Compilação e Uso de Corpora Paralelos

Ana Frankenberg-Garcia[*]

Este trabalho procurará, inicialmente, dar a conhecer as especificidades de corpora paralelos e discutir os processos decisórios por trás da compilação de um corpus deste gênero. A seguir, será mostrado o caso de um corpus paralelo em particular - o COMPARA. Por fim, à guisa de exemplo, serão apresentados alguns trabalhos no âmbito da lexicografia bilíngüe e dos estudos de tradução realizados com base em corpora paralelos.

Palavras-chave: corpora, corpus paralelo, lexicografia bilíngüe, tradução, português, inglês

Introdução

Um corpus é basicamente uma coleção extensa de textos naturais, selecionados de acordo com critérios específicos e armazenados em formato digital. Um corpus paralelo é, por sua vez, uma combinação de pelo menos dois sub-corpora alinhados entre si. Na sua acepção mais simples, podemos ter, de um lado, um sub-corpus composto de textos originais numa determinada língua (L1) e, do outro, um sub-corpus com os mesmos textos traduzidos para uma outra língua (L2). Os dois são então alinhados para que se possa extrair concordâncias paralelas, tornando possível pesquisar originais e traduções em simultâneo.

[*] Instituto Superior de Línguas e Administração (ISLA) e Fundação para a Computação Científica Nacional (FCCN), Lisboa.

Além de tudo o que um corpus monolíngüe faz, um corpus paralelo permite ainda uma série de outras análises que seriam impossíveis numa estrutura não paralela. É preciso ter em conta, porém, que a sociedade só traduz uma ínfima parte daquilo que se diz e escreve numa determinada língua, limitando seriamente o número de textos paralelos ou bi-textos a partir dos quais é possível coligir um corpus paralelo. Os corpora paralelos são, por isso, em geral bastante menores do que os corpora monolíngües. Convém levar este fato em consideração quando se utiliza um corpus em pesquisas monolíngües que dispensam a estrutura paralela. Nestes casos, um corpus não paralelo de maior dimensão e variedade de textos pode ser mais adequado. Ou seja, a grande vantagem de um corpus paralelo reside justamente no aproveitamento da sua disposição em paralelo, sobre qual falarei a seguir.

Em termos estruturais, existem corpora paralelos unidirecionais, bidirecionais ou mistos. A configuração unidirecional permite analisar textos traduzidos de L1 para L2, conforme indica a figura 1. A estrutura bidirecional, por sua vez, tem como base textos originais em duas línguas diferentes (L1 e L2) alinhados com as suas traduções recíprocas para L2 e L1. Assim sendo, a estrutura bidirecional permite analisar traduções de L1 para L2 e de L2 para L1, conforme mostra a figura 2. Os corpora mistos, por sua vez, combinam segmentos unidirecionais e bidirecionais.[1]

O alinhamento L1-L2 dos corpora paralelos unidirecionais faz com que estes possam ser utilizados diretamente como dicionários e gramáticas bilíngües ou para aprimorar

[1] No caso dos corpora paralelos multilíngües, a relação entre os diversos pares de línguas que os integram também pode ser uni ou bidirecional.

Figura 1 - Estrutura de um corpus paralelo unidirecional

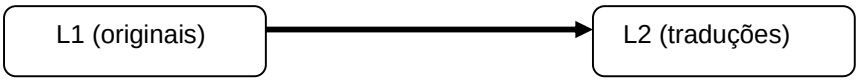
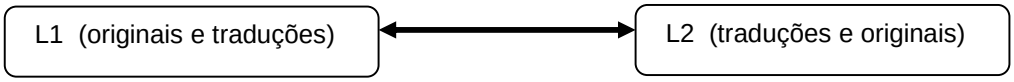


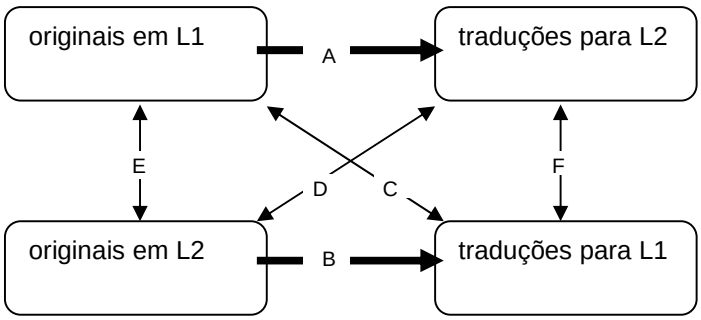
Figura 2- Estrutura de um corpus paralelo bidirecional



dicionários bilíngües já existentes ou ainda para desenvolver léxicos computacionais para serem usados na tradução automática e na tradução assistida por computador. Já o alinhamento L1-L2 acrescido do alinhamento L2-L1 dos corpora paralelos bidirecionais dá acesso a um leque muito mais amplo de opções.

Conforme mostra a figura 3, a estrutura bidirecional permite, em primeiro lugar, desenvolver estudos bilíngües não só de L1 para L2 como também de L2 para L1 (setas A e B).

Figura 3 - Análises possíveis num corpus paralelo bidirecional



A bidirecionalidade é importante na medida em que as correspondências tradutórias entre duas línguas nem sempre são biunívocas. Ou seja, a tradução de X para Y não garante que a tradução de Y seja sempre X. Desta forma, não basta inverter um dicionário de L1 para L2 para se desenvolver um léxico bilíngüe de L2 para L1: é preciso analisar as correspondências lingüísticas em cada direção separadamente.

Se dispensarmos o alinhamento, a estrutura paralela de um corpus bidirecional também é passível de ser aproveitada nos estudos comparativos entre textos numa língua em versão original e textos na mesma língua em versão traduzida (setas C e D). As diferenças observadas entre estes sub-corpora comparáveis podem ajudar a identificar características próprias da língua traduzida e, eventualmente, revelar alguns traços do fenômeno designado de "tradutês", que ocorre devido à influência da língua do original no texto da tradução.

Note-se que é igualmente possível realizar estudos comparativos em que o plano indicado pela seta C atua como controle do plano representado pela seta D (ou vice-versa). Quando as mesmas transformações de língua original para língua traduzida são observáveis tanto em C como em D, poderemos estar mediante evidência empírica de alguns dos universais da tradução.

Já os sub-corpora ligados pela seta E podem ser aproveitados como corpora comparáveis bilíngües, em estudos de terminologia e lingüística contrastiva pura, onde ficam de fora

as eventuais influências do processo tradutório. Por fim, os sub-corpora sinalizados pela seta F representam só traduções, servindo de base para análises do texto traduzido independentes do texto de origem.

Compilação de corpora paralelos

A compilação de um corpus paralelo passa por uma série de processos decisórios. Numa primeira etapa, é necessário determinar quantas e quais línguas serão representadas no corpus e decidir se a estrutura do corpus será uni ou bidirecional. A seguir, temos de optar pelos textos que iremos incluir, levando em conta variáveis tais como gênero (textos literários, científicos, técnicos, etc.), modo (oral ou escrito), estilo (culto ou popular) e época (textos antigos ou contemporâneos). Ainda no plano da seleção de textos, temos de especificar que tipo de traduções queremos (traduções profissionais? informais? publicadas? feitas por tradutores diferentes?).

É importante notar que a conjugação dos fatores acima não é simples e fica sempre condicionada pelas traduções disponíveis. Ao selecionarmos as línguas a serem representadas no corpus, temos de ter em mente que não existem bi-textos em número suficiente para todos os pares de línguas. Por exemplo, há muito mais opções para o par português-inglês do que para português-húngaro. Também não nos devemos esquecer de que poderá não haver traduções para todos os gêneros, estilos, modos e épocas e, mesmo que haja, nem sempre existem traduções bidirecionais. Por exemplo, existe informação turística traduzida de português para inglês em abundância, mas há muito pouco deste gênero traduzido de inglês para português.

Ainda nesta primeira etapa, é necessário decidir se o corpus será para uso privado ou público. Um corpus público é sempre mais trabalhoso, porque carece de autorizações para textos protegidos por direitos de autor. No entanto, um corpus público não só pode ser aproveitado por muito mais pessoas, como permite a realização de estudos passíveis de verificação por terceiros, o que raramente ocorre quando o corpus é particular.

Num corpus paralelo, para cada bi-texto necessitamos sempre de autorizações duplas. De nada vale obter autorização para utilizar um original se não conseguirmos permissão para incluir a sua tradução no corpus. Mesmo se utilizarmos originais que estejam no domínio público[2], convém lembrar que as suas traduções podem ainda estar protegidas.

Para se obter autorização para utilizar um texto num corpus, é preciso, em primeiro lugar, identificar quem, entre autores, tradutores, editoras e eventuais herdeiros, detém os direitos sobre o texto. Feito isto, a grande dificuldade reside em explicar a estas pessoas o que é um corpus. A maioria delas desconhece o tipo de uso que se pretende fazer do texto e teme a sua utilização indevida. O sucesso na obtenção das autorizações dependerá, em grande medida, do esclarecimento de que não será dado acesso a textos completos, mas somente a concordâncias e listas de frequências para análise lingüística. Este acesso parcial pode eventualmente até ser uma maneira de despertar o interesse dos utilizadores do corpus pela aquisição da obra completa.

[2] Encontram-se no domínio público as obras de autores falecidos há setenta anos ou mais.

Numa segunda etapa, a compilação de um corpus paralelo envolve decisões acerca do seu alinhamento, ou seja, é preciso determinar que nível de correspondências iremos estabelecer entre originais e traduções. As obras podem ser alinhadas por texto, parágrafo, frase, oração e até por palavra. Quanto mais fino o alinhamento, mais complexa será a sua automatização. O grau de alinhamento escolhido irá naturalmente depender do uso posterior que se fará do corpus, sendo que a maioria dos corpora paralelos existentes encontra-se alinhada por frase.

Quanto à etiquetagem, convém deixar claro que as únicas etiquetas obrigatórias nos corpora paralelos são as de alinhamento. É a partir destas etiquetas que o processador de corpora extrai concordâncias paralelas. Como em qualquer outro corpus, todas as outras etiquetas são opcionais. Nesta fase, é necessário refletir sobre quais processos queremos automatizar no corpus, pois é muito mais simples automatizar aquilo que estiver etiquetado. Por exemplo, se quisermos procurar notas de tradução de maneira automática, convém etiquetá-las previamente; se quisermos ter a possibilidade de fazer pesquisas gramaticais, precisaremos de inserir etiquetas de anotação morfossintática.

Tomadas todas as decisões relevantes para definir o tipo e estrutura do corpus, passa-se à sua montagem propriamente dita. Se o corpus for público, convém começar pelos pedidos de autorização. De nada valerá o trabalho braçal de digitalização, limpeza, etiquetagem e alinhamento de textos se a autorização dos mesmos não estiver garantida.

Por fim, será necessário optar por uma ferramenta de processamento de corpora paralelos e adaptar o corpus às suas especificidades, lembrando sempre que é relativamente simples adaptar um corpus de uma ferramenta a outra, para que este possa ser utilizado com programas diferentes[3].

Na próxima seção, à guisa de exemplo, serão discutidas algumas das opções tomadas pelo corpus paralelo COMPARA (Frankenberg-Garcia e Santos 2001, 2003). Mais informações sobre este e outros corpora paralelos públicos podem ser obtidas através dos seus respectivos *sites,* fornecidos na figura 4.

Figura 4 - Alguns corpora paralelos acessíveis em rede

COMPARA (português-inglês; inglês-português) http://www.linguateca.pt/COMPARA/
Multisemcor (italiano-inglês, romeno-inglês) http://multisemcor.itc.it/frameset2.php
OPUS, EUROPARL (11 línguas usadas na União Européia) http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=pt
Parte pública do OSLO MULTILINGUAL CORPUS (Vários mini-corpora com diferentes combinações de línguas, incluindo alemão, francês, espanhol, inglês e norueguês) http://khnt.hit.uib.no/webtce.htm
HUNGLISH CORPUS (inglês-húngaro) http://szotar.mokk.bme.hu/hunglish/search/corpus
CORPUS PARALELO CLUVI (Vários mini-corpora com diferentes combinações de línguas, incluindo galego, espanhol, catalão, basco, português, inglês) http://sli.uvigo.es/CLUVI/

[3] Existem processadores de corpora, tais como o Multiconcord (Wools 2000) e o ParaConc (Barlow 2002), concebidos de raiz para corpora paralelos. Também é comum adaptar à estrutura paralela uma ferramenta de processamento de corpora como o IMS-CWB (Christ et al. 1999). A interface DISPARA, concebida por Santos (2002), é um exemplo.

Opções do corpus COMPARA[4]

Estrutura

O COMPARA é um corpus paralelo e bidirecional de português e inglês. A sua estrutura bidirecional permite comparar: (a) originais em português com suas traduções para inglês; (b) originais em inglês com as suas traduções para português; (c) português original com português traduzido; (d) inglês original com inglês traduzido; (e) português não traduzido com inglês não traduzido; (f) traduções em português com traduções em inglês.

Textos

Estão representadas na versão 9.0 do corpus obras publicadas entre 1837 e 2002, oriundas de Portugal, Angola, Moçambique, Brasil, Reino Unido, Estados Unidos e África do Sul. O COMPARA aceita tanto textos antigos como novos, de todas as variantes do português e do inglês. O corpus é extensível e contém atualmente cerca de 3 milhões de palavras provenientes de excertos de 72 textos escritos, alinhados com 75 traduções. A discrepância entre originais e traduções ocorre devido a três dos textos estarem alinhados com mais de uma tradução. Só são admitidos no corpus originais e traduções publicados, sujeitos, portanto, a um processo de seleção e revisão editorial. Com isso, procura-se uma maior garantia de qualidade e menor probabilidade de erros lingüísticos, tipográficos e de tradução. Também só se admite no corpus inglês traduzido diretamente do português e

[4] O COMPARA é (parcialmente) financiado pela Fundação para a Ciência e Tecnologia, co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001, e pelo POSC através do projecto POSC 339/1.3/C/NAC, ambos executados pela FCCN. Mais informações sobre as opções do corpus disponíveis em Frankenberg-Garcia et al (2006).

português traduzido diretamente do inglês, excluindo-se assim a influência de eventuais línguas intermediárias.

Optou-se por iniciar a constituição do corpus a partir de textos literários, uma vez que, ao contrário de outros gêneros, existe um número razoável de obras literárias em língua portuguesa com traduções inglesas publicadas, garantindo assim a bidirecionalidade do corpus. O gênero literário é ainda tido como sendo um dos mais ricos, permitindo, em tese, uma grande variedade lexical e morfossintáctica num corpus relativamente pequeno.

Digitalização

No corpus não se preservam elementos extra-textuais, tais como numeração das páginas, figuras e diagramas, corrigem-se os eventuais erros tipográficos e actualiza-se a ortografia dos textos digitalizados a partir de edições antigas.

Alinhamento

O alinhamento adotado no COMPARA é direcional, baseando-se sempre numa frase completa do texto original. 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 unidades vazias. As frases introduzidas pelo tradutor sem texto correspondente no original são, por sua vez, inseridas na unidade de alinhamento imediatamente precedente. Quando há frases reordenadas na tradução, o alinhamento segue as regras anteriores e a mudança na ordem é codificada separadamente.

O fato de o alinhamento do corpus ser baseado exclusivamente na divisão frásica do texto original simplifica o alinhamento de um mesmo original com várias traduções e possibilita, indiretamente, a comparação entre duas (ou mais) traduções, usando como denominador comum o original de que ambas derivam. Este tipo de alinhamento permite ainda expandir o corpus com um número ilimitado de traduções (para uma ou mais línguas) sem necessidade de se reprocessar o texto fonte.

Etiquetagem e anotação gramatical

Além de identificarem quais partes da tradução correspondem a cada frase do original, as etiquetas de alinhamento do corpus incluem informações que permitem localizar a adição, supressão, junção, separação e o reordenamento de frases automaticamente.

O corpus também possui etiquetas semânticas que facilitam a recuperação de notas de tradução, títulos, palavras e expressões estrangeiras, palavras e expressões com ênfase e entidades mencionadas.

Os textos em língua portuguesa do COMPARA foram recentemente acrescidos de anotação gramatical, com a introdução de etiquetas morfossintáticas através do analisador PALAVRAS (Bick 2000). A anotação dos textos em língua inglesa será feita através do CLAWS (Garside & Smith 1997) e está prevista para muito breve.

Codificação e acesso

O COMPARA encontra-se codificado para ser pesquisado com o processador de corpora IMS-CWB (Christ et al. 1999) através da interface DISPARA (Santos 2002). O corpus é público e, de acordo com as autorizações obtidas, pode ser utilizado gratuitamente para pesquisa e fins educacionais. Para que um maior número de pessoas possa usá-lo, o corpus possui uma interface em rede no endereço fornecido na figura 2. Na interface pode-se optar por interagir com o corpus através de um serviço em português ou em inglês. O corpus foi concebido para utilizadores experimentados e leigos ao mesmo tempo e, como tal, possui uma interface de pesquisa simples, com funcionalidades básicas, e outra complexa, bastante mais sofisticada. As interfaces encontram-se em constante desenvolvimento, com o objetivo de melhorar cada vez mais a usabilidade do corpus.[5]

Na próxima seção, serão dados alguns exemplos de utilização do COMPARA em lexicografia bilíngüe e estudos de tradução.

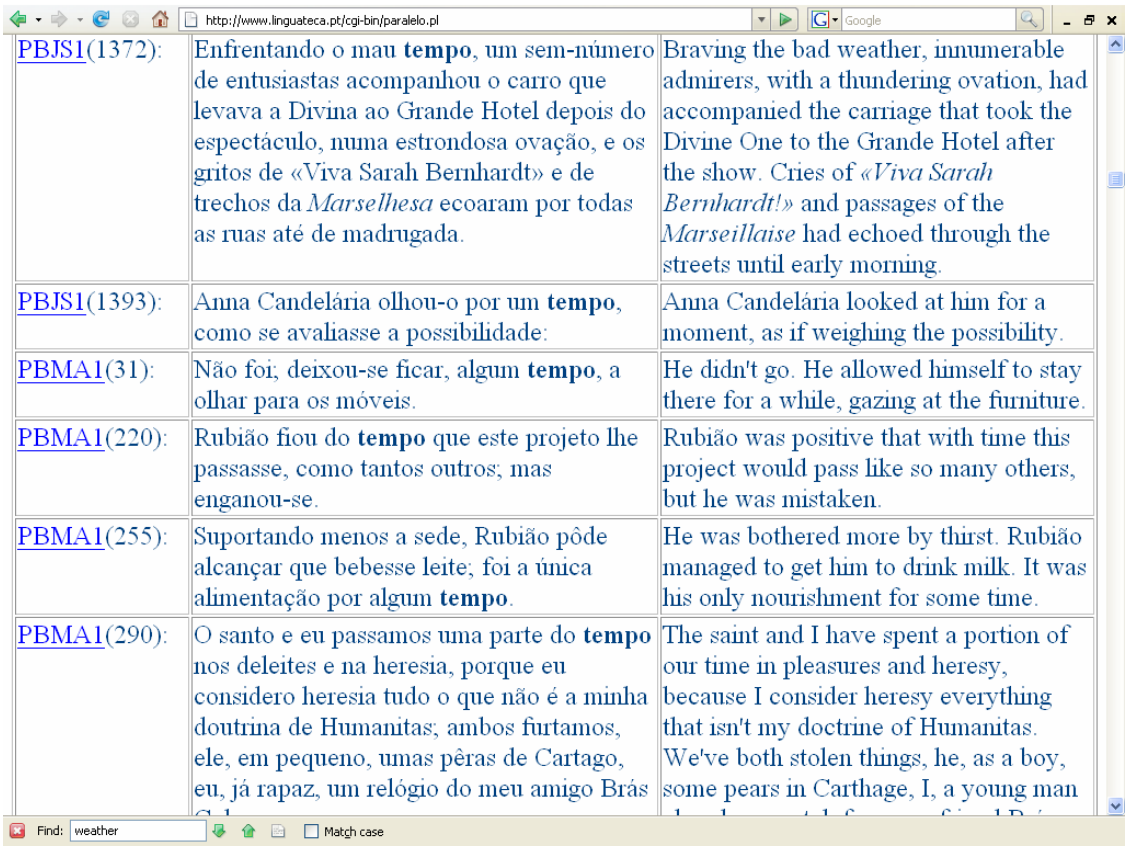
Exemplos de utilização de um corpus paralelo

Conforme referi na introdução, o alinhamento L1-L2 de um corpus paralelo faz com que este possa ser utilizado no aperfeiçoamento de dicionários bilíngües já existentes ou no desenvolvimento de léxicos computacionais para serem usados na tradução automática e na tradução assistida por computador. Uma das utilizações mais comuns é o estudo das diferentes traduções de uma palavra polissêmica. Ribeiro e Dias (2005), por exemplo, comparam as traduções humanas do adjetivo *grande* no COMPARA com os resultados

[5] Mais detalhes em Santos e Frankenberg-Garcia (2007)

apresentados pelo tradutor automático Babel Fish, identificando, desta forma, algumas das limitações da tradução automática e sugerindo maneiras de as superar. Já Specia et al. (2005) e Oliveira-Netto (2005) utilizam o COMPARA em estudos de voltados para o desenvolvimento de módulos de desambigüação lexical de sentido, para serem incorporados em sistemas de tradução automática inglês-português.

Figura 5 - Concordâncias paralelas para a palavra *tempo* no COMPARA 8.2

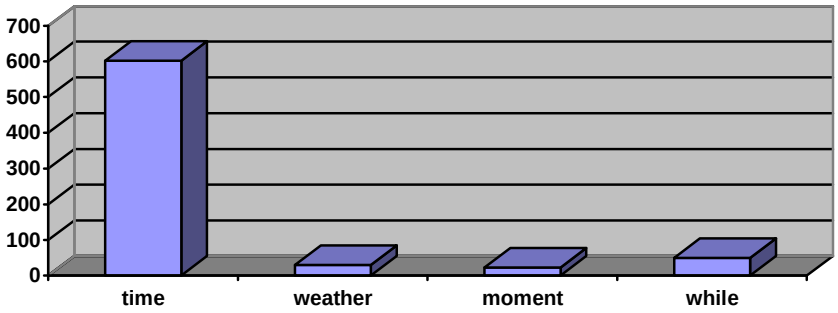


Para se descobrir quais são as traduções inglesas mais prováveis de uma determinada palavra portuguesa no COMPARA, podemos começar a pesquisa restringindo o corpus

de modo a pesquisar a palavra só de originais para traduções[6]. Se pesquisarmos então uma palavra como *tempo*, veremos, como mostram as concordâncias paralelas apresentadas na figura 5, que existe mais de uma tradução possível. Nos resultados parciais ilustrados pela figura, a palavra *tempo* aparece três vezes traduzida por *time*, uma vez por *weather*, outra por *moment* e outra por *while*.

Para obter um quadro mais completo das correspondências entre *tempo* e estas quatro traduções, podemos a seguir pesquisar cada uma delas enquanto restrição de alinhamento de *tempo* e pedir distribuições combinadas das expressões de busca em português e inglês[7]. No COMPARA 9.0, obteve-se os resultados apresentados na figura 6:

Figura 6 - Concordâncias paralelas com a palavra *tempo* em português original e *time*, *weather*, *moment* e *while* em inglês traduzido.



Note-se que, para interpretar os resultados apresentados como se fossem equivalentes tradutórios, temos que ter em conta uma pequena margem de erro, atribuível ao fato de o

[6] Faz-se a restrição do corpus no passo 3.3 da Pesquisa Avançada. A restrição é necessária porque, como referido na introdução, as traduções de uma língua para outra podem não ser biunívocas. Ao restringir o corpus da maneira indicada, estaremos excluindo o português das traduções e trabalhando exclusivamente com português original traduzido para inglês.

[7] A restrição de alinhamento é introduzida no primeiro passo da Pesquisa Avançada, na janela ao lado da expressão de pesquisa; o pedido de distribuição combinada é, por sua vez, feito no passo 4. Para mais informações sobre a Pesquisa Avançada do COMPARA, ver Frankenberg-Garcia (2005).

corpus não se encontrar alinhado à palavra. Ou seja, os resultados dizem respeito à concordâncias paralelas onde constam *time*, *wheather*, *moment* e *while* numa mesma unidade de alinhamento que a palavra *tempo*. Embora na grande maioria das vezes isto indique que uma palavra tenha sido traduzida pela outra, não é garantido que isto aconteça sempre. Por exemplo, na unidade de alinhamento abaixo, aparecem as palavras *tempo* no lado português e *weather* no lado inglês da concordância, mas uma não é a tradução da outra.

"PPJSA2(291):
Já se percebeu que a casa é antiga, sem conforto, de um **tempo** espartano e bronco, quando sair para a rua, na altura dos frios maiores, ainda era o melhor remédio para quem não dispusesse senão de um corredor gélido onde aquecer o corpo em pequenos exercícios de marcha.

Easy to see that the house is old and lacking in comfort, dating from more spartan and primitive times, when to go outdoors with the **weather** at its coldest was still the best solution for anyone who had nothing better than a freezing corridor where he could march up and down in an effort to keep warm."

No entanto, em 16 das 18 vezes em que *tempo* e *weather* coincidem numa concordância paralela, *weather* é de fato a tradução de *tempo*. Sabendo que a margem de erro é pequena, os resultados indicam que *time* é uma tradução muito mais provável para *tempo* do que qualquer um dos outros equivalentes tradutórios apresentados.

Deixando a lexicografia bilíngüe de lado, passemos agora a um exemplo de utilização do corpus em estudos de tradução. Como foi referido na introdução, se dispensarmos o alinhamento, podemos aproveitar a estrutura paralela de um corpus paralelo bidirecional para comparar uma língua em versão original e esta mesma língua em versão traduzida.

O corpus COMPARA, por exemplo, pode ser usado para contrastar inglês original com inglês traduzido, ou para contrastar português original com português traduzido. Existem já diversos estudos sobre as diferenças entre inglês original e inglês traduzido com base em corpora. Olohan e Baker (2000), por exemplo, constataram que, seguido de verbos dicendi, o pronome relativo opcional *that* é muito mais freqüente no inglês das traduções do Translational Corpus of English do que nos textos escritos originalmente em inglês do British National Corpus. Resultados muito semelhantes foram também conseguidos por Frankenberg-Garcia (2002), usando os originais e as traduções inglesas do COMPARA. Em ambos os estudos, concluiu-se que a explicitação do pronome relativo *that* após um verbo dicendi é uma das caracterísitcas que distingue o inglês original do inglês traduzido.

Apesar de estarmos plenamente conscientes de que o português original é bastante diferente do português das traduções, existem muito poucos estudos empíricos a contrastar os dois. Num dos poucos estudos com corpora publicados na área, Frankenberg-Garcia (2005) mostra que, no COMPARA, o uso de palavras estrangeiras é completamente distinto nos originais em língua portuguesa e nos textos traduzidos para português. As diferenças não se ficam pela quantidade de palavras e expressões estrangeiras utilizadas, mas incidem também sobre o número de línguas de empréstimo adotadas.

Um dos tipos de pesquisas mais simples no âmbito dos estudos de tradução com corpora envolve comparar a distribuição de uma palavra em textos traduzidos e não traduzidos.

Em Frankenberg-Garcia (2002), estudou-se a distribuição do verbo *nod* em inglês original e inglês traduzido numa das primeiras versões do COMPARA e verificou-se que a utilização deste verbo é muito mais comum em inglês original. Em raríssimas vezes utilizou-se o verbo *nod* em textos traduzidos de português para inglês. Inversamente, o advérbio *already* aparenta ser muito mais freqüente em inglês traduzido do que em inglês original (Frankenberg-Garcia 2004).

Podemos facilmente efetuar estudos semelhantes no lado português do corpus. Tomemos como exemplo o lema *diferente*, que possui 311 ocorrências no COMPARA 9.0. Se pedirmos uma distribuição deste lema em português original e português traduzido, veremos que ele ocorre 210 vezes em 796 889 palavras de texto traduzido e 101 vezes em 639 660 palavras de texto original[8]. Ou seja, em cada 100 mil palavras, o lema *diferente* ocorre 26,4 vezes em português traduzido e apenas 15,8 vezes em português original. É portanto possível que o uso deste lema nas traduções seja excessivo e contribua para tornar a leitura de uma tradução e a leitura de um texto escrito originalmente em português distintas.

Comentários finais

Existem diferentes tipos de corpora paralelos. As opções prévias à constituição de um corpus paralelo influenciam a forma como ele poderá ser utilizado, por quem e com que finalidades. Neste trabalho pretendeu-se discutir os processos decisórios por trás da compilação de um corpus paralelo, mostrar as opções tomadas por um corpus em

[8] Solicita-se a distribuição em texto original e traduzido no passo 4 da Pesquisa Avançada do COMPARA.

particular - o COMPARA - e apresentar alguns trabalhos no âmbito da lexicografia bilíngüe e dos estudos de tradução realizados com base neste corpus.

CHRIST, O., 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).

BARLOW, M. (2002) ParaConc: Concordance software for multilingual parallel corpora. *Language Resources for Translation Work and Research*, pp 20-24.

BICK, E. (2000) *The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press.

FRANKENBERG-GARCIA, A. (2002) "Using a parallel corpus to analyse English and Portuguese translations", *Translation (Studies): a crossroads of disciplines*, Faculdade de Letras, Universidade de Lisboa, 14-15 de Novembro de 2002.

FRANKENBERG-GARCIA, A. (2004) "Lost in parallel concordances" In G. Aston, S. Bernardini & D. Stewart (eds.) *Corpora and language learners*, Amsterdam: John Benjamins, pp 213-229.

FRANKENBERG-GARCIA, A. (2005) "A corpus-based study of loan words in original and translated texts" In P. Danielsson & M. Wagenmakers (eds.) *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, 14-17 July 2005, ISSN: 1747-9398.

FRANKENBERG-GARCIA, A. (2007) "COMPARA - Aula Prática em Português". Disponível em http://www.linguateca.pt/COMPARA/AulaPratica.doc

FRANKENBERG-GARCIA, A. & SANTOS, D. (2001) "COMPARA, um corpus paralelo de português e inglês na Web" *Cadernos de Tradução IX*, Universidade Federal de Santa Catarina, Brasil.

FRANKENBERG-GARCIA, A. & SANTOS, D. (2003) "Introducing COMPARA, the Portuguese-English parallel translation corpus". In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translation Education*, Manchester: St. Jerome Publishing, pp. 71-87.

FRANKENBERG-GARCIA, A, SANTOS, D., & SILVA, R. (2006) Construção do COMPARA. Disponível em http://www.linguateca.pt/COMPARA/Construcao.html

GARSIDE, R., AND SMITH, N. (1997) "A hybrid grammatical tagger: CLAWS4" In R. Garside, G. Leech & A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora,* Longman: London, pp. 102-121

OLOHAN, M. & BAKER, M. (2000) "Reporting *that* in translated English: Evidence for subconscious processes of explicitation?" *Across Languages and Cultures* 1(2): 141-158.

RIBEIRO, G. & DIAS, M.C. (2005) "Two Corpus-based Studies on the Translation of Adjectives in English and Brazilian Portuguese", In P. Danielsson & M. Wagenmakers (eds.) *Proceedings of the Corpus Linguistics 2005 conference,* Birmingham, UK, 14-17 July 2005, ISSN: 1747-9398.

SANTOS, D. (2002) "DISPARA, a system for distributing parallel corpora on the Web" In E. Ranchhod & N. Mamede (eds.) *Advances in Natural Language Processing*, LNAI 2389, Springer, pp. 209-218.

SANTOS, D. & FRANKENBERG-GARCIA, A. (2007)"The corpus, its users and their needs: a user-oriented evaluation of COMPARA", *International Journal of Corpus Linguistics* 12, pp. 335-374.

SPECIA, L., NUNES, M.G.V. & STEVENSON, M. (2005) "Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation", *Recent Advances in Natural Language Processing (RANLP-2005)* (Borovets, Bulgaria, 21-23 September 2005)

WOOLS, D. (2000) "From purity to pragmatism; user-driven development of a multilingual parallel concordancer", In S. Botley, A. McEnery & A. Wilson (eds.) *Multilingual Corpora in Teaching and Research,* Amsterdam: Rodopi.