A ocorrência de metáforas é previsível?

Is metaphor occurrence predictable?

Tony Berber Sardinha
Pontifícia Universidade Católica de São Paulo

Resumo

Um dos grandes obstáculos à disseminação da pesquisa sobre o uso de metáforas em *corpora* tem sido a grande demanda de tempo exigida na identificação das ocorrências metafóricas. Esta pesquisa tenta verificar se é possível prever a frequência das metáforas com acuidade, diminuindo ou eliminando a identificação manual. Para tanto, foi usado um *corpus* já anotado manualmente por inteiro, o VUAMC (*Vrije Universiteit Amsterdam Metaphor Corpus*), e, a partir dele, foram construídos modelos matemáticos por meio de Análise de Regressão. Esses modelos geraram previsões da frequência de metáforas em cada texto do *corpus*, que foram contrastadas com as frequências obtidas pela análise manual. Os resultados mostram que os valores estimados são de boa qualidade, com correlação em torno de 80% e média de 14% de discrepância. Os resultados sugerem, portanto, que a previsão das frequências de metáforas por meio de Análise de Regressão é uma alternativa viável e deve ser contemplada nos estudos de metáfora com *corpus*.

Palavras-chave

Linguística de *Corpus*, Metáfora, Análise Multi-Dimensional.

Abstract

One of the major impediments to *corpus*-based research on metaphor has been the great amount of time needed to code metaphors manually. The main goal of the research reported here

is to verify to what extent metaphor frequency for whole texts can be predicted by using metaphor frequency data from small text samples or from no metaphor frequency data at all. To meet this goal, a *corpus* that had been fully annotated for metaphor by hand was used, namely VUAMC (Vrije Universiteit Amsterdam Metaphor Corpus), and statistical models were derived from it through Regression Analysis, having metaphor frequency as a dependent variable. The models obtained generated predictions about the frequency of metaphor in the *corpus*, which were then contrasted with the frequencies based on the manual analysis. Results showed that the predicted values were generally accurate, with correlations around 80% and metaphor count differences between observed and predicted values around 14%. Results therefore suggest that predicting metaphor frequencies through Regression Analysis is a viable alternative and should be taken into consideration in *corpus*-based metaphor research.

Keywords

Corpus Linguistics, Metaphor, Multi-Dimensional Analysis

Introdução

Os estudos da metáfora em uso têm se aproximado da Linguística de *Corpus* (BERBER SARDINHA, 2004; BIBER, CONRAD; REPPEN, 1998; SHEPHERD, BERBER SARDINHA; VEIRANO PINTO, 2012) há pelo menos 13 anos, com a publicação de '*corpus-based research into metaphor*', de Alice Deignan (DEIGNAN, 1999), na coletânea de Cameron e Low (CAMERON; LOW, 1999). Atualmente, a pesquisa em metáfora com *corpora* tem crescido muito e se tornado influente na área como um todo (GIBBS, 2008a).

Metáfora é entendida aqui como o uso de uma palavra que pode 'ser potencialmente explicado por alguma forma de mapeamento entre domínios a partir de um sentido mais básico da palavra (STEEN; DORST; HERRMANN *et al.*, 2010, p. 40). Esse conceito informa o procedimento MIPVU (*Metaphor Identification Procedure Vrije University*), usado para identificar as metáforas no *corpus* empregado neste trabalho.

A pesquisa em metáfora baseada em *corpora* tem uma grande contribuição a fazer aos estudos da metáfora (BERBER SARDINHA, 2011; DEIGNAN, 2005; GIBBS, 2008b), em várias frentes, que incluem as seguintes:

- Forma das metáforas na língua: uma vez que as metáforas existem na língua em uso, é necessário sabermos que forma elas tomam, ou seja, quais são seus padrões de uso. Com base nisso, podemos saber se há preferências léxico-gramaticais associadas ao uso metafórico da língua, em oposição ao uso não metafórico.
- Ocorrência de metáforas: na medida que as metáforas existem na língua em uso e se materializam em determinadas formas nos textos, é legítimo perguntar qual é sua frequência nesses textos, escritos e falados. A sua presença ou ausência é significativa para a caracterização da língua em uso em contextos determinados.

- Variação de uso de metáforas: havendo metáforas que assumem certas formas e que se materializam ou não em textos, a partir de coerções diversas, surge a questão de como se dá essa variação, quais fatores a influenciam, e até que ponto essa variação é sistemática.

Todas essas questões (além de outras não levantadas aqui) são prementes e exigem resposta para que possamos melhorar nosso entendimento de como, quando, por quem e por que as metáforas são usadas. Questões que demandam o uso de *corpora* no estudo de metáforas enfrentam um grande empecilho: a quantidade de trabalho envolvida na identificação manual de metáforas (BERBER SARDINHA, no prelo-a). Para se ter ideia da extensão do problema, basta dizer que o *corpus* de estudo utilizado aqui, o VUAMC (*Vrije Universiteit Amsterdam Metaphor Corpus*), uma amostra de 84 arquivos de texto que somam 205 mil palavras (*tokens*), levou cinco anos para anotar (STEEN; DORST; HERRMANN *et al.*, 2010, p. 7), por uma equipe com seis pesquisadores (Gerard Steen, Lettie Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr e Trijntje Pasma). O fator tempo, portanto, é sem dúvida um dos grandes obstáculos à disseminação dos estudos de metáforas com base em *corpora*.

Há pelo menos três maneiras de reduzir o trabalho manual de anotação, quais sejam, o emprego de amostragem, a utilização de métodos automáticos de identificação de metáfora e a aplicação de métodos estatísticos de previsão da frequência. A primeira opção implica a seleção prévia de apenas alguns casos para serem analisados manualmente. Essa opção é bastante usada na literatura (p.ex. DEIGNAN, 2005) e tem como vantagem o foco em um certo aspecto do *corpus* (p.ex. léxico relacionado a um determinado tópico), mas sua desvantagem principal é o viés que a seleção de antemão introduz na pesquisa, pois serão levadas em conta apenas as metáforas existentes na amostra definida *a priori*, e não a totalidade de metáforas do *corpus*. A segunda opção, por sua vez, ainda é pouco difundida na literatura, devido à falta de programas de computador que automatizem a identificação de metáforas com segurança. Ela tem como principais vantagens a economia de tempo na anotação e a possibilidade de levar em conta o *corpus* inteiro, mas seu principal problema é a baixa precisão da identificação, em comparação à anotação feita por seres humanos. Por fim, a terceira opção tem como principal atrativo a possibilidade de calcular a frequência de metáforas no *corpus* inteiro com base em amostras desse *corpus*. Ou seja, ela pode combinar a abordagem por amostragem com a do censo, ou seja, o

pesquisador pode ter uma estimativa confiável da frequência das metáforas do *corpus* inteiro a partir da análise manual de amostras. Essa opção, que ainda não foi testada na literatura, é o objeto de investigação deste trabalho.

Assim, neste artigo, enfocamos a segunda das frentes de pesquisa apontadas acima, isto é, a frequência de metáforas na língua em uso, com o auxílio do procedimento estatístico Análise de Regressão Múltipla. Mais precisamente, utilizamos esse procedimento para verificar até que ponto podemos prever a frequência de metáforas no VUAMC. Caso a resposta seja positiva, isto é, exista a possibilidade de prever de modo confiável a frequência de uso de metáforas nesse *corpus*, podemos aproveitar os modelos matemáticos (equações) baseados nele para prever a frequência de metáforas em outros *corpora*, sem que haja necessidade de anotá-las todas à mão.

Metodologia

Perguntas de pesquisa

1. Quais modelos de regressão conseguem estimar melhor a frequência de metáforas?
2. Qual o melhor conjunto de modelos de regressão em termos da economia do trabalho de anotação manual das metáforas?
3. Qual a quantidade de metáforas estimada a mais ou a menos pelos modelos?
4. Qual é a precisão dos modelos que não envolvem análise manual de metáforas?
5. Se as frequências estimadas pelos modelos de regressão fossem resultado da identificação manual de um segundo analista, qual seria o grau de confiabilidade da análise estatística?
6. Como esses modelos podem ser aplicados na prática?

Essas perguntas serão respondidas na seção de Resultados.

Corpus usado nesta pesquisa

Conforme dito acima, o *corpus* usado nesta pesquisa foi o *Vrije Universiteit Amsterdam Metaphor Corpus* (VUAMC), que é uma amostra do BNC-Baby (versão reduzida do *British National Corpus*). O VUAMC é distribuído gratui-

tamente no Oxford Text Archive (OTA) e vem etiquetado morfossintaticamente pelo etiquetador CLAWS. Para esta pesquisa, essa etiquetagem foi removida e o *corpus* foi etiquetado novamente pelo *Biber Tagger*. Essa versão etiquetada pelo *Biber Tagger* do VUAMC tem a seguinte composição:

TABELA 1
Composição do VUAMC anotado com o *Biber Tagger*

Registro	Palavras (*tokens*)	Textos
Acadêmico	68.276	15
Conversação	48.768	11
Ficção	45.663	12
Jornalístico	46.208	46
Total	208.915	84

Anotação de metáfora: o procedimento MIPVU

Por anotação de metáfora, entende-se o processo de identificar e sinalizar no texto (em papel ou em arquivo de computador) a presença de metáforas. A anotação pode ser feita manualmente ou por máquina, total ou parcialmente. Conforme dito acima, o VUAMC foi inteiramente anotado manualmente, por uma equipe liderada por Gerard Steen. O método de identificação empregado foi o MIPVU (*Metaphor Identification Procedure Vrije Universiteit*) (STEEN; DORST; HERRMANN *et al.*, 2010, p. 25), uma variação do MIP (*Metaphor Identification Procedure*) (PRAGGLEJAZ GROUP, 2007). O MIPVU consiste nos seguintes passos:[1]

1. Encontre palavras relacionadas à metáfora (MRW) examinando o texto palavra a palavra.
2. Quando uma palavras for usada indiretamente e esse uso puder ser potencialmente explicado por alguma forma de mapeamento entre domínios a partir de um sentido mais básico da palavra, anote a palavra como sendo metaforicamente usada (MRW).
3. Quando uma palavra for usada diretamente e seu uso puder ser potencialmente explicado por alguma forma de mapeamento entre domínios a partir de um sentido mais básico da palavra, anote a palavra como sendo metaforicamente usada (MRW, *direct*).

4. Quando palavras forem usadas com o propósito de substituição léxico-gramatical, como pronomes de terceira pessoa ou na ocorrência de elipse, em que algumas palavras podem ser entendidas como se estivessem faltando, como em algumas formas de coordenação, e quando um sentido direto ou indireto for produzido por essas substituições ou elipses que podem ser potencialmente explicadas por alguma forma de mapeamento entre domínios a partir de um sentido mais básico, ou de um referente ou tópico, insira o código de metáfora implícita (MRW, *implicit*).
5. Quando uma palavra funcionar como um índice de um mapeamento entre domínios, anote-a como um sinalizador de metáfora (MFlag).
6. Quando uma palavra for uma cunhagem de uma nova forma, examine as palavras que a formam, de acordo com os passos 2 a 5. (STEEN; DORST; HERRMANN *et al.*, 2010, p.25)

Os três tipos de metáfora indicadas no procedimento são ilustrados abaixo.

a. Metáfora direta: palavras cujo estatuto metafórico é sinalizado diretamente (STEEN; DORST; HERRMANN *et al.*, 2010, p. 39), por palavras como *resembling*, *like* ou *as*. Por exemplo:[2] *words started as a coat-hanger to hang pictures on* (KRENNMAYR, 2011, p. 31) [palavras começaram como um cabide para pendurar quadros].
b. Metáfora indireta: palavras cuja metaforicidade não é sinalizada explicitamente (STEEN; DORST; HERRMANN *et al.*, 2010, p. 33). Este é o modo *default* de manifestação das metáforas nessa perspectiva. Por exemplo: *high wages* (KRENNMAYR, 2011, p. 31). [Altos salários]
c. Metáfora implícita: palavras cujo estatuto metafórico é realizado por substituição (e.g.: *it* em *to embark on such a step is not necessarily to succeed immediately in realizing it* [tomar tal passo não significa necessariamente ter sucesso em realizá-lo], em que *it* refere-se à palavra usada metaforicamente *step*) or elipse (*but he is (an ignorant pig)* [mas ele é (um porco ignorante)], em que *is* recebe o código de metáfora implícita, pois refere-se a 'porco ignorante', que é metafórico (STEEN; DORST; HERRMANN *et al.*, 2010, p. 40).

Nas análises aqui apresentadas, considerou-se metáfora apenas o tipo 'indireto', isto é, aquela referente ao passo 2 do procedimento, pois as demais não tinham frequência alta o bastante para serem tratadas estatisticamente pela Análise de Regressão.

Como se percebe, o MIPVU é altamente laborioso, pois exige atenção a cada palavra do texto e, para cada uma, a definição de seu sentido básico e contextual. Em muitos casos, para a determinação do sentido básico, é preciso consultar dicionários, contemporâneos ou históricos. Tudo isso faz com que o tempo necessário para a aplicação do procedimento seja muito elevado, quando se trata de análise de *corpora* com dezenas ou centenas de milhares de palavras. O VUAMC levou cerca de seis anos para ser anotado por completo pelo MIPVU. Foram identificadas 26.686 metáforas nesse *corpus* (STEEN; DORST; HERRMANN *et al.*, 2010, p. 177).

O MIPVU pode ser potencialmente aplicado a qualquer língua (por exemplo, PASMA, 2011, que o empregou na análise de um *corpus* de holandês), embora tenha sido originalmente desenvolvido com dados do inglês.

Análise de Regressão: uma brevíssima introdução

A metodologia desta pesquisa baseia-se na Análise de Regressão Múltipla. A Análise de Regressão é um procedimento estatístico cuja aplicação visa a prever ocorrências futuras por meio de um conjunto de observações efetuadas. A Análise de Regressão é muito empregada em diversos campos de atuação humana, desde áreas acadêmicas como a Economia, Ciências da Saúde e Educação, até campos profissionais, como os segmentos de seguros, segurança pública e previsão do tempo. Até onde pudemos descobrir, esta é a primeira aplicação deste procedimento no estudo de metáforas.

Para explicar a Análise de Regressão, é útil começar com um exemplo. Suponhamos que queiramos verificar a relação entre anos de escolaridade, renda dos pais e renda dos filhos adultos. Nossa hipótese é a de que indivíduos com mais escolaridade e cujas famílias tenham mais renda terão renda individual mais alta. Coletamos dados referentes a uma amostra de indivíduos e para cada um anotamos quantos anos de ensino cursou, qual a renda dos pais e qual a sua renda, resultando em uma tabela como a TAB. 2.

TABELA 2
Anos de escolaridade, renda dos pais e renda dos filhos adultos (dados fictícios)

Indivíduo	Anos de escolaridade do filho	Renda dos pais	Renda do filho adulto
1	0	1	1
2	8	8	3
3	12	10	4
4	16	12	8
5	20	22	12

Em seguida, criamos dois gráficos de dispersão, um para escolaridade e renda individual, e outro para escolaridade e renda familiar, conforme mostra a FIG. 1.

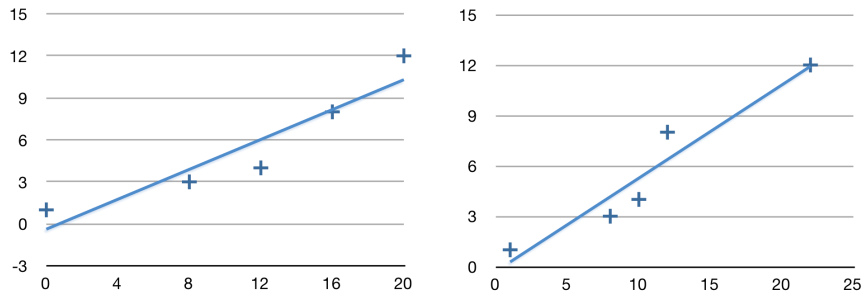

FIGURA 1: Gráficos de dispersão: esolaridade e renda individual (à esquerda) e renda dos pais e renda individual (à direita)

Notamos, nesses gráficos, que os pontos formam uma linha ascendente, que pode ser visualizada por uma reta que passa na menor distância entre os pontos. Ela mostra uma relação quase perfeita entre a variável dependente (renda) e as independentes (renda dos pais e escolaridade). As variáveis independentes são chamadas de estimadoras, pois servem para estimar um valor previsto para a variável dependente. Esses valores estimados são calculados por um modelo, i.e. uma equação, que cria uma reta que se ajusta da melhor maneira possível aos dados observados. Na Análise de Regressão Simples, há apenas uma variável independente, portanto apenas um plano, mas na Análise de Regressão Múltipla, há mais de uma variável independente, portanto múltiplos planos.

Nesse exemplo, há duas variáveis independentes, portanto para visualizar o gráfico de dispersão correspondente, precisamos de um plano tridimensional, o que não pode ser representando numa página de papel bidimensional. O leitor tem de imaginar uma reta que cruza esse espaço tridimensional e se ajusta o melhor possível para mostrar a relação ascendente entre as variáveis, chegando o mais perto possível dos pontos existentes. Havendo mais de duas variáveis independentes, precisamos de dimensões extras, que não podemos representar fisicamente, mas que a matemática é capaz de conceituar. Essa reta é criada por um modelo matemático, uma equação, que tem a seguinte forma:

$$Y' = á + â_1x_1 + â_2x_2 + â_nx_n$$

Onde:

- Y' = Variável dependente
- á = Intercepto
- â = coeficiente angular
- x= valor observado

O intercepto é o ponto onde a reta cruza o eixo Y, sendo assim um valor constante. O coeficiente angular é um parâmetro calculado pela Análise de Regressão que deve ser multiplicado pelo valor observado.

Nesse exemplo, o modelo é:

- á = 0,413
- $â_{escolaridade}$ = 0,135
- $â_{renda\ dos\ pais}$ = 0,425

Substituindo esses valores na fórmula, temos:

Renda prevista = -0,413 + anos de escolaridade X 0,135 + renda dos pais X 0,425

Aplicando esse modelo ao segundo indivíduo, temos:

$$Renda\ prevista_{indivíduo\ 2} = -0,413 + 8\ X\ 0,135 + 8\ X\ 0,425 = 4,06$$

A renda observada é 3, portanto houve uma discrepância de 1,06 entre os valores observado e previsto. A distância entre os valores observados e previstos

é denotada pelo Erro Padrão da Estimativa, que nesse exemplo é 1,688. A proximidade entre os valores observados e previstos é representada por um coeficiente de correlação. Na Análise de Regressão Múltipla, emprega-se o coeficiente de correlação múltipla, R, que mede a precisão dos valores estimados a partir do conjunto de variáveis estimadoras, indo de 0 a 1. Também se utiliza R^2, que indica a quantidade de variância em comum entre a variável dependente e as estimadoras, que também vai de 0 a 1. Nos dados usados como exemplo, obtivemos R = 0,962, que indica uma relação quase perfeita entre renda dos filhos, renda dos pais e escolaridade, e R^2 = 0,926, que mostra que perto de 93% da variação da renda dos filhos pode ser prevista sabendo sua escolaridade e a renda dos pais. Com isso, a hipótese inicial, que previa uma relação entre essas variáveis, foi mantida. Além disso, o modelo obtido pode ser usado para estimar a renda de uma amostra da população, sabendo apenas a escolaridade e a renda da família.

Fazendo a ponte entre esse exemplo e o problema de pesquisa tratado neste trabalho, a renda individual seria a frequência de metáforas, enquanto anos de escolaridade e renda da família seriam outras variáveis de cunho linguístico e textual teoricamente associadas à presença de metáforas na língua.

Há ainda muitos outros pontos a tratar sobre Análise de Regressão, mas, nesta seção, detivemo-nos nos pontos estritamente necessários para que, em seguida, o leitor possa entender o funcionamento básico do procedimento.

Análise de Regressão para estimar a frequência de metáforas

Nesta seção, são detalhados os procedimentos seguidos na aplicação da Análise de Regressão para o cálculo da frequência de metáforas. As variáveis dependentes são:

1. total de metáforas: a quantidade de metáforas de um texto, normalizada para mil palavras. Por exemplo, se um texto tem 2000 palavras e 10 metáforas, o valor desta variável será 5 (i.e. 10/2000 x 1000).
2. pacotes lexicais metafóricos: a quantidade de pacotes lexicais (*bundles*) metafóricos em um texto, normalizada por mil palavras. Um pacote lexical (cf. BIBER; CONRAD; CORTES, 2004) é definido aqui como um trigrama (i.e. uma sequência de três palavras) que (1) contém uma metáfora e (2) aparece na lista de trigramas do COCA (*Corpus* of Contemporary American

English, corpus.byu.edu/coca) (BERBER SARDINHA, no prelo-c). Essa lista inclui trigramas com frequência igual ou maior a 25 na língua inglesa. Essa variável representa o grau de convencionalidade das metáforas, pois os pacotes lexicais são unidades léxico-gramaticais convencionais.

3. agrupamentos metafóricos: a quantidade de agrupamentos (*clusters*) metafóricos de um texto, normalizada por mil palavras. Um agrupamento metafórico é uma sequência de pelo menos três metáforas separadas por no máximo oito palavras, que é a distância média entre as metáforas no VUAMC (BERBER SARDINHA, no prelo-c).
4. adjetivos metafóricos: a quantidade de adjetivos metafóricos de um texto, normalizada por mil palavras.
5. Substantivos metafóricos: a quantidade de substantivos metafóricos de um texto, normalizada por mil palavras.
6. Preposições metafóricas: a quantidade de preposições metafóricas de um texto, normalizada por mil palavras.
7. Verbos metafóricos: a quantidade de verbos metafóricos de um texto, normalizada por mil palavras.

As variáveis independentes ou estimadoras usadas nesta pesquisa são de dois tipos: referentes a (1) dimensões de variação e (2) amostragem de metáforas.

As variáveis referentes a dimensões de variação baseiam-se nas dimensões de variação do inglês de Biber. Uma dimensão de variação é um parâmetro subjacente de variação linguística de registros (BERBER SARDINHA, 2000; BIBER, 1988). Registros, por sua vez, são variedades textuais definidas situacionalmente, podendo ser bastante amplas (como escrita acadêmica) ou bastante específicas (como editoriais de jornal). As dimensões de variação são calculadas segundo a metodologia da Análise Multidimensional, uma vertente da Linguística de *Corpus* que visa a caracterizar textos ou variedades textuais a partir de uma análise exaustiva de sua composição linguística e da consequente interpretação funcional dos elementos linguísticos coocorrentes nos textos. São cinco as dimensões de variação do inglês, a saber:

TABELA 3
Dimensões de variação do inglês (BIBER, 1988; 2009)

#	Rótulo em inglês	Tradução para o português
1	Involved *versus* Informational Production	Produção marcada por envolvimento *versus* informacional
2	Narrative *versus* Non-narrative discourse	Discurso narrativo *versus* não narrativo
3	Situation-dependent *versus* elaborated reference	Referência dependente de situação *versus* elaborada
4	Overt expression of argumentation	Argumentação explícita
5	Abstract *versus* non-abstract style	Estilo abstrato *versus* não-abstrato

Um escore de dimensão, por sua vez, é um valor que mede a posição do texto em relação a uma determinada dimensão. Cada texto do *corpus* recebe, assim, cinco escores, um para cada dimensão. O escore é composto pela contagem das características linguísticas coocorrentes nos textos do *corpus*, determinadas estaticamente por meio de Análise Fatorial. Nesta pesquisa, não foi realizada uma Análise Fatorial para determinar quais características linguísticas compõem cada dimensão; ao contrário, a composição linguística das dimensões reflete aquela determinada em Biber (1988). Os escores de dimensão podem ser positivos, negativos ou zero. Um escore positivo indica que o texto pode se caracterizado segundo o primeiro parâmetro constante na dimensão, mas se for negativo, o texto reflete o segundo parâmetro da dimensão. Por exemplo, se um texto possuir escore positivo na dimensão 1, ele pode ser considerado como tendo 'produção marcada por envolvimento', ao passo que se tiver escore negativo, pode ser visto como 'marcado por produção informacional'; se seu escore for zero, ele não é marcado por nenhum parâmetro dessa dimensão. Do mesmo modo, se tiver escore positivo na dimensão 2, é tido como 'narrativo', e se tiver escore negativo, 'não-narrativo', e zero, não marcado para narratividade. Ao contrário das demais, a dimensão 4 não é bipolar, dessa forma ela caracteriza os textos como tendo 'argumentação explícita', se tiverem escore positivo, 'argumentação inexplícita', se tiverem escore negativo, ou não marcado para argumentação, se tiverem zero de escore.

O intuito da adoção dessas variáveis de dimensão nesta pesquisa é de capturar a relação entre metáfora e variedades textuais, mais especificamente no nível

de registro. Por registro, entende-se uma forma socialmente convencionalizada de expressão textual, escrita ou falada, situacionalmente definida (BIBER, 1988). Nas diferentes vertentes de estudos linguísticos, há vários termos que concebem, cada um a seu modo, as variedades textuais, como gênero, registro e tipo textual (cf. BERBER SARDINHA, 2009). Na Análise Multidimensional, o termo usado para definir uma variedade textual situacionalmente definida é registro, por isso adotamos essa terminologia neste trabalho. Muitas pesquisas sobre uso metafórico deixam claro que a ocorrência de metáforas na língua é influenciada pelo registro em que a metáfora é veiculada. Assim, em registros como a conversação informal e narrativas de vida, há tipicamente escassez de metáforas (BERBER SARDINHA, 2008c; KAAL, 2012), ao passo que, no jornalismo e nos negócios, há profusão de metáforas (BERBER SARDINHA, 2008a; b; KRENNMAYR, 2011).

Os escores de texto foram calculados pelo *TagCount*, que por sua vez exige textos etiquetados pelo *Biber Tagger*. O *Biber Tagger* etiqueta cada texto morfossintaticamente, anotando de modo automático a ocorrência de centenas de características linguísticas. O *TagCount*, por sua vez, processa esses textos etiquetados, contando e sintetizando as etiquetas em torno de 128 características. Ele calcula o escore de dimensão de cada texto, com base em 46 características linguísticas (cf. BERBER SARDINHA, no prelo-b). Ambos os programas foram desenvolvidos por Doug Biber e estão disponíveis aos membros do Grupo de Estudos de Linguística de *Corpus* (GELC) da PUCSP.

São cinco as variáveis independentes relativas a dimensões de variação:

1. dim1: O escore do texto na dimensão de variação 1.
2. dim2: O escore do texto na dimensão de variação 2.
3. dim3: O escore do texto na dimensão de variação 3.
4. dim4: O escore do texto na dimensão de variação 4.
5. dim5: O escore do texto na dimensão de variação 5.

O segundo tipo de variáveis independentes são as de amostragem de metáfora, assim chamadas porque refletem a frequência de metáfora em amostras do texto. As frequências são obtidas a partir da identificação manual das metáforas. Como o VUAMC já havia sido anotado manualmente por completo, não houve necessidade de analisar manualmente cada amostra para saber qual a frequência de metáforas. No caso de variáveis baseadas em classes gramaticais

(substantivos, adjetivos, preposições e verbos), o procedimento adotado foi o seguinte. Primeiramente, foram identificadas todas as ocorrências dessa classe gramatical nos textos, a partir da etiquetagem produzida pelo *Biber Tagger*. Em seguida, foi feito um recorte das ocorrências desejadas, e por fim, efetuada a contagem das ocorrências de metáforas dentre essas ocorrências. Por exemplo, para obter o valor da variável mets100j, foi preciso primeiro selecionar os 100 primeiros adjetivos do texto, e em seguida contar quantos adjetivos entre esses 100 eram metafóricos. Por fim, esse valor foi digitado em um banco de dados. O mesmo procedimento foi executado para as demais variáveis gramaticais. Já no caso das variáveis baseadas em amostras de palavras corridas, o procedimento adotado foi o seguinte. Primeiramente, foi selecionada a amostra desejada (as primeiras 100, 200, 300, 400 ou 500 palavras de cada texto) e em seguida foram contadas as metáforas existentes na amostra. A contagem resultante foi inserida no banco de dados.

As variáveis independentes de amostragem de metáfora são 11:

1. mets100j: A frequência de metáforas entre os 100 primeiros adjetivos do texto.
2. mets100p: A frequência de metáforas entre as 100 primeiras preposições do texto.

3, 4. mets100n, mets200n: A frequência de metáforas entre os 100 e 200 primeiros substantivos do texto, respectivamente.

5, 6. mets100v, mets200v: A frequência de metáforas entre os 100 e 200 primeiros verbos do texto, respectivamente.

7, 8, 9, 10, 11. mets100w, mets200w, mets300w, mets400w, mets500w: A frequência de metáforas entre as 100, 200, 300, 400 e 500 primeiras palavras (*tokens*) do texto, respectivamente.

Desse modo, as variáveis independentes somam 16.

Uma vez feita a contagem dessas variáveis e a digitação dos seus respectivos valores, os dados foram preparados para serem processados pelo programa SPSS 20 para Macintosh. O comando para realizar uma Análise de Regressão no SPSS é *Analyze* > *Regression* > *Linear*. O método de análise foi *Stepwise*, que seleciona as variáveis independentes que farão parte do modelo por meio de critério estatístico, deixando apenas aquelas variáveis estatisticamente significativas. O comando executado para cada Análise de Regressão, em sintaxe SPSS, foi o seguinte:

```
REGRESSION
/MISSING LISTWISE
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT all_METMET
/METHOD=STEPWISE dim1 dim2 dim3 dim4 dim5 mets100w
mets200w mets300w mets400w mets500w mets100p mets100j
mets100n mets200n mets100v mets200v
```

Esse comando especifica como variável dependente (/DEPENDENT) o total de metáforas (cujo código é all_METMET) e emprega o método *Stepwise* (METHOD=STEPWISE), o qual inclui uma variável por vez no modelo e calcula a significância dessa variável; se a significância for menor que 0,5 (PIN(.05)), a variável é mantida, caso contrário (PIN(.10)), ela é removida. As demais variáveis listadas também sofrem o mesmo processamento e a cada uma que entra é calculada sua contribuição para aumentar o valor de R^2 (a variância em comum entre a variável dependente e as independentes), de tal forma que restarão no modelo apenas aquelas que mais contribuam para sua robustez. Assim, esse método baseia-se no princípio da parcimônia, criando modelos com o menor conjunto de variáveis independentes que melhor estime o valor da variável dependente. Uma decorrência positiva disso é a ausência de colinearidade (correlação entre variáveis independentes), pois caso haja relação mútua entre duas variáveis independentes, apenas uma será mantida. Cada comando de análise pode gerar muitos modelos, caso haja muitas combinações de variáveis independentes estatisticamente significativas. O método *Stepwise* requer a existência de pelo menos cinco vezes mais casos do que variáveis independentes; nesta pesquisa, há 16 variáveis independentes e 84 textos (casos), o que dá uma razão aceitável de 5,25.

Algumas variáveis independentes de amostragem são mutuamente excludentes, como mets100n e mets200n, pois a análise de 200 substantivos pressupõe a de 100 substantivos. Por isso, tais variáveis nunca entraram juntas no comando de regressão, pois uma delas seria redundante, prejudicando o modelo. Assim, a quantidade de comandos de análise para cada variável dependente foi de 196, e o total geral, de 1.372.

Essa quantidade de comandos de análise exigiu processamento por lote, que no SPSS 20 leva o nome de *Production Facility*. Essa opção, por sua vez, exige arquivos de texto (.spj) com uma sintaxe particular, os quais foram gerados por um *script* especialmente produzido para esta pesquisa. Esses arquivos .spj foram então rodados no SPSS e os resultados salvos em arquivos texto. Esses arquivos de texto foram posteriormente processados por outro *script* especial que localizou os modelos gerados e seus parâmetros, os quais foram então importados em um banco de dados, para que fossem encontrados os melhores modelos de estimação. O total de modelos gerados foi de 56.771.

Resultados

Pergunta 1: Melhores modelos individuais

Para responder a essa pergunta, foi identificado o modelo com a maior correlação múltipla (R) para cada variável dependente.

TABELA 4
Maiores correlações para cada variável dependente

variável dependente	Correlação (R)	R^2
total de metáforas	0,787	62%
pacotes lexicais metafóricos	0,758	57%
agrupamentos metafóricos	0,821	67%
adjetivos metafóricos	0,813	66%
substantivos metafóricos	0,791	63%
preposições metafóricas	0,827	68%
verbos metafóricos	0,752	57%
Média	0,792	63%

As correlações obtidas são todas altas (acima de 0,7), chegando ao máximo de 0,827, com média de 0,792. A quantidade de variância em comum (R^2) vai de 57% a 68%, com média de 63%, o que indica que a maioria da variação da frequência das variáveis dependentes pode ser prevista pelos modelos de regressão. Em outras palavras, 63% da frequência de metáforas em um texto é previsível com base em sua ocorrência em uma parcela desse texto.

Os modelos são os seguintes:

- Total de metáforas = 63,009 + dim1 x -1,093 + mets100j x 1,460 + mets100n x 1,027
- Pacotes lexicais metafóricos = 97,854 + dim1 x -0,555 + mets100j x 1,017 + dim3 x 3,036 + dim4 x 5,291
- Agrupamentos metafóricos = -14,987 + dim1 x -0,791 + mets100j x 1,308 + mets100n x 0,871 + mets100p x 1,329
- Adjetivos metafóricos = 8,333 + dim1 x -0,134 + mets100w x 0,653 + mets200n x -0,083 + dim2 x -0,638
- Substantivos metafóricos = 12,341 + dim1 x -0,336 + dim2 x -1,236 + mets200v x 0,662
- Preposições metafóricas = 15,752 + dim1 x -0,611 + mets100n x 1,010 + mets200v x -0,396
- Verbos metafóricos = 16,662 + mets100j x 0,843 + mets500w x -0,090

Em termos do conjunto de modelos, são 11 as variáveis estimadoras presentes:

1. dim1
2. dim2
3. dim3
4. dim4
5. mets100j
6. mets100n
7. mets100p
8. mets100w
9. mets200n
10. mets200v
11. mets500w

Em termos do trabalho de anotação de metáforas, esse conjunto de modelos exige a identificação manual de 1.100 palavras por texto, ou seja: 100 adjetivos, 100 preposições, 200 substantivos, 200 verbos e as 500 primeiras palavras corridas do texto.

Tomando o *corpus* VUAMC como base, a economia média de anotação representada por esses parâmetros aparece na TAB. 5.

TABELA 5
Economia de anotação manual

Variável	Anotação exigida	Economia de anotação
Adjetivos	71,2%	28,8
Substantivos	54,1%	45,9%
Preposições	61%	39%
Verbos	67,7%	32,3%
Palavras corridas	39,3%	60,7%
Média	59%	41%

A quantidade de palavras necessária para anotar no *corpus* VUAMC é de 59%, o que significa uma economia média de anotação de 41%.

Pergunta 2: Melhor conjunto de modelos

Com essa pergunta, estamos interessados em obter um conjunto de modelos que diminua ao máximo a quantidade de trabalho de anotação manual envolvido. Para tanto, foram buscados, entre todos os modelos obtidos, aqueles que atendessem a três critérios: ter pelo menos uma correlação de 0,7, exigirem a menor quantidade de anotação em si, e não aumentarem a quantidade de anotação total do conjunto. Por exemplo, se para uma mesma variável dependente houvesse dois modelos, ambos com correlação igual ou maior a 0,7, sendo que um deles exigisse a anotação de 200 substantivos e outro 100 substantivos, o segundo seria preferido, pois a quantidade de dados a serem anotados seria menor. Ou se ambos os modelos exigissem a mesma quantidade de anotação, mas um deles não adicionasse um parâmetro novo ao conjunto, esse último seria preferido – por exemplo, se um modelo exigisse a anotação manual de 200 substantivos, e outro 100 adjetivos mais 100 preposições, mas outros exigissem 100 adjetivos e 100 preposições, mas não 200 substantivos, o modelo baseado nos 100 adjetivos mais 100 preposições seria escolhido, pois ele não causaria aumento de palavras a serem anotadas, já que suas variáveis independentes já fazem parte de outros modelos.

Com a aplicação desse procedimento, houve alteração nos modelos referentes a adjetivos, preposições, substantivos e verbos metafóricos (os demais permaneceram inalterados, tal como apresentados nos resultados da pergunta anterior), conforme detalha a TAB. 6.

TABELA 6
Modelos de regressão mais econômicos

Variável dependente	Modelo menos econômico*		Modelo mais econômico	
	Parâmetros (anotação exigida no conjunto)	Correlação	Parâmetros (anotação exigida)	Correlação
Adjetivos metafóricos	mets100w**, dim1, dim2, mets200n (200)	0,813	mets100w, dim1, dim2 (100)	0,771
Preposições metafóricas	mets100n**, mets200v, dim1 (200)	0,827	dim1, mets100n (100)	0,794
Substantivos metafóricos	mets200v**, dim1, dim2 (0)	0,791	mets100j, dim1, mets100v, dim2 (200)	0,786
Verbos metafóricos	mets100j, mets500w (600)	0,752	mets100j** (0)	0,732
Total de anotação	1100		400	

* As variáveis dim1 e dim2 não envolvem anotação manual, portanto não foram computadas na soma da anotação exigida pelo conjunto.

** Parâmetro não incluído na soma porque já está pressuposto em outro modelo do conjunto.

Como mostra a tabela, houve uma redução de 600 palavras de anotação manual, que é considerável. A perda do poder de previsão dos modelos foi pequena: apenas 0,069 nos coeficientes de correlação, ou 0,48% da variância em comum. Os modelos são os seguintes:

- Adjetivos metafóricos = 5,921 + mets100w x 0,522 + dim1 x -0,146 + dim2 x -0,636
- Preposições metafóricas = 13,409 + dim1 x -0,539 + mets100n x 0,760
- Substantivos metafóricos = 7,724 + mets100v x 0,703 + mets100j x 0,469 + dim1 x -0,297 + dim2 x -1,263
- Verbos metafóricos = 14,647 + mets100j x 0,812

São nove os parâmetros exigidos pelo conjunto:

1. dim1
2. dim2
3. dim3
4. dim4
5. mets100j
6. mets100n
7. mets100p
8. mets100v
9. mets100w

O total de anotação manual necessária é de 500 palavras por texto (100 adjetivos, 100 substantivos, 100 preposições, 100 verbos e 100 palavras corridas), ante 1.100 do conjunto anterior. A economia de anotação aparece na TAB. 7.

TABELA 7
Economia de anotação com modelos mais econômicos

Variável	Anotação exigida	Economia de anotação
Adjetivos	71,2%	28,8
Substantivos	30,9%	69,1%
Preposições	61%	39%
Verbos	45,2%	54,8%
Palavras corridas	8,5%	91,5%
Média	43%	57%

A quantidade necessária de anotação do *corpus* VUMAC diminuiu, passando de 59% para 43%, e a economia aumentou, de 41% para 57%.

Pergunta 3: Discrepância entre as frequências observadas e estimadas

Para responder essa pergunta, foi calculada, para cada texto e para o *corpus* inteiro, a diferença entre as frequências observadas e as estimadas pelo modelo abaixo:

Total de metáforas = 63,009 + dim1 x -1,093 + mets100j x 1,460 + mets100n x 1,027

Por exemplo, para o texto a1h, a quantidade de metáforas observadas (por mil palavras) é 173,3, enquanto a quantidade estimada é 194,5. A discrepância entre os dois é de 21,2 metáforas, o que representa 12% do total de metáforas observadas.

A discrepância mínima foi de 0,1, a máxima, de 160,4 e a média no *corpus* inteiro foi de 19,7 (14% da média observada de 143,6 metáforas por mil palavras).

Pergunta 4: Melhores modelos sem anotação manual

Para responder a essa pergunta, foram selecionados apenas os modelos cujas variáveis independentes eram dimensões de variação (dim1, dim2, dim3, dim4, dim5). As correlações obtidas aparecem na TAB. 8.

TABELA 8
Maiores correlações para cada variável dependente dos modelos sem anotação manual

variável dependente	Correlação (R)	R^2
total de metáforas	0,574	33%
pacotes lexicais metafóricos	0,711	51%
agrupamentos metafóricos	0,558	31%
adjetivos metafóricos	0,511	26%
substantivos metafóricos	0,553	31%
preposições metafóricas	0,591	35%
verbos metafóricos	0,457	21%
Média	0,565	33%

As correlações obtidas são, na maioria, medianas (entre 0,49 e 0,7), com exceção de uma, que é alta (pacotes lexicais). A quantidade de variância em comum (R^2) vai de 21% a 51%, mas quase todas ficam abaixo de 50%, o que significa que não explicam a maior parte da variação da frequência de metáforas. Em comparação aos modelos obtidos em resposta às perguntas anteriores, esses são menos confiáveis. Por outro lado, esses índices podem ser considerados expressivos na medida que foram obtidos apenas com informações acerca de características linguísticas e funcionais dos textos, sem nenhum conhecimento acerca de metáforas. O fato de tais modelos terem explicado 1/3 da variação na frequência de metáforas indica que essa parcela do uso metafórico está ligada a características do texto apenas.

Os modelos são os seguintes:

- Total de metáforas = 136,116 + dim1 x -1,663
- Pacotes lexicais metafóricos = 125,478 + dim3 x 3,788 + dim4 x 5,717 + dim1 x -0,716
- Agrupamentos metafóricos = 67,410 + dim1 x -1,474
- Adjetivos metafóricos = 14,696 + dim1 x -0,236
- Substantivos metafóricos = 30,886 + dim1 x -0,588
- Preposições metafóricas = 39,306 + dim1 x -0,658
- Verbos metafóricos = 36,264 + dim1 x -0,378 + dim4 x 1,364

Pergunta 5: Confiabilidade

Nos estudos de metáfora com base em *corpora*, o grau de sucesso do processo de identificação das metáforas, quando realizado por mais de um pesquisador, é algumas vezes medido pelo nível de concordância entre anotadores (*inter-rater agreement*), calculado por meio da estatística Kappa de Cohen. Concordância total é atingida quando duas análises são idênticas, ou seja, dois pesquisadores marcaram as mesmas metáforas; falta de concordância, por sua vez, acontece quando as metáforas marcadas por um dos pesquisadores não foram identificadas por outro. O Kappa de Cohen quantifica o grau de concordância, entre os dois extremos (total concordância e total falta de concordância) com um valor que vai de 0 a 1. A concordância entre anotadores é interpretada como o nível de confiabilidade da análise, pois se dois ou mais pesquisadores identificam os mesmos casos, o sistema de anotação é tido como compreensível, executável e replicável, em suma, não arbitrário; assim, o produto da anotação é visto como confiável, já que reflete um procedimento criterioso.

Embora este caso particular não seja de fato uma análise produzida por um ser humano, mas por um modelo matemático, a ideia é considerar a frequência estimada pela Análise de Regressão como se tivesse sido fruto da análise de outro pesquisador. Com isso, torna-se possível calcular Kappa e estimar o sucesso da aplicação dos modelos como se fosse mais um pesquisador.

Para calcular a estatística Kappa de Cohen, é preciso obter os seguintes dados:

- 1, 1: Quantidade de vezes em que ambos os analistas consideraram uma palavra como sendo metafórica.
- 1, 0: Quantidade de vezes que o primeiro analista considerou uma palavra metafórica e o segundo não metafórica.
- 0, 1: Quantidade de vezes que o primeiro analista considerou uma palavra não metafórica e o segundo metafórica.
- 0, 0: Quantidade de vezes que ambos os analistas consideraram uma palavra como sendo não metafórica.

Para obter os dados acima, parte-se do pressuposto de que a frequência estimada sempre concorda com a observada, ou seja, que os dois 'analistas' teriam encontrado as mesmas metáforas. Quando o modelo estima uma frequência de 100 e a frequência obtida na análise manual é 150, pressupõe-se nessa simulação 100 metáforas anotadas em comum. Numa situação real de anotação, isso não

poderia ser presumido, visto que o primeiro analista poderia ter encontrado qualquer número de metáforas entre zero e 100 em comum com o segundo (desde que houvesse mais de 100 casos para anotar): a frequência em si não seria garantia de concordância. Porém, nessa simulação, considera-se que as frequências indicam análises coincidentes, isto é, identificação mútua das metáforas.

Os valores observados e estimados foram convertidos para o formato exigido da seguinte maneira:

- 1,1: menor valor entre as frequências observadas e previstas. Por exemplo, se a frequência observada foi de 143,8 e a prevista de 120,7, o valor aqui será de 120,7;
- 1,0: diferença entre os valores observados e previstos, quando o valor observado é maior que o previsto, caso contrário zero. Se a frequência observada foi de 143,8 e a prevista, de 120,7, o valor será de 23,1;
- 0,1: diferença entre os valores observados e previstos, quando o valor observado é menor que o previsto, caso contrário zero. No exemplo anterior, o valor será zero;
- 0,0: a diferença entre 1.000 e os valores anteriores somados. A base 1.000 é usada, pois as frequências do *corpus* são normalizadas por mil. Usando o exemplo anterior, o valor é de 856,2 (i.e. 1.000 - (120,7 + 23,1 + 0)).

O modelo de regressão testado foi o referente ao total de metáforas, apresentado na resposta à pergunta 1.

Os dados foram coletados para cada texto e para o *corpus* como um todo e submetidos ao SPSS, primeiramente com a opção *Data* > *Weight Cases*, e, em seguida, por meio do procedimento *Crosstabs*, com a opção Kappa ativada.

Os resultados mostraram um valor de Kappa de Cohen de 0,961 (t=278,77, p=0,000) para o *corpus* inteiro, o que significa 96,1% de concordância. Houve um caso de baixa concordância (texto b1g) com 23,3%, mas trata-se de um *outlier*, isto é, um texto cuja densidade metafórica foge totalmente do esperado. Tanto assim que a segunda menor concordância foi do texto a1g, com 72,3%. O texto com maior concordância foi a1m, com 100%. Caso a previsão dos modelos fosse uma anotação de metáfora, ela seria considerada altamente confiável, pois seu Kappa se situa acima de 0,9.

O valor de Kappa na anotação manual do VUAMC feita pela equipe de Amsterdã variou entre 0,7 (conversação), 0,8 (ficção) e cerca de 0,9 (jornalismo e escrita acadêmica) (STEEN; DORST; HERRMANN *et al.*, 2010 , p.157). Esses valores não são diretamente comparáveis aos atingidos pela Análise de Regressão, pois são análises diferentes.

Pergunta 6: Demonstração de uso

Para ilustrar o processo de aplicação dos modelos obtidos em uma pesquisa, tomemos o caso dos substantivos metafóricos, cujo modelo apresentado em resposta à pergunta 2 é:

Substantivos metafóricos = 7,724 + mets100v x 0,703 + mets100j x 0,469 + dim1 x -0,297 + dim2 x -1,263

Para o texto a1e do *corpus*, os valores das variáveis independentes são os seguintes:

- mets100v: 17
- mets100j: 22
- dim1: 0,1
- dim2: 0,21

Substituindo esses valores na fórmula, temos:

Substantivos metafóricos = 7,724 + (17 x 0,703) + (22 x 0,469) + (0,1 x -0,297) + (0,21 x -1,263) = 7,724 + 11,951 + 10,318 - 0,297 – 0,26523 = 29,7

O modelo prevê a frequência de 29,7 substantivos metafóricos por mil palavras no texto. A frequência observada por meio da anotação manual é de 33,4, uma discrepância de apenas 3,8 metáforas.

Conclusão

Este trabalho apresentou uma pesquisa de cunho quantitativo para tentar descobrir até que ponto é possível prever a frequência de metáforas em *corpora*. Tomando como base o *Vrije Universiteit Amsterdam Metaphor Corpus* (VUAMC), foi utilizada a Análise de Regressão Múltipla, um procedimento estatístico que visa criar modelos de previsão. Esses modelos previram as frequências das

metáforas nesse *corpus*, que foram então comparadas às frequências efetivas, obtidas pela análise manual do *corpus*. Esses modelos utilizaram as frequências de metáfora em pequenas porções dos textos (nas 100 ou 200 primeiras palavras), para testar a ideia de que seria possível saber quantas metáforas há no *corpus*, analisando apenas essas porções menores de cada texto, a fim de descobrir qual é a menor quantidade de análise manual exigida para se obter uma contagem confiável de metáforas. A intenção de diminuir ao máximo a análise manual de metáforas é movida pela constatação de que o enorme trabalho manual envolvido na análise de metáforas em *corpora* tem prevenido a disseminação desse tipo de pesquisa. Isso, por sua vez, prejudica o campo como um todo, visto que os estudos de metáfora com base em *corpora* já são reconhecidos como de primordial importância para o entendimento do uso metafórico.

Os resultados foram positivos, sugerindo que seja possível prever a ocorrência de metáforas na língua em uso com boa precisão. Isso indica que a frequência de metáforas em um texto parece estar associada a aspectos linguísticos e funcionais e à ocorrência de metáforas nas 100 ou 200 primeiras palavras do texto.

Tendo mostrado que a previsão da frequência de metáforas é estatisticamente factível, este estudo também buscou descobrir qual seria a quantidade mínima de anotação manual necessária para prever a ocorrência das metáforas. Foi extraído um conjunto econômico de modelos, que implicasse a menor quantidade de anotação manual possível, sem comprometer a qualidade da previsão. O conjunto mais eficiente (que combina maior precisão com menor trabalho manual) exige a anotação de apenas 500 palavras por texto, sendo capaz de prever 63% da variação na frequência das metáforas, com altas correlações que chegam a 0,821.

Caso se abra mão por completo da anotação de metáforas, os resultados mostraram que é possível prever apenas algo em torno de 1/3 da variação na frequência de metáforas. Para que a estimação da frequência seja confiável, portanto, é imprescindível que seja feita a anotação manual de metáforas pelo menos em 100 ou 200 palavras de cada texto. Por outro lado, o fato de podermos prever uma a cada três metáforas 'às cegas', apenas levando em conta aspectos linguísticos e textuais, revela uma faceta até então pouco conhecida do uso metafórico, qual seja, a relação estreita entre sua recorrência e a estrutura linguística.

A discrepância entre a frequência observada e estimada foi baixa, em média 19 metáforas (por mil palavras). E numa simulação em que as frequências previstas foram tidas como o trabalho de um segundo analista, o grau de concordância entre as duas 'análises' foi alto (96% para o *corpus* inteiro).

Em suma, os resultados são favoráveis, indicando que a aplicação da Análise de Regressão ao problema da contagem de metáforas em *corpora* é um caminho promissor. Entretanto, todos os modelos apresentados aqui devem ser validados em outros *corpora*, para saber até que ponto suas previsões são precisas, antes de serem aplicados em outros projetos de pesquisa. Também é necessário construir modelos para *corpora* de outras línguas, como o português, por exemplo. Essas são tarefas para pesquisas futuras.

Temos consciência da dificuldade em transferir esses resultados para outras pesquisas, devido ao *know-how* exigido (conhecimento de estatística e informática, incluindo etiquetagem morfossintática e cálculo dos escores de dimensão, por exemplo) e à infraestrutura necessária (disponibilidade de etiquetador, especialmente o *Biber Tagger* e de outros *software*). Parece-nos que um caminho para vencer essas barreiras seja a colaboração entre grupos de pesquisa, para que sejam garantidas as condições necessárias para implementar este tipo de pesquisa, em particular, e para disseminar as pesquisas em metáfora com base em *corpora*, em geral.

Este trabalho pretende ter contribuído com os estudos de metáfora, mostrando uma alternativa para a custosa tarefa de identificação manual de metáforas em grandes *corpora*.

Notas

[1] No original:

> "1. Find metaphor-related words (MRW) by examining the text on a word-by-word basis.
> 2. When a word is used indirectly and that use may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word, mark the word as metaphorically used (MRW).
> 3. When a word is used directly and its use may potentially be explained by some form of cross-domain mapping to a more basic referent or topic in the text, mark the word as direct metaphor (MRW, direct).
> 4. When words are used for the purpose of lexico-grammatical substitution, such as third person personal pronouns, or when ellipsis occurs where words may be seen as missing, as in some forms of co-ordination, and when a direct or indirect meaning is conveyed by those substitutions or ellipses that may potentially be explained by some form of cross-domain mapping from a more basic meaning, referent, or topic, insert a code for implicit metaphor (MRW, implicit).

5. When a word functions as a signal that a cross-domain mapping may be at play, mark it as a metaphor flag (MFlag).
6. When a word is a new-formation coined, examine the distinct words that are its independent parts according to steps 2 through 5." (STEEN; DORST; HERRMANN *et al.*, 2010, p. 25)

[2] Nos exemplos, as palavras sublinhadas são aquelas usadas metaforicamente.

Referências

BERBER SARDINHA, T. Análise Multidimensional. *Delta*, v. 16, n. 1, p. 99-127, 2000.

BERBER SARDINHA, T. *Lingüística de Corpus*. São Paulo: Manole, 2004. 410 p.

BERBER SARDINHA, T. Metáforas de Teleconferências de Negócios. *Cadernos de Estudos Lingüísticos*, v. 50, p. 171-188, 2008a.

BERBER SARDINHA, T. Metaphor probabilities in corpora. In: ZANOTTO, M. S. *et al* (Org.). *Confronting Metaphor in Use: An Applied Linguistic Approach*. Amsterdam/Atlanta, GA: Benjamins, 2008b. p. 127-148.

BERBER SARDINHA, T. Telling one's life stories with metaphors: A corpus-driven investigation. Paper. In: International Conference on Researching and Applying Metaphor (RaAM 7), Cáceres, Spain, 2008c.

BERBER SARDINHA, T. *Pesquisa em Lingüística de Corpus com WordSmith Tools*. Campinas: Mercado de Letras, 2009. 272 p.

BERBER SARDINHA, T. Metaphor and Corpus Linguistics. *Revista Brasileira de Linguística Aplicada*, v. 11, p. 329-360, 2011.

BERBER SARDINHA, T. An assessment of metaphor retrieval methods. In: MACARTHUR, F. *et al* (Org.). *Metaphor in Use:* Context, Culture, and Communication. Amsterdam: John Benjamins, no prelo-a.

BERBER SARDINHA, T. Perspectiva multidimensional de linguagens da Internet. In: SHEPHERD, T. M. G.; SALIES, T. (Org.). *Linguística da Internet*. São Paulo, SP: Contexto, no prelo-b.

BERBER SARDINHA, T. Register variation and metaphor use: A multi-dimensional perspective. In: HERRMANN, J. B.; BERBER SARDINHA, T. (Org.). *Metaphor in Specialist Discourse: Investigating metaphor use in technical, scientific and popularized discourse contexts*. Amsterdam / Philadelphia, PA: John Benjamins, no prelo-c.

BIBER, D. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BIBER, D. Multi-dimensional approaches. In: LÜDELING, A.; KYTÖ, M. (Org.). *Corpus Linguistics – An International Handbook*. Berlin / New York: Walter de Gruyter, 2009.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, v. 25, n. 3, p. 371-405, 2004.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics – Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

CAMERON, L.; LOW, G. *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, 1999.

DEIGNAN, A. Corpus-based research into metaphor. In: CAMERON, L.; LOW, G. (Org.). *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, 1999. p. 177-199.

DEIGNAN, A. *Metaphor and corpus linguistics*. Amsterdam; Philadelphia: J. Benjamins Pub., 2005. viii, 235 p.

GIBBS, R. W. (Org.). *The Cambridge Handbook of Metaphor and Thought*. New York: Cambridge University Press. 2008a

GIBBS, R. W. Metaphor and thought – The state of the art. In: GIBBS, R. W. (Org.) *The Cambridge Handbook of Metaphor and Thought*. New York: Cambridge University Press, 2008b. p. 3-13.

KAAL, A. A. *Metaphor in conversation*. (PhD dissertation), Vrije University, Amsterdam, 2012.

KRENNMAYR, T. *Metaphors in newspapers*. (PhD dissertation), Vrije University, Amsterdam, 2011.

PASMA, T. *Metaphor and register variation*: The personalization of Dutch news discourse. (PhD dissertation), Vrije University, Amsterdam, 2011.

PRAGGLEJAZ GROUP. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, v. 22, n. 1, p. 1-39, 2007.

SHEPHERD, T.; BERBER SARDINHA, T.; VEIRANO PINTO, M. (Org.). *Caminhos da Linguística de Corpus*. Campinas, SP: Mercado de Letras. 2012

STEEN, G. *et al. A Method for Linguistic Metaphor Identification*: From MIP to MIPVU. Amsterdam: John Benjamins, 2010.

Submissão do artigo: 27/07/2012
Aprovação do artigo: 29/09/2012