LITERATURA ATRAVÉS DO COMPUTADOR: FEIXES LEXICAIS E CÂNONE[1]

Vander Viana (PUC-Rio)
Fabiana Fausto (UFRJ)
Sonia Zyngier (UFRJ)

RESUMO

O presente trabalho utiliza-se do referencial teórico da Lingüística de *Corpus* de forma a tentar mapear características canônicas ou populares na linguagem de duas obras literárias específicas. A escolha dos autores investigados foi realizada após uma consulta a leitores reais, que indicaram suas preferências por obras de Machado de Assis e Dan Brown. Em um estágio posterior, escolheram-se duas obras principais destes autores, a saber, *Dom Casmurro* e *O Código da Vinci*. Os textos foram, então, digitalizados, formatados e compilados em dois *corpora* de modo que pudessem ser lidos pelo *WordSmith Tools* (Scott, 1999), um programa computacional de análise textual. A investigação realizada foi calcada no conceito de feixe lexical (cf. Biber et al., 2004). Decidiu-se aqui trabalhar com feixes compostos por quatro palavras à semelhança do estudo citado. Uma vez tendo realizado o levantamento dos feixes lexicais em ambos os *corpora* de pesquisa, eles puderam ser classificados estrutural e funcionalmente de acordo com as taxonomias propostas por Biber et al. (2004). Os resultados desta análise sugerem que a existência de uma maior variedade de estruturas e funções nos feixes encontrados na obra de Machado de Assis quando comparada às escolhas lexicais na obra de Dan Brown.

PALAVRAS-CHAVES: Lingüística de *Corpus*; Literatura; Cânone

INTRODUÇÃO

Em um estudo anterior, Deus et al. (2006) investigaram o posicionamento de leitores perante o objeto literário em duas comunidades do Orkut, aleatoriamente escolhidas. Porém, não se oferece nenhuma justificativa para tal procedimento e/ou para a opção por duas comunidades em vez de um número maior e talvez mais representativo.

[1] Trabalho apresentado na I Semana Nacional de Crítica Textual e Edição de Textos.

O trabalho de Deus et al (2006) parece apontar alguns resultados que necessariamente precisam ser melhor investigados. Por exemplo, quando da análise da comunidade 'Preponderância erudita', as autoras afirmam que os participantes mencionam tanto autores canônicos (Machado de Assis e José de Alencar) como não-canônicos (Paulo Coelho e Dan Brown). Argumenta-se que

> Os feixes lexicais com maior número de ocorrências – *livro(s) de auto-ajuda, o Código da Vinci, (Dan) Brown ou Paulo Coelho* – indicam que a discussão prioriza autores e obras de literatura popular. Entretanto, logo após estes itens, nota-se a menção à Machado de Assis, indicando um interesse dos participantes em discutir obras canônicas. Desta forma, ao falar sobre Literatura, os participantes não mencionam somente os clássicos, mas mostram-se atentos às obras populares. Claro que o fato de os participantes terem se agrupado em um tópico intitulado *Prepotência erudita* leva a esse tipo de posicionamento. No entanto, pode-se notar que as afirmações feitas e a filiação ao grupo são de livre escolha. (Deus et al., 2006: 104-105)

Apesar de haver liberdade de asserção dentro da comunidade, o que as autoras deixam de perceber é que a menção a ambas as obras ocorrem muito provavelmente por causa do tópico da comunidade. Em outras palavras, o resultado encontra-se diretamente relacionado à escolha aleatória das comunidades a serem investigadas. De forma a mapear a preferência de leitores, objetivo principal do estudo delas, outras comunidades mais neutras deveriam ter sido investigadas.

Contudo, a existência de uma comunidade denominada 'Preponderância erudita' já aponta inicialmente para o fato de que leitores reais nem sempre seguem a opinião de críticos literários na hora de decidirem que obras lerão. O presente estudo toma, então, por pressuposto que leitores inseridos no sistema literário nem sempre optam pela leitura de textos canônicos, mas fazem escolhas com base em suas preferências individuais de leitura.

De forma a tentar identificar quais seriam os autores canônicos e não-canônicos preferidos pelos participantes investigados, lançou-se uma pergunta em 25 comunidades diferentes no Orkut relacionadas ao ato de ler.[2] Apesar de indicar aos participantes que aque-

[2] Ressalta-se aqui que o objetivo desta consulta a leitores através do Orkut não teve o objetivo de mapear a preferência de leitura da população como um todo assim como no estudo de

le tópico era parte integrante de uma pesquisa, informou-se somente que a mesma relacionava-se às preferências de leitura dos participantes da comunidade. Solicitou-se, então, que fossem indicadas uma obra clássica e outra popular que lhes tivesse proporcionado prazer.[3] Após a compilação das respostas e a subseqüente análise, notou-se não haver um consenso em relação às obras, mas os participantes indicaram obras de Machado de Assis e Dan Brown mais freqüentemente.

Com a informação de quais autores foram indicados como os preferidos pelos participantes investigados, optou-se por selecionar uma obra de cada um deles para mapear lingüisticamente as diferenças entre as mesmas. Decidiu-se trabalhar com uma das obras mais populares de cada um dos autores citados: *Dom Casmurro* e *O Código da Vinci*. Analisou-se a obra de Dan Brown em sua tradução para a língua portuguesa uma vez que os participantes das comunidades do Orkut não mencionaram ler a mesma em sua versão original em inglês.

O objetivo maior desta pesquisa é verificar de que forma as construções lexicais empregadas nas duas obras contribuem para a diferença entre as mesmas. Têm-se, então, duas perguntas de pesquisa:

> (a) Levando-se em consideração a noção de feixes lexicais (cf. BIBER et al., 2004), como se estrutura a linguagem em *Dom Casmurro* e em *O Código da Vinci*?
>
> (b) Quais são as diferenças e/ou semelhanças entre as obras investigadas?

O presente artigo se estrutura em quatro partes principais. Primeiramente, revisa-se a literatura pertinente ao estudo em tela, a saber, a Ciência Empírica da Literatura (doravante CEL) e a Lingüís-

---

Deus et al. (2006), o que seria novamente incongruente já que tal mapeamento demanda um número muito maior de respondentes. O foco era somente obter nomes de obras canônicas e não-canônicas que são efetivamente lidas por agentes do sistema literário para que a pesquisa principal – comparação das opções lexicais em cada uma delas – pudesse ser realizada.

[3] Optou-se intencionalmente pelo emprego dos adjetivos 'clássico' e 'popular' em vez de 'canônico' e 'não-canônico' para que a pergunta fosse a mais clara possível para todos os participantes das 25 comunidades investigadas.

tica de Corpus (doravante LC). Posteriormente, são relatados os procedimentos metodológicos adotados nesta pesquisa. Em um terceiro momento, os resultados são apresentados e discutidos. Finalmente, algumas considerações finais são traçadas e desdobramentos futuros são delineados.

REVISÃO DA LITERATURA

São detalhados aqui os principais pressupostos teóricos que guiam este estudo. Para tanto, divide-se esta seção em duas. Inicialmente, discorre-se a respeito da CEL, focalizando principalmente os quatro agentes do sistema literário. Em um segundo estágio, enfoca-se a LC, ressaltando-se a necessidade de pesquisas lingüísticas serem *empíricas*, isto é, baseadas em linguagem em uso.

Ciência Empírica da Literatura (CEL)

O presente estudo se baseia nos pressupostos teóricos da CEL na medida em que as observações sobre as obras analisadas foram baseadas em dados reais, e não em interpretações individuais e subjetivas dos textos. Nesta seção, serão descritos os principais conceitos desta área do conhecimento.

A CEL é definida por Schmidt (1983), seu principal fundador, como uma ciência da literatura calcada em uma rede de elementos tanto teóricos quanto empíricos. Esta área do conhecimento se desenvolveu como ciência a partir da década de 70, com a formação do grupo NIKOL (*Nicht Konservativ Literaturwissenschaft* – Ciência da Literatura Não-Conservadora). O principal objetivo de seus participantes era construir fundamentos teórico-metodológicos mais cientificamente consistentes para os estudos literários, até então baseados em interpretações hermenêuticas. Desta forma, o objeto de investigação dos estudos empíricos não é o texto isolado de seu contexto histórico, mas as ações que possibilitam sua criação, leitura e distribuição na sociedade.

Sob esta ótica, a literatura é vista como um sistema social, no qual quatro papéis podem ser delineados: o de produtor de comunicados literários; o receptor, que constrói sentido a partir da leitura; o

mediador, responsável por possibilitar o acesso às obras, e o pós-processador, que constrói novos comunicados literários a partir da base comunicativa estabelecida pelo autor. Estas ações formam o sistema LITERATURA[4], que, por sua vez, se relaciona a outros sistemas sociais, influenciando-os e sendo influenciados por estes (Schmidt, 1983).

Desta forma, considerar um texto literário não depende apenas das características formais deste texto, mas das ações tomadas a partir da leitura dele. Durante esta leitura, o significado se constrói a partir das ações cognitivas dos quatro agentes supracitados, quando o texto se posiciona social e culturalmente dentro de um contexto específico. Para tanto, Schmidt (1982) faz uma distinção entre o *texto*, visto apenas como uma base comunicativa, do *comunicado*, um construto mental desenvolvido a partir da leitura.

A pesquisa aqui descrita está em conformidade com a CEL, uma vez que se procurou investigar obras canônicas e não-canônicas preferidas por leitores reais, agentes do sistema literário. Tendo em vista que a construção do significado de uma obra se dá através do desenvolvimento de comunicados literários que os agentes processam interagindo socialmente, foi considerado na escolha dos textos o que leitores reais consideram como canônico e não-canônico, ao invés de basear a seleção na preferência dos pesquisadores.

Lingüística de Corpus (LC)

Diferentemente da Lingüística tradicional, em voga nos anos 60 e 70, a LC privilegia a investigação baseada em dados reais, isto é, em instâncias de linguagem em uso criteriosamente compiladas em um *corpus*, que precisa ser representativo de uma língua específica ou de um de seus traços lingüísticos. O *corpus* no âmbito da LC precisa ser legível por computador de forma que possa ser investigado automaticamente, diminuindo sobremaneira a possibilidade de erros e incorreções.

---

[4] O termo 'literatura' é grafado aqui com letras maiúsculas, em conformidade com Schmidt (1982), com o objetivo de ressaltar o conceito de literatura como um sistema.

Na base de conceitos importantes na LC como, por exemplo, colocação e coligação, há uma visão de linguagem como probabilidade de uso. Esta visão difere-se da Lingüística tradicional que objetiva mapear os universais lingüísticos que são comuns a todas as línguas sem se importar com o uso que é feito da mesma por seus usuários.

Parece haver duas possibilidades na produção lingüística: a de utilizar combinações lexicais novas e a de empregar combinações lexicais já conhecidas e utilizadas por outros usuários da língua. Para explicar estes fenômenos, Sinclair (1991) faz uso do que ele nomeia de 'princípio da livre escolha' e 'princípio idiomático'. De acordo com o primeiro, seria possível escolher toda e qualquer palavra que integra um discurso. Em outras palavras, o falante/escritor selecionaria palavra por palavra à medida que a necessidade para tal se apresentasse. Porém, Sinclair (1991) argumenta que o princípio mais produtivo é o idiomático segundo o qual as escolhas lexicais são realizadas em um nível superior ao das palavras, a saber, as seqüências de palavras. O falante/escritor lançaria mão em seu discurso de seqüências lexicais, processadas como blocos de palavras, que ele já leu ou ouviu anteriormente.

Seguindo a mesma linha de argumentação, Biber et al (2004) defendem que a linguagem não é estritamente composicional. Em outras palavras, as escolhas feitas por falantes/escritores seriam operadas na base de seqüências de palavras. De acordo com os pesquisadores, estas seqüências já foram estudadas inúmeras vezes e receberam denominações distintas: "'sintagmas lexicais, 'fórmulas', 'rotinas', 'expressões fixas' e 'padrões pré-fabricados'"[5] (Biber et al., 2004: 372). Eles, no entanto, adotam o termo 'feixe lexical' (ou '*lexical bundle*' no original em inglês) para se referir às "seqüências lexicais recorrentes mais freqüentes em um registro"[6] (Biber et al.,

[5] No original, lê-se: "'lexical phrases', 'formulas', 'routines', 'fixed expressions', 'prefabricated patterns'".

[6] Tradução livre do seguinte fragmento: "the most frequent recurring lexical sequences in a register".

2004: 376).[7] Um feixe lexical não corresponde a uma seqüência de palavras listada pelo computador. Nos termos de Biber et al. (2004), uma seqüência qualquer de palavras deve ocorrer, no mínimo, 40 vezes em grupos de 1.000.000 de palavras e em cinco textos distintos para que seja considerada um feixe lexical e, conseqüentemente, incluída na análise.

Em relação à classificação de feixes lexicais, Biber et al. (2004) propõem duas taxonomias baseadas em suas estruturas e funções. De acordo com a classificação estrutural, um feixe pode ser de três tipos distintos a depender do (fragmento de)[8] sintagma que ele incorpora: nominal e/ou preposicional, verbal simples, ou verbal com uso de estrutura subordinativa. A Figura 1 exemplifica a classificação com feixes retirados dos *corpora* analisados.

	Dom Casmurro	*O Código da Vinci*
Tipo 1	era a primeira vez	ele fez uma pausa
	a alma é cheia	busca o orbe da
Tipo 2	por que é que	tenho certeza de que
	e disse-me que	que o Santo Graal
Tipo 3	na sala de visitas	do outro lado da
	a denúncia de José	o Priorado de Sião

Figura 1: Exemplificação da taxonomia estrutural

Os feixes do Tipo 1 são aqueles que incorporam fragmentos de sintagmas verbais simples nos quais não há o indício de orações subordinadas. Estes feixes podem, por exemplo, começar com o sintagma verbal propriamente dito ('busca o orbe da') ou por um sintagma nominal seguido do sintagma verbal ('a alma é cheia'). Os feixes do Tipo 2, que abarcam as instâncias de orações subordinadas, podem começar com a oração subordinada ('que o Santo Graal') ou com um sintagma verbal simples seguido de algum indício de oração subordinada como em 'e disse-me que' no qual só há o pronome 'que' indicando o início da estrutura subordinada. Já os feixes do Ti-

[7] Deve-se ressaltar que o conceito de 'feixe lexical' remonta ao trabalho de Biber et al. (1999). Porém, é no artigo de 2004 que as taxonomias estrutural e funcional serão propostas mais claramente.

[8] Muitas vezes o feixe incorporará somente um fragmento de um sintagma dado o seu tamanho reduzido, contendo apenas quatro palavras.

po 3 incorporam fragmentos de sintagmas nominais (‘a denúncia de José’) e/ou preposicionais (‘do outro lado da’).

Em termos funcionais, há quatro categorias distintas: atitudinal, referencial, discursiva e conversacional. A Figura 2 exemplifica as categorias com os feixes dos *corpora* aqui investigados.

	Dom Casmurro	*O Código da Vinci*
Atitudinal	não posso ser padre	tinha certeza de que
	quis saber o que	não podia deixar de
Referencial	de um lado para	os olhos de Langdon
	dous ou três meses	com todo o cuidado
Discursivo	uma vez que não	à medida que o
Conversacional	que me disse isto	está me dizendo que

Figura 2: Exemplificação da taxonomia funcional

Os feixes do tipo atitudinal indicam o posicionamento do falante/escritor e sinalizam como a proposição seguinte deve ser interpretada. No caso de ‘tinha certeza de que’ em *O Código da Vinci*, por exemplo, não se deixa a possibilidade de discordar da afirmação que é feita a seguir. Os feixes referenciais apontam, em sua grande maioria, para o mundo exterior, fazendo menção a entidades físicas ou abstratas, ou a alguma de suas características. Enquanto o feixe ‘os olhos de Langdon’ se refere a algo concreto, ‘dous ou três meses’ serve ao propósito de marcar a referência temporal. A categoria discursiva engloba os feixes que indicam, neste caso, ao leitor qual é a relação entre o que já foi lido e o que será lido. Por exemplo, o emprego do feixe ‘uma vez que não’ tem o propósito de explicitar ao leitor a razão de algo (não) ter sido feito. Finalmente, os feixes conversacionais servem para indicar o relato de um relato, ou seja, indicam o discurso indireto nas obras analisadas: ‘que me disse isto’ e ‘está me dizendo que’.

METODOLOGIA

De forma a investigar as obras de Machado de Assis e Dan Brown através do computador, foi necessário compilar dois *corpora* de pesquisa. No caso de *Dom Casmurro*, a obra já se encontrava em formato digital na Biblioteca Virtual do Estudante de Língua Portuguesa gerenciada pela Escola do Futuro da USP. *O Código da Vinci*, no entanto, teve que ser digitalizado manualmente com auxílio de

um scanner. Em ambos os casos, os *corpora* foram formatados adequadamente de acordo com certos princípios definidos pelos pesquisadores.[9] Com vistas a redução da ocorrência de erros, principalmente na obra digitalizada manualmente, os dois *corpora* foram verificados por dois pesquisadores em momentos distintos.[10]

Após a etapa de compilação, os *corpora* foram investigados com o auxílio do programa *WordSmith Tools* (Scott, 1999). O *corpus* que corresponde à obra de *Dom Casmurro* (doravante DmC) contém 66.881 itens e 8.689 formas. O *corpus* contendo *O Código da Vinci* (doravante CdV) totaliza 148.214 itens e 14.774 formas.

O escopo do presente trabalho; porém, relaciona-se ao emprego de feixes lexicais. Antes de proceder a análise, foi preciso decidir o que era um feixe lexical já que este não corresponde a toda e qualquer seqüência de palavras listada pelo computador. Optou-se neste estudo, à semelhança ao estudo de Biber et al. (2004), por um ponto de corte arbitrário e baseado em freqüência.[11] Em outras palavras, para que uma seqüência de palavras fosse considerada um feixe lexical, ela deveria ocorrer, no mínimo, três vezes em DmC e seis vezes em CdV.[12]

Além disto, decidiu-se também que os feixes deveriam marcar o início de sintagmas nominais, preposicionais ou verbais. Assim sendo, toda e qualquer seqüência que não cumprisse esta exigência não seria incluída na análise como, por exemplo, 'alma é cheia de' em DmC e 'outro lado da sala' em CdV. Deve-se ressaltar, contudo,

---

[9] Por causa da limitação de espaço, não são especificados aqui os princípios adotados. Pode-se, no entanto, a título de exemplificação, mencionar que foram inseridos pontos após os nomes de cada capítulo de forma que o programa computacional pudesse ler adequadamente o término daquela seqüência de palavras.

[10] Agradecemos à Suzana de Deus pelo auxílio no tratamento dos dados no estágio inicial desta pesquisa.

[11] No entanto, o critério de dispersão empregado por Biber et al. (2004) não foi aqui empregado visto que cada *corpus* corresponde a uma única obra.

[12] Apesar de as freqüências absolutas serem discrepantes, o ponto de corte foi decidido com base em freqüência relativa. Em outras palavras, decidiu-se que uma seqüência de palavras deveria ter freqüência igual ou maior do que 4 vezes por grupos de 100.000 itens. A freqüência relativa é calculada dividindo a freqüência absoluta pelo tamanho do *corpus* (em itens) e multiplicando o resultado, neste caso específico, por 100.000.

que feixes semelhantes a estes (‘a alma é cheia’ e ‘do outro lado da’) foram incluídos na análise. A adoção deste critério, na verdade, significou a exclusão de muitas seqüências iguais e sobrepostas, evitando assim que os resultados aqui encontrados fossem inflacionados.

RESULTADOS E ANÁLISES

Após a compilação da lista de feixes lexicais em ambos os *corpora*, verificou-se a distribuição de feixes lexicais de acordo com a taxonomia estrutural de Biber et al. (2004). Esta taxonomia, já descrita e ilustrada na Subseção 2.2, prevê que os feixes podem ser de três tipos a depender da natureza do sintagma que eles incorporam. A Figura 3 indica a distribuição dos feixes nos *corpora* aqui investigados.

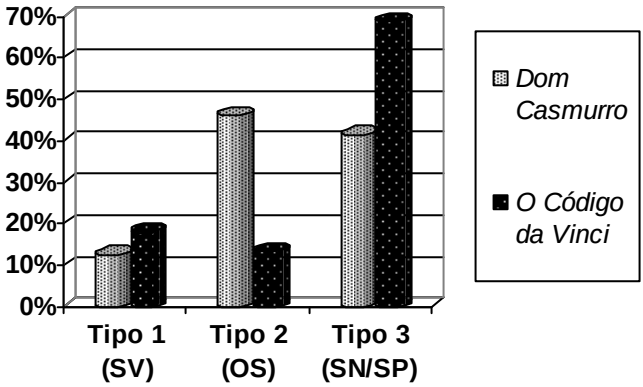

Figura 3: Distribuição estrutural de feixes lexicais

O *corpus* DmC contém mais feixes com sintagmas verbais sejam eles coordenados ou subordinados (Tipos 1 e 2) do que feixes que contém sintagmas nominais e/ou preposicionais (Tipo 3). Quando se observa a distribuição de feixes nos Tipos 1 e 2, conclui-se que os sintagmas verbais em DmC são partes de estruturas subordinadas em sua maioria (46,04%).

A distribuição estrutural em CdV é distinta. Apesar de a obra ser uma narrativa, os padrões lexicais mais recorrentes não são os

que indicam processos como seria esperado, mas aqueles que descrevem coisas. Feixes do Tipo 3 são os mais freqüentes com 68,71% das instâncias analisadas. No que tange à utilização de feixes dos Tipos 1 e 2, conclui-se que, diferentemente da obra de Assis, a obra de Brown utiliza-se mais de feixes incorporando sintagmas verbais simples do que aqueles que estão inseridos em orações subordinadas.

Os feixes também foram classificados funcionalmente, como explicitado na Subseção 2.2. A Figura 4 ilustra a distribuição de feixes nestas categorias.

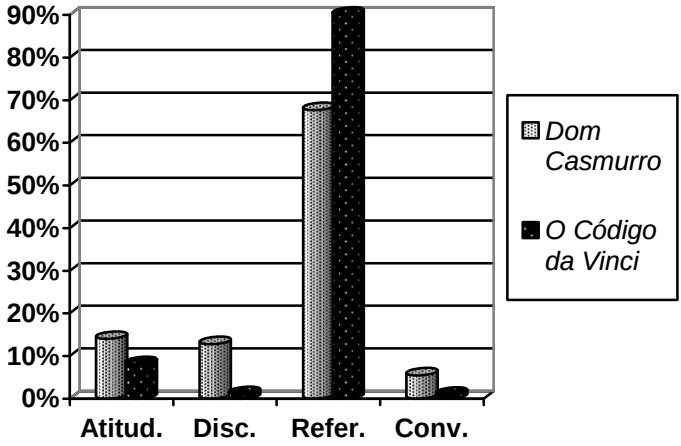

Figura 4: Distribuição funcional de feixes lexicais

Ambos os *corpora* apresentam uma grande concentração de feixes referenciais, que, de uma forma geral, apontam para entidades físicas ou abstratas, ou alguma de suas características.

Porém, esta concentração é especialmente maior em CdV uma vez que tal categoria corresponde à 89,97% das instâncias analisadas. Este resultado corrobora a classificação estrutural anteriormente descrita. *O Código da Vinci* contém mais feixes nominais e/ou preposicionais, que desempenham a função referencial, o que aponta que a descrição padronizada desempenha um papel importante no âmbito desta obra.

No *corpus* DmC, apesar de haver uma concentração de feixes referenciais (67,68%), há mais feixes desempenhando outras funções. Em outras palavras, há mais feixes de naturezas distintas no *corpus* DmC do que no *corpus* CdV. O que se ressalta aqui é a diferença de feixes discursivos (11,76%), atitudinais (5,94%) e conversacionais (4,59%) entre os dois *corpora*. Parece que, em *Dom Casmurro*, o texto aponta de forma mais clara para os seus leitores quais são as relações entre as diferentes partes da obra. Há também na obra de Assis uma variedade maior de tipos de feixes atitudinais, indicando não somente habilidade e modalização epistêmica como em CdV, mas também desejo, intenção/predição e obrigação/diretiva. A ocorrência de mais feixes conversacionais também indica que há nesta obra diferentes níveis de relato se for considerado que os feixes conversacionais neste *corpus* servem ao propósito do discurso indireto, ou seja, o relato de um relato.

CONCLUSÕES

Com base nos pressupostos teóricos da CEL, o presente estudo investigou as seqüências lexicais em duas obras distintas: *Dom Casmurro* de Machado de Assis e *O Código da Vinci* de Dan Brown. Pode-se argumentar que a comparação entre as duas obras não deveria ser realizada haja vista que os contextos de produção das mesmas diferirem em termos temporais, geográficos e lingüísticos. Contudo, a seleção das obras seguiu a resposta fornecida por leitores, agentes do sistema literário, que identificaram obras canônicas e não-canônicas de sua preferência. Desta forma, evitou-se que a escolha das mesmas fosse uma decisão totalmente arbitrária tomada pelos pesquisadores.

A LC forneceu as diretrizes fundamentais para que este estudo empírico pudesse ser realizado. As obras foram analisadas não com base em interpretações pessoais e subjetivas, mas nas seqüências lexicais que ambos os autores empregaram em suas respectivas obras. Desta forma, a LC parece se integrar harmonicamente aos objetivos da CEL já que permite deixar de lado a hermenêutica, investigando obras literárias com base em seu material textual.

Em relação à comparação realizada, nota-se que há dois padrões lingüísticos distintos nas obras analisadas. Estruturalmente, os feixes em *Dom Casmurro* incorporam fragmentos de sintagmas verbais com concentração de orações subordinadas, o que pode indicar o uso de uma linguagem mais complexa que exige maior concentração por parte do leitor. Em termos funcionais, os feixes, apesar da concentração de feixes referenciais, encontram-se distribuídos em outras categorias. Este resultado pode indicar a variedade no emprego de feixes se for considerado que os mesmos desempenham funções diversas.

Por outro lado, *O Código da Vinci*, de Dan Brown, parece lançar mão de feixes que incorporam principalmente sintagmas nominais e/ou preposicionais desempenhando a função referencial. Parece haver nesta obra uma grande recorrência de descrições padronizadas.

Explicar as diferenças aqui encontradas ainda não é possível. Contudo, levantam-se algumas hipóteses para as mesmas. Pode ser que o fato de *O Código da Vinci* ter sido escrito originalmente em língua inglesa tenha influenciado a sua tradução para a língua portuguesa. Pode ser que as diferenças estejam relacionadas com as épocas nas quais estas obras foram produzidas: *Dom Casmurro* remonta ao século XIX enquanto *O Código da Vinci* é datada do século XXI. Por fim, pode-se supor que estas diferenças estejam relacionadas ao que se considera canônico e não-canônico, ou seja, a canonicidade (ou a não-canonicidade) de uma obra, um construto teórico, estaria marcada também na linguagem empregada pelo autor.

De forma a tentar responder às hipóteses levantadas, novos estudos comparativos precisam ser realizados. É importante que outras obras, produzidas em contextos semelhantes, sejam analisadas para que se possa afirmar qual fator é o preponderante nas diferenças mapeadas neste estudo.

REFERÊNCIAS

BIBER, D. et al. *Longman grammar of spoken and written English*. London: Longman, 1999.

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

DEUS, S. de; FAUSTO, F.; TELES, C. O conceito de literatura para leitores: uma investigação em redes sociais na internet. **In**: ZYNGIER, S.; VIANA, V.; SPALLANZANI, A. M. (Org.). *Linguagens e tecnologias: estudos empíricos*. Rio de Janeiro: Publit, 2006. p. 99-110.

SCHMIDT, S. J. *Foundations for the empirical study of literature: the components of a basic theory*. Hamburg: Helmut Buske, 1982.

———. The empirical science of literature ESL: a new paradigm. *Poetics*, 12, p. 19-34, 1983.

SCOTT, M. *WordSmith tools 3.0*. Oxford: Oxford University Press, 1999.

SINCLAIR, J. (Ed.). *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.