APLICAÇÕES LÉXICO-TERMINOGRÁFICAS DA LINGÜÍSTICA DE CORPUS: RELATO DA ELABORAÇÃO DE UM GLOSSÁRIO BILÍNGÜE DE COLOCAÇÕES NA ÁREA DE NEGÓCIOS

Adriane ORENHA (Universidade de São Paulo)
adrianeorenha@terra.com.br

ABSTRACT: *Specialized corpora have recently constituted an extremely valuable source for lexicographers and terminographers. Based on the report of an experience, this article aims at discussing some lexico-terminological applications of Corpus Linguistics to the compilation of a corpus-based bilingual Glossary of Business Collocations.*

KEYWORDS: *terminology, collocation; glossary; corpus; phraseology.*

0. Introdução

A *compilação* de um *glossário* ou *dicionário* é uma tarefa bastante árdua. Entretanto, essa tarefa se torna ainda mais complicada quando se tenciona compilar uma obra fraseológica, como no caso, de *colocações*, haja vista a dificuldade que se tem para buscar, extrair e, posteriormente, verificar o uso das *colocações* levantadas (Orenha, 2002).

Num passado não muito remoto, essa busca era feita de maneira manual. Com o advento da *Lingüística de Corpus* e o uso e exploração de *corpora* como metodologia de pesquisa, o levantamento de palavras e *colocações* se tornou maior e muito mais rápido. Além disso, disponibiliza ao *lexicógrafo/ terminógrafo* um contexto muito mais amplo, imprescindível para uma pesquisa sobre *colocações*.

Segundo Pearson (2000), *glossários* produzidos a partir de recursos eletrônicos são, em linhas gerais, mais confiáveis em relação àqueles compilados utilizando-se métodos mais convencionais.

Tendo em vista tais considerações e, baseado no fato de haver uma grande escassez de obras fraseológicas no Brasil, decidimos compilar um *glossário* bilíngüe (Inglês-português/ Português-inglês) de *colocações*, baseado em um *corpus especializado comparável bilíngüe* da área de *negócios*.

Neste artigo, vimos relatar nossa experiência em compilar o *glossário* proposto, assim como ressaltar as vantagens de fazer uso da *Lingüística de Corpus* como metodologia para o levantamento e seleção das *colocações*.

1. Perspectiva Teórica

1.1. A relevância das c*olocações* para a fluência na tradução e em uma língua estrangeira

Quando aprendemos uma língua estrangeira (doravante LE), nos é dito que as regras gramaticais são essenciais para o aprendizado e, quanto ao vocabulário, este é ensinado de maneira segmentada e artificial. Não é à toa que o discurso de muitos dos aprendizes é cortado e interrompido a todo o momento, sendo sua comunicação nem rápida, nem efetiva. Esse problema de comunicação pode até mesmo comprometer transações de negócios, pelo simples motivo de não seguir regras composicionais e convencionais ditadas pela língua.

Acreditamos que, se um aprendiz de uma LE desejar que seu discurso seja mais rápido e fluente, esse discurso tem que estar de acordo com as regras de aceitabilidade da comunidade na qual essa língua está inserida, ou seja, seu discurso tem que estar em conformidade com as convenções ditadas por essa comunidade. Achar que o discurso seja baseado apenas em um conjunto de regras sintáticas e gramaticais, e que o aprendiz, tendo domínio dessas regras gramaticais, apenas escolhe, de maneira aleatória, os elementos lexicais para compor, no caso, as *colocações*, é no mínimo lacunoso.

Segundo Bénjoint (1994), "*as palavras têm 'características embutidas' e, assim, a escolha de uma palavra, ou até um sentido específico de uma palavra, acarreta necessariamente na escolha obrigatória ou preferencial de outras palavras, ou de alguma construção sintática*".

Dessa maneira, julgamos que não devamos fazer o ensino do léxico de maneira descontextualizada e, no caso das *colocações*, de maneira deslexicalizada, isolando seus constituintes que, a rigor, só têm valor no conjunto da *colocação*. Devemos sim, ensiná-las em blocos, pois é muito mais fácil ter que desmembrar esses blocos futuramente, do que combinar essas palavras para formar esse mesmo bloco. De acordo com McCarthy, "a maioria das palavras na língua vem em séries pré-embaladas, que mostram um número limitado de padrões, em oposição à clássica noção lingüística de que a língua consiste de uma série de 'aberturas/brechas' sintáticas ('*slots*') dentro das quais, itens lexicais podem ser inseridos (McCarthy, 1988b:56, apud Bénjoint, 1994:211, tradução minha).

Recentes pesquisas em lingüística, mais especificamente em *Fraseologia* (Fillmore, 1979; Kjellmer, 1991; Lewis, 1997; dentre outros), têm mostrado que a produção em uma LE, seja ela oral ou escrita, e a fluência nessa mesma língua, está intimamente ligada à capacidade do aprendiz de produzir *unidades fraseológicas* cada vez mais complexas. Na tipologia dessas unidades, podemos incluir as *colocações* (Orenha, 2002:1).

A essa fluência no aprendizado também acrescentamos a fluência na tradução. Para que um tradutor seja fluente e para que suas traduções e versões sejam um reflexo da linguagem em uso, é necessário que o mesmo tenha consciência do fator convencionalidade na língua e transponha esse conhecimento em suas traduções.

1.2. Definição e taxonomia das *colocações*

Dentro da esfera da *fraseologia* inserem-se as *colocações*, uma de suas partes mais relevantes, e também consideradas a maior classe *de unidades multi-palavras*.

O termo '*collocation*' foi pela primeira vez empregado por J. R. Firth para designar casos de co-ocorrência léxico-sintática, ou seja, palavras que usualmente andam juntas (Apud Tagnin, 1989:30).

Tagnin (1999) afirma que, para uma combinação ser qualificada como *colocação*, ela precisa apresentar as seguintes características:

1. Recorrência;
2. Não-idiomaticidade, ou seja, seu significado tem que ser composicional;
3. Coesão – é necessário que haja uma ligação muito forte entre seus elementos, muito mais forte do que se esperaria de uma combinação qualquer;
4. Restrição contextual – deve haver uma probabilidade de que ocorram dentro de um contexto específico;
5. Co-ocorrência arbitrária entre seus elementos, ou seja, não há razão semântica que explique tal ocorrência.

Em relação a sua taxonomia, as *colocações* podem ser classificadas da seguinte maneira:

- **Verbal**
  - Verbo $_{\textbf{colocado}}$ + Substantivo $_{\textbf{base}}$: *set up a business*
  - Substantivo $_{\textbf{base}}$ + Verbo $_{\textbf{colocado}}$: *the economy shrinks*
  - Verbo $_{\textbf{colocado}}$ + Preposição + Substantivo $_{\textbf{base}}$: *Drop out of a business*
  - Verbo $_{\textbf{colocado}}$ + Adjetivo $_{\textbf{base}}$: *grow serious*
- **Nominal**
  - Substantivo $_{\textbf{base}}$ + Substantivo $_{\textbf{colocado}}$: *stock brokers*
  - Substantivo $_{\textbf{base}}$ + Preposição + Substantivo $_{\textbf{colocado}}$: *a decline in value*
- **Adverbial**
  - Advérbio $_{\textbf{colocado}}$ + Adjetivo $_{\textbf{base}}$: *outrageously high*
  - Verbo $_{\textbf{base}}$ + Advérbio $_{\textbf{colocado}}$: *to rise sharply, to invest heavily*
  - Advérbio $_{\textbf{colocado}}$ + Verbo $_{\textbf{base}}$: *sharply attack*
- **Adjetiva**
  - Adjetivo $_{\textbf{colocado}}$ + Substantivo $_{\textbf{base}}$: *a declining economy; a low cost*

2. As contribuições da *Lingüística de Corpus* para a *Lexicologia* e *Terminologia*

Conforme já mencionamos, a *Lingüística de Corpus* (doravante *LC*) veio a revolucionar a prática lexico-terminográfica. Segundo Teubert (2001), a *Lexicografia* é o segundo maior campo no qual a *LC* não apenas introduziu novos métodos, mas também estendeu o escopo de pesquisa.

Graças à *LC*, podemos mais facilmente, e de maneira mais rápida e eficiente, levantar e selecionar não apenas palavras, mas também combinações de palavras. Por meio de um *corpus*, podemos analisar essas combinações em seus contextos naturais e, principalmente, com um contexto muito maior à disposição do lexicógrafo/terminógrafo, que antes dependia de uma árdua busca manual. De acordo com Teubert (2001), "a tradicional prática lexicográfica de descontextualização e isolamento de palavras nos impedia de saber o significado de unidades mais amplas". Esse fato pode ser explicado simplesmente por restrições de espaço no passado: era impossível listar todas as *colocações* e frases fixas mesmo em um dicionário de vários volumes. Quando se tratava de *colocações*, por exemplo, a pesquisa se tornava inviável, posto que o significado de unidades múlti-palavras é bem mais específico do que palavras simples e isoladas.

Uma outra enorme vantagem advinda da *LC* para as áreas de *Lexicografia* e *Terminografia* foi as análises estatísticas, a saber *T-score*, *Mutual Information*, dentre outras, permitidas pelo uso de ferramentas de pesquisas específicas, como por exemplo o *WordSmith Tools* (Scott, 1997), também utilizado nesta pesquisa. Sem a possibilidade de contar com uma ferramenta como essa, a tarefa de compilação de um *glossário* ou *dicionário* seria muito maçante, sem dizer inviável. Ademais, como garantir a confiabilidade do resultado? Como saber se as combinações, tidas como *colocações*, não são simplesmente a somatória de um elemento A + B?

Neste caso, a *Lexicologia* tradicional fracassaria em chegar a uma definição plausível de *colocação*. Recorrendo à definição já acima mencionada, para que uma combinação seja considerada uma *colocação*, é necessário que a mesma seja recorrente e que a combinação de seus elementos seja co-ocorrente. Isso somente é possível com o uso de um *corpus* suficientemente grande, já que há a possibilidade de, por meios estatísticos, detectar uma co-ocorrência significativa desses elementos.

3. Metodologia

Para fazer o levantamento e extração das *colocações*, foi construído um *corpus* especializado comparável bilíngüe da área de *negócios*, o qual passou a ser nosso *corpus de estudo* e que, no presente momento, possui 560.000 palavras.

Foram criados dois *subcorpora*:

- Um de inglês, com textos dos jornais *The New York Times* e *Financial Time*, e das revistas *Time* e *Businessweek*;
- Um de português, com textos dos jornais *Gazeta* e *Estado de São Paulo* e das revistas *Veja* e *Exame*.

Para auxiliar na extração das *colocações* foi utilizado o programa *WordSmith Tools*.

A partir de uma lista de palavras-chave, obtida por meio da comparação da freqüência da lista de palavras de nosso *corpus de estudo* com a freqüência da lista de palavras de nosso *corpus de referência* (*British National Corpus Sampler*, doravante *BNC*, com aproximadamente 2 milhões de palavras), demos

início ao levantamento das palavras-chave de conteúdo para, posteriormente, extrair as *colocações* por meio da ferramenta *Concord* do *WS Tools* e pela análise da lista de *Collocates* e *Clusters*.

A princípio, foram selecionadas apenas as *colocações* com freqüências acima de 4, pelo motivo de ser esta a freqüência mínima para que o *Mutual Information* possa ser calculado. Porém, começamos a nos deparar com determinadas combinações que, de acordo com as análises estatísticas, não seriam consideradas como tais, haja vista sua baixa freqüência no *corpus* de estudo (inferior a 4). No entanto, não desejávamos descartá-las, dada sua importância para a área de *negócios*. Em razão disso, tomamos a resolução de buscá-las em um *corpus* bem maior – o *BNC Online* – considerando a impossibilidade de se aumentar nosso *corpus* de estudo por motivo de tempo. Neste caso, o *BNC* só foi utilizado como forma de comprovação de que se tratava realmente de uma *colocação*.

A título de exemplificação, analisemos a *colocação 'business acumen'*. A experiência nos dizia que se tratava de uma combinação bastante usual na área de *negócios*. Porém, ela só havia aparecido três vezes em nosso *corpus de estudo*:

1 Rempt and partner Hans van Bennekom want to use their **business acumen** and contacts to save the companies in venture firms' portfolios.
2 it's part of a multibillion-dollar entertainment conglomerate whose programming decisions are based on sober **business acumen**.
3 You'll need to convince the admissions officers that the road you chose will add to the school's culture and more than make up for any quantitative skills or **business acumen** you may lack.

Entretanto, ao buscarmos somente a palavra '*acumen*', notamos que ela ocorria 4 vezes, e destas 4 ocorrências, três exemplos são com a combinação '*business acumen*'. Ao buscarmos essa mesma combinação no *BNC Online*, observamos que ela ocorre 47 vezes, sendo que somente a palavra '*acumen*' ocorre 111 vezes e '*business*' 35.141.

Assim, quando constatamos a recorrência dessas *colocações* no *BNC*, optamos por inseri-las no *glossário*. Dessa forma, não acreditamos estar invalidando nossa pesquisa que afirma ser norteada pela *LC* e que, portanto, deve ser empiricamente comprovada. Estamos selecionando *colocações* recorrentes, seja por meio de nosso *corpus de estudo* - que é a base de onde os dados são retirados – seja por nosso *corpus de referência*. Enfatizamos que nosso *corpus de estudo* é o provedor dessa seleção.

Vale lembrar que foram criadas fichas terminológicas e que as *colocações* selecionadas serão inseridas no programa de banco de dados *Access*, da Microsoft (2002).

4. Projeto de Iniciação Científica

Com o propósito de agilizar nossa pesquisa, decidimos por elaborar um projeto de iniciação científica envolvendo um grupo de alunos do curso de Tradutor e Intérprete de uma dada faculdade onde lecionamos. Essa decisão se deu por ter consciência de que, apesar de a metodologia aplicada facilitar muito o trabalho do terminógrafo, conforme anteriormente mencionado, o trabalho a ser realizado é muito extenso e, além disso, requer muita análise e interpretação dos

dados coletados e, dessa forma, uma equipe viria a contribuir muito para atingir seu êxito. Ademais, não podemos nos esquecer de que, por estarmos compilando um *glossário* bilíngüe, temos o problema da equivalência na língua de chegada, tornando a pesquisa ainda mais trabalhosa.

Um outro objetivo desse projeto é engajar os alunos em uma pesquisa a qual utilizará *corpora on-line* como instrumento de pesquisa para a busca de possíveis traduções, contribuindo, assim, para sua fluência na tradução e preparando-os melhor para o mercado de trabalho, seja qual for a área de conhecimento que terão que atuar. Para tanto, os alunos envolvidos têm como tarefa buscar possíveis traduções para as *colocações* previamente levantadas pela professora-pesquisadora com o auxílio da ferramenta de busca *WS Tools*, a partir de seu *corpus de estudo* da área de *Negócios*. Para realizar essa tarefa, contarão com a ajuda do *BNC*, do emulador de concordenciado *Webcorp* e dicionários monolíngües e bilíngües indicados pela coordenadora do projeto.

Com esse projeto, também visamos à divulgação de uma metodologia bastante útil à área de Tradução e que poderá ser compartilhada por outros colegas do curso de Tradutor e Intérprete.

Além disso, como resultado concreto desse projeto, objetivamos que as traduções sugeridas para as *colocações* levantadas pela professora-pesquisadora sejam inseridas no *Glossário Bilíngüe de Colocações, na área de Negócios, baseado em Corpus Comparável Bilíngüe Eletrônico*, de sua própria autoria.

5. A macro e micro estruturas do *glossário*

A definição da macroestrutura e, principalmente da microestrutura, é de fundamental importância para a compilação de um *glossário*, pois segundo Barbosa (1989), uma vez adotado um programa para a microestrutura de uma obra lexicográfica, teremos que sustentá-lo ao longo de toda obra, caso contrário, correremos o risco de empobrecer a qualidade da obra lexicográfica. Entretanto, cabe lembrar que teceremos apenas algumas considerações acerca desses itens nesse artigo, tendo em vista a complexidade do assunto.

No intuito de assegurar um tratamento sistemático das *colocações* em nosso *glossário*, e levando em consideração que se trata de um *glossário* destinado à codificação, as *colocações* serão inseridas pela *base*, subtendendo-se "aquilo já conhecido do consulente" (Haussman, 1985, apud Tagnin, 1998).

Sendo assim, como o *glossário* pretendido trará todos os tipos de *colocações* – verbais, nominais, adjetivas e adverbiais –, a base ora será (Orenha, 2002):

- ✓ um **substantivo** (nas *colocações nominais*, *adjetivas* e *verbais*);
- ✓ um **adjetivo** (em um dos dois tipos de *colocação adverbial* – *increasingly irrelevant; to grow serious*);
- ✓ um **verbo** (em um dos dois tipos de *colocação adverbial* – to *grow enormously; sharply attack*)

A fim de facilitar a busca por parte do consulente, elas serão arranjadas da seguinte maneira:

- ✓ **Quando a *base* for um substantivo**:
  - Verbo (preposição) + Base (*colocação verbal*)
  - Base + Verbo (*colocação verbal*)
  - Substantivo + Base (*colocação nominal*)
  - Base + Substantivo (*colocação nominal*)
  - Quantificador + Substantivo (*colocação nominal*)
  - Adjetivo + Base (*colocação adjetiva*)

- ✓ **Quando a *base* for um adjetivo**:
  - Advérbio + Base (*colocação adverbial*)
  - Verbo + Base (*colocação verbal*)

- ✓ **Quando a *base* for um verbo:**
  - Base + Advérbio (*colocação adverbial*)
  - Advérbio + Base (*colocação adverbial*)

Quando a tradução não for uma *colocação*, o verbete trará um símbolo ☞ que precederá uma tradução apenas explicativa (Tagnin 1998). Ex.:

- Face $_{n}$.
  set one's ~ against ☞ Opor-se a, desaprovar

Quanto ao paradigma definicional, quando o termo tiver mais de um semema/significado, ou seja, quando se tratar de um termo polissêmico, ele terá uma definição entre parênteses, havendo, assim, entradas distintas para cada semema:

- Share $_{n}$ [$_{usu\ pl}$.] (units of capital stock) Ações $_{subst.}$ (de empresas)
- Share $_{n}$ (person's part in something done, received, etc. by several people) Parte $_{subst}$, porção $_{subst}$

6. Conclusão

São unânimes as opiniões quanto à enorme contribuição que a *LC* trouxe para a *Lexicografia* e *Terminografia* (Teubert 2001*;* Sinclair, *1991;* Pearson, 2000 *etc)*. De acordo com Verlinde and Selva (2001, apud Grefenstette, 2002): "it's commonly admitted that corpus-based lexicography gives a strong and necessary empirical evidence to the lexicographer's personal intuition".

Nossa experiência na compilação de um *glossário* bilíngüe de *colocações* a partir de *corpora eletrônicos* vem a ratificar essas opiniões. O uso da referida metodologia tornou possível e real a extração de um número bastante grande de *colocações*, não apenas pelo uso de análises estatísticas, mas também pelo acesso a contextos bastante amplos e somente possíveis devido ao uso e exploração de um *corpus eletrônico*.

Segundo Ooi (1998), "não é mais possível ou desejável considerar lingüística computacional, *lexicografia computacional* e *lingüística de corpus* isoladas uma da outra".

Assim sendo, não podemos mais supor realizar a *compilação* de um dicionário ou *glossário* sem a ajuda da *lingüística de corpus*.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, M. Aparecida. Da Microestrutura dos Vocabulários Técnico-Científicos. *Anais do IV Encontro Nacional da ANPOL*: 567-578, 1989.

BÉNJOINT, Henry. Dictionaries and the Dictionary. In *Tradition and Innovation in Modern English Dictionaries*: 6-41, 1994.

FILLMORE, C. J. *On Fluency*. In Fillmore, C. J. et al (eds.) *Individual Differences In Language Ability and Language Behavior*. New York: Academic Press: 85-99, 1979.

GREFENSTETTE, G. The WWW as a resource for lexicography. In *Lexicography and Natural Language Processing*. In Honour of B. S. T. Atkins. Euralex, 2002.

OOI, Vincent B. Y. *Computer corpus lexicography*. Edinburgh: Edinburgh University Press, 1998.

ORENHA, Adriane. As colocações e a compilação de um glossário bilíngüe de colocações, na área de Negócios, baseado em corpus comparável bilíngüe eletrônico. Projeto de Pesquisa para a elaboração da dissertação de Mestrado, visando ao Exame Geral de Qualificação, USP, 2002.

PEARSON, Jeniffer. Teaching terminology using electronic resources. In Botley, Mc Enery & Wilson: 223-243, 2000.

SCOTT, M. *WordSmith Tools 3.0*. Oxford: Oxford University Press, 1997.

SINCLAIR, J. M. *Corpus, concordance and collocation*. Oxford: Oxford University Press, 1991.

TAGNIN, Stella E. O. *Expressões Idiomáticas e Convencionais*. São Paulo: Ática, 1989.

TAGNIN, Stella E. O. Convencionalidade e Produção de Texto: um dicionário de Colocações Verbais Inglês/Português; Português/Inglês. Tese de Livre-Docência. Universidade de São Paulo, 1998.

TAGNIN, Stella E. O. Collecting data for a bilingual dictionary of verbal collocations: from scraps of paper to corpora research. In *PALC '99 Practical applications in Language Corpora*. Lodz: Lodz University Press, 1999.

TEUBERT, Wolfgang. Corpus linguistics and lexicography. *International Journal of Corpus Linguistics*, vol. 6 (Special Issue): 125-153, 2002.