Discurso literário e linguística de *corpus*: uma visão empírica

Sonia Zyngier

Vander Viana

Natália Giordani Silveira

1. Introdução

Na virada do século XX, as consequências da Revolução Industrial e o progresso da tecnologia modificaram em muito as formas de se ver e de se pensar no campo da ciência e acabaram por provocar a formação de áreas cada vez mais segmentadas. Se isso, por um lado, resultou no aprofundamento de cada uma dessas áreas, por outro, causou o isolamento de cientistas, que passaram a habitar territórios estanques com fronteiras bem definidas e sem que houvesse diálogo entre os diferentes campos do saber. Na área de Letras, o pesquisador solitário de literatura tornava-se especialista em A ou B, sem estabelecer canal de comunicação qualquer com seus colegas de língua, e muito menos com os de outras áreas. A codificação das áreas em números (vide as listas fornecidas pelas agências de fomento) e a organização de universidades em departamentos é um exemplo deste paradigma.

Passado um século, esta situação já não se sustenta e a necessidade de trabalhos inter- e transdisciplinares realizados por grupos de pesquisadores se torna cada vez mais premente. O pensar coletivo, e a contribuição de visões e metodologias diversas cada vez mais se tornam essenciais para o fazer científico contemporâneo. Este artigo mostra alguns exemplos da interação entre a Literatura e a Linguística de *Corpus* em solo brasileiro.

2. Análise textual em estudos literários

Passado quase um século desde que os Formalistas Russos e os Estruturalistas Tchecos investiram na sistematização dos estudos literários em busca de uma abordagem mais científica para a literatura, ainda hoje a questão não foi resolvida a contento, e talvez jamais seja. No entanto, pequenos avanços podem ser sentidos nesta direção. Um dos conceitos mais importantes continua sendo o de estranhamento, proposto por Shklosvky (1917), cujo efeito é o de provocar o prolongamento da percepção do leitor. Em outras palavras, quando algo se desvia até um determinado ponto das normas e das convenções linguísticas e alonga o processo de percepção do leitor, este pode vir a perceber a arte verbal de um texto e classificá-lo como literário.

A função da arte, portanto, seria a de fazer com que os indivíduos renovassem sua percepção do mundo. Este processo de desfamiliarização, que desautomatiza reações reflexivas e automatizadas, leva o leitor a apreciar a beleza de um objeto de arte. Como consequência, ao invés da fluidez de comunicação, a obra apresenta uma densidade de formas que retarda a apreensão do leitor, prolongando a experiência estética. O resultado é o efeito de *foregrounding*, termo que remete à distinção entre os conceitos de figura e fundo nas artes visuais. Aqui ele se refere à reação do leitor ao processo psicolinguístico de colocar em primeiro plano o objeto desfamilizariado[1]. No caso da literatura, a desfamiliarização é obtida através da manipulação de formas linguísticas de maneira a produzir desvios e paralelismo.

Os desvios aparecem em construções linguísticas que se afastam do esperado, daquilo que o leitor, como falante da língua, já naturalizou. São quebras com relação a regras, convenções, e expectativas normalmente associadas ao contexto em que ocorrem, como no que se convenciona chamar de 'licença poética'. Ao fundo, fica o que é linguisticamente considerado padrão e, por isso, chama menos atenção; a figura, em primeiro plano, é a forma difícil, as partes do texto desviantes da norma, cujo processamento prolongado dilata a experiência da percepção.

Já o paralelismo é uma forma controlada de repetição, em que são mantidos constantes determinados traços de uma estrutura, enquanto outros variam. É possível que se repitam todos os traços, e nesse caso temos a repetição simples; mas as formas mais ricas de paralelismo, naturalmente, são aquelas em que variações e constâncias são manipuladas de maneira a criar efeitos estilísticos.

Segundo os Formalistas Russos, a literariedade distingue os textos que chamamos de literários. Tais textos seriam marcados por inovação e imprevisibilidade no uso da língua. Essa postura fortaleceu-se no decorrer do século XX como uma possibilidade promissora no estudo da materialidade do texto no âmbito da Ciência da Literatura. Atualmente, a literariedade não é considerada um traço exclusivo da literatura, podendo ser percebida em outros tipos de textos (CARTER E NASH 1990). Hoje a noção de uma gradação (em oposição a uma divisão rígida entre o literário e o não-literário) é mais aceita. Nesta gradação, as formas literárias ocupam um pólo e se opõem às formas de linguagem mais padronizadas e previsíveis.

A contribuição dos Formalistas Russos, no entanto, ainda não contava com uma metodologia mais sistematizada e precisa para a verificação de seus pressupostos. As análises eram feitas tomando-se por base a visão e o conhecimento linguístico do analista. Um exemplo disto encontra-se no trabalho desenvolvido em parceria entre Jakobson e Lévi-Strauss ([1962] 1987). Esta análise veio a se tornar um modelo para os estruturalistas. Ainda estavam para nascer um paradigma de linguagem e uma metodologia que iriam mudar este quadro.

3. Linguística de *Corpus*

Entre as décadas de 50 a 70, o pensamento linguístico foi marcado pela influência da teoria gerativa, segundo a qual a linguística deveria se restringir, como aponta Sinclair (1991: 1), à intuição de um indivíduo. Os estudos da linguagem eram pautados pela introspecção e o enfoque dependia da competência do falante. Em outras palavras, "antes da disponibilização de grandes quantidades de dados, muitas das generalizações eram feitas por suposições intuitivas; não era possível verificar tais noções na era anterior à Linguística de *Corpus*" (SINCLAIR 2003: ix). O objetivo

residia na obtenção de universais linguísticos, isto é, estruturas gerais que fossem válidas em todas as línguas existentes.

A mudança neste panorama só ocorreu na década de 80, com a popularização dos computadores e a disponibilização de um grande número de dados (BIBER, CONRAD E REPPEN 1998: 4). Na verdade, a introdução da ferramenta computacional na investigação linguística acabou com uma das maiores críticas a estudos baseados em dados reais, a saber, a falta de confiança na análise manual. O computador passou, então, a substituir o analista na realização de tarefas repetitivas e laboriosas, auxiliando-o, por exemplo, na identificação e contagem de itens lexicais. Em suma, pode-se dizer que a Linguística de *Corpus* constitui uma nova abordagem nos estudos da linguagem, favorecendo uma análise objetiva de dados reais.

Com a possibilidade de se utilizar o computador na investigação linguística, o pesquisador passa a ter acesso a vários dados em curto espaço de tempo. Os mais diversos programas de análise de *corpus* – como, por exemplo, o *WordSmith Tools* (SCOTT 1999) – são capazes de fornecer o número de itens, as formas e a razão entre os dois. O termo 'item' refere-se a cada uma das palavras existentes em um dado corpus. Já o termo 'forma' corresponde a cada uma das palavras diferentes que podem ser encontradas neste mesmo corpus. A razão forma/item, como o nome indica, é resultado da divisão entre o número de palavras distintas pelo número de palavras existentes em um mesmo conjunto de textos. Como o resultado final é expresso em termos percentuais, o valor encontrado após a divisão é multiplicado por 100. A razão é, em última instância, uma medida de variedade lexical em termos de palavras. Quanto mais variados forem os itens lexicais, maior será a razão encontrada. Contudo, como se pode antecipar, o valor encontrado é altamente dependente do tamanho do corpus, isto é, quanto maior o corpus, maior é a probabilidade de sua razão forma/item ser menor. Daí se utiliza a razão forma/item padronizada. Esta é obtida a partir da média entre as várias razões calculadas para intervalos regulares de palavras no corpus.[2]

Um importante conceito na área de Linguística de *Corpus* é a noção de que a linguagem é um sistema de probabilidades. Esta perspectiva implica que, apesar de várias estruturas léxico-gramaticais poderem ocorrer em um determinado contexto, elas não ocorrem com a mesma frequência ou distribuição em registros distintos. Neste sentido, a Linguística de *Corpus* contraria o postulado de Saussure (1916, p. 28) de que "nada existe [...] de coletivo na fala; suas manifestações são individuais e momentâneas". Ao admitir a possibilidade de se estudar as frequências de determinado traço linguístico, considera-se a linguagem como um fenômeno que contém padrões. Torna-se possível, assim, estudar as probabilidades de ocorrência de determinados traços linguísticos. Ressalta-se, aqui, que a ausência de traços é igualmente relevante.

Alinhando-se à noção de padronização, Sinclair (1991) explica que há dois princípios em ação na produção linguística: o idiomático e o da livre escolha. O último relaciona-se ao uso de palavras isoladas, escolhas estas feitas pelo falante ou pelo escritor à medida que a necessidade se apresenta, tendo como única restrição a gramaticalidade da sentença. No entanto, Sinclair (1991) identifica o princípio idiomático como o mais produtivo. Este se relaciona à escolha de sequências de palavras. Assim sendo, "um usuário da língua tem à sua disposição um grande número de sintagmas parcialmente pré-construídos que constituem escolhas únicas, mesmo que eles possam aparentar ser analisáveis em segmentos" (SINCLAIR 1991: 110)

De forma análoga à proposta de Sinclair (1991), Biber, Conrad e Cortes (2004) também tomam por base a noção de que o uso da linguagem é formulaico e discutem as sequências de palavras que são utilizadas por falantes de inglês como primeira língua em quatro registros distintos (conversa, aulas, livros didáticos e prosa acadêmica). Argumentando que os usuários de uma língua recorrem frequentemente às sequências lexicais, eles optam por estudar o que denominam de 'feixes lexicais' (em inglês, '*lexical bundles*'), definidos como "sequências lexicais recorrentes mais frequentes em um registro" (BIBER, CONRAD E CORTES 2004: 376). Ressalta-se aqui que, segundo a proposta destes pesquisadores, nem toda sequência listada pelo computador é considerada um feixe lexical. Para ser considerado como tal, eles adotam critérios de identificação baseados em frequência e em dispersão. Mais especificamente, uma sequência deveria ocorrer pelo menos 40 vezes em cada grupo de um milhão de palavras e em cinco textos diferentes.

A Linguística de *Corpus*, portanto, pode fornecer ao crítico literário as ferramentas e os princípios necessários para que a literatura seja analisada sob uma nova perspectiva, mais objetiva do que a que tradicionalmente caracteriza a análise de base hermenêutica. Por exemplo, esta abordagem pode consubstanciar (ou não) as impressões intuitivas de um leitor, a partir do uso que o autor faz da linguagem.

4. Estudos literários baseados em *corpora*

Vários estudos do discurso literário de base computacional já começam a despontar no Brasil. Zyngier (2002) investiga *Macbeth* de Shakespeare em comparação a um *corpus* de tragédias shakespeareanas por meio do programa *MicroConcord*. A metodologia envolve duas abordagens distintas: ascendente e descendente. No primeiro caso, a pesquisadora parte da lista de palavras mais frequentes para então poder chegar a certas generalizações provisórias e válidas somente para peça em análise. Ao examinar uma lista de verbos, nota-se que '*enter*' [entrar] é mais frequente em *Macbeth* do que no *corpus* de referência, totalizando 0,42% no primeiro caso em face de apenas 0,29% no segundo. Tal item lexical é o 5º verbo mais frequente em *Macbeth*, mas aparece somente na 9ª colocação no *corpus* de referência. A pesquisadora percebe que a alta frequência deste verbo se dá principalmente nas rubricas da peça teatral e interpreta tal resultado como uma indicação do dinamismo da peça, que apresenta muitas entradas e saídas, especialmente por parte de Macbeth. Com relação aos substantivos mais frequentes, a análise aponta que o uso de '*thane*' [título honorífico anglo-saxão mais próximo à noção de barão] na peça leva o espectador a ver tal título sob a perspectiva de Macbeth, isto é, como o objeto de seu desejo. A partir da análise de '*blood*' [sangue], afirma-se que em *Macbeth* este item lexical ocorre de forma diferenciada das outras tragédias. Aqui, seu uso é literal e não metafórico, o que contribui para a violência desta peça. Além disso, a verificação dos colocados desta palavra revela abundância na segunda ou terceira posição à esquerda; ocorre, também, uma associação de '*blood*' com '*gold*' [ouro], apontando para o fato de que o primeiro seria algo valioso. No caso específico de '*sleep*' [sono], o estudo revela uma ambiguidade no uso deste item lexical. Há instâncias do vocábulo com prosódia semântica negativa ("*curtained sleep*", "*thralls of sleep*", "*swinish sleep*", "*equivocates him in a sleep*") enquanto outros apresentam prosódia semântica positiva ("*innocent sleep*", "*the benefit of sleep*"). Por fim, valendo-se de uma abordagem descendente, isto é, partindo da

crítica que é geralmente feita à peça para a confrontação da mesma com a linguagem empregada pelo escritor, a pesquisadora verifica até que ponto é verdadeira a afirmação de que a questão do medo é central em *Macbeth*. O resultado encontrado é que a palavra ocorre 35 vezes na peça, sendo que em aproximadamente 50% dos casos o medo é negado como ocorrem em "*Fear not*", "*What need I fear?*", "*Hang those that talk of fear*" e "*nor shake with fear*". Apesar de a autora afirmar que a análise realizada é inicial, ela aponta para novos caminhos na análise de textos literários sob a perspectiva da Linguística de *Corpus*.

Por sua vez, Gonçalves (2007) também privilegia uma abordagem ascendente assim como a primeira parte do estudo de Zyngier (2002). Neste caso, a pesquisadora analisa *Dubliners*, uma coletânea de contos de James Joyce. Outra diferença deste estudo, quando comparado ao de Zyngier (2002), é que aquele se utiliza de um programa de análise textual, chamado *WordSmith Tools*, que oferece uma ferramenta para extrair palavras-chave de um *corpus* de estudo quando comparado a um *corpus* de referência. Segundo a autora, as palavras-chave são especialmente interessantes por poderem indicar temas ou estilos. Na pesquisa realizada, contrastou-se o *corpus* de estudo de 67.936 formas com um *corpus* de referência, contendo contos de autores contemporâneos a Joyce, num total de 212.591 formas. Assim sendo, afirma-se que o *corpus* de referência elimina as questões referentes à diacronia e aos gêneros textuais. Uma primeira questão analisada pela pesquisadora foi o emprego do pronome 'she' [ela] na obra de Joyce. A partir da análise dos cotextos no qual a palavra se encontra, nota-se que o mesmo ocorre, em sua maioria, com verbos na voz ativa e que estes são classificados como 'volitivos', ou seja, verbos que denotam ações deliberadas. Além disto, a pesquisadora verificou que os verbos expressavam conotação positiva (48,76%) ou neutra (42,29%) majoritariamente. A primeira é compreendida como ações independentes enquanto a segunda categoria englobava os verbos não-indicativos de independência nem de servilismo. Apenas em 8,95% dos casos analisados, 'she' se combinava com verbos indicativos de inferioridade. A autora conclui afirmando que "Temos portanto um argumento objetivo e estatístico para discordar dos críticos e mostrar como a mulher em Dubliners, longe de ser 'a oprimida dos oprimidos', tem voz e poder deliberativo sobre sua vida e muitas vezes sobre as situações em que se encontra" (GONÇALVES 2007: 4).

Um campo semântico definido pela pesquisadora a partir da análise das palavras-chave foi o da música, no qual há 17 termos distintos. O artigo ressalta a importância da música em *Dubliners* a partir dos contos "*An Encounter*", "*Two Gallants*", "*A Mother*", "*A Painful Case*" e "*Araby*". Segundo Gonçalves (2007), de forma contrária à crítica feita a *Dubliners* de que a música teria apenas uma participação periférica no conto, ela é de fato parte integrante da estrutura da obra. Em suas palavras, "a música tem um papel mais fundamental na narrativa joyceana. Ela, entre outras atuações, define personalidades e estados de espírito" (GONÇALVES 2007: 5).

Apesar de utilizar o programa *WordSmith Tools*, o estudo de Viana, Fausto e Zyngier (2007) adota outra perspectiva metodológica: utilizando-se da ferramenta *WordList*, são investigados os feixes lexicais de quatro palavras mais comumente encontrados em *Dom Casmurro* e *O Código da Vinci*. A escolha das obras analisadas deveu-se a uma consulta preliminar acerca dos hábitos de leitura dos integrantes de diversas comunidades relacionadas à leitura de uma rede social on-line. Os respondentes apontaram Machado de Assis e Dan Brown como seus autores canônico e popular favoritos. Os pesquisadores adotaram não só o conceito de feixe lexical,

mas também as taxonomias estrutural e funcional propostas por Biber, Conrad e Cortes (2004). Os resultados indicam que estruturalmente os feixes em *Dom Casmurro* incorporam fragmentos de sintagmas verbais, especialmente em orações subordinadas enquanto o texto de *O Código da Vinci* faz uso de muitos feixes que incorporam fragmentos de orações nominais e/ou preposicionais. Em termos funcionais, ambos os *corpora* apresentam alta ocorrência de feixes referenciais; porém, é em *Dom Casmurro* que se verifica uma maior variedade de funções exercidas pelos feixes. Os autores explicam este resultado, afirmando que "indica uma abordagem mais formulaica ao uso da linguagem na obra de Dan Brown, que parece ser consequentemente menos inovativa e criativa. Tal uso de linguagem mais fácil pode ser responsável pela popularidade de *O Código da Vinci* quando comparado a *Dom Casmurro*" (VIANA, FAUSTO E ZYNGIER 2007: 254).

Em mais outro estudo baseado em *corpus*, Camargo (2006) investiga *O Sumiço da Santa* escrito por Jorge Amado. Apesar de o foco do artigo recair na comparação entre o texto original e a sua respectiva tradução para a língua inglesa, a pesquisadora dedica uma parte de seu estudo à comparação do texto sob análise a um *corpus* geral de língua portuguesa, a saber, o Banco de Português, sediado na Pontifícia Universidade Católica de São Paulo. Como recurso computacional, a autora lança mão do *WordSmith Tools* para calcular índices de variedade lexical. Verifica que a razão forma/item do Banco de Português (0,26%) é bem menor do que a observada no texto de Jorge Amado (15,39%). O resultado obtido no estudo já era, de certa forma, esperado, uma vez que o *corpus* de referência contém mais de 230 milhões de itens enquanto o texto literário totaliza apenas 126.443. Ao recorrer à razão forma/item padronizada, verifica-se novamente que o resultado obtido pelo texto literário (54,99%) é superior ao do Banco de Português (46,08%). A autora justifica os resultados encontrados afirmando ser o texto de Jorge Amado "de cunho regionalista, o qual, pela sua natureza, passa a requerer uma diversidade de termos culturalmente marcados. Também o estilo do autor normalmente associado ao pitoresco e mesmo ao exotismo, mostra-se como um fator que contribui para o uso de uma linguagem mais rica e variada" (CAMARGO 2006: 44).

Ainda utilizando metodologia semelhante aos dois últimos estudos acima, a investigação de Zyngier et al. (2007) acerca das versões original e traduzida de *O Código da Vinci* divide-se em duas partes. Inicialmente, os pesquisadores relacionam o conceito de literariedade a medidas de inovação lexical nos *corpora* analisados. Tais medidas incluem a razão forma/item padronizada e a frequência relativa de feixes lexicais. O primeiro fator relaciona-se ao uso de palavras distintas enquanto o segundo liga-se ao uso de palavras em sequências diversas. Ao contrastar as duas versões da obra sob análise com outras obras canônicas e não-canônicas em suas versões originais (*Grande Sertão: Veredas*, *Memórias Póstumas de Brás Cubras* e *O Alquimista*) e traduzidas (*The Devil to Pay in the Backlands*, *Epitaph of a Small Winner* e *The Alchemist*), os pesquisadores percebem haver dois grupos de obras diferentes. Um englobaria as obras consideradas canônicas (*Grande Sertão: Veredas* e *Memórias Póstumas de Brás Cubas*) em ambas as versões. Por sua vez, o outro consistiria da obra não-canônica de Paulo Coelho tanto em português como em inglês. O resultado inesperado do estudo é que a obra de Dan Brown nas duas línguas analisadas se insere no grupo de obras canônicas e não no de não-canônicas como se esperaria. O estudo também aponta para o fato de que as obras em língua portuguesa apresentam valores superiores aos seus correspondentes em língua inglesa, não obstante a língua em que a obra foi originalmente escrita. Na segunda parte da

investigação, os pesquisadores analisam a obra de Dan Brown a partir de outro ponto de vista, enfocando os feixes lexicais presentes na mesma à semelhança do procedimento adotado no estudo de Viana, Fausto e Zyngier (2007). Os achados indicam que os feixes distribuem-se estrutural e funcionalmente de forma semelhante no par de obras analisadas. Contudo, nota-se uma pequena diferença, havendo mais feixes conversacionais e atitudinais na versão original em inglês, e mais feixes referenciais e discursivos na obra traduzida para o português. Os pesquisadores interpretam este resultado como indicativo de maior envolvimento na obra original e, consequentemente maior distanciamento por parte do tradutor na obra em português. Como explicado no artigo, por não ser o autor, talvez o tradutor tenha se preocupado em fazer escolhas mais informacionais do que o próprio escritor, lançando mão também de sequências lexicais indicativas da relação entre as diferentes partes do discurso.

Como encaminhamento do estudo de Zyngier et alii (2007), a investigação de Zyngier, Viana e Silveira (2007) objetiva o desenvolvimento de um índice de literariedade. Para tanto, entende-se a literariedade como restrita a características de uso da linguagem que podem ser descritas como inovadoras e surpreendentes. Com auxílio computacional, os pesquisadores mapeiam a repetição lexical em diferentes obras da literatura em língua inglesa, atribuindo um valor a cada uma das variáveis selecionadas para o índice de literariedade. Tal índice contempla três variáveis, dando conta tanto do uso de palavras isoladas como de sequência de palavras tal como já havia ocorrido de certa forma no estudo anterior (cf. ZYNGIER ET AL. 2007). Uma das variáveis corresponde à frequência relativa de feixes lexicais, indicando em que quantidade as sequências de palavras são encontradas em uma dada obra. Porém, como não basta somente contar a ocorrência de feixes, a segunda variável – razão forma/item de feixes lexicais – indica a diversidade destas sequências. Por fim, a terceira variável relaciona-se à razão forma/item padronizada de palavras calculada automaticamente pelo *WordSmith Tools*. Atribui-se peso 2 às duas primeiras variáveis, que lidam com feixes lexicais, e peso 1 à variável que se relaciona ao uso de palavras. O resultado final do índice varia em uma escala de cinco pontos – de 0 a 5 – indicando a literariedade de uma obra. Ao testar o índice em 46 obras literárias, sendo 27 canônicas e 19 não-canônicas, os pesquisadores observam uma concentração de obras canônicas entre os 15 primeiros resultados e, conversamente, uma recorrência de obras não-canônicas ou de obras canônicas destinadas ao público infantil nos últimos 15 resultados. A primeira obra da lista é *Ulysses* de James Joyce, cujo texto é reconhecidamente inovador em termos de uso da linguagem. Em último lugar, tem-se *Alice in Wonderland* de Lewis Carroll, posição esta que pode ser explicada pelo fato de a mesma usar uma linguagem mais formulaica e acessível ao público infanto-juvenil.

5. Conclusão

Os estudos descritos acima indicam que o uso da ferramenta computacional para análise de textos literários já começa a abrir caminho no Brasil. Estas são pesquisas iniciais que têm como característica a disposição de desbravar um campo desconhecido e tornado possível devido ao avanço tecnológico. Como tal, ainda precisam ser criticadas e observadas sob diversos ângulos antes que se possa afirmar que de fato contribuem para modificar de vez paradigmas existentes. No sentido

kuhniano, são esforços que, se não se definem ainda como centrais, questionam concepções tradicionais e, no melhor espírito científico, buscam oferecer novas alternativas para os estudos na interface entre língua e literatura.

Sonia Zyngier
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Vander Viana
QUEEN'S UNIVERSITY

Natália Giordani Silveira
STANFORD UNIVERSITY

Notas

[1] Embora este seja um emprego comum, também se encontram definições de *foregrounding* como os próprios artifícios linguísticos que levam à desfamiliarização (VAN PEER E HAKEMULDER 2006).

[2] O padrão no *WordSmith Tools* (SCOTT 1999) calcula a razão para grupos de 1.000 palavras.

Referências

BIBER, Douglas; CONRAD, Susan; CORTES, Viviana. "If you look at...: lexical bundles in university teaching and textbooks." In: *Applied Linguistics*, 25-3/2004, 371-405.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randy. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press, 1998.

CAMARGO, Diva Cardoso de. "Uso do conjunto léxico por um tradutor literário em The War of the Saints". In: *Matraga*, 19/2006, 33-48.

CARTER, Ronald; NASH, Walter. *Seeing through language: a guide for styles of English writing*. Oxford: Basil Blackwell, 1990.

GONÇALVES, Lourdes Bernardes. "Lingüística de Corpus e Análise Literária". In: *Anais do I Congresso Internacional da ABRAPUI*, 2007, 1-9.

JAKOBSON, Roman; LEVI-STRAUSS, Claude. "Charles Baudelaire's 'Les Chats.'" . In : POMORSKA, Krystyna; RUDY, Stephen (Eds.).*Language in Literature*. Cambridge, MA: Belknap, (original publicado em *L'Homme* 2, 1962) 1987, 180-197.

MCENERY, Tony; WILSON, Andrew. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1996.

SAUSSURE, Ferdinand de. *Curso de lingüística geral*. 26. ed. São Paulo: Cultrix, 1916 (impressão 2004).

SCOTT, Mike. *WordSmith Tools 3.0*. Oxford: Oxford University Press, 1999.

SHKLOVSKY, Victor. "Art as technique". In: LEMON, Lee Thomas; REIS, Marion J. (Eds.). *Russian Formalist criticism: four essays*. Lincoln, NE: University of Nebraska Press, (original publicado em 1917) 1965, 3-24.

SINCLAIR, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

_____. *Reading concordances: an introduction*. London: Longman, 2003.

VAN PEER, Willie; HAKEMULDER, Jèmeljan. "Foregrounding". In: BROWN, Keith (Ed.). *Encyclopedia of language and linguistics*. Oxford: Elsevier, 2006, 546-551.

VIANA, Vander; FAUSTO, Fabiana; ZYNGIER, Sonia. "Corpus linguistics & literature: a contrastive analysis of Dan Brown and Machado de Assis". In: ZYNGIER, Sonia; VIANA, Vander; JANDRE, Juliana (Eds.). *Textos e leituras: estudos empíricos de língua e literatura*. Rio de Janeiro: Publit, 2007, 233-256.

ZYNGIER, Sonia. "Smudges on the canvas: a corpus stylistics approach to Macbeth". In: BIERMANN, Ina; COMBRINK, Annette. (Eds.). *Poetics, linguistics and history: discourses of war and conflict*. Potchefstroom: University of Potchefstroom Press, 2002, 529-546.

ZYNGIER, Sonia; VIANA, Vander; SILVEIRA, Natália Giordani. "Por uma literatura de corpus: literariedade através do computador". In: *Caderno de resumos do VI Encontro de Lingüística de Corpus*, 2007, 77-78.

ZYNGIER, Sonia et alii. "Firing the canon through the computer: lexical bundles, literary genre and the translation of a best-seller". In: *Anais do 4º Simpósio Internacional de Estudos de Gêneros Textuais*, 2007, 354-366.

Resumo

O presente artigo tem por objetivo demonstrar como estudos na interface entre língua e literatura podem utilizar a Linguística de *Corpus* para desenvolver análises pautadas em dados reais, evitando, até certo ponto, o subjetivismo de interpretações hermenêuticas. A partir das noções de estranhamento e *foregrounding*, e da descrição da linguagem proposta pela Linguística de *Corpus*, os estudos aqui descritos indicam apenas alguns dos caminhos a serem trilhados por aqueles que desejam trabalhar na confluência entre Literatura e Linguística de Corpus.

Palavras-chave: discurso literário; Linguística de *Corpus*; análise textual; abordagem empírica.

Abstract

This article demonstrates how studies which work on the interface between language and literature can resort to Corpus Linguistics to develop evidence-based analyses. In this sense, they avoid – to a certain extent – the subjectivism of hermeneutic interpretations. From the notions of foregrounding and defamiliarization and from the description of language proposed by Corpus Linguistics, the studies here described indicate but a few paths to be taken by those who intend to work on the meeting-ground between Literature and Corpus Linguistics.

Keywords: literary discourse; Corpus Linguistics; text analysis; empirical approach.