Do texto ao dado: debates sobre leitura distante nas humanidades

From texto to data: debates about distant reading in the humanities

https://doi.org/10.21814/h2d.3569

Suemi Higuchi, Fundação Getúlio Vargas, Brasil

Como citar

Higuchi, S. . (2021). Do texto ao dado: debates sobre leitura distante nas humanidades. *H2D|Revista De Humanidades Digitais*, *3*(2). https://doi.org/10.21814/h2d.3569

ISSN: 2184-562X

Do texto ao dado: Debates sobre leitura distante nas humanidades

From text to data: Debates about distant reading in the humanities

https://doi.org/10.21814/h2d.3569

Suemi Higuchi, Fundação Getúlio Vargas, Brasil

Resumo

As áreas das humanas, em especial a literatura e a história, sempre legaram aos registros textuais grande parte da sua razão de ser e de seu modo de fazer. O presente artigo busca refletir e ampliar o horizonte das relações entre as humanidades e o uso das tecnologias disponíveis, focando prinicpalmente nos métodos de leitura distante para estudos literários, alicerçados na linguística com corpus. De que forma a prática de pesquisa nestas áreas tem sido impactada com o uso de ferramentas digitais? Que desafios precisam ser enfrentados e que oportunidades se abrem neste cenário potencialmente inovador? Estas são algumas das questões discutidas no texto.

Palavras-chave

leitura distante; linguística; corpus; linguagem; semântica

Abstract

The humanities, especially fields like literature and history, have always assigned to textual records a great part of their reason for being and their way of doing. This paper aims to reflect and broaden the horizon of the relationship between the humanities and the use of available technologies, focusing mainly on methods of distant reading for literary studies, based on corpus linguistics. How has research practice in these areas been impacted by the use of digital tools? What challenges need to be faced and what opportunities open up in this potential innovative scenario? These are some of the issues discussed in the text.

Keywords

distant reading; linguistics; corpus; language; semantics

1. Leitura distante: conceitos, métodos e questões

O imbricamento entre as práticas tradicionais de registro do conhecimento e as novas tecnologias é a marca indelével do movimento das Humanidades Digitais (HDs). Elas incorporam os métodos e questões desenvolvidas pelas ciências humanas e sociais, ao mesmo tempo em que mobilizam as ferramentas e perspectivas únicas abertas pela tecnologia digital (Schnapp et al., 2009).

Na área mais intimamente ligada à linguagem e literatura, temos observado de forma crescente o uso de métodos quantitativos para analisar grandes coleções digitais (Jockers, 2013; Santos, 2019). Há pouco mais de vinte anos, Franco Moretti, teórico literário italiano, identificou essa prática com o conceito de leitura distante ou leitura à distância (*distant reading*), onde o distanciamento:

> é uma condição de conhecimento: permite que você se concentre em unidades que são muito menores ou muito maiores do que o texto: dispositivos, temas, tropos – ou gêneros e sistemas. E se, entre o muito pequeno e o muito grande, o próprio texto desaparece, bem, é um daqueles casos em que se pode dizer justificadamente, Menos é mais. (Moretti, 2000, p. 57)

A leitura distante seria uma inversão deliberada do termo mais familiar "leitura atenta", que significa um exame cuidadoso e minucioso das particularidades de um texto. Em contraste, a leitura distante envolve o uso de automação para fazer generalizações sobre vastos corpora. A partir desse novo posicionamento, Moretti defende uma forma não usual de análise literária, pois, para ele, a compreensão da literatura não deveria se restringir apenas aos livros que conseguimos consumir na nossa limitada capacidade de leitura, mas abranger quantidades muito maiores, incluindo aqueles que a maioria nunca acessará: *the great unread*[1]. É evidente que essa preocupação não se iniciou com Moretti, mas foi ele quem efetivamente propôs usar métodos computacionais e de estatística para estudar a história da literatura, substituindo a leitura atenta por modelos abstratos emprestados das ciências. Ao transformar os textos em dados, um mundo de possibilidades de análises se abre, permitindo descobertas sobre a literatura particular ou geral que soam muito mais realistas, porque dizem respeito não apenas aos cânones, mas a um universo maior das obras que um certo período da história ou uma determinada região do globo produziu.

Naturalmente, muitas críticas foram levantadas contra essa abordagem generalista de Moretti, afinal, não se trocam as lentes de observação usadas até então impunemente. Katherine Bode, em seu artigo intitulado "*Literary studies in the digital age",* apresenta uma discussão centrada na adoção de métodos quantitativos para estudos literarios e cita alguns exemplos de argumentos usados

por autores contrários à abordagem preconizada por Moretti. Para uns, nada justificaria a violação da individualidade de um texto, pois como seria possível quantificar sem perder os detalhes disruptivos e separar as significações que aprendemos a atender? Para outros, os experimentos quantitativos de Moretti encerram uma tentativa de controlar a indecidibilidade inerente da cultura literária, criando padrões autoritários totalizantes que reduzem a complexidade do campo literário a modelos por demais simplistas (Bode, 2014, p. 10).

2. Novas escalas de observação

A despeito dessas discussões paradigmáticas, é inegável que os computadores de fato inverteram algumas ordens estabelecidas, ao proporcionar maneiras inéditas de leitura e permitir *insights* interessantes sobre grandes corpora. Embora as máquinas não possam ler e entender um romance da maneira que as pessoas podem, elas são muito boas em procurar informações específicas e identificar padrões, tanto linguísticos como estruturais, que não seriam visíveis no ato da leitura a 'olho nu' (Jockers, 2013; Freitas, 2015).

Não é um método novo, a abordagem interpretativa de textos apoiada em distribuições e frequências, por exemplo, existe há séculos (Santos et al, 2020, p. 281). Linhas de concordância com palavras e expressões da bíblia datam do século XIII pelas mãos de estudiosos que as indexavam manualmente em arranjos alfabéticos, juntamente com as passagens em que elas ocorriam (McCarthy & O'Keeffe, 2010). Entretanto, a evolução da tecnologia abriu novas possibilidades de manuseio, permitindo o emprego de múltiplas escalas de observação. As tarefas de quantificação ganharam nova dimensão, números e estatísticas emergem do próprio texto em uma relação direta com a linguagem. Passa a ser possível, por exemplo, medir e comparar o comprimento de sentenças, observar padrões sintáticos e quantificar as variações lexicais dentro de um conjunto de obras, sintetizando dados para outras investigações (Santos, 2014).

Em *The best seller code,* escrito pelos pesquisadores Jodie Archer e Matthew Jockers em 2016, uma teoria baseada em padrões e algoritmos é apresentada para explicar por que certos livros alcançam sucesso e outros não. Para chegar a ela, os autores desenvolveram um programa que "leu" mais de 20 mil romances buscando identificar características típicas dos títulos que figuraram na lista de best-sellers do New York Times. Informações detalhadas que vão desde os temas abordados até os altos e baixos emocionais dos personagens foram extraídas a partir de pistas linguísticas, seja a nível lexical, seja de estrutura sintática, semântica, entre outros. Assim, analisaram-se aspectos como variedade de vocabulário, uso de pronomes e pontuações, frequência de advérbios, natureza desses advérbios, emprego de adjetivos, dentre outros. Mais de vinte mil características foram extraídas, mas apenas cerca de 2.800 consideradas relevantes para diferenciar as histórias que são fenômeno de venda daquelas que não saem da prateleira, independentemente do gênero. Utilizando técnicas de aprendizado de máquina, construíram um classificador que, segundo eles, é capaz de devolver para cada

obra o seu índice de sucesso, com uma média de acerto de 80% (Archer & Jockers, 2016, p. 26).

Algo desse experimento pode nos levar a crer que em breve os primeiros "leitores" de muitas obras de ficção não serão mais pessoas, mas máquinas que estarão realizando o mesmo tipo de atividade que pessoas contratadas por editoras e agências literárias realizam. Considerando que diariamente centenas de manuscritos originais são despejados nesses balcões e que, muitas vezes, a decisão de publicar ou não recai nas mãos de um ou outro indivíduo sujeito a circunstâncias tão particulares como crenças, gostos e, até mesmo, humores, o uso de algoritmos é uma alternativa que já deve estar sendo pensada seriamente no mundo editorial. O quanto esse cenário impactará no papel do crítico literário e outros especialistas da área? Mudanças de métodos e adoção de tecnologias diferentes sempre virão acompanhadas de debates e pontos de vista longe de serem consensuais em qualquer área, e especialmente nas humanidades.

A combinação entre dados quantitativos para conclusões qualitativas e dados qualitativos para conclusões quantitativas pode ser de fato valiosa na análise de grandes volumes de texto, mas desde que o conhecimento linguístico e o conhecimento estatístico se façam presentes (Santos, 2014, p. 205). Inspirado pela micro e macroeconomia, Matthew Jockers utiliza o termo macroanálise para descrever os métodos estatísticos aplicados nesse tipo de análise em textos (Jockers, 2013). Segundo o autor, a nova abordagem irrompe para o mundo textual tal como este é encontrado atualmente: na forma digital e em larga escala. Assim como Moretti, Jockers deixa claro que coisas importantes que somente são percebidas através da leitura atenta correm o risco de escapar ao macro, e, por isso, as duas escalas de observação devem coexistir como abordagens complementares, conforme discutiremos adiante.

3. Promessas e desconfianças do campo

As ideias de Moretti foram acolhidas por muitos acadêmicos e estudiosos que acreditavam que métodos computacionais poderiam representar uma alternativa sólida às abordagens hermenêuticas tradicionais, não apenas nos estudos literários, mas em todas as disciplinas que lidam com grandes volumes de texto em formato digital (Jockers, 2013; Bonfiglioli, 2015). Todavia, conforme mencionado anteriormente, muitos críticos foram contrários à essa proposta, não enxergando aí resultados que pudessem ser considerados verdadeiramente interessantes, ou ainda, confiáveis (Araújo, 2016; Hammond, 2017).

Desse modo, embora seja apenas um dos inúmeros métodos forjados sob a alcunha das humanidades digitais, a leitura distante tem sido vista como um bom exemplo das promessas, e sobretudo, desconfianças, que emanam desse novo campo. O crítico Adam Kirsch (2014) aponta que os humanistas digitais costumam alardear sobre a grande capacidade de processamento que suas pesquisas passaram a oferecer, porém são eles incapazes de realizar análises inéditas propriamente

ditas. Essa visão pessimista destaca que a adoção indiscriminada da computação, muitas vezes, não pressupõe a compreensão exata acerca do que acontece aos dados (Higuchi, 2021; Ribeiro et al, 2020), e ao usarem as ferramentas, os pesquisadores depositam uma confiança cega nos algoritmos e acabam por não examinar os resultados com o devido cuidado (Dobson, 2015). Do rol de desconfortos, há o que diz respeito à ideia de que os métodos computacionais parecem mover-se na direção de tornarem o trabalho do humanista irrelevante para a produção de conhecimento original ou inédito, porque agora ele poderia ser obtido bastando apenas o emprego de estatística e aprendizado de máquina. Ademais, a abordagem reduziria todos os aspectos dos estudos à busca pela quantificação de evidências presentes no corpus, como se fosse este o objetivo maior da pesquisa (Buonfiglioli, 2015, p. 4).

4. Leitura distante x leitura atenta: abordagens complementares

Respeitando os diferentes viéses suscitados pelo uso da computação no processamento e análise de fontes textuais, é importante, contudo, tecer razões para a conexão íntima que deve existir entre leitura distante e leitura atenta. A experimentação de novas ferramentas e a incorporação do componente tecnológico às investigações históricas e literárias evidencia o quão fundamental é lançar um olhar qualitativo sobre os dados quantitativos, evitando-se cair numa espécie de "fetichismo" dos recursos computacionais que imaginasse que eles funcionariam por si só, revelando verdades ocultas no corpus analisado (Castro et al, 2021). Muito pelo contrário, fica evidente que a presença do especialista é imprescindível para a pesquisa, desde a elaboração das questões iniciais, passando pela tomada de decisões mais técnicas, até a análise intelectual dos resultados.

Em geral, a primeira etapa de um trabalho de processamento de texto deve lidar com os seguintes aspectos: a formalização da tarefa de pesquisa, a adaptação da técnica computacional escolhida e a definição das variáveis em uma representação que os algoritmos possam entender. Outro momento consiste em abstrair conclusões úteis a partir das saídas geradas pelas análises. Embora um método computacional possa capturar relacionamentos adicionais no corpus, ainda será função do *expert* humano identificar os corretos e, em seguida, validá-los e interpretá-los, ou então colocá-los de lado. Nesta etapa, o especialista deve entender se há causalidade por trás das correlações ou decidir ajustar as *features* para executar novamente o modelo em busca de resultados melhores (Bonfiglioli, 2015). Mais uma vez, um forte conhecimento do domínio é claramente importante neste processo.

O esquema da figura acima ilustra a proposta de interação entre leitura distante e leitura atenta para pesquisas em HDs. Podemos ver que cada escala de observação tem seu lugar e momento de realização, mas a complementaridade entre ambas é a chave do sistema.

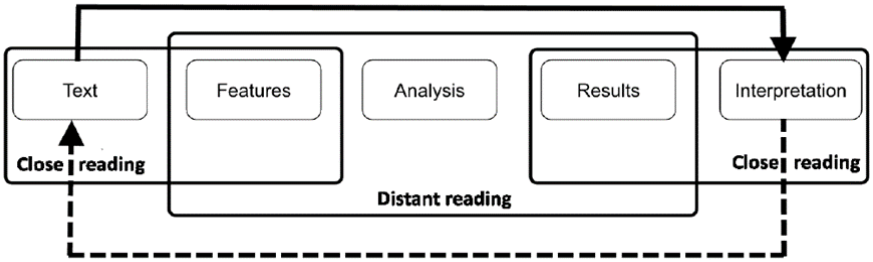

Figure 1: Modelo de interação entre leitura distante e leitura aproximada (extraído de Bonfiglioli, 2015)

Pesquisas recentes no campo do aprendizado de máquina têm se debruçado no chamado *deep learning* ou aprendizado profundo, uma família de algoritmos que, dentre outras coisas, visa aprender automaticamente o que seriam boas *features* para extrair informação útil em sistemas de extração de informação, na tentativa de substituir o trabalho feito manualmente pelo especialista e, por consequência, "tornando as máquinas independentes do conhecimento humano" (Najafabadi, 2015, p. 4). Trabalhos com corpora que adotam técnicas de aprendizado profundo se converteriam então em algo como "leitura distante profunda" (Bonfiglioli, 2015, p. 9). Vale a reflexão sobre o quanto iremos delegar para as máquinas das tarefas consideradas muito especializadas e/ou qualificadas, e o impacto que isso tem nas análises resultantes.

5. As dimensões teóricas e metodológicas sobre o estudo com corpus

Derivar uma análise textual não significa simplesmente lançar palavras e frases em uma engrenagem esperando que resultados prontos saiam do outro lado da máquina. Para além dos recursos informáticos disponíveis, existem questões sensíveis às formações teórico-filosóficas do campo e as investigações próprias sobre linguagem que ajudarão na compreensão do material e na interpretação dos dados analisados. Em suma, as coisas não se resumem a contar frequências ou identificar padrões automaticamente.

Antes, é preciso fixar o conceito de corpus. Um corpus é uma coleção de textos na forma eletrônica, selecionada de acordo com critérios externos para representar, tanto quanto possível, uma língua ou variedade de uma língua como recurso de dados para pesquisa linguística, qualquer que seja o domínio (Sinclair, 2005). Nesse sentido, estudos com corpus fazem uso de uma abordagem empirista e tem como eixo a noção de linguagem enquanto sistema probabilístico, sendo possível evidenciar e quantificar regularidades ou padrões no seu uso (Manning & Schutze,

1999). Vamos ver agora a correlação existente entre os traços linguísticos e os contextos situacionais que terão efeito nas análises produzidas.

5.1 Linguagem e sentido

Para o historiador britânico E. H. Carr (1892-1982), o processo de reconstituição que governa a seleção e interpretação dos fatos são ações derivadas da mente de quem escreve a história, imbuído das suas próprias experiências (Carr, 1978). Por outro lado, a leitura e interpretação destes registros são fenômenos diretamente ligados à capacidade de abstração conceitual – humana ou computacional – que possibilita circunscrever palavras aos seus significados. Grosso modo, significação associa um objeto, um ser, uma coisa, uma noção ou um acontecimento a um signo capaz de evocá-los, e está vinculada, entre outras coisas, às representações que fazemos dos conceitos e das construções nas quais estes participam.

Cara para a Linguística, a questão do signo já foi muito debatida, e ainda o é, por vários autores seminais, como Ferdinand de Saussure e Jacques Lacan. Para o primeiro, o significado é escorregadio: "se um objeto pudesse, onde quer que seja, ser o termo sobre o qual é fixado o signo, a linguística deixaria instantaneamente de ser o que ela é, do topo até a base (Saussure, 2002, p. 197). Para o segundo, mais que isso, a linguagem entendida a partir do signo fica aprisionada nessa relação necessária entre significado e significante, provocando, "a ilusão de que o significante responde à função de representar o significado" (Lacan, 1998, p. 501).

Para além do sentido isolado das expressões linguísticas, Wittgenstein (1979) observa que as funções práticas da linguagem precisam ser levadas em consideração. Segundo o autor, a linguagem não é uma coisa morta em que cada palavra representa algo de uma vez por todas, mas ao contrário, precisa considerar as influências que participam do processo de construção do pensamento socialmente instituído, e, por conseguinte, do significado das coisas. Nesse processo, não haveria como se instituir verdades, apenas construir certezas situacionais (Gracioso e Saldanha, 2010); é somente no contexto de uma sentença que a palavra tem significado (Rorty, 1997).

No âmbito da computabilidade, Almeida e Souza (2011, p. 36) afirmam que "a especificação semântica não corresponde precisamente ao significado dos termos, mas sim ao significado de sentenças de acordo com uma função interpretação sobre expressões sintaticamente bem formadas de um dado modelo de mundo". Assim, "saber o significado de uma sentença equivale a conhecer suas condições-verdade, o que não é o mesmo que saber o seu valor-verdade, ou seja, se o fato é verdadeiro ou não" (Almeida & Souza, 2011, p. 36).

No contorno geral de um sistema de processamento da linguagem natural (PLN), a ambiguidade do sentido da palavra se caracteriza como um importante estímulo desde os anos 1950 (Ide & Véronis, 1998). Soluções são buscadas apoiadas em técnicas que combinam recursos lexicais genéricos, como dicionários, tesauros e redes semânticas como as wordnets (Rademaker et al, 2017), além de classifi-

cadores estatísticos supervisionados ou não (Manning & Schutze, 1999; Nadeau et al, 2006).

Quando se trata do reconhecimento das entidades mencionadas nos textos, tarefa fundamental para análises de redes de atores, por exemplo, Santos (2007) defende uma forma de fazer PLN mais dirigida pelo contexto e menos pelo léxico. As palavras somente podem ser consideradas unidades de sentido quando se encontram em contexto, operacionalizadas nas citações do corpus, ou seja, os sentidos das palavras somente serão definidos em relação a um conjunto de interesses; dessa forma, o conjunto de sentidos definido pelo dicionário pode ou não corresponder ao conjunto que é relevante para uma determinada aplicação de PLN (Manning & Schutze, 1999, p. 230).

Tais incursões favorecem reflexões teóricas e metodológicas importantes e minimizam os desassossegos movidos por um campo que não é regido por leis universais e invariáveis, mas sobretudo interpretativas, e ajudam a lidar de forma mais adequada com as informações disponíveis nestas narrativas. Na sua Introdução à Análise Estrutural da Narrativa, Roland Barthes (1976) afirma que:

> diante da infinidade de narrativas, da multiplicidade de pontos de vista pelos quais se podem abordá-las (histórico, psicológico, sociológico, etnológico, estético, etc.), o analista encontra-se na mesma situação que Saussure, posto diante do heteróclito da linguagem e procurando retirar da anarquia aparente das mensagens um princípio de classificação e um foco de descrição [...] O discurso tem suas unidades, suas regras, sua 'gramática' (Barthes, 1976, p. 20).

Para o autor, todas as unidades possuem um significado dentro da narrativa, e estas unidades se correlacionam de várias formas, dando sentido ao todo. Já Fairclough propõe uma análise do discurso que reúna a análise linguística e a teoria social como método para revelar conexões e causas ocultas nos textos, levando em consideração o contexto ao qual estão ligadas (Fairclough, 2008).

Não obstante todas estas manifestações, a escolha de como um conceito é expresso pode revelar informação sobre as ideologias contidas em uma narrativa ou a relação entre participantes da conversa (Wilson e Thomas, 1997). Uma ilustração disso é o texto do verbete sobre o movimento deflagrado em 31 de março de 1964, incluído no Dicionário Histórico-Biográfico Brasileiro, da Fundação Getulio Vargas (FGV). Nele, o evento é denotado de duas formas canônicas. Defensores e participantes referem-se a ele como "Revolução de 1964", por considerarem que o seu objetivo era produzir uma reformulação completa na vida política do país, eliminando a corrupção e os mecanismos de poder que estariam sendo utilizados para favorecer a subversão comunista no Brasil; seus opositores e adversários, no entanto, definem-no como "Golpe de 1964", por tratar-se da deposição de João Goulart, um presidente que foi eleito legitimamente pelo povo (Abreu et al, 2001). O que há é uma situação em que precisamos ser capazes de identificar tanto os termos que se relacionam um ao outro semanticamente quanto os sentidos das

palavras em determinados contextos, a fim de podermos expandir as conexões (Garside et al, 1997). O ideal é que os sistemas reconheçam que "golpe de 64", “movimento de 64”, "regime militar" e "ditadura" são conceitos conexos, que se relacionam para mapear um mesmo campo semântico, neste caso, um período da história política brasileira.

O conceito de gramaticalidade, tão caro aos estruturalistas, se preocupa fundamentalmente com a boa formação das sentenças a partir de propriedades finitas da língua. Sendo verdade ou não que “todas as gramáticas vazam” (Sapir, 1949), hoje fenômenos linguísticos complexos, como metáforas, colocações, vaguezas e ambiguidades podem ser observados e explicados através de modelos estatísticos que medem a distribuição das palavras e expressões nos contextos em que aparecem. Já dizia Wittgenstein (1979), que o significado de uma palavra é definido pelas circunstâncias de seu uso.

Por fim, é importante destacar o que Helena Martins diz sobre o propósito da linguagem: ela “se manifesta patentemente em sua própria estrutura; caso suas partes (os nomes) não estejam em conformidade com o seu propósito, a linguagem não funciona; para utilizá-la corretamente, precisamos conhecer e respeitar sua arquitetura e seu propósito” (Martins, 2005, p. 459).

5.2 Tradição e prática

Em 1957, Noam Chomsky publicou *Syntactic Structures*, trabalho que veio a se tornar um divisor de águas na linguística do século XX. A partir dele, desenvolve o conceito de uma gramática gerativa, que se distanciava do estruturalismo e do behaviorismo das décadas anteriores, traçando uma distinção fundamental entre o conhecimento que uma pessoa tem das regras de uma língua – ‘competência’ – e o uso efetivo desta língua em situações reais – ‘performance’ (Weedwood, 2002, p. 132; Marcondes, 2009). A linguística, para Chomsky, deveria ocupar-se com o estudo da competência e não do desempenho, em uma clara crítica aos linguistas que buscavam sustentar seus trabalhos baseados no uso de amostras ou corpora. Para ele, tais amostras seriam inadequadas porque representavam apenas uma fração ínfima dos enunciados que é possível dizer numa língua, ou seja, o importante mesmo era focar na descrição das regras que governam a estrutura da competência (Sardinha, 2000; Weedwood, 2002, p. 133).

Em oposição a Chomsky, o linguista de tradição empirista Michael Halliday, acena a partir dos anos 1960, com uma outra abordagem que ficou conhecida como linguística sistêmica (Weedwood, 2002, p. 137), em que a linguagem é vista enquanto sistema probabilístico dependente dos contextos sociais de uso pelos falantes. Esta visão significa dar primazia aos dados provenientes da observação da linguagem, geralmente reunidos sob a forma de corpora. Assim, Sardinha destaca duas considerações importantes da abordagem defendida por Halliday:

> A primeira é a importância primordial de um corpus como fonte de informação, pois ele registra a linguagem natural realmente utilizada por falantes e escritores da língua em situações reais. A segunda é a

> não-trivialidade da investigação da frequência de ocorrência de traços linguísticos de várias ordens (lexicais, sintáticos, semânticos, discursivos etc.), pois é através do conhecimento da frequência atestada que se pode estimar a probabilidade teórica. (Sardinha, 2000).

Dessa forma, teoricamente falando, a utilização de corpus nos estudos linguísticos e demais investigações representa um deslocamento das premissas originais chomskianas, onde o foco passa a ser eminentemente na performance ao invés da competência. O objetivo seria mais o de descrever o uso da linguagem do que identificar universais linguísticos, e o elemento quantitativo (frequência de ocorrências) considerado relevante e, dependendo da abordagem, ser usado para determinar as categorias da descrição da língua (Bonelli, 2010).

Ferramentas estatísticas e recursos sofisticados para exploração de corpora permitem estudos que confirmam ou não hipóteses linguísticas preestabelecidas. Em Santos et al (2015), os autores descrevem o trabalho da Gramateca*2* e afirmam que a intenção do recurso é contribuir com a metodologia científica no campo da linguística, isto é, não só permitir a repetição de uma experiência – que é uma das propriedades exigidas à metodologia científica –, mas também partilhar diferenças de interpretação de um mesmo corpus: "enquanto nas ciências naturais se espera que a mesma experiência leve aos mesmos resultados, nas ciências humanas é não só esperável, mas provável, que haja diferenças na interpretação quando algo é repetido por outros pesquisadores" (Santos, 2015, p. 13).

Descobertas interessantes podem ser encontradas quando o pesquisador se debruça sobre dados reais. Um estudo estatístico sobre estruturas de oração realizado com o corpus Lancaster-Leeds Treebank (compilado a partir de fontes reais) demonstrou que, ao contrário do que os livros de linguística pregavam para o inglês, não era verdade que sentenças do tipo "sujeito – verbo intransitivo" são as construções mais encontradas ao lado de "sujeito – verbo transitivo – objeto". Apesar de exemplos como *the lazy child slept* serem a forma básica daquele primeiro tipo, o que se vê na realidade é que, na falta de um objeto seguindo o verbo, quase sempre há algum constituinte ocupando este espaço, como um elemento adverbial ou outro complemento qualquer (Sampson, 2001, p. 92). É um engano achar que exemplos criados nas gramáticas tradicionais são o espelho da vida real, o que definitivamente não é.

Nos estudos literários, ainda que as narrativas sejam construídas ao nível ficcional, suas estruturas são analisadas não com a intenção de julgamento sobre se determinada construção está sintaticamente correta ou seguem a norma culta, mas para identificar estilos de escrita capazes de se repetirem sob determinada autoria, por exemplo.

Em suma, a pesquisa baseada em corpus compreende dimensões tanto teóricas quanto metodológicas, e diferentes maneiras de entender e estudar a língua se renovam com as ferramentas computacionais disponíveis nos seus diferentes campos de aplicação. Enquanto um texto é para ser lido horizontalmente prestando-se atenção às cláusulas, sentenças e parágrafos, um corpus é varrido

verticalmente, buscando-se por padrões presentes em janelas de contexto. Na linguística forense, são as características e padrões de tipicidade encontradas e demonstradas estatisticamente, que corroboram com a evidência (ou não) de unicidade ou genuinidade de autoria dos textos. Na pragmática, uma área fértil tem sido o uso de corpora para comparar características como vagueza, ironia, humor, hipérbole e metáfora entre diferentes línguas. Na sociolinguística, os contextos do discurso político e debates parlamentares, assim como coberturas de notícias políticas, levam os linguistas a explorar de forma criativa informações lexicais e morfossintática existentes nos corpora para fazer análise de palavras-chave, análise crítica do discurso e comparações, procurando expor as ideologias subjacentes aos textos (McCarthy & O'Keeffe, 2010).

Todos estes exemplos de exploração em cima de fenômenos linguísticos presentes nos textos vêm sendo cada vez mais apropriados por outros campos, permitindo experimentações e *insights*, apoiando-se tanto em técnicas quantitativas quanto qualitativas.

6. Considerações finais

O artigo buscou explorar questões que têm sido endereçadas no horizonte das relações entre as humanidades e o uso de ferramentas computacionais, focando principalmente nos métodos de leitura distante para estudos literários. As investigações nos levam a crer que estas novas ferramentas podem sim suscitar a ampliação da pesquisa acadêmica nas ciências aplicadas, sob diversos aspectos, tanto em termos de renovação de métodos quanto de produção de conhecimento. Mas a disposição para experimentar novos caminhos, especialmente no campo das humanidades digitais, deve vir acompanhada desta compreensão: não se trata de renunciar ou se opor às abordagens tradicionais de investigação, mas de perceber que a complementaridade entre os métodos computacionais e os mais artesanais ou humanos tem potencial tanto para ajudar a responder questões antigas quanto para produzir novas questões.

As áreas das ciências humanas – em especial a literatura e a história –, sempre legaram aos registros textuais grande parte da sua razão de ser e de seu modo de fazer. No escrutínio das leituras e na produção da escrita, o saber é retido, somado e construído com a ajuda inestimável das fontes e dos recursos disponíveis. Aprende-se que o ofício é regido por métodos e técnicas condizentes a cada época, afinal, "cada sociedade se pensa historicamente com os instrumentos que lhe são próprios" (Certeau, 1988, p. 28).

Finalmente, as dificuldades existem e os desafios são muitos, mas as possibilidades abertas por cenários de inovação levam a um contínuo e merecido esforço para tentar superá-los.

Notas

1) A expressão "o grande não lido" (*the great unread*) é tomada emprestada por Moretti da especialista em Literatura Margaret Cohen para designar a vasta produção literária – passada e presente – que não figura na leitura que a grande maioria dos estudiosos realiza (Moretti, 2000, p. 57).

2) Recurso voltado para estudos da língua disponível na Linguateca, em https://www.linguateca.pt/Gramateca.

Submetido: 2021-09-30 | Publicado: 2021-12-31

Referências Bibliográficas

Almeida, M. B., & Souza, R. R. (2011). Avaliação do espectro semântico de instrumentos para organização da informação. Encontros Bibli: Revista eletrônica de biblioteconomia e ciência da informação, 16(31), 25-50. http://dx.doi.org/10.5007/1518-2924.2011v16n31p25

Abreu, A. A. D., Beloch, I., Lattman-Weltman, F., & Lamarão, S. (2001). Dicionário histórico-biográfico brasileiro. CPDOC/Fundação Getúlio Vargas.

Araújo, N. (2016). Vista de longe, a literatura é o que desaparece (Acerca de um fracasso programático em Franco Moretti). In A. Werkema, M.V.N. Soares, & N. Araújo (Eds.), Variações sobre o romance (pp. 259-272). Edições Makunaima.

Archer, J., & Jockers, M. L. (2016). The bestseller code: Anatomy of the blockbuster novel. St. Martin's Press.

Barthes, R. (1976). Introdução à análise estrutural da narrativa. In R. Barthes, T. Todorov, A. J. Greimas, C. Bremond, U. Eco,

J. Gritti., V. Morin, C. Metz, & G. Genette (Eds.), Análise estrutural da narrativa (pp. 19-60) Editora Vozes.

Bode, K. (2014). Reading by numbers: Recalibrating the literary field. Anthem Press. https://doi.org/10.7135/UPO9780857284563

Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe, & M. McCarthy (Eds.), The Routledge handbook of corpus linguistics (pp. 14-28). Routledge.

Bonfiglioli, R., & Nanni, F. (2015). From close to distant and back: How to read with the help of machines. In M. Gadducci, & M. Tavosanis (Eds.), History and philosophy of computing. Third International Conference Hapoc 2015, Pisa, Italy, October 8-11, 2015. Revised Selected Papers (pp. 87-100). Springer.

Carr, E. H. (1978). Que é história? Conferências George Macaulay Trevelyan proferidas por E. H. Carrna Universidade de Cambridge, janeiro-março de 1961. Paz e Terra.

Castro, C., Higuchi, S., & Monnerat, S. (2021). A obra de Gilberto Velho: Uma leitura distante para observar o familiar. CPDOC.

Certeau, M. (1988). A operação histórica. In P. Nora, & J. Le Goff (Eds.), História: Novos problemas. Editora F. Alves.

Dobson, J. E. (2015). Can an algorithm be disturbed? Machine learning, intrinsic criticism, and the Digital Humanities. College Literature: A Journal of Critical Literary Studies, 42(4), 543-564. https://doi.org/10.17613/M6QW2C

Fairclough, N. (2008). Discurso e mudança social. Editora UnB.

Freitas, C. (2015). Corpus, Linguística Computacional e as Humanidades Digitais. In M. Leite, & C. T. Gabriel (Eds.), Linguagem, Discurso, Pesquisa e Educação (pp 18-46). DP Et Alii Editora.

Garside, R., Leech, G. N., & Mcenery, A. M. (1997). Corpus annotation: Linguistic information from computer text corpora. Taylor & Francis.

Gracioso, L., & Saldanha, G. S. (2010). Ciência da Informação e Filosofia da Linguagem: Da pragmática informacional à web pragmática. Junqueira & Marin Editores.

Hammond, A. (2017). The double bind of validation: Distant reading and the digital humanities"trough of disillusionment". Literature Compass, 14(8). https://doi.org/10.1111/lic3.12402

Higuchi, S. (2021). Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico do Brasil [Tese de doutoramento, Pontifícia Universidade Católica do Rio de Janeiro]. ETDs @PUC-Rio. https://doi.org/10.17771/PUCRio.acad.54623

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. Computational linguistics, 24(1), 1-40. https://aclanthology.org/J98-1001

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.Kirsch, A. (2014, 2 de maio). Technology is taking over English Departments: The false promise of the Digital Humanities. The New Republic. https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch

Lacan, J. (1998). Escritos. Zahar.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT Press.

Marcondes, D. (2009). Textos básicos de linguagem: De Platão a Foucault. Zahar.

Martins, H. (2005). Três caminhos na filosofia da linguagem. In F. Mussalin, & A. Bentes (Eds.), Introdução a linguística – Fundamentos Epistemológicos (Vol. 3). Cortez Editora.

McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe, & M. McCarthy (Eds.), The Routledge handbook of corpus linguistics (pp. 3-13). Routledge

Moretti, F. (2013). Distant reading. Verso Books.

Moretti, F. (2000). Conjectures on world literature. New Left Review, 1. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21. https://doi.org/10.1186/s40537-014-0007-7

Rademaker, A., Chalub, F., & Freitas, C. (2017). Two Corpus Based Experiments with the Portuguese and English Wordnets. In J. P. McCrae et al. (Eds.) Proceedings of the {LDK} 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries {\&} Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge {(LDK} 2017), Galway, Ireland, June 18, 2017 (134-145). CEUR-WS.org. http://ceur-ws.org/Vol-1899/CfWNs/_2017/_proc2-paper/_4.pdf

Ribeiro, C. J. S., Higuchi, S., & Ferla, L. A. C. (2020). Aproximações ao cenário das humanidades digitais no Brasil. Digital Humanities Quarterly, 14(2). http://www.digitalhumanities.org/dhq/vol/14/2/000453/000453.html

Sampson, G. (2001). Empirical Linguistics. Continuum.

Santos, D. (2007). Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. IST Press.

Santos, D. (2014). Podemos contar com as contas? In S. M. Aluísio, & S. E. O. Tagnin (Eds.), New Language Technologies and Linguistic Research: A Two-Way Road (pp. 194-213). Cambridge Scholars Publishing.

Santos, D. (2019). Literature studies in Literateca: Between digital humanities and corpus linguistics. In M. Doerr, Ø. Eide, O. Grønvik, & B. Kjelsvik (Eds.), Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore. Novus Forlag.

Santos, D., Alves, D., Amaro, R., Branco, I. A., Fialho, O., Freitas, C., ... & Terra, P. (2020). Leitura distante em português: Resumo do Primeiro Encontro MatLit (Centro de Literatura Portuguesa da Universidade de Coimbra), 8(1), 279-298.

Santos, D., Marques, R., Freitas, C., Simões, A., & Mota, C. (2015). Comparando anotações linguísticas na Gramateca: Filosofia, ferramentas e exemplos. Domínios de Lingu@ gem, 9(2), 11-26.

Sapir, E. (1949). Language, an introduction to the study of speech. Harcourt.

Sardinha, T. B. (2000). Lingüística de Corpus: Histórico e problemática. DELTA, 16(2), 323-367. https://doi.org/10.1590/S0102-44502000000200005

Saussure, F. de (2002). Curso de linguística geral. Organizado por Charles Bally e Albert Sechehaye. Prefácio de Isaac Nicolau Salum. (24.ª ed.). Cultrix.

Schnapp, J., Presner, T., & Lunenfeld, P. (2009). Digital humanities manifesto 2.0. https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Sinclair, J. (2005). "Corpus and text – Basic principles". In M. Wynne (Ed.), Developing Linguistic Corpora: a Guide to Good Practice (pp. 1-16). Oxbow Books. http://ota.ox.ac.uk/documents/creating/dlc/

Weedwood, B. (2002). História concisa da linguística. Parabola Editora.

Wittgenstein, L. (1979). Investigações Filosóficas (J. C. Bruni, Trad., 2.ª ed.). Abril Cultural (Os Pensadores).

Wilson, A., & Thomas, J. (1997). "Semantic Annotation". In R.G. Garside, G. Leech, & A. M. Mcenery (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman.Pensadores).Wilson, A., & Thomas, J. (1997). "Semantic Annotation". In R.G. Garside, G. Leech, & A. M. Mcenery (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman.