OS *CORPORA*: INSTRUMENTOS DE AUTO-AJUDA PARA O TRADUTOR[1]

Stella E. O. Tagnin
USP

1. Introdução

Bowker (1998) salientou que o uso de *corpora* especializados em língua materna teve um papel significativo na melhora da qualidade da tradução de seus alunos quanto à "escolha correta do termo e redação idiomática" (1998: 648).

Se entendermos "escolha correta do termo" como colocação e "redação idiomática" como linguagem natural, perceberemos que esses dois aspectos são os pilares do que denominamos "convencionalidade (ou fraseologia) na língua", cuja falta de conhecimento caracteriza o "falante ingênuo" (Fillmore 1979). Veremos também que, em situações semelhantes, um tradutor pode ser igualmente "ingênuo".

2. O falante ingênuo

Fillmore cunhou esse termo para se referir a um aprendiz de língua estrangeira que desconhece as convenções da língua. Por exemplo, ele não saberia que *prisoner* (prisioneiro) e *jailer*

(carcereiro) significam coisas diferentes. Por que deveriam ser diferentes? Afinal, ambas são formadas por uma base *prison* (prisão) e *jail* (cárcere) que significam "casa de detenção" ou, em inglês, "a building where wrong-doers are locked up", mais acrescidas de um sufixo agentivo *–er*. Como se explica, então, que um *prisoner* é "a person kept in a prison" (uma pessoa mantida numa prisão), enquanto um *jailer* é "a man in charge of a jail" (um homem encarregado de uma prisão)?

Da mesma forma, um falante ingênuo desconhece, entre outras coisas, a ordem preferencial de binômios como *cats and dogs, bed and breakfast, knife and fork,* cujos correspondentes em português seriam *cães e gatos, cama e mesa, garfo e faca*. Também não sabe que há determinadas combinações fixas ou semi-fixas conhecidas como colocações, constituídas por substantivo + substantivo (*credit card, quality control, cost of living),* substantivo + adjetivo (*nursing home, silent movie, elementary school)*, substantivo, como sujeito, + verbo (*a river flows, a volcano erupts*) ou verbo + substantivo como objeto (*pay a visit, ask a question, make a decision*), verbo + advérbio (*pay dearly, cry loudly, hurt badly*), e adjetivo + advérbio (*deeply hurt, happily married, lavishly illustrated)*.

Também desconhece as fórmulas da língua, principalmente as fórmulas de rotina (*Good evening, Have a nice day, I'm really sorry*) e as fórmulas situacionais (*Break a leg, It takes one to know one, Have it your way)*.

Em resumo, o falante ingênuo não tem consciência de que grande parte da língua é formada de partes pré-fabricadas, de expressões prontas, das chamadas unidades fraseológicas, unidades que não precisam ser geradas a cada vez que são empregadas.

Observe-se também que, dependendo da situação, podemos todos ser falantes ingênuos em nossa própria língua materna. Como poderia um leigo conhecer os termos técnicos (na maioria, *colocações*) de certas profissões como medicina ou direito? Ou, como saberíamos o que dizer (usar as *fórmulas* corretas) em situações desconhecidas como, por exemplo, um velório, se jamais fomos a um?

Mas é no confronto de duas línguas que essas convenções se tornam mais evidentes. E é quando o tradutor entra em cena.

2.1 O Tradutor Ingênuo

Basicamente, a ingenuidade do tradutor se configura numa compreensão composicional do significado e numa falta de consciência do quanto uma língua é constituída dessas partes pré-fabricadas.

A ingenuidade do tradutor pode transparecer tanto na sua habilidade de compreensão, quando na de produção. Em termos de compreensão, ele pode não entender expressões idiomáticas como *a hard nut to crack, put one's best foot forward,* ou *cut corners*, por serem não-composicionais, ou seja, uma expressão cujo significado total não corresponde à soma dos significados individuais de seus componentes. Ele pode não compreender muitas fórmulas discursivas por não conhecer as convenções sociais que determinam seu uso na língua alvo. Pode também não compreender referências humorísticas que resultem da manipulação das categorias convencionais da língua. Por exemplo, não compreenderá um trocadilho como *fish and chimps* (de uma tirinha de Frank & Earnest em que um deles vê esse "prato" no cardápio e comenta "It's probably a typo, but why take a chance?" Deve ser erro de grafia, mas pr'á quê arriscar?) a menos que conheça o binômio *fish and chips*. Ou outro como *Ear today, gone tomorrow* (calcado em *Here today, gone tomorrow*), num artigo sobre a luta de boxe em que Mike Tyson arrancou um pedaço da orelha de seu adversário.

Por estranho que pareça, mesmo como falante nativo da língua alvo, o tradutor pode ter problemas no nível da produção para conseguir soluções naturais, caso se atenha tanto ao texto de partida a ponto de não perceber que, entre formas igualmente gramaticais, uma delas é de uso mais corrente. Em outras palavras, pode não se dar conta de que, dentro de uma gama de formas gramaticalmente *possíveis*, há certas formas que têm uma *probabilidade* maior de ocorrerem. Caso o tradutor selecione uma dessas formas *possíveis*,

em detrimento da mais *provável***,** produzirá uma tradução não natural, não fluente. Esse problema certamente se agravará quando o tradutor não estiver traduzindo para sua língua materna.

Nesse aspecto, as colocações e fórmulas são as categorias que apresentam maior dificuldade. No caso das colocações – palavras que co-ocorrem em freqüência maior do que se se tratasse de uma combinação aleatória – , a dificuldade pode residir no fato de, em geral, não constituírem problema de compreensão, de modo que tendem a passar despercebidas. Em outras palavras, por serem em grande parte composicionais, as colocações são de fácil compreensão. Entretanto, quando se trata de produzi-las, não são facilmente buscadas na memória, uma vez que não houve um esforço consciente para memorizá-las.

Uma idéia bastante "ingênua" seria acreditar que um dicionário poderia resolver todos os problemas do tradutor em termos de convencionalidade. É verdade que há algumas obras de referência que abordam essas categorias, principalmente dicionários mono- e bilíngües de expressões idiomáticas (para o inglês, Boatner & Gates 1975, Spears 1988, Spears 1989 , entre outros; para o par inglês-português Serpa 1982, Camargo & Steinberg 1989, 1990). Entretanto, há poucos dicionários de fórmulas em inglês, (especialmente Partridge 1977, Spears et allii. 1995, e Spears 1996), e talvez menos ainda de colocações (principalmente Cowie et allii. 1983, Benson, Benson & Ilson 1986 e Hill & Lewis 1997).

Para efeitos deste artigo, será considerado tradutor o profissional que traduz textos escritos, pois o problema das colocações será abordado com relação a sua freqüência no discurso escrito. Já as fórmulas fazem parte principalmente do discurso falado, objeto da interpretação, dublagem e legendagem, que requerem habilidades especiais, das quais não trataremos aqui.

3. As Colocações e o Tradutor

Vejamos os dois tipos principais de colocações e os caminhos do tradutor para encontrar uma tradução adequada para eles. A terminologia aqui usada é sintática, não funcional. Segundo Hausmann (1985), as colocações são formadas por uma *base* – a palavra de maior carga semântica –, geralmente um substantivo, mais um *colocado*. O nome da colocação será derivado do colocado. Assim, uma colocação de verbo + substantivo será uma *colocação verbal*, um adjetivo + substantivo será uma *colocação adjetiva*, e assim por diante

3.1 Colocações Nominais e Adjetivas

Esses dois grupos certamente constituem a maior parte do inventário fraseológico. Há milhares delas e a cada dia surgem outras, pois são empregadas para nomear novas tecnologias, processos, teorias etc., (por exemplo, *computer aided design, computer graphics, Computer Assisted Language Learning, Corpus Linguistics, Translation Studies, data storage)*, e novos objetos e produtos (*mouse pad, video game, food processor, video camera, London Eye, RealPlayer, RealJukebox*). Essas colocações só aparecem em dicionários bastante especializados e, mesmo assim, só quando seu uso já estiver bastante disseminado.

3.2 Colocações Verbais

Apesar de serem em número bem menor, raramente são encontradas em dicionários da língua geral. Quando o são, vêm em geral listadas no verbete do verbo, que é justamente a *incógnita* da colocação. Em português, por exemplo, dizemos *marcar uma consulta* ("make a doctor's appointment") ou *marcar um encontro* ("make an appointment with someone"). Mas também dizemos *marcar uma reunião*, que corresponde ao inglês "call a meeting". Em congressos, podemos *fazer uma comunicação* ou *apresentar*

um trabalho. enquanto em inglês temos a opção de "give a paper" (**dar um trabalho* é inaceitável em Português!).

4. Dicionários vs. *Corpora*

É evidente que há escassez de recursos lexicográficos fraseológicos. Os poucos dicionários desse tipo no mercado oferecem uma lista restrita de ocorrências. A título de exercício, busquei *computer* em três dicionários diferentes e dois *corpora*.

Por exemplo, o *BBI* (1993), no verbete *computer* lista as seguintes colocações nominais e adjetivas:

> 6. an *analog; desktop; digital; electronic; general-purpose; handheld; home; laptop; mainframe ~ ; […] parallel; personal; serial ~* (p. 72)

O *LTP Dictionary of Selected Collocations* (1997) lista apenas *home, laptop, mainframe, palmtop, personal ~* (p. 51).

A edição atualizada do *Longman's Dictionary of English Language and Culture* (1993) lista as seguintes colocações que começam com *computer*: *computer-aided design, computer dating agency, computer game, computer graphics, computer hacker, computer modelling, computer programmer, computer science* and *computer virus*.

Outra fonte foi o *English Collocations on CD-ROM* da Collins Cobuild (1995), uma ferramenta já pronta, baseada no corpus Bank of English, que apresenta 10.000 palavras pré-selecionadas com aproximadamente 20 colocados para cada uma . Uma rápida busca ofereceu a seguinte tabela para *computer***:**

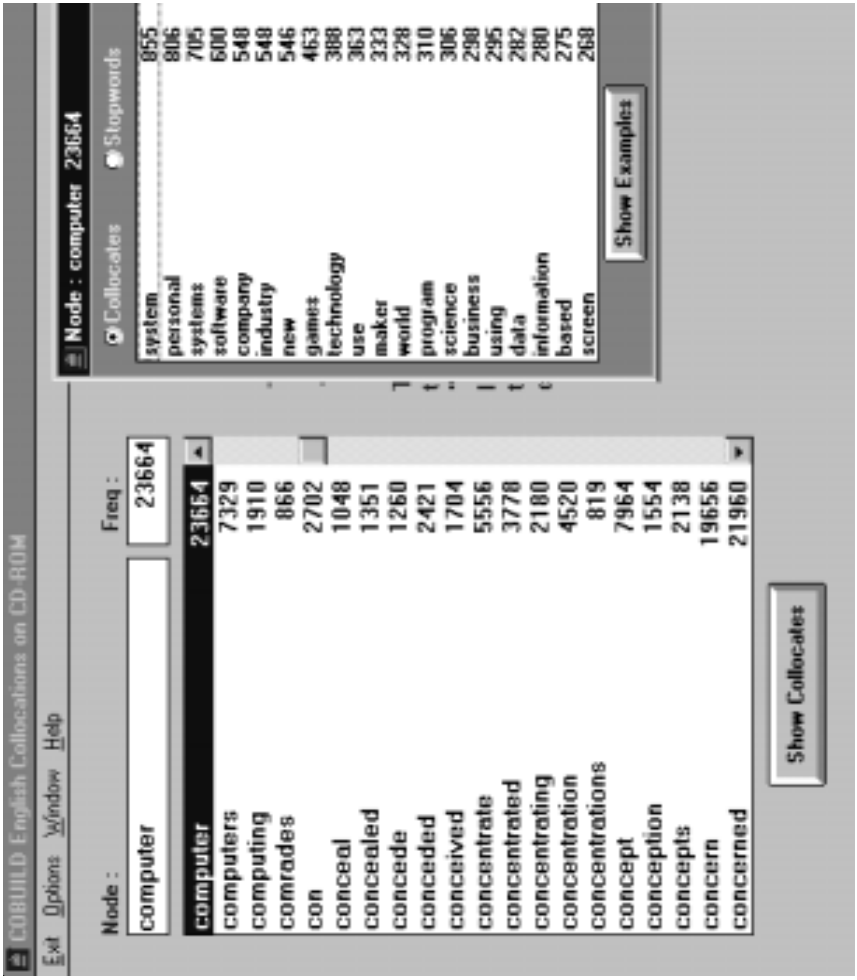

Figura 1: Tela para *computer* no English Collocations on CD-ROM

Ao acessar os exemplos para cada colocado, surgiram mais algumas colocações: *computer hardware, computer manufacturers, computer-products company, computer marketing research company, computer services company, computer software company, computer-security industry, computer video games, computer systems, computer workstations, computer-driven programs, computer-reservation, computer store, computer service business, computer-based information system, computer databases, computer information system, computer information network, computer-based graphics package, computer-based system* e *computer-based service*.

O que é de se estranhar é que, à exceção de *personal computer*, nenhuma dessas ocorrências é listada nos dicionários mencionados. Talvez não seja tão estranho se considerarmos que um dicionário somente incluirá palavras – e principalmente colocações – que já estiverem em uso há algum tempo. Um dicionário jamais conseguirá acompanhar o ritmo em que são criadas novas colocações.

Como último passo recorri ao WebCorp (http://webcorp.connect.org.uk/cgi-bin/webcorp[2]), um instrumento de busca *online* que usa a Web como corpus. Na época da busca para *computer* (2000), a ferramenta acessou 60 sites no Altavista e produziu 134 concordâncias. As colocações mais freqüentes foram: *computer systems* (9), *host computer* (8), *computer service* (7), *digital computer* (4), *electronic computer* (3), *computer hardware* (3), *computer store* (3) e 2 ocorrências para cada uma das colocações *computer keyboard, computer design, computer center, computer field, computer products, computer dealers* and *computer software*. Há também algumas ocorrências únicas, dentre as quais apenas as seguintes são listadas nos dicionários consultados:

BBI (1996)	LTP (1997)	LDELC (1993)
digital computer	*mainframe computer*	*computer game*
electronic computer		*computer programmer*

general-purpose computer *computer science*
mainframe computer

Vale salientar que quando algumas colocações são registradas nos dicionários já podem ter caído em desuso, como parece ser o caso de *personal computer*, que foi substituído por *PC*, ou então *desktop computer* reduzido simplesmente a *desktop* (plural *desktops*). Não foi encontrada nenhuma ocorrência para essas duas colocações entre as 134 concordâncias geradas pelo Webcorp. Isso demonstra que uma busca num corpus produzirá colocações que estão em uso, ou seja, um corpus não fornecerá apenas a forma correta, mas principalmente a forma mais usual na língua sob investigação.

A *língua* inglesa tem o privilégio de poder contar com dicionários de colocações como o BBI e o LTP, mas, para a maioria das línguas, ainda não existe essa fonte de referência. O quadro se agrava quando se trata de dicionários bilíngües. Segundo meu conhecimento, existe uma versão japonesa (Akimoto et allii 1993) e uma chinesa (Longman 1995) do BBI. Há também o Russian-English Dictionary of Verbal Collocations compilado por Benson & Benson (1993). Um dicionário semelhante de colocações verbais de inglês e português do Brasil, nas duas direções está sendo compilado no âmbito da Universidade de São Paulo (Tagnin 2000).

5. O projeto fraseológico

Do panorama acima descrito parece ficar claro que o recurso a *corpora* é um instrumento fundamental para garantir uma tradução em linguagem natural. Por essa razão, após meus alunos do Curso de Especialização em Tradução da Universidade de São Paulo (1º. semestre de 2000) serem apresentados ao componente convencional/ fraseológico da língua e conscientizados sobre os problemas que as

unidades fraseológicas podem causar na tradução, foi-lhes solicitado construírem um pequeno corpus de onde extrairiam todas as unidades fraseológicas possíveis (colocações, binômios etc.) e as apresentariam como trabalho final.

5.1 O processo de construção dos corpora

Os 48 alunos foram divididos em 11 grupos; cada um, por razões práticas, escolheu uma área de pesquisa. Essas áreas cobriam desde assuntos mais gerais como moda, culinária e beleza até áreas altamente especializadas como biotecnologia, finanças e computação. Dentro de cada área, optaram por um tópico mais específico, pois assim que começaram a coletar os textos, perceberam que era imprescindível delimitar melhor o campo de pesquisa, caso contrário, não teriam como manipular o vasto material que coletaram no seu entusiasmo inicial.

A partir daí, cada grupo escolheu um texto típico de sua área, no qual deveria identificar todas as ocorrências fraseológicas e tentar traduzi-las. Cada semana um grupo diferente apresentava seus resultados, em primeiro lugar para discutir se as unidades identificadas eram realmente fraseológicas, em segundo para garantir que tivessem encontrado uma tradução confiável. Por "confiável" entendíamos "natural" (ou "idiomática", como muitos estudiosos, a exemplo de Bowker, a denominam), uma tradução aceitável no sentido de ser a combinação que de fato era usada naquela área. Em outras palavras, caso o termo tivesse sido apenas encontrado num dicionário, teria de ser validado por uma ocorrência em contexto autêntico.

É nesse momento que se faz necessário um corpus. Cada grupo passou a construir um corpus de aproximadamente 200.000 palavras, 100.000 em cada língua. Durante o processo, ficou evidente que se tratava de um número muito ambicioso, embora alguns grupos tenham chegado bem perto desse objetivo. Os textos deveriam ser originais ou traduções e cada texto deveria ser identificado quanto

à fonte, língua e o fato de tratar-se de original ou tradução.[3]

5.2 Um desdobramento natural

Embora o curso não pretendesse enfocar especificamente a linguagem técnica, a maior parte das unidades fraseológicas caracterizava-se como termos técnicos dentro da área investigada. Foi isso que fez com que fosse sugerido aos alunos organizarem essas unidades sob a forma de glossário. Assim, além de construírem um corpus bilíngüe comparável, de proporções bem menores do que inicialmente proposto, cada grupo apresentou um glossário de 50 a 200 termos em cada língua. Os glossários apresentaram os termos equivalentes com exemplos autênticos em ambas as línguas. Não havia definições, pois não pretendia ser um recurso terminológico propriamente dito, isto é, um glossário definitório. Pretendia ser uma fonte de referência para o tradutor, oferecendo-lhe os termos técnicos, seus equivalentes e, acima de tudo, contextos de uso em ambas as línguas.

5.3 Avaliação geral

Como o projeto estendeu-se de forma não prevista, discutirei cada tarefa em separado.

5.3.1 Coleta dos equivalentes fraseológicos

Em relação ao objetivo inicial do projeto, houve consenso entre os grupos de que o experimento era extremamente válido por conscientizá-los para um aspecto da língua até então desconhecido para eles: o significado nem sempre é composicional; com freqüência as palavras adquirem seu sentido pela "companhia com que andam", conforme salientou Firth, ou seja, das palavras com que co-ocorrem. Em outras palavras, os alunos

1. conscientizaram-se da presença da convencionalidade/fraseologia

na língua, ou seja, perceberam que a língua tem um número muito grande de itens como as colocações;
2. aprenderam a identificar unidades fraseológicas, principalmente devido a sua recorrência;
3. compreenderam que as unidades fraseológicas do texto de partida deveriam, sempre que possível, ser traduzidas por unidades fraseológicas na língua alvo a fim de garantir uma linguagem natural;
4. deram-se conta de que os dicionários bilíngües são fontes de referência deficientes quando se trata de encontrar unidades fraseológicas equivalentes;
5. descobriram que mesmo um corpus de pequenas proporções, mas composto de textos criteriosamente selecionados, pode ser muito útil como fonte de equivalentes usuais.

5.3.2 Técnicas de construção de corpus

Devido às correções de percurso, os alunos sentiram falta de instruções mais detalhadas. Na realidade, a maioria teve de aprender as técnicas de construção de corpus da forma mais difícil: fora das aulas, com equipamento e recursos próprios, uma vez que nosso Departamento não conta com instalações informatizadas adequadas. No entanto, durante o processo

1. aprenderam que era preciso delimitar melhor a área de pesquisa e ser mais criteriosos na seleção dos textos. Como a maioria começou a coleta do material sem qualquer critério, logo perceberam que grande parte não era adequada para seus propósitos;
2. deram-se conta de que textos "tradicionais", caso fossem incluídos no corpus, deveriam ser digitados ou escaneados. Digitar exigia muito tempo e escanear apresentava problemas técnicos em termos de equipamento (quase ninguém tinha acesso a um *scanner*) ou de *software* para transformar a imagem em texto;

3. descobriram a riqueza da Web como fonte de textos em formato eletrônico, mas também perceberam que eram muito mais numerosos em inglês do que em português, o que, por vezes, os obrigou a recorrer a textos tradicionais;
4. deram-se conta, logo no início do projeto, de que era imprescindível saberem usar um computador, dominarem o Word e o Excel, e saber navegar na Internet. Além disso, tiveram de aprender a usar programas de busca como o Simple Concordance Program (que acharam lento demais), ou o IntraText, um serviço gratuito através do qual se envia um texto por e-mail, que é devolvido, dentro de poucos minutos, com diversas informações lexicais, inclusive colocações. O problema que se colocava, no entanto, é que os arquivos eram devolvidos compactados, o que exigia um *software* específico que poucos alunos tinham;
5. aprenderam a fazer buscas *online* nos *sites* do WebCorp e do BNC (British National Corpus) para confirmar certas colocações ou encontrar outras;
6. finalmente, aprenderam que era preciso identificar seus textos e dispô-los numa estrutura hierárquica para que pudessem consultá-los de acordo com suas necessidades. A estrutura foi-lhes fornecida.

Em suma, apesar de se queixarem de que o projeto era complexo demais para ser completado no período de um semestre, concordaram que foi uma experiência valiosa. Acima de tudo, porém, estavam certos de que as habilidades recém-adquiridas de busca e de construção de *corpora* eram instrumentos que poderiam evitar que atuassem como "tradutores ingênuos".

6. Projeto terminológico para tradutores

No ano seguinte (1º. semestre de 2001), foi ministrada a disciplina

Tradução Técnica. Parte da turma era composta pelos mesmos alunos que haviam trabalhado no projeto fraseológico. Foi sugerido que os grupos retomassem suas áreas de pesquisa, dessa feita com o objetivo explícito de elaborarem um glossário bilíngüe a partir de corpora de 200.000 palavras em cada língua que deveriam construir.

Houve uma diferença fundamental em relação ao desenvolvimento do projeto anterior. Face aos problemas ocorridos então

1. houve uma introdução explícita à noção de corpus, às etapas de sua construção e às ferramentas de busca. No caso, houve uma detalhada explicação sobre o uso da versão demo do Wordsmith Tools, ferramenta que fornece, a partir de textos pré-selecionados, concordâncias para a palavra de busca, *clusters* (agrupamentos freqüentes), listas das palavras mais freqüentes num texto, bem como palavras-chave de um texto. Por falta de um laboratório de informática, essas informações foram transmitidas por meio de transparências, cabendo aos alunos "porem a mão na massa" em casa;
2. como a maioria dos grupos se manteve, foram incentivados a melhor delimitar sua área de pesquisa. Assim, a área de Culinária ficou restrita ao tema de Temperos, a de Informática concentrou-se no aspecto da Segurança na Internet e assim por diante (vide abaixo lista completa);
3. foram discutidos os critérios de seleção dos textos: as fontes deveriam ser idôneas, de preferência acadêmicas, associações profissionais, revistas especializadas etc. Como a maioria dos textos seria extraída da Internet, onde já são encontrados em formato eletrônico, deveriam ser evitados *sites* comerciais porque, em geral, apresentam uma linguagem descuidada. Os textos deveriam ser completos, sem restrição quanto à extensão, para permitirem, no futuro, também pesquisas textuais e não apenas lexicais. Poderiam ser originais (de preferência) ou traduções;

4. com relação aos direitos autorais, foi redigido um pedido, em inglês e em português (anexos 1 e 2), que os alunos deveriam encaminhar aos autores dos textos que coletaram para obter permissão de inclui-los num corpus destinado exclusivamente à pesquisa. Nem sempre foi possível contatar o autor ou uma pessoa responsável, mas, mesmo assim, o resultado foi bastante promissor;
5. foi introduzido um cabeçalho para identificar cada texto quanto ao título, autor, tipo (original ou tradução), língua, local de publicação, data etc.;
6. foi definida uma estrutura para montagem do corpus de modo a facilitar a seleção de textos no momento da busca. Os textos foram organizados em três grupos: Inglês, Português e Paralelos. Dentro dos dois primeiros foram subdivididos em Originais e Traduções. Observe-se que essas traduções não eram traduções dos originais; eram textos independentes, que foram encontrados apenas na forma traduzida. O terceiro grupo, o Paralelo, que constava de originais e suas respectivas traduções, foi subdividido de acordo com a direção da tradução: inglês-português ou português-inglês. Cada texto foi gravado num arquivo. Para sua correta identificação o nome deveria ser imediatamente seguido de uma das seguintes siglas:

   IO = inglês – original
   IT = inglês – tradução
   PO = português – original
   PT = português – tradução
   PIPIO = paralelo inglês-português: inglês – original
   PIPPT = paralelo inglês-português: português – tradução
   PPIPO = paralelo português-inglês: português – original
   PPIIT = paralelo português-inglês: inglês – tradução.

   Assim, por exemplo, SafetyIO é um texto original em inglês identificado com o nome de Safety, SoyPIPIO é o original em

inglês de um texto identificado como Soy que, por constar de um corpus paralelo, possui uma tradução para o português identificada como SoyPIPPT;

7. todos os alunos usaram o Wordsmith Tools para suas buscas, ao contrário do que ocorreu no projeto anterior, em que apenas alguns conseguiram utilizar ferramentas de busca informatizadas como o Simple Concordance Program ou o IntraText;
8. os termos que comporiam o glossário não se restringiram a unidades fraseológicas, no caso, colocações. Foram também incluídos termos monolexêmicos, justamente por se tratar de um glossário técnico. No entanto, ao contrário da maioria dos glossários, que tende a se restringir a substantivos ou sintagmas nominais, os glossários em questão, sempre que possível, incluíram colocações verbais específicas da área. Isso se deve ao fato de que nosso objetivo era compilar um glossário para o tradutor, ou seja, um glossário de produção, não apenas de compreensão.

   Talvez essa diferença mereça uma explicação. Em primeiro lugar, o glossário não era definitório, ou seja, não constava de um termo com sua respectiva definição. Na realidade, apresentava apenas o termo e um exemplo autêntico em cada língua. Como na maioria dos casos os termos não eram extraídos de textos paralelos (um original e sua respectiva tradução), os exemplos não eram equivalentes. Eram apenas ilustrativos do uso do termo em seu contexto usual. Isso é de suma importância para o tradutor, pois lhe fornece o ambiente natural de ocorrência do termo, por exemplo, se ocorre com ou sem artigo, com ou sem preposição, se co-ocorre regularmente com outra palavra etc.;
9. para assegurar uma apresentação padronizada, foi elaborado um programa na plataforma Access, cujo produto final tem o seguinte formato:

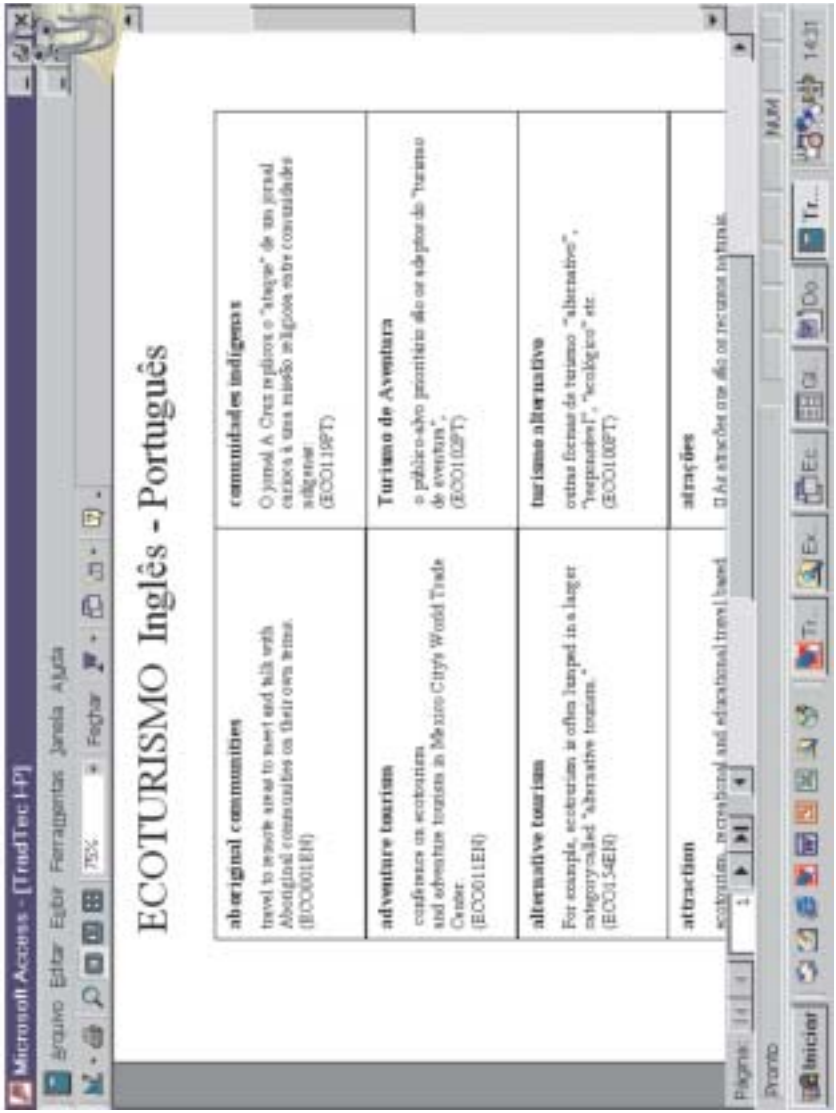

Figura 2: Glossário inglês-português para a área de Ecoturismo, gerado pelo Access

Ao final do projeto obtivemos os seguintes corpora especializados com aproximadamente 200.000 palavras em cada língua:

Biotecnologia: alimentos transgênicos
Moda: roupas
Culinária: temperos
Informática: segurança na Internet
Veterinária: doenças dos bovinos
Ecologia: biodiversidade
Odontologia: ortodontia
Automação Industrial: sensores
Finanças: mercado de ações
Turismo: ecoturismo

Cada um desses corpora produziu um glossário bilíngüe. Esses glossários estão sendo disponibilizados no *site* do CITRAT (Centro Interdepartamental de Tradução e Terminologia) http://www.fflch.usp.br/citrat/, na seção de Glossários, Tradução.

Para que os alunos pudessem compartilhar os *corpora*, esses foram gravados num CD-ROM, juntamente com a versão demo do Wordsmith Tools. Por questões de direitos autorais, já que ainda não foi possível obter autorização para todos, o CD-ROM destina-se apenas ao "uso interno", ou seja, para as pesquisas dos alunos do CETRAD.

Nosso próximo passo é incorporar todos esses textos num corpus maior intitulado COMET – Corpus Multilíngüe para Ensino e Tradução, que está sendo construído no âmbito do CITRAT, na Universidade de São Paulo. O COMET abrigará todos os corpora construídos, tanto pelos alunos do CETRAD, quanto por pós-graduandos dos diversos departamentos da Faculdade, nas várias línguas lá oferecidas. Dentre as várias áreas técnicas, três merecerão atenção especial: Direito Comercial, Informática e Odontologia, o que significa que serão ampliadas de forma sistemática. Esse corpus, quando construído, será disponibilizado via Internet (Tagnin 2002).

6.1. Avaliação

Ao final do segundo projeto, os alunos estavam perfeitamente conscientes e convencidos da relevância de *corpora* técnicos para o trabalho do tradutor, principalmente face à falta de obras de referência especializadas em diversas áreas. Vale ressaltar, no entanto, que, mesmo que essas obras existam no mercado, ainda assim o corpus oferecerá uma visão mais atualizada da linguagem em questão, fornecendo, quando conveniente, indicações quanto à freqüência de uso de determinado(s) vocábulo(s) e, principalmente, apresentando a palavra buscada num contexto de uso real, juntamente com as palavras com que usualmente co-ocorre, isto é, seus colocados. Essa informação poderá ativar o conhecimento passivo do consulente, confirmando suas intuições e permitindo-lhe produzir um texto mais natural, o que evidencia também o aspecto didático de uma consulta a corpus. Além do mais, a experiência conferiu aos alunos alto grau de autonomia uma vez que podem, sempre que necessário, construir, em relativamente pouco tempo, um corpus que atenda a suas necessidades.

7. Conclusão

Relatamos dois experimentos de construção de corpora por alunos de tradução como fonte de referência para suas tarefas tradutórias. O primeiro – projeto fraseológico – configurou-se como um experimento oportunista, mas produziu pequenos corpora e glossários fraseológicos em diversas áreas técnicas. No segundo – um projeto de glossários técnicos para tradutores - , o desenvolvimento foi mais metódico, com uma apresentação formal das noções básicas de corpus e das ferramentas de busca, e resultou em trabalhos mais consistentes. Além dos produtos obtidos, o processo vivenciado pelos alunos conscientizou-os para a relevância do uso de corpora na tradução, principalmente pela possibilidade

de encontrarem o "termo correto" num contexto autêntico de uso, o que lhes fornece dados para empregá-lo, em suas traduções, de forma natural e fluente.

Notas

1. Esta é uma versão ampliada e atualizada de um artigo que deverá ser publicado, em inglês, sob o título "*Corpora* and the Innocent Translator: how can they help him" nos anais de *The Lodz Session of the 3rd International Maastricht-Lodz Duo Colloquium on „Translation and Meaning"* realizado em Lodz (Polônia), 22 – 24 de setembro de 2000.

2. A URL atual é www.webcorp.org.uk .

3. Trabalhos semelhantes foram desenvolvidos por Maia (2000) e Varantola (2001). Veja também artigos das autoras neste volume.

Referências

Akimoto, S., A. Baba & T. Ogura (eds.) (1993). *BBI Eiwa Rengo Katsuyo Jiten*, Toquio: Maruzen.

Benson, Morton & Evelyn Benson (1993). *Russian-English Dictionary of Verbal Collocations*, Amsterdam/Philadelphia: John Benjamins.

Benson, Morton, Evelyn Benson & Robert Ilson (1993 [1986]). *The BBI Dictionary of English Word Combinations*, Amsterdam/Philadelphia: John Benjamins.

Boatner, Maxine Tull & John Edward Gates (1975). *A Dictionary de American Idioms*, Woodbury, N.Y.: Barron's Educational Series.

Bowker, Lynn (1998). "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: a Pilot Study", *Meta XLIII*, 4, pp 631-651.

Camargo, Sidney & Martha Steinberg (1989). *Dicionário de Expressões Idiomáticas Metafóricas Português-Inglês*, São Paulo: Editora Pedagógica e Universitária.

Camargo, Sidney & Martha Steinberg (1990). *Dictionary of Metaphoric Idioms English-Portuguese*, São Paulo: Editora Pedagógica e Universitária.

Cowie, A.P., R. Mackin & I. R. McCaig (1983). *Oxford Dictionary of Current Idiomatic English*, Oxford: Oxford University Press.

Fillmore, Charles J. (1979). "Innocence: A Second Idealization for Linguistics", *Berkeley Linguistic Society 5*, pp 63-76.

Hausmann, Franz Josef (1985). "Kollokationen im deutschen Wörterbuch - ein Betrag zur Theorie des lexikographischen Beispiels". In Bergenholtz, Henning & Joachim Mugdan (eds.). *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch* (Lexicographica Series Maior 3), pp 118-129.

Hill, Jimmie & Michael Lewis (1997). *LTP Dictionary of Selected Collocations*, Hove: Language Teaching Publications.

Longman Dictionary of English Collocations. (1995). Hong Kong: Longman Asia.

*Longman Dictionary of English Language and Culture***,** (1993 updated [1992])**,** London: Longman.

Maia, Belinda (1997)**. "**Making corpora – a learning process"**.** Comunicação apresentada no *CULT 97*, disponível no site www.sslmit.unibo.it/cultpaps/paps.htm.

Maia, B. (2000). "Making corpora – a learning process", in Bernardini, S. & F. Zanettin, (eds). 2000, pp 46-7.

Partridge, Eric (1977). *A Dictionary of Catch Phrases*, London: Routledge & Kegan Paul.

Serpa, Oswaldo (1982). *Dicionário de Expressões Idiomáticas Inglês-Português / Português-Inglês*, Rio de Janeiro: MEC/FENAME.

Spears, Richard A. (1988). *NTC's American Idioms Dictionary*, Lincolnwood, Ill.: National Textbook Company.

Spears, Richard A. (1989). *NTC's Dictionary de American Slang and Colloquial Expressions*, Lincolnwood, Ill.: National Textbook Company.

Spears, Richard A. (1996). *Common American Phrases in Everyday Contexts***,** Lincolnwood, Ill.: National Textbook Company.

Spears. R. A. et allii. (1995) *NTC's Dictionary of Everyday American English Expressions*, Lincolnwood, Ill.: National Textbook Company.

Tagnin, Stella E. O. (2000) "Collecting data for a bilingual dictionary of verbal collocations: From scraps of paper to corpora research". In Lewandowska-Tomaszczyk, B & Melia, P.J. (eds.*) PALC '99: Practical Applications in Language Corpora*. *Articles from the International Conference at the University of Lodz, 15-18 April 1999***;** Frankfurt am Main: Peter Lang GmbH, 399-407.

Tagnin, Stella E. O. (2002). "Taking off in Brazil: COMET – A multilingual corpus for teaching and translation", comunicação apresentada em The 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English - ICAME 2002 – The Theory and Use of Corpora, Gotemburgo, Suécia, 22-26/05/2002.

Varantola, Krista (2001) "Disposable" corpora as translation tools" , palestra proferida no II Seminário sobre Estudos de *Corpora* – Perspectivas para a Tradução, USP/SP, 31/07-02/8/2001.

ANEXO 1

Profa. Dra. Stella E. O. Tagnin

Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Letras Modernas
Tel. 3091-4296
Av. Prof. Luciano Gualberto, 403
Fax: 3032-2325
05508-900 – São Paulo – SP
E-mail: seotagni@usp.br

< Data >

< Endereço >

Prezados Senhores

Na qualidade de coordenadora do projeto COMET, um Corpus Multilíngüe para Ensino e Tradução, para o qual conto com Bolsa-Pesquisa do CNPq (Processo 301020/91-4), venho solicitar-lhe autorização para incluir o texto ..... no referido corpus.

O corpus constará de textos de especialidade em inglês e português, tanto originais quanto traduções, para servirem de fonte de pesquisa para estudos nas áreas de tradução e ensino de línguas. Os textos serão tratados de forma a possibilitar uma busca eletrônica, ou seja, acesso às informações via computador.

Um dos principais objetivos desse corpus é constituir-se como uma base de textos fidedigna para ensejar trabalhos que possam vir

a contribuir para melhorar o ensino e a aprendizagem de línguas (tanto materna quanto estrangeiras), a prática da tradução e a confecção de obras lexicográficas específicas (dicionários e glossários).

A informação bibliográfica completa dos textos selecionados para o corpus será facultada aos seus usuários e na apresentação do corpus haverá um agradecimento individual a todos os autores, tradutores e editoras participantes.

Sua colaboração é importante para este projeto. Se concordarem com a utilização do texto acima referido, peço a gentileza de assinar a carta de autorização em anexo e remetê-la ao endereço indicado.

Este corpus é um projeto acadêmico e não tem quaisquer fins lucrativos. Caso desejem maiores informações, coloco-me a sua inteira disposição para quaisquer esclarecimentos que se façam necessários.

Esperando poder contar com sua colaboração, subscrevo-me

atenciosamente,

Profa. Dra. Stella E.O. Tagnin

COMET
CORPUS MULTILÍNGÜE PARA ENSINO E TRADUÇÃO
Coordenação científica: Profa. Doutora Stella E. O. Tagnin
Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Letras Modernas
Av. Prof. Luciano Gualberto, 403
05508-900 – São Paulo – SP

AUTORIZAÇÃO

Texto: _____

Concedo autorização a Stella E. O. Tagnin para introduzir o(s) texto(s) em epígrafe no COMET - CORPUS MULTILÍNGÜE PARA ENSINO E TRADUÇÃO, que está sendo criado no âmbito da Universidade de São Paulo, na condição de ser disponibilizada aos usuários do corpus a informação bibliográfica completa do(s) mesmo(s) e de constar, na apresentação do corpus, um agradecimento à minha colaboração.

Data:

Assinatura: _____

(nome por extenso)

ANEXO 2

Profa. Dra. Stella E. O. Tagnin

Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Letras Modernas
Tel. 3091-4296
Av. Prof. Luciano Gualberto, 403
Fax: 3032-2325
05508-900 – São Paulo – SP E-mail: seotagni@usp.br

< Date >

< Address >

Dear Sir/Madam

I am a professor at the Department of Modern Languages of the University of São Paulo, in São Paulo, Brazil, where I am currently coordinating the construction of a multilingual corpus for teaching and translation purposes, the **COMET** (**Co**rpus **M**ultilingüe para **E**nsino e **T**radução) with a research grant from one of our leading state funding agencies (CNPq Process 301020/91-4). In that capacity I would like to ask for your permission to include the following text in our corpus:

< Title of text >

The **COMET** will consist of technical texts in English and Portuguese, in their original or translated forms, to be used as research sources for studies in language teaching and translation. The texts will be stored electronically so that they can be used in the automatic retrieval of information for Portuguese-English language contrasts.

It is meant to be a reliable source of authentic texts to be used in research that will eventually improve language teaching and acquisition, the practice of translation, as well as furnish material for compiling specific dictionaries and glossaries.

Complete bibliographical information of the selected texts will be available to all users and there will be special mention to the contributing authors, translators and publishers in the preface to the corpus. This is an academic research project and is not being undertaken for commercial gain.

Your cooperation will be greatly appreciated. If you are able to give permission to include the above text(s) in our corpus, I would be grateful if you could sign the attached permission form and return it to me at the address indicated.

Finally, please do not hesitate to contact me if you require any further clarification about this project.

Yours sincerely,

Dr. Stella E. O. Tagnin
Coordinator

COMET

CORPUS MULTILÍNGÜE PARA ENSINO E TRADUÇÃO

Coordenação científica: Profa. Doutora Stella E. O. Tagnin
Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Letras Modernas
Av. Prof. Luciano Gualberto, 403
05508-900 – São Paulo – SP

PERMISSION

< Text >

Permission is hereby granted to include the above text(s) in the Multilingual Corpus for Teaching and Translation purposes - **COMET** - on condition that the users of the corpus are provided with the full bibliographical reference to the texts in question and that a personal acknowledgement to the author is included in the preface to the corpus.

Date:

Signature: