CANÔNICO OU POPULAR? QUALIDADE LITERÁRIA ATRAVÉS DO COMPUTADOR

Vander Viana (PUC-Rio)

Fabiana Fausto (UFRJ)

Sonia Zyngier (UFRJ)

1) Introdução

Em um estudo anterior, Deus et al. (2006) analisaram o conteúdo de dois tópicos de discussão de uma comunidade no Orkut[1] com vistas a mapear como leitores reais se posicionam acerca do objeto literário. Para a realização da análise, baseada nos princípios da Lingüística de *Corpus* (doravante LC), utilizou-se o programa computacional *WordSmith Tools* (SCOTT, 1999).

O trabalho de Deus et al. (2006) apresenta algumas lacunas tanto no procedimento metodológico quanto nos resultados encontrados. Afirma-se no artigo que os dois tópicos analisados foram escolhidos aleatoriamente; porém, não se oferece nenhuma justificativa para tal opção. De forma semelhante, também não se é sustentada a decisão de analisar dois tópicos em vez de um número maior e talvez mais representativo. Em relação aos resultados, as autoras afirmam que os participantes mencionam tanto autores não-canônicos (Dan Brown e Paulo Coelho) como canônicos (José de Alencar e Machado de Assis) quando da análise do tópico 'Prepotência erudita'. Apesar de a liberdade de expressão ser assegurada dentro da comunidade, nota-se que a menção a ambas as obras ocorrem muito provavelmente por causa do tópico da mesma. Desta forma, o resultado encontra-se diretamente relacionado à escolha, dita aleatória, dos tópicos a serem investigados. Uma possível solução para este efeito

[1] O Orkut é um serviço on-line oferecido pelo Google no qual é possível construir uma página pessoal e relacioná-la às de outros usuários, formando redes sociais.

tendencioso seria investigar comunidades com tópicos mais neutros, que não guiassem a discussão gerada por seus participantes.

No entanto, a própria existência de um tópico de discussão nomeado 'Prepotência erudita' aponta, inicialmente, para o fato de que leitores reais nem sempre seguem a recomendação da crítica literária quando decidem as obras que lerão. Este estudo tem, então, como pressuposto que o leitor, um dos agentes do sistema literário, também faz suas escolhas com base em preferências individuais.

Com o propósito de identificar quais seriam os autores canônicos e não-canônicos preferidos pelos sujeitos investigados, lançou-se uma pergunta para 25 comunidades diferentes no Orkut relacionadas ao ato de ler.[2] Apesar de saberem que a pergunta era parte integrante de uma pesquisa, informou-se unicamente aos participantes que esta objetivava levantar preferências de leitura. Pediu-se, desta forma, que fossem indicadas uma obra clássica e outra popular cuja leitura tivesse sido prazerosa.[3] Após a análise, notou-se uma maior freqüência de indicação das obras de Machado de Assis e Dan Brown.

Optou-se a seguir pela seleção de uma obra de cada um dos referidos autores para mapear as diferenças de escolha lexical 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*. Ressalta-se que a obra de Dan Brown foi analisada em sua tradução para a língua portuguesa já que os participantes não mencionaram ler a mesma em sua versão original em inglês.

---

[2] 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 Deus et al. (2006), o que seria novamente incongruente visto 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 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.

O objetivo desta pesquisa é verificar de que forma as seqüências lexicais empregadas nas duas obras contribuem para possíveis semelhanças e/ou diferenças entre as mesmas. Têm-se, então, duas perguntas de pesquisa explicitadas a seguir:

(a) Considerando-se 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) Que diferenças e/ou semelhanças podem ser levantadas entre as obras investigadas?

O presente artigo se estrutura em três quatro principais. Primeiramente, revisa-se a literatura sobre LC, e explicita-se o conceito de feixe lexical. Em seguida, é relatada a metodologia adotada nesta pesquisa. Em um terceiro momento, os resultados são apresentados e discutidos. Por fim, considerações finais são tecidas e possíveis encaminhamentos são delineados.

2) Lingüística de Corpus (LC)

Em contraste com a Lingüística tradicional, em voga, por exemplo, nas décadas de 60 e 70, a LC confere importância à investigação da linguagem baseada em dados reais. Em outras palavras, a LC investiga 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. Ademais, este *corpus* precisa ser formatado de forma que possa ser lido por computador, possibilitando sua investigação (semi-)automática e minimizando a possibilidade de incorreções e erros.

De acordo com a LC, há duas possibilidades na produção lingüística: o emprego de combinações lexicais novas ou de seqüências lexicais já conhecidas e utilizadas por outros usuários da língua. Sinclair (1991) nomeia estas possibilidades de ‘princípio da livre escolha’ e ‘princípio idiomático’. O primeiro se relaciona com a escolha de palavras individuais, ou seja, o falante/escritor selecionaria palavra por palavra. O

segundo abrange as escolhas lexicais realizadas em um nível superior ao das palavras. Nesta visão, o falante/escritor utiliza seqüências lexicais, processadas como blocos de palavras, que ele já leu ou ouviu anteriormente. De acordo com Sinclair (1991), é este segundo princípio o mais produtivo.

Biber et al. (2004) também defendem que a linguagem não é estritamente composicional. As escolhas feitas por falantes/escritores seriam operadas na base de seqüências de palavras, que eles denominam de 'feixe lexical' (ou '*lexical bundle*' no original em inglês) para se referir às "seqüências lexicais recorrentes mais freqüentes em um registro"[4] (BIBER et al., 2004: 376).[5] Porém, um feixe lexical não corresponde a qualquer seqüência de palavras listada pelo computador. No estudo citado, era preciso que uma seqüência ocorresse, no mínimo, 40 vezes em grupos de 1.000.000 de palavras e em cinco textos distintos para que fosse considerada um feixe lexical.

Além do conceito de feixe lexical propriamente dito, Biber et al. (2004) propõem duas taxonomias para os mesmos. A classificação estrutural compreende três tipos diferentes a depender do (fragmento de)[6] 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	é cheia de mistérios	Langdon sacudiu a cabeça
	as leis são belas	meteu a mão no
Tipo 2	a verdade é que	tinha certeza de que
	que Deus lhe dera	que a pedra-chave
Tipo 3	depois de alguns instantes	do Priorado de Sião
	os olhos de ressaca	os olhos de Langdon

Figura 1: Classificação estrutural – alguns exemplos

[4] Lê-se no original: "the most frequent recurring lexical sequences in a register".
[5] O conceito de 'feixe lexical' foi primeiramente proposto no trabalho de Biber et al. (1999). Porém, as taxonomias estrutural e funcional dos feixes foram propostas no artigo de 2004.
[6] Dado o seu tamanho reduzido, o feixe freqüentemente incorporará somente um fragmento de um sintagma.

O Tipo 1 abarca feixes que incorporam fragmentos de sintagmas verbais simples. Estes feixes podem, por exemplo, começar com o sintagma verbal (‘meteu a mão no’) ou com um sintagma nominal seguido do sintagma verbal (‘as leis são belas’). O Tipo 2 agrupa as instâncias de orações subordinadas. Os feixes deste tipo podem começar com uma oração subordinada (‘que Deus lhe dera’) ou com um sintagma verbal simples seguido de algum indício de oração subordinada (‘tinha certeza de que’). O Tipo 3, por sua vez, engloba os feixes que contêm fragmentos de sintagmas nominais (‘os olhos de ressaca’) e/ou preposicionais (‘do Priorado de Sião’).

Funcionalmente os feixes podem ser atitudinais, discursivos, referenciais ou conversacionais. A Figura 2 apresenta exemplos provenientes dos *corpora* aqui investigados.

	Dom Casmurro	O Código da Vinci
Atitudinal	que não era preciso	tenho certeza de que
	você o que quer	não podia deixar de
Referencial	na sala de visitas	do outro lado da
	depois de alguns instantes	a pedra-chave do
Discursivo	para o fim de	à medida que o
Conversacional	disse-me que era	está me dizendo que

Figura 2: Classificação funcional – alguns exemplos

Os feixes do tipo atitudinal indicam o posicionamento do falante/escritor. No caso de ‘tenho certeza de que’ em *O Código da Vinci*, por exemplo, não há a possibilidade de discordar da afirmação que se segue. Os feixes referenciais fazem menções a entidades físicas ou abstratas, ou a alguma de suas características. Enquanto o feixe ‘a pedra-chave do’ se refere a algo concreto, ‘na sala de visitas’ estabelece uma referência espacial. Por sua vez, a categoria discursiva engloba feixes que indicam a relação entre diferentes partes do discurso. Por exemplo, o uso do feixe ‘para o fim de’ tem o propósito de explicitar ao leitor a finalidade de algo (não) ter sido realizado.

Finalmente, os feixes conversacionais são utilizados, nestas duas obras específicas, para indicar o discurso indireto: 'disse-me que era' e 'está me dizendo que'.

3) Procedimentos metodológicos

Com a finalidade de investigar *Dom Casmurro* e *O Código Da Vinci* com o uso do programa *WordSmith Tools* (SCOTT, 1999), tornou-se necessário compilar dois *corpora* de pesquisa. No caso da obra de Machado de Assis, o texto já se encontrava em formato digital na Biblioteca Virtual do Estudante de Língua Portuguesa. *O Código da Vinci*, no entanto, teve que ser digitalizado manualmente. Em ambos os casos, os *corpora* foram formatados adequadamente de acordo com princípios definidos pelos pesquisadores.[7] Com vistas à redução da ocorrência de erros, os *corpora* foram verificados por dois pesquisadores em momentos distintos.[8]

O *corpus* que corresponde à obra de *Dom Casmurro* (doravante DmC) contém 66.881 itens e 8.689 formas. Já 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. Portanto, foi necessário definir o que era um feixe lexical, já que este não corresponde a toda e qualquer seqüência de palavras listada pelo computador. Assim sendo, optou-se por um ponto de corte arbitrário e baseado em freqüência, à semelhança ao estudo de Biber et al. (2004).[9] Para que uma seqüência de palavras fosse considerada

[7] Devido à limitação de espaço, não são especificados aqui os princípios adotados. Porém, a título de exemplificação, menciona-se um dos critérios adotados: a inserção de pontos após os nomes de cada capítulo de forma que o programa computacional pudesse ler adequadamente o término daquela seqüência de palavras.

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

[9] O critério de dispersão empregado por Biber et al. (2004), contudo, não foi utilizado nesta pesquisa uma vez que *corpus* corresponde a uma única obra.

um feixe lexical, ela deveria ocorrer, no mínimo, três vezes em DmC e seis vezes em CdV.[10]

Ademais, decidiu-se também que os feixes deveriam marcar o início de sintagmas nominais, preposicionais ou verbais. Deste modo, toda e qualquer seqüência que não estivesse em conformidade com esta exigência não foi incluída na análise, como, por exemplo, 'denúncia de José Dias' em DmC e 'banco de custódia de' em CdV. Ressalta-se, contudo, que feixes semelhantes a estes ('a denúncia de José' e 'do banco de custódia') foram incluídos na análise. A adoção deste critério resultou na exclusão de muitas seqüências iguais e sobrepostas, evitando assim que os mesmos feixes fossem contabilizados mais de uma vez nos resultados finais.

4) Discussão dos resultados

Primeiramente, verificou-se a distribuição de feixes lexicais de acordo com a taxonomia estrutural de Biber et al. (2004). Segundo os autores, os feixes podem ser de três tipos, dependendo das características do sintagma que eles incorporam. A Figura 1 indica a distribuição dos feixes nos *corpora* investigados nesta pesquisa.

[10] Apesar de as freqüências absolutas serem diferentes, 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 por 100.000 neste caso específico.

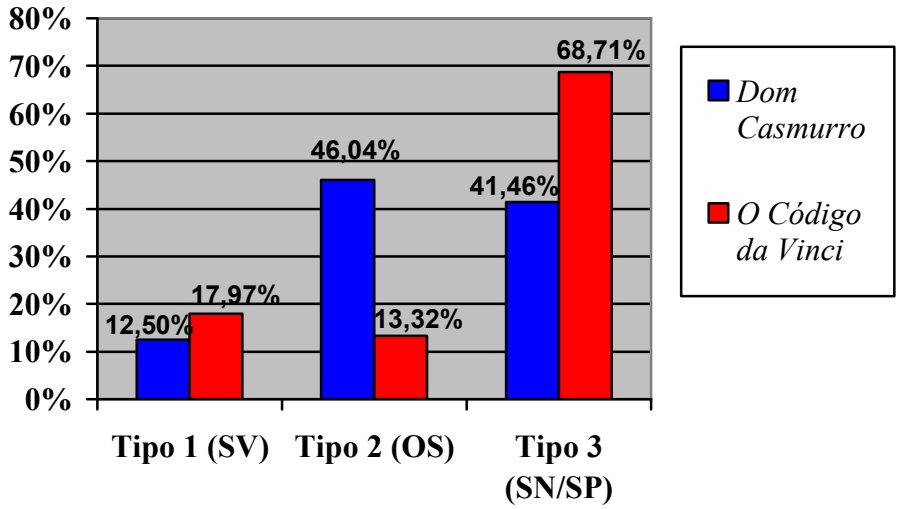

Figura 3: Distribuição estrutural de feixes lexicais

Nota-se que 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). Ao analisar a distribuição de feixes nos Tipos 1 e 2, observa-se que os sintagmas verbais em DmC são partes de estruturas subordinadas em sua maioria (46,04%), como pode-se observar na Figura 3.

Já a distribuição estrutural dos feixes em CdV é distinta na medida em que os padrões lexicais mais recorrentes são os do Tipo 3, com 68,71% das instâncias analisadas. Ressalta-se que, apesar de a obra ser uma narrativa, os padrões lexicais mais freqüentes não são os que indicam processos, como seria esperado para este tipo textual, mas aqueles que descrevem coisas. Em relação à utilização de feixes dos Tipos 1 e 2, conclui-se que, em contraste com a 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.

Em seguida, os feixes também foram classificados funcionalmente, como descrito na Subseção 2.2. A Figura 2 abaixo ilustra a distribuição de feixes nestas categorias.

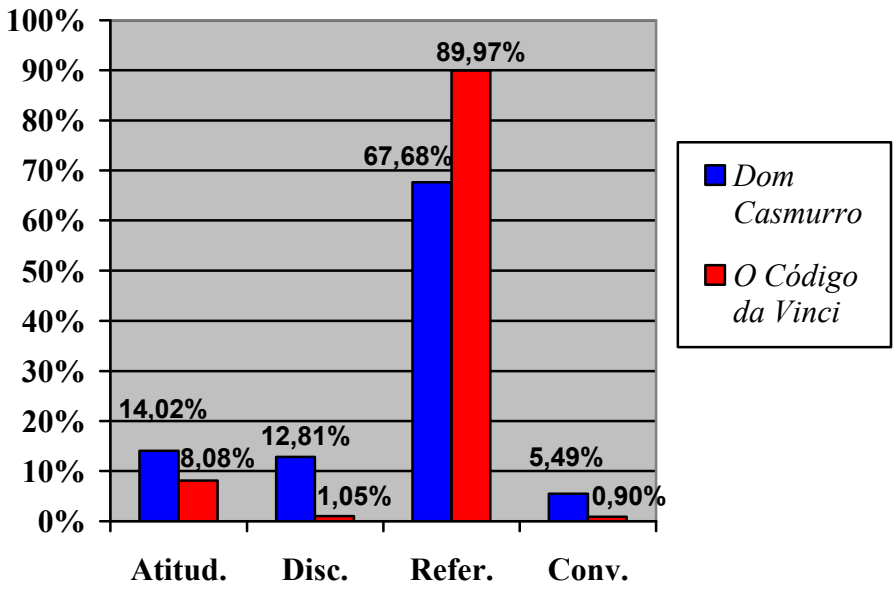

Figura 4: Distribuição funcional de feixes lexicais

Observa-se que os feixes com função referencial são os mais utilizados em ambos os *corpora*. No entanto, nota-se que esta concentração é consideravelmente maior em CdV. A categoria em questão corresponde a 89,97% das instâncias analisadas. O fato de CdV conter mais feixes referenciais se alinha ao fato de o mesmo conter mais feixes nominais e/ou preposicionais. Isto indica que a descrição padronizada desempenha um papel importante nesta obra.

No *corpus* DmC, apesar de haver uma concentração maior de feixes referenciais (67,68%), há um índice maior de feixes desempenhando outras funções. Ressalta-se aqui a diferença de feixes discursivos (11,76%), atitudinais (5,94%) e conversacionais (4,59%) na comparação entre o *corpus* DmC e o CdV. Nota-se que em *Dom Casmurro*, o texto aponta mais claramente para os seus leitores quais são as relações entre as diferentes partes da obra. Há também na obra de Assis uma diversidade maior de tipos de feixes atitudinais, indicando não somente modalização epistêmica e habilidade 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,

levando-se em consideração que os feixes conversacionais neste *corpus* servem ao propósito do discurso indireto.

5) Conclusões e encaminhamentos

O presente estudo investigou as seqüências lexicais nas obras *Dom Casmurro* de Machado de Assis e *O Código da Vinci* de Dan Brown com base nos pressupostos teóricos e metodológicos da LC. Desta forma, a análise aqui descrita baseou-se em seqüências lexicais que ambos os autores utilizaram nas referidas obras, não em interpretações individuais e subjetivas.

Em relação à comparação realizada, notou-se o delineamento de 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 ativa esquemas de leitura mais complexos na mente do leitor. No que tange à análise funcional, observou-se houve uma maior concentração de feixes referenciais; porém, notou-se também uma maior distribuição de feixes em outras categorias. Por outro lado, *O Código da Vinci* emprega feixes que incorporam principalmente sintagmas nominais e/ou preposicionais desempenhando função referencial. Este resultado pode indicar uma grande recorrência de descrições mais padronizadas nesta obra.

Por fim, o presente trabalho levanta algumas questões relevantes para o estudo de textos literários sob a luz da LC. É possível que o fato de *O Código da Vinci* ter sido escrito originalmente em língua inglesa tenha influenciado a sua tradução para a língua portuguesa. Além disso, as diferenças aqui apresentadas podem estar relacionadas às épocas nas quais estas obras foram produzidas: *Dom Casmurro* foi escrita no século XIX, enquanto *O Código da Vinci* foi produzida no século XXI. Uma terceira hipótese

aponta para o fato de que a canonicidade (ou não) de obras literárias possa estar relacionada ao uso da linguagem pelo autor, ou seja, seria possível identificar características literárias por intermédio da análise de aspectos lingüísticos das obras.

De forma a tentar responder às questões levantadas neste artigo, novas investigações comparativas deverão ser realizadas. Desta forma, torna-se necessário que outras obras, produzidas em contextos semelhantes, sejam analisadas para que se possa afirmar quais aspectos são relevantes dentre as diferenças ressaltadas nesta pesquisa.

6) 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.

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

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