

A Construção (e alguns usos) do corpus Compara

Ana Frankenberg-Garcia (ISLA, Lisboa)

1. O corpus Compara

O Compara é um corpus paralelo e bidireccional de português e inglês que pode ser consultado gratuitamente em <http://www.linguateca.pt/COMPARA/Bem-vindos.html>. O corpus é extensível e, na actual versão 3.5, contém mais de um milhão de palavras provenientes de excertos de 23 textos de ficção alinhados com 26 traduções¹. Estão nele representados autores e tradutores de Portugal, Angola, Moçambique, Brasil, Reino Unido, Estados Unidos e África do Sul, com textos publicados entre 1865 e 2000².

2. A construção do corpus

A concepção do Compara teve início no final de 1999. Estabeleceu-se desde o princípio que o corpus seria um recurso de acesso público, disponibilizado através da Internet pela Linguateca (<http://www.linguateca.pt>), então projecto Processamento Computacional do Português. O corpus encontra-se codificado no sistema IMS Corpus Workbench (Christ et al. 1999), sendo este, conforme Santos e Ranchhod (1999), o sistema de corpora que melhor se adapta às necessidades da Linguateca.

As particularidades técnicas de programação por trás do Compara encontram-se descritas em Santos (2002). A presente comunicação pretende realçar outros aspectos da construção do corpus e alguns de seus usos.

2.1 Selecção de textos

Apesar da dificuldade notória em se obter autorização para a utilização, na Internet, de obras protegidas por direitos de autor, optou-se por iniciar a constituição do corpus a partir de uma colecção de textos de ficção. A decisão deveu-se, em primeiro lugar, ao facto de existir um número razoável de obras de ficção em língua portuguesa com traduções inglesas publicadas, ao contrário do que acontece com outros géneros linguísticos, como, por exemplo, o jornalístico, o científico e o técnico³. Em segundo lugar, entrou em linha de conta o facto de as obras de ficção passarem por um processo de selecção e revisão editorial antes de serem publicadas, havendo assim uma melhor garantia de qualidade e menos hipóteses de erros linguísticos e tipográficos. O género ficção é ainda tido como sendo bastante rico em termos linguísticos, permitindo, em tese, uma grande variedade lexical e morfo-sintáctica num corpus relativamente pequeno.

Para além da opção por uma colecção inicial de obras de ficção, não houve nenhuma

outra restrição no que diz respeito à selecção de textos para o corpus. Aceitaram-se tanto textos antigos como novos, de todas as variantes do português e do inglês.

2.2 Pedidos de autorização

Nos pedidos de autorização para a utilização de obras de ficção no Compara, o mais importante talvez tenha sido explicar, numa linguagem acessível e não técnica, o que é, para que serve e como funciona um corpus. Desde os primeiros contactos com autores, tradutores e editoras, foi importante esclarecer que os textos utilizados no Compara – geralmente excertos de 30% de uma obra – apesar de disponíveis para consulta na Internet, não poderiam ser recuperados integralmente pelos utilizadores do corpus. Um outro factor que pode ter contribuído para uma boa aceitação dos pedidos de autorização foi mencionar, sempre que se fazia um novo pedido, quais editoras, autores e tradutores já estavam a colaborar com o projecto. Apesar de o processo de contactar os detentores de direitos de autor e obter autorização para o uso dos textos no corpus ser bastante lento e trabalhoso, os pedidos feitos foram na generalidade bem recebidos, e temos hoje autorização para utilizar 40 originais e 64 traduções com direitos reservados. Por outro lado, alguns detentores de direitos de autor nunca chegaram a nos responder, mas apenas dois autores e uma editora apresentaram recusas explícitas.

2.3 Digitalização dos textos⁴

No processo de digitalização dos textos do corpus, não se preservam elementos extra-textuais, tais como numeração das páginas, figuras e diagramas, corrigem-se os erros tipográficos encontrados, actualiza-se a ortografia dos textos digitalizados a partir de edições antigas do português e, para distinguir travessões de hífen, os segundos são grafados como uma sequência de dois hífen: --. O objectivo do corpus não é recuperar a forma original do texto impresso, e apenas as correcções tipográficas se encontram identificadas (com etiquetas <corr>) para uma eventual referência no futuro.

As notas de autor são preservadas, identificadas com a etiqueta <anote>, e introduzidas imediatamente após à frase onde aparece o sinal que as identifica no texto. No lugar do símbolo que remete à nota, insere-se <marca num=x>. Por exemplo:

PPJP1.po

Filomena. Ou Mena. Filomena Joana Vanilo <marca num=1> Athaide (segundo os arquivos daquele Depósito Disciplinar) de 23 anos, solteira, que por autorização superior visitou o major Dantas Castro nas datas tais e tais e nas condições de vigilância determinadas pelo Regulamento, Elvas, Forte da Graça, tantos de tal. <anote> *Van Niel*, e não *Vanilo*. A mãe de Mena, já falecida, era filha de comerciantes sul africanos (correção, a lápis, do inspector Otero). </anote> Otero: A que propósito é que uma merda destas vem em ofício confidencial?

As notas de tradução também são mantidas, mas identificadas com a etiqueta <tnote> e inseridas no lugar do símbolo a que se reportam. Por exemplo:

EBJB1.po

ele revelou-me o seu interesse por Gosse <tnote> Edmund William Gosse (1849-1928), crítico inglês </tnote> e pela sociedade literária inglesa dos finais do século passado.

O corpus não se encontra anotado gramaticalmente, mas, para as partes do texto sublinhadas, ou salientadas em itálico, em negrito, com fontes diferentes ou por indentação, existem:

a. Etiquetas <title> em torno de títulos de livros, jornais, revistas, filmes, programas de televisão, canções, poemas, obras de arte, etc. (sejam eles verdadeiros ou fictícios). Por exemplo:

EBDL2T1.en

When we sat on the sofa together to watch <title>News at Ten</title>

PPEQ1.po

Há, porém, um ponto de <title>Jerusalém Passeada</title> que não posso deixar sem enérgica contestação.

b. Etiquetas <named> em torno de nomes próprios utilizados para designar estabelecimentos comerciais, hotéis, empresas, produtos, doutrinas, etc. Por exemplo:

EBJB1.en

He stayed at the <named>Hotel Paris</named>

EBDL1T1.po

me puseram a alcunha de <named> Bolinha </named> quando estava na tropa

EBDL1T1.po

passou-me uma receita de <named> Valium </named>

c. Etiquetas <foreign> em torno de palavras numa língua diferente da língua principal do texto. Por exemplo:

EBJB1.en

But the white bear, <foreign> thalassarctos maritimus </foreign>, is the aristocrat of bears...

Note-se que, no que diz respeito aos nomes próprios, só se utiliza a etiqueta <foreign> no caso deles incluírem substantivos comuns. Por exemplo:

EBJB1.po

Além disso, lembro-me do fim de <foreign><title>L' Education sentimentale</title></foreign>.

PPEQ2.po

como uma senhora da <foreign><named> Belle Époque </named></foreign> ajustaria seu vestido

Os exemplos acima mostram também que as etiquetas <foreign> podem ser utilizadas concomitantemente com <named> e <title>.

d. Etiquetas <emph> em torno de palavras ou expressões isoladas dentro de uma frase indicando ênfase linguística. Por exemplo:

```
EBDL1T1.po
acaba por se esquecer de ter medo, até que acaba por verificar que não há <emph>
de que </emph> ter medo.
```

```
EBDL1T1.en
It rarely happened when I might have expected it, like when I was playing golf
or tennis, but it could happen just <emph> after </emph> a game,
```

Note-se que não pode haver sobreposição das etiquetas <emph> com <foreign>, uma vez que é demasiadamente subjectivo avaliar se o texto se encontra salientado por ênfase linguística, pelo uso de palavras estrangeiras, ou por ambos. Nos casos ambíguos, a etiqueta <foreign> prevalece. Por exemplo:

```
EBDL1T1.en
"<foreign>Au contraire</foreign>, as Amy would say..."
```

e. Etiquetas <voice> em torno de partes salientes do texto que marcam citações ou indicam uma mudança de voz na narrativa, o que normalmente acontece quando o narrador se põe a pensar, recordar, ou quando a voz de uma outra personagem se intromete no texto principal. Por exemplo:

```
The station is plastered with notices saying that platforms will be closed one
minute before the advertised departure times of trains <voice> « in the
interests of punctuality and customer safety » </voice>, but he could have let
me through without endangering either.
```

```
Eu disse que queria o carro. <voice> Eu tinha de ter o carro. </voice> O
vendedor disse que podia conseguir um outro em duas ou três semanas
```

Note-se que os segmentos <voice> por vezes abrangem parágrafos inteiros. Os sinais de pontuação pertencentes a estes segmentos são mantidos dentro das etiquetas. Note-se também que apesar de estes segmentos poderem conter etiquetas <title>, <foreign>, <named> e <emph>, não pode haver sobreposição completa uma vez que, nestes casos, é impossível determinar a que se deve o destacamento do texto. Nos casos ambíguos, a etiqueta <voice> prevalece. O exemplo abaixo, potencialmente ambíguo, deve portanto ser marcado com etiquetas <voice>:

```
"To understand a message is to decode it. Language is a code. <voice>But every
decoding is another encoding. </voice> If you say something to me..."
```

É importante notar que os títulos, os nomes próprios, as palavras estrangeiras, a ênfase linguística e as mudanças de voz na narrativa só são etiquetados se o texto se

encontra de alguma maneira grifado. Isso inclui palavras e expressões em fonte normal que estejam inseridas em trechos mais longos em itálico, onde se usa a fonte normal justamente para indicar uma parte saliente do texto. Os elementos apontados acima que o autor ou tradutor não tiver salientado tipograficamente não são anotados no corpus⁵.

2.4 Opções estruturais

Os textos do Compara não se encontram anotados em termos de divisão de capítulos e parágrafos, uma vez que não se obteve autorização para os redistribuir. As opções estruturais de codificação do corpus têm como objectivo a análise da tradução de frases completas na língua de origem. Entende-se por frase completa uma sequência de palavras iniciada por letra maiúscula e terminada em ponto final, reticências, ponto de exclamação ou ponto de interrogação, seguida de uma nova sequência de palavras iniciada por letra maiúscula, ou sem seguimento nenhum, como nos casos de fim de parágrafo, capítulo ou texto⁶. A título de exemplo, veja-se, no parágrafo abaixo, a separação de frases segundo os critérios adoptados:

PPJP1.po

```
<s> Elias está sem óculos, tem pálpebras pisadas e rugosas como as dos perus.  
<s> Mastiga em seco fitando sempre (através das pálpebras? por uma réstea  
sumida?) aqueles retratos desfalecidos em sépia de antepassado. <s> Depois  
levanta-se e atravessa o corredor, há aqui um cheiro que não engana: ratos?
```

Nos casos de discurso directo, é de notar que podem haver palavras iniciadas por letra maiúscula a seguir aos sinais de pontuação definidos acima sem que isso implique em separação de frase. Por exemplo:

EBJT1.en

```
<s>'You OK?' Robin's daughter said, standing close to him, but not touching.
```

Também no discurso directo, existem frases em que o hífen é usado de maneira idiossincrática para indicar uma interrupção da fala. Nesses casos, considerou-se-o como separador de frase:

EBJT2.en

```
<s> It's your baby -- '  
<s> ` Yes, but you're my niece and we've always been particular friends.
```

As sequências de palavras terminadas em dois pontos só são consideradas frases separadas nos casos coincidentes com os de fim de parágrafo:

PMMC1.po

```
<s>De repente, gritou-se num desespero:  
<s>-- Mulher, ajuda-me.
```

Se não houver fim de parágrafo, não há separação de frase, quer a sequência seguinte comece com letra maiúscula ou não:

PPEQ2.po

<s>Até eu disse ao Padre Eugénio : "O Eugenininho, o Senhor hoje tem desgosto!"

PPSC1.po

<s> De Paris, amo tudo com igual amor : os seus monumentos, os seus teatros, os seus bulevares, os seus jardins, as suas árvores...

É de notar também que podem existir casos de mudança de linha sem que haja separação de frase, como é notório nos excertos que incluem poesia, mas não só. Nesses casos, assinala-se a mudança de linha com
 para uma melhor visualização das concordâncias:

PPCP1.po

<s> Mas vamos à casa, é o que interessa,

 <voice> DESCOBERTO O COVIL DO CRIME

 Onde Teria Estado Sequestrada

 UMA JOVEM ENLOUQUECIDA, </voice>

 os jornais embandeiraram-na em títulos e fotografam-na em todas as posições, de frente e de lado e voltada para o pinhal.

2.5 Alinhamento

O sistema de alinhamento adoptado no Compara é direccional, baseando-se 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 entidades vazias. As frases introduzidas pelo tradutor sem texto correspondente no original são, por sua vez, inseridas na unidade de alinhamento imediatamente precedente e identificadas com a etiqueta <add>. Por exemplo:

a. Frase mantida na tradução (1-1)

EBJT21.en (original)

<s> He still said, though less angrily now, that she had deceived him.

EBJT2.po (tradução)

<s> Ele ainda afirmava, embora menos encolerizado, que ela o tinha desiludido

b. Frase dividida na tradução (1-2)

EBDL3T1.en (original)

<s> "Spare me the narrow misses, Bill, what have you got?"

EBDL3T1.po (tradução)

<s2> "Não me fale do que perdi, Bill. <s2> O que é que ainda tem? "

c. Frases unidas na tradução (1-1/2, 1-1/2)

PBPM1.po (original)

<s> Muito bem.

<s> O casal vem chegando, dentro do automóvel.

PBPM1.en (tradução)

<s½> So then,

<s½> the couple arrives in the automobile.

d. Frase suprimida na tradução (1-0)

PBAD1.po (original)

<s> A cara impenetrável, os olhos não diziam nada.

<s> **Não estava mais ali quem falou.**

<s> Ele agora atendia uma freguesa que queria três metros de morim.

PBAD1.en (tradução)

<s> Zito's face was inscrutable, his eyes said nothing.

<s>

<s> Now he was serving a customer who wanted three metres of cambric.

e. Frase acrescentada na tradução (1-1+1ad)

PPCP1.po (original)

<s> "Porquê, acha que é assim de deitar fora?"

PPCP1.en (tradução)

<s> But why should we waste them? <add>Why?</add>

Quando há frases reordenadas na tradução, o alinhamento segue as regras anteriores, e a mudança na ordem é codificada separadamente. No exemplo abaixo, <reord> identifica a frase cuja ordem foi alterada na tradução e <place> identifica o ponto onde o tradutor a inseriu:

EBOW1.en (original)

<s>The picture had to be concealed.

<s>There was no help for it.

EBOW1.po (tradução)

<s><reord 3>Era preciso esconder o retrato.</reord>

<s> Não havia remédio. <place 3>

As opções tomadas fazem com que o alinhamento do corpus não possa ser totalmente automático. Conforme descrito em Frankenberg-Garcia e Santos (2001) e Santos (2002), o alinhamento grosso fornecido pelo Easy Align – o alinhador do IMS Corpus Workbench – tem que ser adaptado manualmente de maneira a que se coadune com os critérios de alinhamento descritos acima. Apesar do trabalho acrescido que isto representa, o sistema de codificação adoptado torna possível pesquisar automaticamente os casos de junção, separação, omissão, adição e reordenamento de frases na tradução. Pelo quanto que sabemos, o Compara é o único corpus paralelo existente que permite fazer isto. Além disso, os critérios de alinhamento do corpus, baseados sempre na divisão frásica do texto original, simplificam o alinhamento de um mesmo original com várias traduções, e permitem,

indirectamente, a comparação entre duas (ou mais) traduções, usando como denominador comum o original de que ambas derivam. O facto de a unidade de alinhamento adoptada basear-se exclusivamente na divisão frásica dos originais permite ainda expandir o corpus com um número ilimitado de traduções (para uma ou mais línguas) sem necessidade de se reprocessar o texto fonte.

2.6 O interface Dispara

O interface Dispara de utilização do Compara na Internet (Santos 2002) permite ao utilizador escolher entre uma pesquisa simples e uma pesquisa avançada. A pesquisa simples dá acesso a concordâncias paralelas na direcção português-inglês ou na direcção inglês-português (ver <http://www.linguateca.pt/COMPARA/BuscaSimples.html>), utilizando-se todos os textos do corpus. A pesquisa avançada permite restringir o corpus em termos de variante linguística, data de publicação dos textos, direcção da tradução (ou seja, só de originais para traduções ou só de traduções para originais) e permite também seleccionar manualmente textos específicos (por exemplo, os textos de um determinado autor).

Além de procurar palavras e expressões em contexto, a pesquisa avançada permite ainda introduzir restrições de alinhamento às expressões de busca. Por exemplo, pode-se procurar a palavra *sim* no lado português do corpus sem que haja a palavra *yes* no lado inglês. Devido à maneira como o alinhamento do corpus se encontra codificado, também é possível, como já foi referido, pesquisar alterações nas traduções relativamente à divisão frásica dos originais. Pode-se ainda procurar automaticamente as notas de tradução, os títulos, a ênfase e as palavras estrangeiras presentes no corpus associando-os ou não a uma palavra ou expressão em contexto⁷.

Em termos de resultados, a pesquisa avançada permite ver concordâncias em contexto com ou sem informações adicionais sobre o alinhamento. Isto é, pode-se pedir, junto com as concordâncias, informação sobre a preservação, divisão, junção, supressão, adição e reordenamento de frases na tradução. Pode-se também pedir para esconder as notas de tradução, caso se queira consultar apenas a localização e não o conteúdo das mesmas. Além de concordâncias, pode-se requisitar uma distribuição das formas. No caso de a expressão de busca abranger mais de uma forma como, por exemplo, as palavras terminadas em *mente*, a distribuição das formas facilita a separação dos advérbios com este sufixo, como *felizmente* e *absurdamente*, das palavras que têm a mesma terminação, mas não são advérbios, como *mente* e *semente*. A opção distribuição das fontes serve, por sua vez, para se ver quantas vezes a sua expressão de pesquisa aparece em cada texto do corpus. Poderá ser útil quando se está a pesquisar a frequência com que um autor ou tradutor a utilizou. Por fim, a opção distribuição combinada pode servir para pesquisas que incluam restrições de alinhamento, pois apresenta um resumo quantitativo do número de ocorrências encontradas nas duas línguas do corpus. Assim, para a expressão *yes* com

a restrição de alinhamento *sim*, conta-se automaticamente o total de ocorrências de *yes* e o número de casos coincidentes com *sim*.

3. Usabilidade

No que diz respeito à usabilidade do corpus, estabeleceu-se desde o princípio que as consultas ao corpus poderiam ser feitas através de um serviço em português e outro em inglês, permitindo assim acesso a utilizadores com poucos conhecimentos de uma língua ou de outra. Procurou-se também reunir condições para assegurar que o corpus não se limitasse a ser útil apenas às pessoas já habituadas a trabalhar com corpora, servindo também a um público menos experiente, incluindo estudantes de línguas e de tradução, professores, autores de materiais didácticos, teóricos da tradução e investigadores na área da literatura comparada.

Desde o primeiro anúncio do corpus em Janeiro de 2001, as funcionalidades do serviço e o desenho das páginas de acesso têm sofrido modificações com base nas mensagens de dúvidas que nos enviam os utilizadores e nalguns registos de pesquisas mal sucedidas. Como resultado directo disto, criou-se uma página de ajuda com hiperligações para cada um dos campos de pesquisa disponíveis. Mais recentemente, um relatório sobre a usabilidade do corpus de autoria de Costa, Mota e Sarmiento (2002) apontou-nos o caminho para mais uma série de melhorias que estão a ser implementadas. Apesar de muito já ter sido feito, temos consciência de que o processo de facilitar ao máximo a utilização do corpus não está terminado.

4. Alguns usos

Sendo este encontro destinado principalmente a pessoas que trabalham com o processamento da linguagem natural, pretendo aqui exemplificar alguns usos do Compara por um público diferente, com menos conhecimentos técnicos, mas nem por isso menos interessado na utilização de corpora. Exemplificarei um dos usos do Compara no ensino-aprendizagem de inglês por falantes nativos de português e outro nos estudos da teóricos tradução.

4.1 O Compara nos estudos teóricos da tradução

Vários exemplos de utilização do Compara nos estudos teóricos da tradução encontram-se descritos em Frankenberg-Garcia (2002a). Um deles é uma análise quantitativa do fenómeno da explicitação. Segundo Vinay e Darbelnet (1958), a explicitação é a introdução, na tradução, de informação que se encontra presente apenas de maneira implícita no original. A explicitação pode manifestar-se de diversas formas. Na tradução da palavra *doctor* para português, por exemplo, o tradutor vê-se forçado a explicitar género: *médico/médica*. Trata-se de um caso de explicitação morfológica obrigatória. Exemplos de explicitação cultural são fáceis de

encontrar nas notas de tradução, como mostra a seguinte tradução num dos textos do Compara:

EBDL3T2.en (original)

" *All's Well That Ends Well?* " he snaps back, quick as a flash.

EBDL3T2.po (tradução)

-- Será que é <title><foreign>All ' s well that ends well</foreign></title> ? --
ele diz rápido como um relâmpago. <tnote> Tudo está bem quando acaba bem é o
título de uma peça de Shakespeare, que nasceu em Stratford-upon-Avon.</tnote>

Segundo Blum-Kulka (1986), a explicitação é uma característica típica do texto traduzido, independentemente da língua utilizada. Ou seja, é considerada um dos fenómenos universais da tradução. Em Frankenberg-Garcia (2002a), o Compara 2.2 foi utilizado para testar esta hipótese com base na análise de um caso específico de explicitação gramatical nas traduções inglesas do corpus. Usou-se, para o efeito, apenas o lado inglês do corpus, subdividindo-se o lado das traduções e o lado dos originais em dois subcorpora comparáveis. A estrutura em análise foi a inserção do pronome relativo opcional *that* a seguir ao verbo *tell*, numa tentativa de reproduzir os resultados obtidos por Olohan e Baker (2000) através do Translational English Corpus (TEC) e do British National Corpus (BNC).

O uso do pronome relativo *that* a seguir a verbos que marcam a utilização de discurso indirecto tais como *tell*, *say*, *ask* e *think* é opcional em inglês. Nos exemplos abaixo, portanto, as duas frases estão correctas:

[1] *He told me he was going to be late.*

[2] *He told me that he was going to be late.*

Enquanto no exemplo [1] o pronome relativo *that* está apenas implícito, no exemplo [2], considerado mais transparente do ponto de vista sintáctico, o pronome relativo está explícito na frase. De acordo com a teoria da explicitação proposta por Blum-Kulka (1986) e com os resultados obtidos por Olohan e Baker (2000), formulou-se a hipótese de que também no Compara o segundo tipo de construção seria mais frequente no inglês das traduções do que no inglês dos textos escritos originalmente nessa língua.

Pesquisou-se primeiro “(*tell/tells/told/telling*)” nos textos originais em inglês do corpus, tomando-se o cuidado de excluir da busca os originais ingleses que se encontram repetidos no corpus devido ao facto de estarem alinhados com mais de uma tradução. A pesquisa gerou 203 ocorrências, mas nem todas eram relevantes para a análise em questão. Concordâncias tais como [3] e [4], em que o verbo *tell* não é utilizado para marcar discurso indirecto, foram desconsideradas e permaneceram 59 concordâncias válidas.

[3] ...you can hardly tell the difference between them...

[4] I took the lift and was told off by a sharp-faced nursing sister...

Aplicou-se então o mesmo procedimento aos textos em inglês traduzido do corpus. De um total de 275 ocorrências de “(tell/tells/told/telling)”, 90 foram identificadas como sendo válidas para a análise da estrutura *tell-that*.

Os dados obtidos através de ambas as pesquisas foram então classificados de acordo com a explicitação ou não do pronome relativo opcional. Os resultados encontram-se resumidos na tabela 1.

Tabela 1

Distribuição da estrutura *tell-that* em inglês original e traduzido no Compara 2.2, de acordo com Frankenberg-Garcia (2002a).

	Inglês original	Inglês traduzido
Explicitação de <i>that</i>	25 (42.4%)	60 (66.7%)
Omissão de <i>that</i>	34 (57.6%)	30 (33.3%)
Total	59	90

Como se pode ver, os resultados indicam que o pronome relativo opcional *that* é muito mais frequente no inglês das traduções do que nos textos escritos originalmente nessa língua, onde o *that* fica mais vezes implícito na frase. Os presentes resultados são, como indicam os números na tabela 2, bastante semelhantes aos totais obtidos por Olohan e Baker (2000) no BNC e no TEC. Ambos os estudos apontam para uma tendência à explicitação de elementos sintáticos opcionais nas traduções para a língua inglesa, fortalecendo, assim, a teoria da explicitação.

Table 2 Distribuição da estrutura *tell-that* em inglês original (BNC) e traduzido (TEC), segundo Olohan e Baker (2000).

	BNC	TEC
Explicitação de <i>that</i>	997 (41.5%)	719 (62.7%)
Omissão de <i>that</i>	1408 (58.5%)	427 (37.3%)
Total	2405	1146

4.2 O Compara no ensino-aprendizagem do inglês

A utilização de corpora paralelos no ensino-aprendizagem de uma língua estrangeira não é consensual. Enquanto alguns apontam para os seus benefícios (por exemplo, Roussel 1991, Barlow 2000, Frankenberg-Garcia 2000, e Johansson & Hofland 2000), outros, como Gellerstam (1996), alertam para os efeitos potencialmente nefastos de se expor os aprendentes de uma língua estrangeira às idiosincrasias típicas das traduções. De facto, a língua usada nas traduções é diferente da língua usada em contextos monolíngues, onde não se aplicam as limitações impostas por originais de um outro idioma. No entanto, como já referi em Frankenberg-Garcia

(2002b), é preciso distinguir as situações em que e as idiossincrasias linguísticas das traduções podem prejudicar os aprendentes de uma língua das situações em que as traduções não interferem e podem até ajudar no desenvolvimento linguístico dos alunos. Antes de se julgar se o uso de corpora paralelos no ensino de línguas faz bem ou mal é preciso saber usá-los, e saber usá-los significa saber tirar partido das diferenças entre língua usada nos originais e a língua usada nas traduções.

Tomemos como exemplo o uso exagerado do advérbio de tempo *already* por parte de falantes nativos de português. É bastante comum ouvir um falante nativo de português dizer [5] quando pretendia dizer [6].

[5] *Have you **already** done your homework?*

[6] *Have you done your homework?*

Enquanto um falante nativo de inglês utilizaria a frase [6] para perguntar simplesmente se o seu interlocutor fez o trabalho para casa, e a frase [5] para dar a entender que o trabalho foi feito antes do tempo previsto, existe uma tendência para os falantes nativos de português utilizarem a frase [5] em ambas as circunstâncias. Uma possível causa disso seria a tradução directa do advérbio de tempo *já*, que em português se utiliza tanto no caso [5] como no [6], marcando-se a diferença entre os dois muitas vezes apenas na entoação.

Se analisarmos a distribuição do advérbio *already* no Compara 3.5, veremos que consta apenas 3,6 vezes em cada 10.000 palavras de inglês original, ao passo que em inglês traduzido a sua frequência é de 6,7 ocorrências por cada 10.000 palavras, ou seja, quase o dobro. Daí conclui-se que o inglês das traduções, que também pode estar influenciado pelo português, pode não ser o mais apropriado para ajudar os alunos a perceberem as situações em que o advérbio *already* está a mais.

Isso não significa, no entanto, que o Compara não possa ser usado para ajudar os alunos a compreenderem melhor esta diferença. Neste caso em particular, pode-se:

- a. restringir o corpus de modo a excluir o inglês das traduções;
- b. procurar *já* no lado português, eliminando *already* do lado inglês através de uma restrição de alinhamento.

Feito isto, é fácil seleccionar concordâncias como as que se seguem na tabela 3, que ensinam justamente que não é preciso dizer *already* em inglês toda a vez que se usa *já* em português.

Assim como as concordâncias abaixo protegem os alunos de eventuais efeitos nefastos do “tradutês”, também existem situações em que os alunos lucram com as traduções (por exemplo, para aprenderem que há certas coisas que não se pode

traduzir à letra), e situações em que as diferenças entre o que se diz nas traduções e o que se diz nos originais de uma mesma língua não afectam aquilo que está a ser ensinado (ver Frankenberg-Garcia 2002b, para mais exemplos).

Tabela 3

Concordâncias paralelas focalizando a ausência de *already* em inglês original e traduções portuguesas contendo *já*

Fonte	Traduções para português	Originais em inglês
EBOW1	Das sombras irreais da noite ressurgue a vida real já de nós conhecida.	Out of the unreal shadows of the night comes back the real life that we had known.
EBJT2	Fosse como fosse, isso já acontecera havia mais de um mês.	Whatever it had been, it had been done more than a month ago now.
EBJB1	E quanto às visitas subsequentes, quando já era o autor da escandalosamente famosa <i>Madame Bovary</i> ?	And what of subsequent visits, when he had become author of the notorious <i>Madame Bovary</i> ?
EBJT1	-- O pai dela já morreu.	' Her father's dead.
ESNG1	Não acha que ele já estudou muito, ficou nisso o dia inteiro, Sonny, ele deveria fechar os livros e ir dormir cedo.	Don't you think he's done enough, he's been at it all day, Sonny, he should close his books and have an early night.
EBDL1T1	Já fiz a artroscopia há um ano e continuo com dores.	It's a year since my arthroscopy, and I'm still getting pain.
EBDL1T1	Sei o que digo: já tive de ir e vir a Londres todos os dias.	I know: I've been a London commuter in my time.
EBDL2	-- Não, já acabei.	' No, I've finished. '
EBDL3T1	Ao cabo de muita prática, Philip já conseguia apanhar bem o sentido e, assim, respondeu seguro:	From long practice Philip was able to follow his drift pretty well, and therefore answered confidently:
EBDL3T2	Já decidi que meu futuro está na mídia.	I've decided my future's in the media."

5. Comentário Final

As opções prévias à constituição de um corpus influenciam a forma como ele poderá ser utilizado, por quem e com que finalidades. Espero ter conseguido demonstrar aqui alguns aspectos da construção do Compara e alguns usos diferentes daqueles que as pessoas que trabalham com o processamento da linguagem natural conhecem melhor. Da minha experiência na elaboração do Compara, fica a certeza de que um corpus ganha em todos os sentidos quando engenheiros da linguagem, linguistas e utilizadores do corpus trabalham em constante diálogo.

Notas

1. A disparidade entre o número de originais e traduções deve-se ao facto de o corpus admitir mais de uma tradução por original. De momento, existem dois originais em língua inglesa alinhados com duas traduções para português cada - uma para a variante portuguesa e outra para a variante

brasileira da língua - permitindo, indirectamente, uma comparação das duas variantes. Encontra-se igualmente disponível um original em língua portuguesa alinhado com uma tradução inglesa do século dezanove e outra do ano 2000, possibilitando um estudo diacrónico das duas traduções.

2. Detalhes sobre a composição do corpus são constantemente actualizados em <http://www.linguateca.pt/COMPARA/Conteudo.html>.

3. Esta limitação não se aplica no caso das traduções de inglês para português, já que existem muito mais traduções nesta direcção linguística.

4. As opções de codificação do corpus aqui alinhavadas representam uma revisão das opções descritas para versões anteriores do Compara em Frankenberg-Garcia e Santos (2000) e Frankenberg-Garcia e Santos (2001). As modificações estão a ser implementadas de raiz em todos os textos novos acrescentados ao corpus a partir da versão 3.0, de Fevereiro de 2003, e incorporadas gradualmente aos textos anteriores a esta data. É importante notar que as novas funcionalidades ainda não estão totalmente operacionais, uma vez que isso só poderá acontecer a partir do momento em que todos os textos antigos do corpus estejam em conformidade com os novos padrões de codificação.

5. Para além dos elementos de formatação já apontados, também foram etiquetadas as partes salientes de texto que, em certas edições impressas mais antigas e nos textos electrónicos obtidos sem formatação, foram grifadas através da utilização de aspas ou letras maiúsculas. Nos casos de utilização de palavras em caixa alta para indicar texto saliente em formato ASCII (como nos textos do projecto Gutenberg, por exemplo), ao introduzirem-se as etiquetas acima mencionadas, as letras maiúsculas são convertidas de volta para minúsculas.

6. A presente definição é na verdade um resumo bastante simplificado dos critérios adoptados, cujos pormenores encontram-se descritos em <http://acdc.linguateca.pt/acesso/atomizacao.html>.

7. As procuras por nota de autor, por elementos etiquetados com <voice> e com <named>, e as sobreposições entre <title> e <foreign> ainda não se encontram operacionais porque uma parte mais antiga do corpus ainda não se encontra em conformidade com os novos critérios de codificação (ver nota 4).

Referências Bibliográficas

Barlow, Michael (2000) "Parallel texts in language teaching". In Simon Botley, Tony McEnery e Andrew Wilson (eds.) *Multilingual corpora in teaching and research* Amsterdam: Rodopi, 106-115.

Blum-Kulka, Shoshana (1986) "Shifts of cohesion and coherence in translation". In Juliane House e Shoshana Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, 17-35. Tübingen: Gunter Narr.

Christ, Oliver, Bruno Schulze, Anja Hofmann e Esther 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/>

Costa, Luís, Cristina Mota e Luís Sarmiento (2002) *Alguns Comentários à Usabilidade do Serviço Compara*. Relatório de 4 de Dezembro de 2002.

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

Frankenberg-Garcia, Ana (2002a) "Using a parallel corpus to analyse English and Portuguese translations" trabalho apresentado na conferência *Translation (Studies): a crossroads of disciplines*, Faculdade de Letras, Universidade de Lisboa, 14-15 de Novembro de 2002.

Frankenberg-Garcia, Ana (2002b) "Lost in Parallel Concordances" trabalho apresentado em *The Fifth International Conference on Teaching and Language Corpora*, Universidade de Bolonha, Bertinoro, Itália, 27-31 de Julho de 2002.

Frankenberg-Garcia, Ana e Diana Santos (2000) "Introducing COMPARA: the Portuguese-English Parallel Corpus". Trabalho apresentado em *The Second International Conference on Corpus Use and Learning to Translate*, Universidade de Bolonha, Bertinoro, Itália, 3-5 de Novembro de 2000. A ser publicado em Silvia Bernardini, Federico Zanettin e Dominic Stewart (eds.), *Corpora in translator education* (título provisório). Manchester: St. Jerome.

Frankenberg-Garcia, Ana e Diana Santos (2001) "COMPARA, um corpus paralelo de português e inglês na Web" *Cadernos de Tradução IX*. Universidade Federal de Santa Catarina, Brasil.

Gellerstam, Martin (1996) "Translations as a source for cross-linguistic studies" In Karin Aijmer, Bengt Altenberg e Mats Johansson (eds.) *Languages in contrast: papers from a symposium on text-based crosslinguistic studies*. Lund Studies in English 88. Lund University Press, 53-62.

Johansson, Stig e Knut Hofland (2000) "The English-Norwegian Parallel Corpus: current work and new directions". In Simon Botley, Tony McEnery e Andrew Wilson (eds.) *Multilingual corpora in teaching and research*. Amsterdam: Rodopi, 106-115.

Olohan, Maeve e Mona Baker (2000) "Reporting *that* in translated English: Evidence for subconscious processes of explicitation?" *Across Languages and Cultures* 1(2): 141-158.

Roussel, Francine (1991) "Parallel concordances and tonic auxiliaries" *ELR Journal* 4, 71-101. Birmingham University Press.

Santos, Diana (2002) "DISPARA, a system for distributing parallel corpora on the Web". In Elisabete Ranchhod & Nuno J. Mamede (eds.) *Advances in Natural Language Processing*, LNAI 2389, Springer, pp.209-218.

Santos, Diana & Elisabete 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.

Vinay, Jean Paul e Jean Darbelnet (1958) *Stylistique Comparée du Français et de l'Anglais: Méthode de Traduction*. Paris: Didier.