

## A relevância da Web como *corpus* para a identificação de padrões de lexicalização: o caso de “bater+SN” no português brasileiro

### The usefulness of the Web as *corpus* for the identification of lexical patterns: The case of “bater+NP” in Brazilian Portuguese

Milena de Uzeda Garrão\*

**RESUMO:** Nesse estudo longitudinal, e inicialmente com base em *corpora* jornalísticos, caracterizamos padrões de lexicalização do tipo Bater+SN, uma vez que: i) identificamos um crescente status de transitividade direta do verbo “bater” tanto no Português Brasileiro, quanto no Português Europeu (como construções do tipo bater o oponente, um caso em que o verbo ganharia uma semântica de “derrotar”); ii) identificamos uma tipologia sintático-semântica do padrão Bater +SN bastante peculiar, como “Bater um medo”, que se distanciava de combinação livre ou de uma coocorrência sintática aleatória (descrita em i); iii) curiosamente, constatamos que esse padrão, que defendemos ser análogo a expressões com verbo-suporte, também se distanciava do fenômeno comumente rotulado como expressão fixa, cristalizada, caracterizado por revelar um alto grau de opacidade semântica (como bater as botas). Também questionamos o nível de previsibilidade da metodologia utilizada para a identificação das expressões fixas, (descritas em iii). Já em 2013, com base na Web como *corpus*, pudemos notar uma mudança da restrição semântica do padrão descrito em ii. Portanto, nesse estudo, caracterizamos o quanto tanto a descrição como a metodologia de identificação de padrões lexicais V+SN merecem ser revistas com base no poder empírico da Web como *corpus*.

**PALAVRAS-CHAVE:** Web como *Corpus*. Lexicologia. Expressões Fixas. Bater+Sintagma Nominal.

**ABSTRACT:** Based on a Brazilian Portuguese journalistic corpus, we described in 2001 and 2003 lexicalization degrees of “Bater+Noun Phrase” pattern. Three reasons motivated our study: i) a new transitivity status of verb “Bater” both in Brazilian and European Portuguese (eg.: “bater o oponente”– “to beat the opponent” – where the verb would take in the meaning of “to defeat”); ii) a particular semantic pattern of “Bater +NP”, such as “bater um medo” (“to get frightened”) which would not conform to a random syntactic cooccurrence, described in i); iii) this pattern, which we claim to behave as support-verb expressions, would also not conform to what is commonly labeled as fixed expressions, which are claimed to have a high degree of semantic opaqueness (such as “bater as botas” or “to kick the bucket”). We also claim that this framework had drawbacks if we rely on corpus evidences. Then, in 2013, in a Web corpus based approach, we could spot a semantic change regarding the pattern described in ii). Therefore, in this longitudinal study, we take a critical stance of the framework chosen along these years for multi-word expressions identification and description, and we portray the Web as corpus as a powerful empirical tool for that purpose.

**KEYWORDS:** Web as Corpus. Lexicology. Multi-word Expressions. “Bater”+Noun Phrase.

\* Doutora em Estudos da Linguagem (PUC-Rio), Professora Adjunta do Instituto Multidisciplinar, UFRRJ.

## 1. Nosso percurso

O interesse por expressões com padrão “Bater+SN” surgiu em função de uma visão inclusiva de um léxico marginalizado para o seu tratamento automático. A motivação inicial do estudo, entre os anos de 2001 (em Garrão e Dias) e 2003 (em Basílio, Oliveira e Garrão), era a de identificação de estruturas lexicalizadas com esse padrão para que pudessem figurar em um dicionário eletrônico. A opção pelo verbo *bater* se justificava em função da detecção do seu alto grau de ocorrência encabeçando o que se convencionou chamar de expressões idiomáticas (como *bater as botas*, *bater perna*, *bater pino*, *bater boca*, *bater os olhos*, *bater o martelo*, *bater ponto*, *bater papo*, entre outras).

Curiosamente, no ano de 2001, em uma das buscas em *corpora* jornalísticos, até então ainda não etiquetados, foi possível identificar o seguinte fato: no Português do Brasil (PB, doravante) havia, além das expressões supracitadas, um tipo de expressão encabeçada por esse verbo, mas que, no entanto, não deveria ser considerada propriamente uma expressão idiomática ou cristalizada, tampouco uma unidade lexical. Tratava-se de construções como *bater um medo*, *bater uma fome*, que, por sua vez, apresentavam um comportamento sintático-semântico diverso, uma vez que seus constituintes pareciam apresentar uma função composicional para o todo. Em outras palavras, as partes da expressão pareciam contribuir para o seu significado global. E o que gerou uma maior motivação na descrição desse padrão foi o fato de figurarem em textos escritos com registro semiformal.

Além desse novo padrão, também notamos, nessa mesma época, que o verbo *bater* vinha ocupando, não somente no PB, mas também no português europeu, um status de transitividade direta muito notório (como sinonímia de superar; ex: *bater o concorrente*). Portanto, dentro desse quadro empírico, passamos a conceber a possibilidade de detecção de níveis de lexicalização de expressões com o padrão Bater+SN.

Como fonte para a identificação desse padrão, utilizamos em 2001 o primeiro jornal brasileiro a ter sua versão online, *o Jornal do Brasil*, e já em 2003, as versões online do jornal *O Globo* e a da revista *Veja*. Nesse mesmo ano, já contávamos com um *corpus* jornalístico etiquetado disponibilizado pelo grupo NILC-São Carlos: o *corpus* da Folha de São Paulo.

O padrão Bater +SN que nos interessou na época, portanto, revelava-se como construções extremamente comuns em PB, encontradas não somente em discurso oral como também em textos escritos com registro semiformal (como *bater uma fome*, *bater um desespero*, *bater uma dúvida*). Na época, propusemos que esse tipo de construção tinha um comportamento análogo

ao de expressões com verbo-suporte (como *dar um riso; fazer compras*), uma vez que pudemos identificar um comportamento sintático-semântico comparável àquelas construções quando submetidas aos testes de níveis de lexicalização. Traçamos, para tanto, um vetor de nível de lexicalização das expressões Bater+SN, com base nos *corpora* e em uma metodologia de detecção de expressões lexicalizadas proposta por Neves (1999).

Como os testes descritos em Neves (adaptados para o português de Radford, 1988:90) se propunham a distinguir níveis de lexicalização tendo como base o padrão V+SN; isto é, visavam contrastar mais especificamente construções livres, construções com verbo-suporte e construções fixas, cristalizadas, lançamos mão dessa mesma metodologia para aferir os diferentes graus de lexicalização de Bater+SN. Portanto, com nítida inspiração no artigo de Neves, pudemos esquematizar o seguinte vetor de nível de lexicalização:

Quadro 1. Identificação de graus de lexicalização do padrão Bater+SN com base em *corpora* jornalísticos.

+ sintaxe		+ léxico
<i>construções livres</i>	<i>construções com verbo-suporte</i>	<i>expressões cristalizadas</i>
(bater transitivo- “derrotar”)	( Bater+SN sentimento de falta)	(Bater+SN fixo)
<i>Bater o concorrente</i>	<i>Bater uma saudade</i>	<i>Bater as botas</i>
	( Bater+SN sentimento negativo)	<i>Bater perna</i>
	<i>Bater um medo</i>	

Fonte: elaboração própria.

Na extrema esquerda, teríamos combinações com verbos plenos e sintagmas nominais complementos, que são completamente livres, onde os dois elementos exercem papéis independentes na estrutura argumental (*bater o concorrente*); na extrema direita, expressões que constituem um significado unitário, em que “nem mesmo parece ser possível postular um SN em posição de objeto” (Neves, 1999:99), (como *bater as botas*); e entre estes dois graus extremos de construção, há aquelas construções intermediárias, constituídas dos chamados verbos-suporte, que, por sua vez, recebem certo grau de esvaziamento do sentido lexical, porém, semanticamente, contribuiriam para o significado total da construção.

### 1.1 Mas o que viria a ser o grau intermediário de lexicalização?

Em relação ao vetor de nível de lexicalização, pudemos constatar que, embora as expressões que se encontrassem na sua extrema esquerda fossem consideradas livres e até imprevisíveis, a distinção entre os dois outros tipos de construções traria uma certa hesitação pelo fato de ambas se situarem no domínio da convencionalidade ou, nas palavras de Neves (1999:103), no domínio “das estruturas recorrentes que o falante escolhe com reduzida liberdade quanto ao modo de composição”.

Portanto, as construções com verbo-suporte foram consideradas por Neves como intermediárias por ora se situarem mais próximas de construções livres, ora mais próximas de expressões cristalizadas. Isto é, por vezes mais próximas de um; por outras vezes mais próximas de outro extremo do vetor. Compõem-se de: um verbo com determinada natureza semântica básica que funciona como instrumento morfológico e sintático na construção do predicado; e um SN que entra em composição com o verbo para configurar o sentido do todo, bem como para determinar os papéis temáticos da predicação.

Em conformidade com essa característica, observamos do ponto de vista sintático-semântico, uma restrição do SN posposto ao verbo bater, a saber: a) determinante intensificador + sentimento de falta (como, por exemplo, *uma fome, a maior saudade, uma dívida*) e b) determinante intensificador + sentimento negativo (como *uma baita tristeza, o maior desespero, um medo*).

Note-se que se a análise for feita sob um aspecto puramente formal, as expressões Bater+SN que figuram no meio do vetor apresentado na Figura 1, não teriam um comportamento exato do que é chamado tradicionalmente de construções com verbo-suporte (como *dar um riso, fazer compras, dar uma olhada*), pois teoricamente a sua construção sintática seria *bater em alguém uma dívida / um medo*. Contudo, propusemos em 2003 que o seu comportamento semântico seria análogo ao dessas construções com base nos seguintes argumentos:

De acordo com Cruse (1986) e Radford (1988), os critérios definidores para uma suposta construção sintática tender a um alto grau de lexicalização, ou seja, se enquadrar na extrema direita do vetor, são a sua resistência aos seguintes rearranjos sintáticos: substituição das partes, posposição, coordenação, inserção de constituinte e elipse. Tendo tais critérios como pressuposto, e se focarmos o meio do vetor, fica bastante evidente que as construções do tipo *bater um medo*, por exemplo, não se comportariam como cristalizadas, já que parecem admitir

coordenação (*bate um medo e uma dúvida*), posposição (*um medo bate*), inserção de constituintes (*bateu uma pontinha de dúvida*), assim como o SN pode funcionar como fragmento da oração ou um constituinte sintagmático (ex.: *bate uma dúvida? Não, um medo*). E, seguindo Neves (1999), é exatamente esse comportamento que revelam as expressões com verbo-suporte, como *fazer compras e dar um riso*.

## 2. Nossos percalços: implicações do método utilizado

Até agora vimos que os instrumentos mais seguros para determinar a estrutura dos constituintes de expressões fixas seriam a resistência aos seguintes movimentos sintáticos: a distribuição, a posposição, a coordenação, a intercalação de advérbios, a elipse.

O semanticista britânico David Cruse (1986) atribui ao grau máximo de lexicalização, ou seja, às expressões que propusemos pertencer à extrema direita do vetor, o rótulo de “expressão idiomática”. O autor afirma que os itens que a compõem não contribuem para o significado total da expressão. A checagem desse nível máximo de opacidade semântica é possível através dos testes descritos acima para se comprovar a impossibilidade de se compreender a expressão sem que todos os seus itens estejam presentes na ordem original (é o caso do exemplo clássico *bater as botas*).

Dentro dessa perspectiva, aplicamos os testes de Neves (1999) para diferenciar as construções cristalizadas das construções com característica de verbo-suporte do tipo *Bater+SN*. Contudo, identificamos que as expressões que seriam consideradas pelos testes como indevassáveis, ou seja, cristalizadas, admitiam, segundo os dados analisados dos *corpora O Globo, Veja e Jornal do Brasil on-line*, inserção de constituinte. Propusemos, portanto, um outro teste a fim de demonstrar que somente os testes de Neves não dariam conta de um padrão revelado pelos *corpora*: era o caso de *bater (muita) perna, bater(um) papo, bater (uma) bola*. Chegamos à conclusão, em contrapartida, de que outras expressões cristalizadas como *bater os olhos, bater as botas* pareciam não admitir a inserção de outros constituintes. Percebemos, na época, que isso parecia estar intimamente relacionado ao perfil semântico destas expressões, que apresentavam um aspecto **pontual**. Portanto, o perfil pontual dessas expressões parecia bloquear a possibilidade de marcação de frequência ou quantificador.

Portanto, já em 2006, argumentamos que mesmo esses casos considerados mais nitidamente impermeáveis do ponto de vista semântico não são tão indivisíveis se recorrermos a evidências de *corpus*. Ou seja, os testes de opacidade semântica não seriam teoricamente

conclusivos porque: 1) os avaliadores são linguistas e não falantes sem pretensões ou interesses em relação a uma teoria de expressões cristalizadas; 2) a noção de *opacidade* versus *transparência semântica* é escorregadia e também carece de uma delimitação teórica precisa e incontroversa.

## 2.1 Quando o *corpus* fala mais alto

Em 2006, em uma busca ao sistema de Recuperação de Informação *Google*<sup>TM</sup>, ou da Web como *corpus*, pudemos constatar um fato curioso. A expressão *bater a caçuleta* também é utilizada no PB (principalmente no nordeste do país) com o mesmo sentido de *bater as botas*, o que se faria supor que o SN da expressão não seria tão fixo quanto se imagina: "E o Doutor Morte finalmente bateu a caçuleta", (*Casseta & Planeta online*, acesso em junho de 2005). O mesmo pudemos questionar em relação às expressões *bater perna* e *dar o braço a torcer*. Se, assim como *bater as botas*, fossem, de fato, expressões indecomponíveis, a inserção de qualquer constituinte tornaria as expressões literais, mas os exemplos a seguir parecem contra-argumentar tal afirmação:

Resumindo, quem quiser economizar, ou fica em casa, ou vai ter que **bater muita perna** para achar onde comer e onde ficar.  
(<http://www.bemtevivrasil.com.br/diarioviagem18.htm>)

Para não dizerem que sou um fanático *apenas* pela aviação militar, **dei meu braço a torcer** e consegui alguns interessantíssimos anúncios de companhias americanas.  
([http://www.jetset.com.br/aviacao\\_mkt.asp](http://www.jetset.com.br/aviacao_mkt.asp))

Ambas as expressões acima teriam comportamento idiomático nos testes, mas sua utilização pelo usuário da língua parece bem mais flexível. No ponto de vista de Cruse (1986), a expressão idiomática é uma unidade lexical elementar: “embora consista em mais de uma palavra, apresenta uma coesão interna de palavras simples” (p. 38). Embora o autor considere Expressões Idiomáticas, Metáforas Cristalizadas e Colocações como tipos de expressões cristalizadas distintas, reconhece que há casos limítrofes. Mas como, então, teorizar sobre um fenômeno que é escorregadio?

Vale (2001, p. 16), numa proposta de tipologia de *expressões cristalizadas* para o PB, também expõe sintomaticamente a arbitrariedade da intuição do pesquisador em relação aos testes de composicionalidade. A sua argumentação deixa clara a falta de força teórica distintiva

entre opacidade e transparência semântica. Ao explicar a aplicação dos testes, recorre ao uso de “asterisco para inaceitabilidade; ponto de interrogação para aceitabilidade duvidosa; dois pontos de interrogação para aceitabilidade ainda mais duvidosa do que a precedente; três pontos de interrogação para aceitabilidade no limite da inaceitabilidade”. Sua tentativa parece ser um sintoma de que não há como teorizar sobre as noções de opacidade/transparência semântica.

Trata-se, portanto, de intuição linguística do pesquisador; mas não de uma regra. Dentro dessa análise, pode-se dizer também que, quando Neves (1999) avalia que a expressão “tomar partido”, em “Valéria *tomou partido* da tia” (p.99), seria uma expressão cristalizada, está, na verdade, desconsiderando o fato de a construção admitir intercalação de advérbio dependendo do teor aspectual da frase em que se insere. Segundo a autora, a expressão não admite inserção de nenhum tipo de constituinte. O mesmo ela afirma em relação à expressão “ter cabeça” em “O capitão Aparício *tem cabeça* para tudo”. O Google, contudo, recupera contra-exemplos para o que a autora propõe:

A Quarta é um meio termo, uma sinfonia que parece não **tomar muito partido** desta relação, pois está montada sobre um afresco extremamente original de ... ([www.mnemocine.com.br/filipe/ensaios.htm](http://www.mnemocine.com.br/filipe/ensaios.htm))

Tem que ser um exame de nível nacional para entrar quem **tem mais cabeça**, ... ([www.museudapessoa.net/MuseuVirtual/hmdepoente/depoimentoDepoente.do?action=ver&idDepoente=63&key](http://www.museudapessoa.net/MuseuVirtual/hmdepoente/depoimentoDepoente.do?action=ver&idDepoente=63&key))

Portanto, ao longo de nosso estudo, notamos que muitas das construções do tipo Bater+SN por nós identificadas, seriam alocadas na extrema direita do vetor de lexicalização por serem consideradas pelos testes de composicionalidade propostos pela autora — substituição, coordenação, posposição, elipse e inserção de constituinte — como uma unidade indevassável. O que nos surpreendeu foi o fato de esse conjunto de testes ser insuficiente para dar conta de uma boa parte dessas construções, já que algumas delas parecem admitir intercalação de intensificador ou marcador de frequência, conforme detectamos até mesmo antes de recorrer à Web como *corpus*, mas ainda com base nos *corpora* etiquetados a) Cetenfolha e b) Portugal Natura Publico, respectivamente:

a) Quem pretende ter peixe à mesa durante a Semana Santa precisa **bater muita**

**perna.**

b) Para os comunistas, o grave é que não estão em condições de **bater demasiado o pé.**

Os exemplos acima demonstram que há expressões supostamente indevassáveis que admitem inserção de advérbio, mais especificamente, de um marcador de frequência ou um intensificador, o que, ao menos, nos faria ampliar os testes geralmente feitos para detectar o teor de fixidez dessas construções. É importante ressaltar, ainda, que a sua alegada opacidade semântica parece não ser definidora do nível de indivisibilidade da expressão, visto que há expressões que admitem marcador de frequência e cujos constituintes não parecem ter o que se costuma chamar de papel composicional (como *bater (muita) perna*, *bater (muita) boca*, *dar (muita) trela*, *fazer (muita) questão*).

Pudemos constatar também que o aspecto verbal da expressão como um todo parece ser muito mais preditivo em relação à sua fixidez do que a sua suposta opacidade, uma vez que tais construções verbais com aspecto pontual tenderiam a um grau de fixidez elevado (*bater o martelo/?bater muito martelo*) e aquelas com aspecto durativo seriam menos rígidas (*bater boca/ bater muita boca*). Mas observamos também que isso parece ser um padrão de comportamento, uma tendência, e não uma regra.

### 3. O que fazer, então?

As conclusões a que chegamos em 2.1 nos impulsionaram a concluir que caracterizar deterministicamente uma expressão como cristalizada ou fixa é elevar um olhar dedutivo, baseado na intuição, a uma supremacia que talvez não mereça. Enquanto se priorizar uma abordagem dedutiva do pesquisador, que se pretende capaz de caracterizar a expressão tendo por base a sua própria intuição, estaremos ignorando o fato de que é o discurso do falante desavisado, sem pretensões nem comprometimentos teóricos, a fonte mais segura para tanto.

Portanto, em Garrão e Dias (2006), tendo como inspiração o artigo “*I don’t believe in word senses*” (2000), de Adam Kilgarriff, um estudioso do léxico do ponto de vista computacional, e seguindo uma visão semântica dependente de *corpus*, discutimos o problema sintomático de profusão de rótulos para identificação de fraseologias do tipo V+SN, como as terminologias recorrentes na área de semântica lexical: colocações, expressões fixas, expressões idiomáticas, metáforas cristalizadas, em função da total dependência dessas rotulações à intuição semântica do pesquisador, muitas vezes, falha. Utilizamos os próprios

contra-exemplos dos *corpora*, apresentados em 2.1, para refutar a abrangência e demonstrar a parcialidade dos testes propostos na literatura fraseológica.

Defendemos, além disso, que o problema de explosão terminológica no domínio multivocabular seria sintoma da impossibilidade de se chegar a um modelo teórico que pudesse dar conta desse conceito (ou não-conceito).

### 3.1. Web como *corpus* e a mudança semântica de construções Bater+SN suporte

Já em 2013, cientes de que não podemos fazer aferições sobre o comportamento dessas expressões sem verificação em *corpora* e com base nos pressupostos de Stefan Diemer em seu contundente artigo *Corpus linguistics with Google?* (2011), discutimos a tão polêmica utilização de *corpora* abertos para considerações de ordem descritiva na língua. Para tanto, levamos em consideração suas implicações e benefícios (cf. Davies, 2011), tendo como foco de discussão agora, o status intermediário do vetor esboçado na seção 1 ou, mais especificamente, o que propusemos como expressão com verbo-suporte com padrão Bater+SN.

Com o intuito de verificar a manutenção do status de restrição semântica do SN posposto ao verbo nessa construção, que foi nosso objeto de estudo em 2001 e 2003, partimos para a observação de sua ocorrência em *corpora* abertos.

Ou seja, apenas com a mais corriqueira e popular ferramenta de busca, identificamos, dessa forma, a importância de uma busca em *corpora* abertos ou da Web como *corpus* para a revisão da restrição sintático-semântica do SN posposto ao verbo “bater” proposta anteriormente, caracterizada como: a) determinante intensificador + sentimento de falta (como, *uma fome, a maior saudade, uma dúvida*) e b) determinante intensificador + sentimento negativo (como *uma baita tristeza, o maior desespero, um medo*).

Observamos, em relação ao que tínhamos detectado em 2001 e 2003, contra-exemplos do SN posposto ao verbo. Ou seja, já existem evidências sólidas de que os falantes ampliaram a restrição em a) para determinante intensificador + sentimento de falta **mas também de plenitude** (com ocorrências como *bater uma certeza, bater um insight*) e b) determinante intensificador + sentimento negativo **mas também positivo** (como *bater uma alegria, bater uma felicidade*). Portanto, os dados recuperados pela Web como *corpus* revelam que a restrição semântica claramente evidenciada pelos *corpora* jornalísticos nos nossos primeiros estudos não procede mais 10 anos depois.

Quadro 2. Uso de *corpora* abertos para ratificação da mudança semântica detectada primeiramente em discurso oral informal.

+ sintaxe	+ léxico	
<i>construções livres</i>	<i>construções com verbo-suporte</i>	<i>expressões cristalizadas</i>
		→
	<p>“bateu uma saudade” (209000 ocorrências) (Bater+SN de fada)</p> <p>“bateu um medo” (9040 ocorrências) (Bater+SN de fada)</p> <p>“bateu uma certeza” (4410 ocorrências) (Bater+SN sentimento de plenitude)</p> <p>“bateu uma alegria” (8080 ocorrências) (Bater+SN sentimento positivo)</p>	

Fonte: elaboração própria.

### 3.2 Em defesa do uso da Web como *corpus*

Alguns resultados da investigação implementada por Diemer (2011, p. 5) em relação às implicações linguísticas do *corpus* aberto foram as seguintes: i) imprecisão do tamanho do *corpus* (um comprometimento quantitativo); ii) organização dos dados (uma vez que existe claramente uma dificuldade de busca por lema ou *POS* – partes do discurso), iii) linguagem lúdica (o que o autor chama de “*playful use of language*”) e iv) facilitação sintática. Estas duas últimas características de escolhas linguísticas inovadoras calcadas na oralidade, encontradas principalmente em ambientes digitais como Blogs, Facebook e Twitter.

Contudo, também concordamos com Diemer (2011; p. 11) quando avalia que o *corpus* mais revelador e confiável é a Web em detrimento a *corpora* fechados sempre que o foco de estudo for o uso real dos falantes. Portanto, para o objeto de estudo dessa etapa da pesquisa, o uso da Web como *corpus* serviu para ratificar que o falante não utiliza a mesma restrição de negatividade ou falta para expressões Bater+SN suporte. A Web foi também fundamental para a ratificação de que a flexão preferencial do verbo que encabeça a expressão em análise é a terceira pessoa do singular no pretérito perfeito, um padrão flexional que havia sido identificado informalmente em linguagem oral.

Dessa forma, passamos a entender que o uso da Web como *corpus* é extraordinariamente mais fiel àquilo que pode ser considerado como dados da intuição de uma comunidade linguística. Diemer (2011, p. 3) ressalta que para muitas buscas, o *corpus* etiquetado é desnecessário, como identificação de novos prefixos. Concluimos, também, que para outras caracterizações empíricas da língua, a Web como *corpus* é valiosa, como:

i) Para identificação de padrão negativo, ou seja, para detecção ou não de um suposto fenômeno linguístico, como implementado em 2.1, em relação às supostas expressões fixas ou cristalizadas do tipo Bater+SN.

ii) Para corroborar evidências esparsas ou isoladas não advindas de *corpus*, como implementado em 3.1, em relação às construções com verbo-suporte do tipo Bater+SN.

Defendemos, portanto, a ideia de que o domínio aparentemente labiríntico da Web como *corpus* é, indubitavelmente, revelador e equivocadamente menosprezado por muitos pesquisadores que se dedicam ao campo da descrição linguística. Acreditamos, assim, que não existem verdades absolutas ou caminhos únicos, irrevogáveis, na Linguística de *Corpus*. É o foco de análise linguística que recorta o caminho mais produtivo para determinado fim; e o nosso foco de análise procurou legitimar o uso da Web como *corpus*.

Portanto, podemos também distinguir aqui, e de forma contundente, a força do pensamento saussureano uma vez que o uso de *corpus* é a concretização da aplicabilidade de duas de suas célebres constatações: a de que “a linguagem é multiforme e heteróclita” e que, portanto, se deixa avaliar de diferentes formas e, primordialmente, a de que “é o ponto de vista que cria o objeto”.

Por fim, ratificamos também que a Linguística de *Corpus* revela claramente a perspicácia do pragmatismo radical da virada linguística de Wittgenstein (1953) quando afirma em *Investigações Filosóficas* que “o significado está no uso”. Conforme argumentamos em Garrão (2006), por uma perspectiva wittgensteiniana da linguagem, podemos enxergar a fusão de dois domínios da Linguística: a Semântica e a Pragmática, ou conhecimento linguístico e conhecimento enciclopédico. E por essa visão negar “a vocação representacionista da linguagem tão defendida por filósofos como Platão, Aristóteles e Locke” (Garrão, 2006:136), ou de uma visão entitativa do significado, podemos supor que o único caminho legítimo para verificação da prática desse uso é o seu registro exaustivo; ou seja, é a Linguística de *Corpus*. E é notável como esse domínio sintetiza de forma curiosa, mas ao mesmo tempo eloquente, um diálogo entre pensadores da linguagem que se debruçaram sobre o fenômeno da significação, mesmo que de forma díspare. Portanto, não seria nenhum exagero afirmar que se trata de um novo paradigma para a ciência da linguagem.

## Referências bibliográficas

AIRES, R. V. X.; ALUÍSIO, S. M. Criação de um *corpus* com 1.000.000 de palavras etiquetado morfossintaticamente. **Série de Relatórios do NILC**, NILC-TR-01-8. 2001

BASÍLIO, M.; OLIVEIRA, C.; GARRÃO, M. U. A Não-Delimitação das Unidades Lexicais. In: Claudio Cezar Henriques. (Org.). **Linguagem, Conhecimento e Aplicação**. 1ed. Rio de Janeiro: Europa, 2003, v. Vol. 1, p. 137-148.

CRUSE, D. **Lexical Semantics**. Cambridge, Inglaterra: Cambridge University Press. 1986.

DAVIES, M. **The Corpus of Contemporary American English (COCA) and Google / Web as Corpus**. 2011 Disponível em <http://view.byu.edu/coca/compare-google.asp> Acesso em 14 de julho de 2013.

DIEMER, S. **Corpus Linguistics with Google?** Saarland University, Alemanha. 2011. Disponível em <http://www.bu.edu/isle/files/2012/01/Stefan-Diemer-Corpus-Linguistics-with-Google.pdf>. Acesso em 20 de julho de 2013.

GARRÃO, M. U.; DIAS, M.C.P. Um Estudo de Expressões Cristalizadas do tipo V+SN e sua Inclusão em um Tradutor Automático Bilíngüe (português/inglês). **Cadernos de Tradução** (UFSC), Florianópolis, v. v.2, n.VIII, 2001, p. 165-182.

GARRÃO, M. U.; DIAS, M.C.P. The *corpus* never lies: a statistical approach for the identification of verbal collocations. In: Marja Nenonen, Simikka Niemi. (Orgs.). **Studies in Language** 41 - Collocations and Idioms 1. Joensuu: University of Joensuu, 2006, v. 41, p. 354-362.

GARRÃO, M. U. Lingüística de Córpus: o lugar da fusão entre semântica e pragmática. **Calidoscópio** (UNISINOS), v. 04, p. 135-140, 2006.

JACKENDOFF, R. **The Architecture of the Language Faculty**. Cambridge, Massachusetts: MIT Press. 1997.

KILGARRIFF, A. **I don't believe in word senses**. 2000. Disponível em [http://www.kcl.ac.uk/humanities/cch/seminar/99-00/seminar\\_kilgarriff.html](http://www.kcl.ac.uk/humanities/cch/seminar/99-00/seminar_kilgarriff.html) Acesso em 8 de outubro de 2005.

NEVES, M. H. M. A delimitação das unidades lexicais: o caso das construções com verbo-suporte. In Basílio, M. (org.) A delimitação de unidades lexicais. **Palavra** n° 5. Rio de Janeiro: Departamento de Letras da PUC, 1999, p.98-114.

RADFORD, A. **Transformational grammar: a first course**. Cambridge, Inglaterra: Cambridge University Press, 1988. **crossref** <http://dx.doi.org/10.1017/CBO9780511840425>

SAUSSURE, F. (1916) **Curso de Linguística Geral**. Cultrix: São Paulo, 1975.

VALE, O. **Expressões Cristalizadas no português do Brasil**: uma proposta de tipologia. Tese de Doutorado, Araraquara: UNESP. 2002.

WITTGENSTEIN, L. (1953) Investigações Filosóficas. **Coleção Os Pensadores**, São Paulo: Abril Cultural, 1979.

Artigo recebido em: 30.03.2015

Artigo aprovado em: 21.06.2015

Domínios de Lingu@gem