Sintagmas terminológicos de núcleo em "*corpus*": identificação de equivalentes terminológicos através da análise de *corpora* comparáveis inglês-português

pg 178-185

Diego Napoleão Viana Azevedo[1]

Resumo

O presente trabalho apresenta um estudo exploratório a respeito da terminologia bilíngue da Linguística de *Corpus*, através da identificação de equivalentes terminológicos, em língua portuguesa, para os sintagmas terminológicos de núcleo em "*corpus*", em língua inglesa. Para a delimitação do conjunto terminológico em língua inglesa, utilizei o glossário de Linguística de *Corpus* de Baker *et al.* (2006). Com relação à identificação dos respectivos equivalentes terminológicos em língua portuguesa, analisei os *corpora* comparáveis, no par de línguas português-inglês, disponibilizados em formato eletrônico pelo projeto *Corpus* Multilíngue para Ensino e Tradução (COMET), da Universidade de São Paulo (USP). A perspectiva teórica da presente pesquisa, no âmbito da Terminologia, está calcada na Teoria Comunicativa da Terminologia (TCT), desenvolvida por Cabré (1999a, 1999b), bem como nas contribuições de Felber (1987), Barros (2004) e Krieger e Finatto (2004) enquanto que, no âmbito da Linguística de *Corpus*, está fundamentada nos aportes de Sardinha (2004), Baker *et al.* (2006) e Tagnin (2010). Como resultado, determinei uma relação de equivalentes terminológicos em língua portuguesa para os sintagmas terminológicos de núcleo em "*corpus*", bem como a menção, quando aplicável de suas não equivalências.

Palavras-chave: Terminologia Bilíngue; Linguística de *Corpus;* Projeto COMET.

TERMINOLOGICAL PHRASES FROM THE TERM "*CORPUS*": IDENTIFICATION OF TERMINOLOGICAL EQUIVALENTS THROUGH THE ANALYSIS OF ENGLISH-PORTUGUESE COMPARABLE CORPORA

Abstract

This paper presents a study on the *Corpus* Linguistics bilingual terminology by identifying terminological equivalents in Portuguese for those emerged from the term "*corpus*" in English. In order to do so, I have analyzed comparable *corpora* in the language pair English-Portuguese made available by the "Corpus *Multilíngue para Ensino e Tradução*" (COMET) Project, of Universidade de São Paulo (USP). This is based on the Communicative Theory of Terminology, suggested by Cabré (1999a, 1999b), as well as on the contributions of Felber (1987), Barros (2004) and Krieger & Finatto (2004), regarding the Terminology field; and also on Sardinha (2004), Baker *et al.* (2006) and Tagnin (2010), on *Corpus* Linguistics. As a result, I have determined a list of terminological equivalents in Portuguese of the terminological phrases from the term "*corpus*", as well as mentioned non-equivalence when such task was not possible.

Keywords: Bilingual Terminology; *Corpus* Linguistics.; COMET Project.

1 Doutorando em Estudos da Tradução (UFSC). Email: diegonapoleao@gmail.com

Introdução

O presente artigo discorre sobre um subconjunto da terminologia da Linguística de *Corpus* com vistas à identificação de equivalentes terminológicos em língua portuguesa dos sintagmas terminológicos de núcleo em *corpus*, encontrados em língua inglesa. Por conseguinte, esta pesquisa fundamenta-se, em termos teóricos, predominantemente nos estudos relacionados à Terminologia e à Linguística de *Corpus* (vide seção 2).

Este trabalho possui pesquisas de cunho exploratório e bibliográfico, baseadas nos fundamentos metodológicos da Linguística de *Corpus* para a análise de *corpora*. Dessa forma, faço uso nesta pesquisa dos *corpora*, disponibilizados gratuitamente em formato eletrônico pelo *Corpus* Técnico (CorTec), o qual está inserido dentro do projeto *Corpus* Multilíngue para Ensino e Tradução (COMET), da Universidade de São Paulo (USP), para a identificação e a extração dos equivalentes terminológicos dos termos em estudo, através do conceito de equivalência utilizado nesta pesquisa (vide seção 3).

A presente pesquisa justifica-se pelo interesse em contribuir, mesmo que minimamente, com a elaboração de recursos terminológicos bilíngues, no que concerne ao par português-inglês para tradutores especializados e estudantes interessados na Linguística de *Corpus*, tendo em vista que grande parte do material desta área ainda se encontra somente em língua inglesa.

Para a elaboração do presente trabalho, foi seccionado em cinco partes distintas: (i) Introdução; (ii) Referencial Teórico; (iii) Método; (iv) Resultados e Discussão; e (v) Considerações Finais, as quais abordam, respectivamente: (i) a contextualização do tema, da justificativa e objetivo do trabalho; (ii) os principais aspectos a respeito de Terminologia e de Linguística de *Corpus*; (iii) as etapas e os procedimentos adotados durante esta pesquisa; (iv) uma breve exposição e análise dos achados da pesquisa; e (v) as suas principais considerações finais.

Referencial teórico

1.1 Terminologia

Considerando-se que a designação "terminologia" possui caráter polissêmico, esta seção inicia-se com a exposição de algumas de suas diferentes acepções. Segundo Cabré (1999a, p. 18), "terminologia" diria respeito, em primeiro lugar, à "disciplina que se ocupa dos termos especializados", ou seja, disciplina que estudaria os termos empregados em áreas específicas do conhecimento, tais como os termos da Física, da Química ou da Informática. Em uma segunda acepção, "terminologia" se referiria ao "conjunto de termos de uma determinada especialidade" (*ibid.*), isto é, ao conjunto de palavras específicas utilizadas em contextos especializados. Por exemplo, pode-se citar a terminologia médica (exemplos de termos: cefaleia, necrose, dorsalgia, etc.), a terminologia jurídica (exemplos de termos: aditivo, *habeas corpus*, comarca, etc.), dentre outras[2].

Para a Terminologia, o "termo" consiste em "unidade lexical com um conteúdo específico dentro de um domínio específico" (BARROS, 2004, p. 40). Os termos se distinguem das palavras (ou unidades léxicas), que são o "conjunto de sons articulados de uma ou mais sílabas com uma significação", pois "as unidades lexicais só se tornam termos quando são definidas e empregadas em textos de especialidade" (KOČOUREK, 1991 *apud* BARROS, 2004, p. 41).

2 A fim de distinguir entre as duas acepções apresentadas, adotou-se "Terminologia" (com inicial maiúscula) para referir-se à disciplina e terminologia (com inicial minúscula), ao conjunto de termos (KRIEGER & FINATTO, 2004).

Em um contexto multilíngue, considerando que todos os sistemas linguísticos possuem suas deficiências e diferenças, nem sempre uma dada unidade terminológica possui sua respectiva equivalência na língua de chegada ou esta mesma unidade pode ser representada de uma maneira diferente daquela apresentada em língua de partida (CABRÉ, 1999b, BARROS, 2004). Felber afirma que:

> Ao se comparar os conceitos existentes de uma dada área em diferentes línguas, nota-se que determinados conceitos coincidem, o que não é o caso da maioria, e que existem graus de equivalência. Esses graus de equivalência dependem do número de características englobadas por compreensão de dois conceitos que coincidem[3] (FELBER, 1987, p. 128).

O autor (*ibid.*, p. 128) aponta ainda que "a compreensão de um conceito consiste no conjunto de características que o constituem" [4]. Dessa forma, o grau de equivalência entre dois conceitos de línguas diferentes seria estabelecido a partir da comparação de suas características. Neste sentido, o autor (*ibid,* p. 129) apresenta, com base na comparação de características, quatro graus de equivalência. São elas: (i) equivalência (=); (ii) interseção (∩); (iii) superordenação (>); e (iv) não equivalência (≠). O autor exemplifica ainda:

- Equivalência: *Machine-outil* [fr] = *Werkzeugmaschine* [de];
- Interseção: *Cricket* [en] ∩ *Schlagball* [de] (os dois jogos utilizam os mesmos instrumentos, porém possuem regras diferentes); e
- Superordenação: *Machine-outil* [fr], *Werkzeugmaschine* [de] >*Machine tool* [en] ("Machine-outil" e "Werkzeugmaschine" são, respectivamente, uma máquina que corta e que molda; enquanto que "Machine tool" é uma máquina que corta somente).

1.2 Linguística de *Corpus*

Quanto à disciplina de Linguística de *Corpus*, segundo Sardinha (2004, p. 3), ela:

> [...] ocupa-se da coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraída por computador.

Em consonância com o exposto acima, cabe apresentar o conceito de *corpus*, que, na ótica de Sardinha (2004, p. 18), em uma definição mais ampla, diz respeito a

> um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.

Em síntese, um *corpus* pode ser entendido como qualquer conjunto de textos compilados em formato digital, para fins de pesquisa linguística. Devido ao seu caráter eletrônico, o *corpus* (ou *corpora*, no plural) permite que um grande número de material compilado possa ser processado facilmente por meio, por exemplo, dos concordanciadores (vide seção 3).

Um fator importante para que uma pesquisa baseada em *corpus* seja bem sucedida é que o repertório possua representatividade. Apesar de não haver critérios objetivos para a determinação da representatividade do *corpus*, acredita-se que ela está associada à sua extensão em números de palavras ou de textos, devendo seus números ser os maiores possíveis. O *corpus* é uma amostra de uma

3 *Lorsqu'on compare les notions qui existent dans un domaine donné dans différentes langues, on constate que quelques notions coïncident mais que ce n'est pas le cas de la plupart d'entre eux et qu'il existe des degrés différents d'équivalence. Ces degrés d'équivalence dépendent du nombre de caractères englobés par compréhension de deux notions qui coïncident* (tradução nossa).

4 *La compréhension d'un notion est l'ensemble des caractères qui constituent cette notion* (tradução nossa).

população cuja dimensão total não se conhece; logo, não se pode afirmar que um *corpus* qualquer seja representativo (SARDINHA, 2004). Embora não haja um consenso neste quesito, Sardinha (*ibid.*), com base em suas análises, sugere uma classificação de *corpus* quanto à sua extensão em número de palavras:

Tamanho em palavras	**Classificação**
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Quadro 1 – Classificação de *corpus* segundo o tamanho em extensão de palavras.

Fonte: Sardinha, 2004, p. 26.

Método

Quanto aos seus procedimentos metodológicos, este trabalho foi elaborado majoritariamente a partir de pesquisas de cunho exploratório e bibliográfico, além de fazer uma pesquisa com base em *corpus*.

Como este trabalho não visa identificar as unidades terminológicas em língua inglesa dos sintagmas terminológicos de "*corpus*", mas sim reconhecer seus equivalentes em língua portuguesa, optei por tomar como ponto de partida, os termos encontrados na obra "*A glossary of Corpus Linguistics*", de Paul Baker *et al.*(2006), por dispor previamente de um repertório de termos da Linguística de *Corpus* amplamente utilizados em língua inglesa. Na referida obra, busquei manualmente por entradas que tivessem como núcleo do sintagma terminológico "*corpus*", porém, sem considerar aquelas que designavam nomes próprios de*corpora*, tais como o *British National Corpus* (BNC), por conta do número elevado de entradas, o que inviabilizaria a execução desta pesquisa.

Para identificar os equivalentes terminológicos em língua portuguesa, conduzi uma pesquisa baseada em *corpus*, escolhendo por utilizar *corpora* já compilados, por questões práticas. Por possuir *corpora* técnicos, inclusive sobre Linguística, tanto em língua portuguesa como em língua inglesa, fiz uso do material coletado pelo projeto *Corpus* Multilíngue para Ensino e Tradução (COMET), subsidiado e disponibilizado gratuitamente pela Universidade de São Paulo (USP). Ao acessar o sítio eletrônico do COMET, dirigi-me à guia CorTec (*Corpus* Técnico-Científico), que possui vinte e um (21) tipos de *corpora* técnicos (dentre eles, os da Culinária, do Turismo e da Linguística). Em geral, cada *corpus* técnico-científico apresenta duas línguas: português e inglês. Especificamente utilizei os corpora da subseção "Linguística" do CorTec. Segundo as informações do COMET, este *subcorpus* é composto de textos acadêmicos coletados da internet e tenta abranger todas as subáreas da Linguística. Apresento o resumo das informações dos *corpora* utilizados, conforme a classificação de tamanho por extensão de palavras, proposta por Sardinha (vide seção 2.2):

Descrição	***Corpus* em língua portuguesa**	***Corpus* em língua inglesa**
Tamanho do *corpus*	Pequeno-médio	Pequeno-médio
Número total de palavras	1.309.967	1.921.811

Quadro 2 – Características dos *corpora* de Linguística, do Projeto COMET.

Fonte: Projeto COMET.

Para o processamento dos *corpora*, o Projeto oferece três opções de ação: o concordanciador, o gerador de listas de palavras e o gerador de n-gramas. Como já havia definido o conjunto terminológico em língua inglesa, segui, portanto, para a identificação dos possíveis equivalentes em língua portuguesa, por meio do uso do concordanciador do *subcorpus*, em língua portuguesa. A partir desta ferramenta, foi possível perceber os usos do termo *corpus*, bem como suas associações, e assim inferir algumas equivalências, como explicado anteriormente (videseção 2.1).

Ancorei, portanto, a pesquisa pela busca do termo "*corpus*" nos *corpora* de estudo, tomando-o como ponto de partida para o reconhecimento dos demais sintagmas terminológicos. Como exemplo, tem-se a Figura 1 abaixo, onde podem ser analisados os fragmentos de texto, dos quais se pode identificar o uso do termo "*corpus* paralelo", derivado da pesquisa de "*corpus*".

Figura 1 – Concordância da palavra de busca "*corpus*" no CorTec.

Fonte: Projeto COMET.

Para Baker *et al.* (2006), "*parallel* corpus" pode ser definido como "*corpus* que contém textos-fonte e suas traduções"[5], que, confrontado com a definição de "*corpus* paralelo" de Tagnin (2010, p. 358) – corpus constituído de originais e suas respectivas traduções –, elucida seu valor equivalente, conferindo ao termo "*parallel* corpus" a condição de equivalente terminológico para "*corpus* paralelo". Dessa forma, realizei o processo de verificação de equivalência entre os termos em língua inglesa e portuguesa.

5 Corpus *that contains source texts and their translations* (tradução nossa).

Resultados e discussões

Com base nos procedimentos metodológicos expostos na seção anterior, coletei 29 unidades terminológicas em língua inglesa, a saber:

balanced corpus; corpus-based; corpus-driven; corpus linguistics; corpus sampler; corpus wizard; diachronic corpus; dialect corpus; dynamic corpus; historical corpus; learner corpus; monitor corpus; multilingual corpus; national corpus; non-standard corpus; on-line corpus; parallel corpus; raw corpus; reference corpus; regional corpus; sample corpus; sample text corpus; specialized corpus; speech corpus; spoken corpus; static corpus; synchronic corpus; training corpus; written corpus.

Quadro 1 – Conjunto terminológico em estudo.

Fonte: Baker *et al.* (2006).

Depois da análise, observei que nem todos os termos extraídos do glossário possuíam equivalentes em língua portuguesa. Após serem analisados em contextos, validaram-se as equivalências terminológicas presentes no quadro abaixo:

Termo em língua inglesa	**Termo em língua portuguesa**
balanced corpus	*corpus* balanceado
corpus-based	baseado(a) em *corpus*
corpus-driven	*equivalente não identificado*
corpus linguistics	linguística de *corpus*
corpus sample	*equivalente não identificado*
corpus wizard	*equivalente não identificado*
diachronic corpus	*corpus* diacrônico
dialect corpus	*corpus* dialeto
dynamic corpus	*corpus* dinâmico
historical corpus	*corpus* histórico
learner corpus	*corpus* de aprendizes
monitor corpus	*corpus* monitor
multilingual corpus	*corpus* multilíngue
national corpus	*equivalente não identificado*
non-standard corpus	*equivalente não identificado*
on-line corpus	*equivalente não identificado*
parallel corpus	*corpus* paralelo
raw corpus	*corpus* cru
reference corpus	*corpus* de referência
regional corpus	*equivalente não identificado*
sample corpus	*corpus* de amostragem
sample text corpus	*equivalente não identificado*
specialized corpus	*corpus* especializado
speech corpus	*corpus* de fala
spoken corpus	*corpus* oral

static corpus	*corpus* estático
synchronic corpus	*corpus* sincrônico
training corpus	*corpus* de treino
written corpus	*corpus* escrito

Quadro 2 – Relação de equivalentes terminológicos para o conjunto de estudo.

Fonte: Dados da pesquisa.

Obtive, como produto final, uma pequena relação de unidades terminológicas em língua inglesa e suas respectivas equivalências (ou menção de não equivalência) em língua portuguesa, no referido *corpus* de estudo. Como previsto inicialmente (vide seção 2.1), nem todos os termos em língua inglesa tinham um equivalente terminológico em língua portuguesa, com base na análise de *corpora.*

Considerações finais

Este artigo tratou de um breve estudo da terminologia bilíngue da Linguística de *Corpus*, com ênfase nos sintagmas terminológicos a partir do termo "*corpus*". Como resultado, obtive uma relação dos termos língua inglesa e as respectivas equivalências de algumas destas listadas em ordem alfabética contínua, pois as demais não foram identificadas nos *corpora* estudados.

É importante ressaltar que este estudo se configura enquanto um estudo exploratório e, tampouco, pretende realizar um estudo aprofundado nessa temática, limitando-se a apresentar uma simulação por meio de um exemplo prototípico de como o estabelecimento de equivalentes terminológicos utilizando *corpora* comparáveis pode ocorrer. Nesse sentido, sugiro um estudo mais aprofundado da área de Linguística de *Corpus*, com a busca de equivalentes em língua inglesa para os demais termos da Linguística de *Corpus*, bem como o desenvolvimento de outros componentes, que possam levar a elaboração de uma obra de referência terminológica, tais como definição e contextos.

Referências

BAKER, P.; HARDIE, A.; MCENERY, T. *A glossary of Corpus Linguistics.* Ediburgh: Ediburgh Universiy Press, 2006.

BARROS, L. A. *Curso básico de terminologia.*São Paulo: Edusp, 2004.

CABRÉ, M. T. *La terminología: representación y comunicación:* elementos para una teoría de base comunicativa y otros artículos. Barcelona: Universitat Pompeu Fabra, 1999a.

_____. *Terminology:* theory, methods and applications. Terminology and Lexicography Research and Practice. Editado por Juan C. Sager. Trad. Janet Ann DeCesaris. Vol. 1. Amsterdã: John Benjamins Publishing, 1999b.

FELBER, H. *Manuel de terminologie*. Paris: UNESCO-INFOTERM, 1987.

KRIEGER, M. G.; FINATTO, M. J. B. *Introdução à terminologia.* São Paulo: Contexto, 2004.

MCENERY, A. M.; XIAO, R. Z. Parallel and comparable corpora: What are they up to? In: *Incorporating Corpora: Translation and the Linguist.* Translating Europe. Multilingual Matters, Clevedon, 2007. Disponível em: http://eprints.lancs.ac.uk/59/1/corpora_and_ translation.pdf. Acessado em: 05/12/2013.

PROJETO *CORPUS* MULTILÍNGUE PARA ENSINO E TRADUÇÃO (COMET). Disponível em: http://www.fflch.usp.br/dlm/comet/. Acessado em 02/12/2017.

SARDINHA, T. B. *Lingüística de corpus*. Barueri: Manole, 2004.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: Vander Viana; Stella E. O. Tagnin. (Org.). In: *Corpora no ensino de línguas estrangeiras*. 1 ed. São Paulo: HUB Editorial, 2010, p. 357-361. Disponível em: http://www.hubeditorial.com.br/ site/recursos/5_glossario/glossario_423.pdf. Acessado em: 01/12/2017.

Submissão: 02 de dezembro de 2017

Aceite: 14 de dezembro de 2018