautiful of fates and that he was a great artist, an artist of genius. |
| PPSC1(277): | Mas o que as fazia mais excitantes era a **saudade** límpida que lembravam de um grande lago azul de água cristalina onde, uma noite de luar, elas se mergulhassem descalças e amorosas. | But what made the dancers so exciting was the limpid nostalgia they evoked for a great blue lake of crystalline water where, on moonlit nights, they would plunge in, barefoot and tender. |
| PPSC1(529): | E a minha **saudade** foi então a mesma que se tem pelo corpo de uma amante perdida... | And my longing then was exactly the same longing you would feel for the body of a lost lover... |

Esperamos que o COMPARA lhe tenha sido útil!


| | | | |
|---|---|---|---|
| Pesquisar novamente | Sobre o COMPARA | Constituição do corpus | Agradecimentos |


*Comentários para compara@informatics.sintef.no*