DOSSIÊ

Comparação de megadados das duas revistas da Anpuh ou Introdução à ciência aberta para historiadores

Comparison of Big Data from the two Anpuh Journals or Introduction to Open Science for Historians

Oldimar Pontes Cardoso*
Marco Aurélio da Costa**
Gustavo Ítalo Freire Martins***
Waleska Maria Lopes Farias+
João Carlos de Melo Silva++

RESUMO

Este artigo analisa o *corpus* completo da *Revista História Hoje* (2012-2020) e o *corpus* da *Revista Brasileira de História* no mesmo período, utilizando-se do método de cotejamento diacrônico de fontes escritas por redes semântico-temporais e de um algoritmo desenvolvido pelos dois primeiros autores. Suas principais conclusões são a ascensão do uso das palavras "aula" e "consciência" na RHHJ e das palavras "nacional" e "América" na RBH e a queda do uso das palavras "novo" e "antigo" na RHHJ e das palavras "trabalho", "processo" e "governo" na RBH. Podemos destacar ainda a ascensão repentina das palavras "direi-

ABSTRACT

This paper analyzes the full *corpus* of *Revista História Hoje* (2012-2020) and the *corpus* of *Revista Brasileira de História* in the same period using the method of diachronic confrontation of written sources by semantic-temporal networks and an algorithm developed by the first two authors. Its main conclusions are the rise of the use of the words "classroom" and "consciousness" in RHHJ and the words "national" and "America" in RBH and the fall of the use of the words "new" and "ancient" in RHHJ and the words "work", "process" and "government" in RBH. We can also highlight the sudden rise of the words

* Netbot Humanidades Digitais Ltda, São Paulo, SP, Brasil. oldimar@gmail.com
** MCBC Informática Ltda, São José dos Campos, SP, Brasil. costa.marco@gmail.com
*** Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brasil. gustavo.italo.freire@gmail.com
+ Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brasil. waleskamlfarias@gmail.com
+ + Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brasil. joaocms.prof@gmail.com

to" e "humano" na RHHJ e das palavras "antigo" e "antiguidade" na RBH. O artigo funciona como uma introdução à ciência aberta para historiadores porque descreve em detalhes os protocolos de ciência aberta utilizados nesta pesquisa histórica, demonstra como outros pesquisadores poderiam utilizar seus dados, métodos, fluxos de trabalho e protocolos para replicá-la ou realizar pesquisas distintas desta, e ainda indica como tudo isso viabiliza reprodutibilidade científica nas humanidades.
Palavras-chave: Anpuh; ciência aberta; reprodutibilidade científica.

"rights" and "human" in RHHJ and the words "ancient" and "antiquity" in RBH. The paper works as an introduction to open science for historians because it describes in detail the open science protocols used in this historical research, it demonstrates how other researchers could use its data, methods, workflows and protocols to replicate it or to conduct other research than this. It also indicates how all of this enables scientific reproducibility in the humanities.
Keywords: Anpuh; open science; scientific reproducibility.

POR QUE PRECISAMOS DE UMA HISTÓRIA ABERTA?

Admitimos que a ciência aberta não tem a mais digna das origens. Segundo a *Organisation for Economic Co-operation and Development* (OECD), ciência aberta é o ato de tornar "os resultados de pesquisa com financiamento público (publicações e dados de pesquisa) acessíveis ao público em formato digital" (OECD, 2015, p. 7). Esta é uma visão tecnocrática de ciência aberta, pautada apenas na otimização do financiamento público, na ideia de reduzir custos com o acesso a publicações científicas e com a reutilização de dados de pesquisa. É melhor não contar aos tecnocratas que ciência é sempre um investimento público a fundo perdido, que significa obrigatoriamente investir em milhares de pesquisas que não vão nos levar a lugar nenhum em troca de algumas dezenas de pesquisas que nos darão resultados significativos. É impossível prever quais pesquisas darão esses resultados significativos, mas, apesar disso, o investimento generalizado em ciência sempre compensou. Somente economizar com acesso a publicações e reutilização de dados não justifica a existência da ciência aberta. Mas apesar de ela significar muito mais do que a OECD consegue enxergar, é importante ter em mente que a ciência aberta só adquiriu

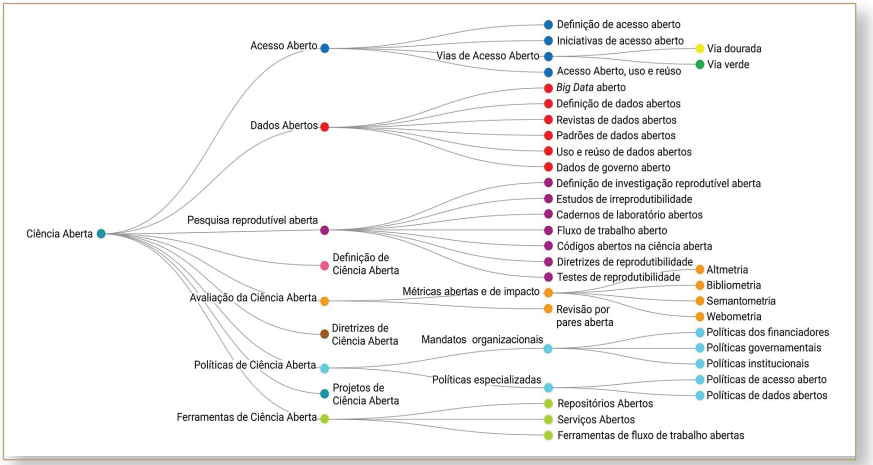
prestígio e acesso privilegiado a financiamento por causa desta visão tecnocrática e não das características mais importantes que descrevemos a seguir.

De acordo com Pontika (2015), para além do acesso aberto e dos dados abertos que bastam à OECD, a ciência aberta também implica pesquisa reprodutível aberta (que é, em nossa opinião, a principal característica), definição de ciência aberta (a discussão metacientífica propriamente dita), avaliação de ciência aberta, diretrizes de ciência aberta, políticas de ciência aberta, projetos de ciência aberta e ferramentas de ciência aberta, como vemos na Imagem 1.

Definimos a seguir como a pesquisa que resulta neste artigo tem relação com o que Pontika (2015) chama de "pesquisa reprodutível aberta" e indicamos a leitura do artigo citado e de Koschtial (2021), Christensen (2019) e Bezjak (2018) para a compreensão do restante da taxonomia da ciência aberta descrita na Imagem 1 e sua relação com as humanidades.

No caso específico da pesquisa histórica, a ciência aberta pode contribuir para combater três tradições que ainda existem no campo e contribuem para reduzir a qualidade de suas pesquisas: acesso privilegiado ou exclusivo a fontes, autoridade moral e ensaísmo. Como se trata de atividade ilegal, não é possível definir com precisão o impacto que o acesso privilegiado ou exclusivo a fontes (via uso de poder político ou suborno de funcionários de baixo escalão) tem sobre a pesquisa histórica atual. Mas é fato que a digitalização cada vez maior das fontes disponíveis aos historiadores reduz sensivelmente este problema. E estabelecer o acesso aberto a todas as fontes o eliminará definitivamente. A ausência de fontes abertas também estimula o recurso à autoridade moral, tirando das fontes o foco da argumentação e projetando-o no poder político do historiador que as utilizou. Abrir as fontes inibe o recurso à autoridade moral e devolve o foco da argumentação às fontes, onde ele sempre deveria estar. A distância ou ocultação das fontes também sempre funcionou como um convite ao ensaísmo e a criação de conceitos vagos e generalizantes que mais atrapalham do que ajudam a análise histórica, o que perde o sentido quanto mais as fontes e o diálogo transparente com elas na comunicação da pesquisa adquirem importância nos protocolos da ciência aberta.

Imagem 1 – Taxonomia da Ciência Aberta.



Fonte: Taxonomia da Ciência Aberta. https://figshare.com/articles/figure/Taxonomiada-CinciaAberta/12124002. Acesso em: 02 jul. 2021. Traduzido por Nivaldo Calixto Ribeiro, Lúcia da Silveira, Sarah Rúbia de Oliveira Santos.

Podemos afirmar que esta pesquisa histórica é reprodutível aberta porque obedece aos seguintes protocolos:

1. Todos os seus dados são abertos e disponíveis em um projeto específico hospedado na plataforma Open Science Framework: https://osf.io/bepqd/;
2. Não apenas os dados originais, mas também os dados tratados, que facilitam futuras pesquisas e viabilizam a reprodutibilidade desta pesquisa, também são abertos;
3. Os seus metadados estão descritos no arquivo _metadata.pdf, na raiz da pasta de dados abertos;
4. Os seus métodos são abertos e descritos no próprio artigo de forma a serem reprodutíveis;
5. Os seus fluxos de trabalho são abertos e descritos no fluxograma da Imagem 2 e no arquivo _workflow.pdf, na raiz da pasta de dados abertos;
6. Todos os seus passos sujeitos a viés no fluxo de trabalho estão explicitados no documento _workflow.pdf;
7. Os seus fluxos de trabalho são automatizados, de forma a viabilizar sua

futura transposição por qualquer outro pesquisador para um *Workflow Management System* (DEELMAN, 2017);

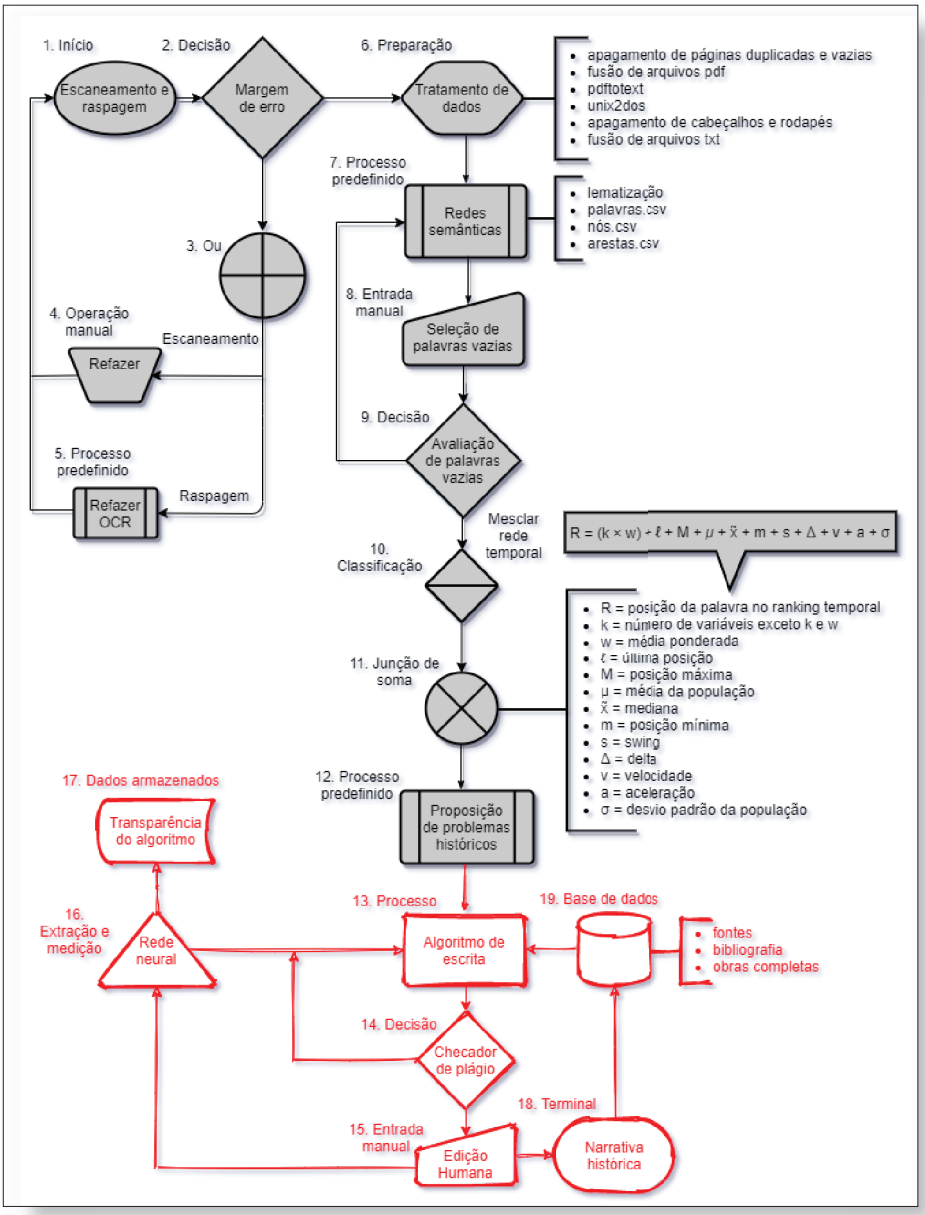
8. Todos os códigos utilizados em seus fluxos de trabalho são abertos e se encontram na pasta "code", na raiz da pasta de dados abertos;
9. Todos os dados, códigos e fluxos de trabalho estão vinculados à Licença CC BY Atribuição 4.0 Internacional (CC BY 4.0) – https://creativecommons.org/licenses/by/4.0/deed.pt_BR;
10. Os resultados desta pesquisa encontram-se nesta publicação de acesso aberto.

DADOS, MÉTODOS, FLUXOS DE TRABALHO E PROTOCOLOS

O que apresentamos aqui é resultado da comparação digital do *corpus* completo da *Revista História Hoje* (2012-2020) e do *corpus* da *Revista Brasileira de História* no mesmo período sob o método de análise de fontes textuais com base em redes semânticas estabelecido por Silva (2016), que fundamentou o método de cotejamento diacrônico de fontes escritas por redes semântico-temporais descrito por Cardoso (2019). Esses métodos e os fluxos de trabalho deles derivados estão organizados em um único algoritmo em desenvolvimento, que é apresentado na Imagem 2 e na Tabela 1. Esta tabela se limita a um relato do algoritmo até a caixa 12 do fluxograma, que foi a parte utilizada nesta pesquisa.

O segundo método citado também é fundamentado na lei de Zipf (ZIPF, 1949; MORENO-SÁNCHEZ, 2016; WILLIAMS, 2016; KANWAL, 2017), que determina que a frequência de qualquer palavra em uma lista ordenada é inversamente proporcional à sua classificação na tabela de frequência. Esta frequência é dada por $f(n) = K/n'$, onde K é uma constante. Uma palavra é menos relevante em um *corpus*, quanto mais avançada é sua posição no *ranking*. A maioria das palavras tem uma frequência muito baixa e desempenha um papel irrelevante. A diminuição da relevância de cada palavra na lista ordenada é frequentemente logarítmica, então as primeiras palavras mais usadas de um *corpus* são sempre relevantes e por vezes suficientes para estabelecer sua síntese. A contagem de palavras não é apenas relevante para o primeiro tratamento das fontes na pesquisa histórica, mas uma ferramenta essencial para deslegitimar a especulação histórica sem qualquer fundamento imposta ao campo por mero poder político.

Imagem 2 – Fluxograma metahistórico.



Fonte: Elaboração dos autores.

O segundo método citado também se utiliza da equação $R = (k \times w) + \ell + M + \mu + \tilde{x} + m + s + \Delta + v + a + \sigma$, criada pelos dois primeiros autores, para reunir redes semânticas sincrônicas numa rede semântico-temporal que nos permite problematizar a fenomenologia da flutuação de cada palavra ao longo do tempo na lista de palavras mais usadas. O algoritmo utilizado nesta pesquisa foi programado com base nessa equação para analisar cinco fenômenos semântico-temporais distintos: ascensão, queda, estabilidade, ascensão repentina e aparecimento repentino. O significado estatístico de cada um desses fenômenos encontra-se detalhado na segunda linha da Tabela 4 e da Tabela 6.

Tabela 1 – Descrição das funções das 12 primeiras caixas do fluxograma.

Caixa	Função
1	Coleta de fontes digitalizando livros impressos com alguma ajuda humana ou copiando dados automaticamente da internet. A maneira mais fácil, barata e eficaz de digitalizar um livro, com melhores resultados em reconhecimento ótico de caracteres (OCR), é cortar suas lombadas e digitalizá-lo em um arquivo em formato pdf como folhas avulsas. O primeiro problema com a obtenção de fontes digitais é retirá-las da internet. Como muitas plataformas acadêmicas têm proteção contra bots, o que é estranho e sintomático, a coleta de dados requer o uso de algumas interfaces de programação de aplicativos (APIs) para contornar essas proteções nas plataformas onde as fontes são armazenadas. Depois de contornar a proteção, o segundo problema é lidar com arquivos digitais com péssimo reconhecimento ótico de caracteres (OCR) porque eles foram feitos há muito tempo, quando essa tecnologia era incipiente. Portanto, precisamos excluir os OCRs e gerá-los novamente com tecnologia melhor.
2	Cálculo da margem de erro do reconhecimento ótico de caracteres (OCR) usado para digitalizar as fontes. Usamos uma ferramenta de verificação ortográfica para saber a margem de erro da digitalização das fontes contando quantas palavras são detectadas como erradas pela ferramenta e comparando essa quantidade com o número de palavras em todo o texto. Se a margem de erro for inferior a 1%, os dados são enviados para a caixa 6 do fluxograma, "Preparação: Tratamento de dados". Se a margem de erro for superior a 1%, os dados são enviados para a caixa 3 do fluxograma, "Ou".
3	A função "Ou" separa as fontes escaneadas, enviadas para a caixa 4 do fluxograma, "Operação manual: Refazer", das fontes raspadas, enviadas para a caixa 5 do fluxograma, "Processo predefinido: Refazer OCR".
4	Operação manual para refazer a digitalização das fontes impressas com margem de erro superior a 1%. Não é possível fazer nada automaticamente se a digitalização de uma fonte impressa foi mal feita, e esta é a única operação manual neste fluxograma que não pode ser substituída por uma automática.

5	O processo predefinido para refazer o reconhecimento ótico de caracteres (OCR) se uma fonte de arquivo em formato pdf apresenta uma margem de erro maior que 1%. Nesse caso, podemos corrigir o problema automaticamente excluindo o OCR antigo e criando um novo.
6	Preparação das fontes através do tratamento de dados, excluindo páginas duplicadas e páginas inúteis (como anúncios e índices), mesclando arquivos em formato pdf (para combinar muitos artigos em apenas um arquivo de edição completo), convertendo esses arquivos em formato txt, excluindo cabeçalhos e rodapés de cada página (a repetição das mesmas palavras em muitas páginas pode distorcer a contagem final de palavras e todos os resultados) e, finalmente, mesclar arquivos em formato txt para criar os *corpora* de cada período selecionado.
7	Processo predefinido para criar uma rede semântica para cada corpus. Esse processo começa com a lematização dos *corpora*, que envolve a diferenciação de substantivos e verbos escritos com as mesmas palavras e a divisão de palavras compostas. Em seguida, as palavras de cada corpus são contadas e classificadas no arquivo words.csv, os nós formados por essas palavras são identificados por um número de identidade no arquivo nodes.csv, e esses números de identidade são usados para estabelecer as arestas entre as palavras.
8	Entrada manual das palavras vazias (*stopwords*), que são as palavras não relevantes para a pesquisa (como o, de, e, ser, para etc.) e eventualmente também as palavras usadas no título das fontes ou do campo de pesquisa. O algoritmo pode ser carregado com uma lista genérica de palavras vazias e pular esta etapa manual, filtrando automaticamente a classificação das palavras mais usadas com sua lista padrão de palavras vazias, mas a entrada manual fornece melhor qualidade até que uma inteligência artificial específica seja desenvolvida para definir que palavras vazias estão em cada corpus.
9	Avaliação se ainda existem palavras vazias entre as palavras do arquivo words.csv. Em caso afirmativo, a lista de palavras é enviada de volta em ciclo para a caixa 7 do fluxograma, "Processo predefinido: Redes semânticas"; caso contrário, ele é enviado para a caixa 10 do fluxograma, "Classificação: Mesclar rede temporal".
10	Mescla das redes semânticas de cada período em uma rede semântico-temporal.
11	Utilização da equação temporalizadora $R = (k \times w) + \ell + M + \mu + \tilde{x} + m + s + \Delta + v + a + \sigma$, descrita na caixa 11 do fluxograma, "Junção de soma", para avaliar o fenômeno de flutuação de cada palavra ao longo do tempo na lista de palavras mais usadas.
12	Proposição de problemas históricos com base em inteligência natural fundamentada na composição de variáveis estatísticas pela equação temporalizadora.

Fonte: Elaboração dos autores.

Uma descrição detalhada dos fluxos de trabalho desta pesquisa pode ser vista no arquivo _workflow.pdf, localizado na raiz da pasta de arquivos do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/). Uma tradução deste arquivo encontra-se abaixo na Tabela 2.

Tabela 2 – Fluxo de trabalho desta pesquisa.

Fluxo de trabalho (versão 01 - 30/05/2021)		
Todo este fluxo de trabalho executa arquivos Python em terminal Linux ou Google Apps Scripts em arquivos Google Planilhas. Todas as etapas sujeitas a viés neste fluxo de trabalho estão em fonte vermelha.		
1	Baixar todos os arquivos pdf (01_pdf)	HH (2012-2020): https://anpuh.org.br/index.php/revistas-anpuh/revista-historia-hoje-i RBH (2012-2020): https://www.scielo.br/scielo.php?script=sci_issues&pid=0102-0188&lng=pt&nrm=iso
2	Converter todos os arquivos pdf para txt (01_pdf)	find -name '*.pdf' -exec pdftotext {} \;
3	Mover todos os arquivos txt para uma nova pasta (01_pdf a 02_txt)	mv *.txt ~/02_txt/2012 mv *.txt ~/02_txt/2013 ...
4	Mesclar todos os arquivos txt em corpora anuais (03_annual_corpora)	for d in $(seq 2012 2020) ; do cat $d/* > $d.txt ; done
5	Mesclar todos os corpora anuais em um corpus completo (04_full_corpus)	cat *.txt > fullcorpus.txt
6	Criar um arquivo words.csv, selecionar palavras vazias manualmente e salvar um novo arquivo stopwords.csv (05_stopwords)	~/filter.py fullcorpus.txt fullcorpus
7	Criar nós, arestas e palavras, arquivos filtrados com palavras irrelevantes (06_words)	for year in $(seq 2012 2020) ; do ~/filter.py $year.txt $year -f stopwords.csv ; done

8	Estabelecer fenômenos semântico-temporais (07_phenomenology)	for linhas in 20 30 40 50 60 70 80 90 100 200 250 300 400 500; do ~/merge_words.py --headers $(seq 2012 2020) --files 20*_words.csv --lines $linhas > phenomenology_${linhas}.csv ; ~/phenomenology.py phenomenology_${linhas}.csv > ${linhas}.txt; done
9	Converter todos os arquivos txt para Google Docs	TXT to GOOGLEDOCS converter V4.gs
10	Tratamento macroscópico (08_specific_words)	BATCH FILE SCANNER V5.gs

Fonte: Elaboração dos autores.

TRATAMENTO DOS DADOS

A Tabela 3 a seguir apresenta as palavras mais usadas na *Revista História Hoje* ao longo do tempo, desde seu surgimento e excetuando-se as palavras vazias, seguidas do número de vezes em que a palavra apareceu naquele *corpus* específico. A tabela destaca a célula de cada palavra em uma cor diferente, para facilitar a visualização dos seus fenômenos semântico-temporais. Apenas as primeiras 50 palavras do *corpus* completo são identificadas com uma cor de célula que se repete na mesma palavra quando ela aparece nas demais colunas. Palavras em branco na Tabela 3 indicam as que não estão entre as 50 primeiras no *corpus* completo. Por uma questão de espaço, apenas as primeiras 19 palavras do *corpus* completo e de cada *corpora* anual são mostradas na Tabela 3. Porém, a planilha completa com todas as 8442 palavras do *corpus* completo e todas as palavras dos *corpora* anuais pode ser vista no arquivo dataset/HH/07_phenomenology/final_words.xlsx do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/).

Tabela 3 – As 19 palavras mais usadas no *corpus* completo da RHHJ e nos *corpora* anuais, seguidas do número de vezes em que cada uma delas aparece em cada *corpus*.

Corpus completo	2012	2013	2014	2015	2016	2017	2018	2019	2020
história, 23377	história, 2044	história, 2995	história, 2081	história, 2880	história, 2501	história, 2622	história, 2039	história, 3298	história, 2917
ensino, 8816	indígena, 1347	professor, 1486	professor, 918	ensino, 1175	histórico, 1146	histórico, 1012	livro, 773	histórico, 1016	histórico, 1244
histórico, 7544	escola, 711	ensino, 1197	ensino, 846	professor, 1127	ensino, 1123	ensino, 906	ensino, 756	ensino, 1004	ensino, 1135
professor, 7423	brasil, 695	histórico, 712	poder, 746	histórico, 895	professor, 895	poder, 750	educação, 686	público, 913	professor, 801
poder, 6199	poder, 674	poder, 694	curso, 724	educação, 722	educação, 815	aluno, 735	didático, 568	poder, 772	educação, 733
educação, 5801	ensino, 674	tempo, 653	educação, 657	curso, 675	aluno, 716	música, 611	poder, 564	professor, 663	poder, 704
aluno, 4548	brasileiro, 645	educação, 648	histórico, 638	escola, 598	poder, 711	professor, 601	professor, 537	educação, 636	direito, 643
escola, 4125	educação, 596	livro, 628	hoje, 503	poder, 584	didático, 505	aula, 555	direito, 489	social, 559	humano, 633
hoje, 4108	negro, 491	aluno, 614	memória, 488	aluno, 571	trabalho, 486	brasil, 508	histórico, 477	escola, 511	aprendizagem, 590
brasil, 3949	índio, 484	didático, 603	aluno, 487	escolar, 562	docente, 479	tempo, 471	indígena, 469	brasil, 505	brasil, 548
social, 3801	hoje, 472	social, 535	museu, 445	tempo, 437	hoje, 454	hoje, 445	escola, 466	passado, 466	aula, 512
tempo, 3637	cultura, 452	escola, 525	público, 401	hoje, 427	social, 431	brasileiro, 442	social, 446	hoje, 457	narrativa, 489
didático, 3474	histórico, 404	hoje, 480	ano, 376	social, 402	livro, 425	livro, 394	político, 435	ano, 412	tempo, 465
livro, 3448	povo, 400	curso, 473	escola, 375	ano, 390	pesquisa, 389	novo, 386	brasil, 429	tempo, 403	aluno, 454
aula, 3356	professor, 395	escolar, 452	pesquisa, 371	avaliação, 385	aula, 389	público, 364	hoje, 427	memória, 384	hoje, 443
ano, 3223	novo, 389	ano, 439	social, 352	pesquisa, 385	cultura, 352	passado, 352	humano, 405	pesquisa, 348	escolar, 433
escolar, 3057	tempo, 385	presente, 415	didático, 347	aula, 385	escola, 349	trabalho, 346	historiar, 384	aluno, 342	brasileiro, 426
público, 3033	cultural, 349	trabalho, 409	trabalho, 336	docente, 376	consciência, 349	social, 340	aluno, 370	historiador, 339	presente, 415
trabalho, 2981	africano, 342	disciplina, 406	brasil, 334	disciplina, 355	presente, 313	presente, 339	aula, 365	político, 331	pesquisa, 402

Fonte: Elaboração dos autores.

O fato de a Tabela 3 ter a palavra "história" em sua primeira linha é irrelevante, já que esta palavra é o nome do campo e usada no título da revista que compõe o *corpus*. Por essas duas razões, nós poderíamos ter optado por incluir a palavra "história" entre as palavras vazias. Não o fizemos por razões didáticas, já que este artigo tem também a função de servir como uma introdução à ciên-

cia aberta para historiadores. Esta escolha explicita que a definição da lista de palavras vazias é um dos passos sujeitos a viés no fluxo de trabalho desta pesquisa, por isso marcado em cor de fonte vermelha no arquivo _workflow.pdf que se encontra na raiz da pasta de arquivos do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/). Nesse mesmo endereço, as palavras vazias podem ser encontradas nos arquivos dataset/HH/05_stopwords/HH_stopwords.csv e dataset/RBH/05_stopwords/RBH_stopwords.csv. Elas foram criadas pelo apagamento das palavras relevantes dos arquivos dataset/HH/04_full_corpus/HH_fullcorpus_words.csv e dataset/RBH/04_full_corpus/RBH_fullcorpus_words.csv. Este apagamento é uma escolha enviesada que impacta o resultado da pesquisa, mas isso não é visto na ciência aberta como um problema. Não há na ciência aberta uma crença na objetividade absoluta da ciência que caracteriza o cientismo. Ciência aberta não é uma ciência sem viés. Toda ciência é enviesada, inclusive as ciências duras (IOANNIDIS, 2005). Em vez de se esforçar para esconder seus vieses, como faz a ciência normal (KUHN, 1996, p. 10; ANDERSEN, 2006, p. 5, 69-72; MARCUM, 2012, p. 42), a ciência aberta transforma em protocolo a explicitação de todos os seus vieses. Uma pesquisa aberta não é reprodutível por ser objetiva, mas por deixar transparentes todas as escolhas subjetivas feitas ao longo do fluxo de trabalho.

A segunda palavra mais usada no *corpus* completo da *Revista História Hoje*, apresentado na Tabela 3, é a palavra "ensino", que flutua entre a segunda, a terceira e a sexta posições ao longo do tempo. Esse fenômeno é compreensível pelo fato de a revista ter se especializado no ensino de História. A terceira palavra mais usada no *corpus* completo da *Revista História Hoje* é "histórico", que eventualmente poderia ser considerada uma palavra vazia. As palavras seguintes são "professor" (que flutua entre a 2ª e a 15ª posição), "poder" (4ª à 8ª posição), educação (4ª à 24ª posição), "aluno" (5ª à 33ª posição) e "escola" (3ª à 27ª). Também é interessante observar por que as palavras "negro", "índio", "povo", "africano", "museu", "avaliação", "consciência", "música", "historiar" e "historiador" aparecem em branco entre as primeiras palavras, ou seja, aparecem de forma significativa apenas em um dos *corpora* anuais.

A observação da flutuação temporal das diversas palavras na Tabela 3 permite muitas análises a olho nu que extrapolam o caráter introdutório deste artigo. Nosso objetivo não é realizar uma análise, mas fornecer dados que possibilitem uma infinidade de análises aos leitores da *Revista História Hoje*.

Cada olho nu poderá enxergar distintos fenômenos na Tabela 3, dependendo de sua formação, área de atuação e interesses específicos. Também é possível aprofundar a análise a olho nu observando a planilha completa com todas as 8442 palavras do corpus completo no arquivo dataset/HH/07_phenomenology/final_words.xlsx do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/).

Para além de tudo que é possível analisar a olho nu na Tabela 3, desenvolvemos uma inteligência natural que analisa os dados dessa tabela e identifica fenômenos semântico-temporais em trechos selecionados das listas de palavras mais usadas, cujos resultados são apresentados a seguir na Tabela 4. Essa inteligência natural se utiliza das variáveis estatísticas identificadas na segunda linha da Tabela 4 para classificar cada palavra diante dos fenômenos de **ascensão** (a palavra passa a ser cada vez mais usada nos *corpora* de cada período), **queda** (a palavra passa a ser cada vez menos usada nos *corpora* de cada período), **estabilidade** (a palavra é usada em quantidades próximas nos *corpora* de todos os períodos), **ascensão repentina** (a palavra já estava entre as mais usadas nos primeiros períodos e passa a ser ainda mais usada nos últimos períodos) e **aparecimento repentino** (a palavra sequer estava entre as mais usadas nos primeiros períodos, mas aparece entre elas nos últimos períodos). A primeira coluna da Tabela 4 indica quantas palavras mais usadas foram consideradas para a definição dos fenômenos descritos nas cinco colunas seguintes. As células em branco na Tabela 4 indicam que o fenômeno descrito naquela coluna não se apresenta naquela linha, ou seja, o fenômeno não se apresenta considerando a quantidade de palavras mais usadas definida naquela linha.

Por uma questão de espaço, não nos aprofundaremos em todas as palavras que aparecem na Tabela 4, mas apenas destacaremos dentre os fenômenos que podem ser observados nela: a **ascensão** das palavras "aula" (por se repetir em cinco linhas da tabela) e "consciência" (por se repetir em duas colunas da tabela), a **queda** das palavras "novo" e "antigo" (pela repetição da primeira em duas linhas e pelo fato de a segunda ser antônimo da primeira e de ambas aparecerem em extremos opostos da tabela), a **ascensão repentina** das palavras "direito" e "humano" (por se repetirem em várias linhas da tabela) e o **aparecimento repentino** da palavra "gênero".

Para observar mais de perto essas palavras selecionadas na Tabela 4, utilizamos dois códigos em Google Apps Script disponíveis na pasta "code" do

projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/). O código "TXT to GOOGLEDOCS converter V4.gs" foi utilizado para converter os corpora da RHHJ (https://drive.google.com/drive/u/0/folders/1psIC_IwfYyO2l6boCBWMNKoWT4pzvWzW) e da RBH (https://drive.google.com/drive/u/0/folders/1pe4HrN30kxPXUgx30EkYPDUbHQ-I8bJP) do formato txt para o formato Google Docs. Em seguida, utilizamos o código "BATCH FILE SCANNER V5.gs" para criar planilhas que listam todos os parágrafos em que uma palavra específica aparece, criando uma aba para cada *corpus* anual e disponibilizando ao lado de cada parágrafo o nome arquivo em que ele se encontra e um *link* direto para este arquivo. O uso desses dois códigos permite transitar da observação microscópica da rede semântico-temporal para uma observação macroscópica ou a olho nu dos mesmos fenômenos. As planilhas criadas pelo código "BATCH FILE SCANNER V5.gs" permitem observar rapidamente os diversos contextos em que uma palavra aparece em cada *corpus* e elaborar hipóteses mais precisas sobre os fenômenos nos quais ela está inserida. As planilhas criadas por este último código para cada uma das palavras listadas no parágrafo anterior (aula, consciência, novo, antigo, direito, humano e gênero) encontram-se disponíveis em formato xlsx na pasta "dataset/HH/08_specific_words" do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/).

Dentre as palavras em **ascensão**, "aula" passou de 83 menções em 2012 para 230 menções em 2020 e "consciência" passou de 17 menções em 2012 para 100 menções em 2020, o que pode ser conferido nos arquivos xlsx citados no parágrafo anterior. As hipóteses para esses fenômenos são o crescimento da preocupação da área de ensino de História com o lugar da sala de aula e a recepção no Brasil do conceito alemão de *Geschichtsbewusstsein*. Dentre as palavras em **queda**, "novo" passou de 84 menções em 2012 para 37 menções em 2020 e "antigo" passou de 29 menções em 2012 para 5 menções em 2020. Observando os arquivos xlsx com todas as menções a cada uma dessas palavras, não conseguimos elaborar uma hipótese para esses dois fenômenos, o que demandaria uma análise macroscópica. Dentre as palavras que apresentaram **ascensão repentina**, "direito" passou de 12 menções em 2015 (ano em que teve o menor número de menções) para 62 menções em 2020 e "humano" passou de 7 menções em 2012 (ano em que teve o menor número de menções) para 35 menções em 2020. Nossa hipótese inicial era a de que essa flutuação correspondia à da ex-

pressão "direitos humanos", mas observando o arquivo "HH direitos humanos. xlsx" percebemos que não há relação entre a flutuação dessa expressão e das duas palavras que a compõem. Formulamos a hipótese de que a ascensão repentina dessas duas palavras poderia estar relacionada aos temas específicos privilegiados nos números em que essa ascensão ocorre. Dentre as palavras que apresentaram **aparecimento repentino**, "gênero" passou de 28 menções em 2016 para 103 menções em 2017, período em que ocorreu o aparecimento repentino. Nossa hipótese inicial era a de que isso poderia estar relacionado a um crescimento da relevância dos estudos de gênero, mas ao analisarmos o arquivo "HH gênero. xlsx", percebemos que a palavra gênero está relacionada sobretudo ao gênero musical, por conta do tema privilegiado pela revista em 2017.

Tabela 4 – Fenomenologia semântico-temporal das palavras mais usadas na RHHJ.

	Ascensão	Queda	Estabilidade	Ascensão repentina	Aparecimento repentino
Palavras	**1. Como explicar a ascensão da(s) palavra(s) x, y…?** • presente em todos os períodos (min > 0); • termina mais alto que começou (swing e delta positivos); • começa entre as 50% últimas palavras (1º valor de data < metade do nº de itens); • termina entre as palavras do primeiro quarto do ranking (primeiras 25%).	**2. Como explicar a queda da(s) palavra(s) x, y…?** • presente em todos os períodos (min > 0); • termina mais baixo que começou (swing e delta negativos); • começa entre as 50% primeiras palavras (1º valor de data > metade do nº de itens); • termina entre as palavras do último quarto do ranking (últimas 25%).	**3. Por que a(s) palavra(s) x, y... é (são) estável(is) na primeira metade das palavras mais usadas?** • presente em todos os períodos (min > 0); • stdev < 2.5; • termina entre as 50% primeiras palavras (último valor de data > metade do nº de itens).	**4. Como explicar a ascensão repentina da(s) palavra(s) x, y...?** • presente em um dos dois últimos períodos (sem 0 no data nos dois últimos períodos); • termina entre as 25% primeiras palavras; •swing e delta > 75%.	**5. Como explicar o surgimento repentino da(s) palavra(s) x, y...?** • presente apenas nos períodos da 2ª metade (sem 0 no data na 2ª metade e com 0 na 1ª metade); • min = 0.
20	professor histórico		história ensino poder		

30		escola	história ensino poder hoje	direito humano	
40			história ensino poder hoje	direito humano aprendizagem	
50		novo	história ensino poder hoje	direito aula humano aprendizagem narrativa	
60	aluno aula	novo	história ensino poder hoje	aula humano aprendizagem narrativa	
70	aula presente		história ensino poder hoje	humano aprendizagem narrativa	
80	aula presente	cultural	história ensino poder hoje	humano aprendizagem narrativa	
90	aula aprendizagem		história ensino poder hoje	humano aprendizagem narrativa estudante	debate
100	aula aprendizagem estudante		história ensino poder hoje	humano aprendizagem narrativa estudante	
200	humano narrativa disciplina sujeito conceito		história ensino poder hoje		gênero
250	docente narrativa disciplina	comunidade mudança	história ensino poder hoje		gênero
300	docente disciplina desenvolvimento		história ensino poder hoje	consciência rüsen	gênero
400	metodologia pensamento metodológico	ambiente publicar	história ensino poder hoje	consciência rüsen funk escravizar narrativo	
500	consciência metodologia investigação pensamento metodológico	antigo publicar	história ensino poder hoje	consciência rüsen funk narrativo turma unilab discente	

Fonte: Elaboração dos autores.

As palavras mais usadas no *corpus* completo da *Revista Brasileira de História*, apresentado na Tabela 5 (na próxima página), são obviamente as três palavras que compõem seu título, que poderiam ser consideradas palavras vazias. Podemos afirmar que entre as palavras mais importantes estão "poder" (que flutua entre a 2ª e a 6ª posição), "político" (2ª à 15ª posição), "social" (4ª à 16ª posição), "estado" (5ª à 18ª posição) e "novo" (9ª à 18ª). Também é interessante observar por que as palavras "indígena", "documento", "escravo", "água", "antigo", "região", "militar", "classe" e "liberdade" aparecem em branco entre as primeiras palavras, ou seja, aparecem de forma significativa apenas em um dos *corpora* anuais.

Por uma questão de espaço, não analisaremos todas as palavras que aparecem na Tabela 6, mas apenas destacaremos dentre os fenômenos que podem ser observados nela: a **ascensão** das palavras "nacional" e "América", a **queda** das palavras "trabalho", "processo" e "governo" e a **ascensão repentina** das palavras "antigo" e "antiguidade" (por se repetirem em várias linhas da tabela).

Para observar mais de perto essas palavras selecionadas na Tabela 4, utilizamos dois códigos em Google Apps Script (de extensão gs) disponíveis na pasta "code" do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/). As planilhas criadas por este último código para cada uma das palavras listadas no parágrafo anterior (nacional, América, trabalho, processo, governo, antigo e antiguidade) encontram-se disponíveis em formato xlsx na pasta "dataset/RBH/08_specific_words" do projeto hospedado na Open Science Framework (https://osf.io/bepqd/files/).

Dentre as palavras em **ascensão**, "nacional" passou de 149 menções em 2012 para 196 menções em 2020 e "América" passou de 69 menções em 2012 para 124 menções em 2020, o que pode ser conferido nos arquivos xlsx citados no parágrafo anterior. As hipóteses para esses fenômenos são o crescimento da preocupação do campo com os temas relacionados a estas palavras. Dentre as palavras em **queda**, "trabalho" passou de 271 menções em 2012 para 88 menções em 2020, "processo" passou de 195 menções em 2012 para 89 menções em 2020 e "governo" passou de 151 menções em 2012 para 89 menções em 2020. A queda no uso da palavra "trabalho" significaria uma redução das produções sobre história do trabalho paralela à precarização do trabalho e às reformas trabalhistas em voga? A queda no uso da palavra "governo" significaria uma redução da produção de história política ou uma ressignificação dessa história para além da história dos governos? Dentre as palavras que apresen-

Tabela 5 – As 19 palavras mais usadas no *corpus* completo da RBH e nos *corpora* anuais, seguidas do número de vezes em que cada uma delas aparece em cada *corpus*.

Corpus completo	2012	2013	2014	2015	2016	2017	2018	2019	2020
história, 12997	história, 1217	história, 1718	história, 1248	história, 1270	história, 1564	história, 1584	história, 1293	história, 1384	história, 1719
brasileiro, 9009	poder, 949	político, 966	brasileiro, 978	brasileiro, 959	brasileiro, 1231	revista, 943	arquivo, 1051	rio, 1335	brasileiro, 1069
revista, 6632	brasileiro, 890	poder, 911	poder, 724	revista, 783	brasil, 966	brasileiro, 938	brasileiro, 1018	brasileiro, 1075	revista, 955
poder, 6560	político, 729	brasileiro, 851	brasil, 688	terra, 749	revista, 794	poder, 690	revista, 787	revista, 798	poder, 654
rio, 5590	brasil, 641	social, 695	político, 656	poder, 606	poder, 718	indígena, 657	brasil, 661	poder, 718	brasil, 484
brasil, 5542	rio, 617	estado, 579	rio, 611	negro, 591	sérgio, 661	político, 641	poder, 590	brasil, 590	paulo, 484
político, 5526	trabalho, 608	brasil, 541	paulo, 541	rio, 588	político, 645	rio, 577	paulo, 568	político, 562	rio, 454
paulo, 4297	revista, 599	revista, 531	trabalho, 505	brasil, 538	estado, 500	índio, 556	documento, 532	paulo, 474	político, 451
social, 4210	primeiro, 546	novo, 525	novo, 475	político, 526	rio, 499	social, 527	rio, 521	social, 451	século, 395
estado, 3745	janeiro, 483	primeiro, 418	janeiro, 451	social, 516	paulo, 489	trabalho, 444	escravo, 489	português, 444	janeiro, 370
novo, 3682	estado, 480	século, 408	revista, 442	paulo, 473	novo, 472	paulo, 435	público, 458	água, 443	social, 348
janeiro, 3577	novo, 467	nacional, 408	estado, 433	estado, 428	social, 449	brasil, 433	janeiro, 437	século, 438	antigo, 322
trabalho, 3385	paulo, 454	governo, 399	social, 411	janeiro, 427	primeiro, 402	guerra, 393	estado, 388	região, 402	nacional, 321
século, 3315	trabalhador, 453	trabalho, 390	trabalhador, 388	novo, 419	janeiro, 385	século, 376	social, 365	público, 397	primeiro, 312
primeiro, 3277	tempo, 450	rio, 388	nacional, 369	trabalho, 392	livro, 380	primeiro, 361	político, 350	novo, 345	estado, 299
nacional, 2768	social, 448	paulo, 379	primeiro, 330	século, 375	trabalho, 375	novo, 356	século, 341	estado, 343	novo, 293
público, 2673	século, 439	janeiro, 346	tempo, 321	trabalhador, 364	histórico, 369	janeiro, 340	nacional, 340	janeiro, 338	obra, 284
histórico, 2521	segundo, 386	dia, 337	política, 319	índio, 334	nacional, 361	estado, 295	novo, 330	primeiro, 325	público, 277
tempo, 2509	militar, 376	classe, 323	governo, 316	nacional, 327	buarque, 360	histórico, 289	liberdade, 318	segundo, 307	josé, 267

Fonte: Elaboração dos autores.

taram **ascensão repentina**, "antigo" passou de 18 menções em 2019 (ano em que teve o menor número de menções) para 74 menções em 2020 e "antiguidade" passou de nenhuma menção em 2019 (ano em que teve o menor número de menções) para 102 menções em 2020. A ascensão do uso das palavras "antigo" e "antiguidade" tem relação apenas com temáticas privilegiadas em

números específicos da RBH ou com uma ascensão da importância da História Antiga no campo?

Tabela 6 – Fenomenologia semântico-temporal das palavras mais usadas na RBH.

Palavras	Ascensão	Queda	Estabilidade	Ascensão repentina	Aparecimento repentino
	1. Como explicar a ascensão da(s) palavra(s) x, y…? • presente em todos os períodos (min > 0); • termina mais alto que começou (swing e delta positivos); • começa entre as 50% últimas palavras (1º valor de data < metade do nº de itens); • termina entre as palavras do primeiro quarto do ranking (primeiras 25%).	**2. Como explicar a queda da(s) palavra(s) x, y…?** • presente em todos os períodos (min > 0); • termina mais baixo que começou (swing e delta negativos); • começa entre as 50% primeiras palavras (1º valor de data > metade do nº de itens); • termina entre as palavras do último quarto do ranking (últimas 25%).	**3. Por que a(s) palavra(s) x, y... é (são) estável(is) na primeira metade das palavras mais usadas?** • presente em todos os períodos (min > 0); • stdev < 2.5; • termina entre as 50% primeiras palavras (último valor de data > metade do nº de itens).	**4. Como explicar a ascensão repentina da(s) palavra(s) x, y...?** • presente em um dos dois últimos períodos (sem 0 no data nos dois últimos períodos); • termina entre as 25% primeiras palavras; • swing e delta > 75%.	**5. Como explicar o surgimento repentino da(s) palavra(s) x, y...?** • presente apenas nos períodos da 2ª metade (sem 0 no data na 2ª metade e com 0 na 1ª metade); • min = 0.
20			história poder brasileiro		
30			história poder brasileiro		
40			história poder brasileiro		
50	nacional	trabalho	história poder brasileiro	antigo	
60	nacional	trabalho processo	história poder brasileiro	antigo	
70	público nacional	processo	história poder brasileiro	antigo	universidade
80	nacional	governo	história poder brasileiro	antigo	
90		governo	história poder brasileiro	antigo	

100		governo	história poder brasileiro	antigo cultural	
200	cultura contexto universidade	lei análise	história poder brasileiro	antigo historiar university history sus saudade esclavo correio grego jazz	área
250	discurso universidade américa	dia jornal sempre	história poder brasileiro	médico historiar university history sus saudade esclavo correio grego jazz antiguidade	área
300	antigo américa debate	terra escravo região dia direito sempre objetivo razão	história poder brasileiro	médico historiar university history sus saudade esclavo correio grego jazz antiguidade clássico doença romano abreu	
400	historiografia debate	força	história poder brasileiro	negro médico historiar university history sus saudade esclavo correio grego jazz antiguidade clássico doença romano abreu wikipédia saúde horácio poeta chile saint	
500	history sistema	arquivo família civil	história poder brasileiro	negro médico historiar university sus saudade esclavo correio grego jazz antiguidade clássico doença abreu wikipédia saúde global horácio poeta chile saint oxford miguel pele bezerro airar not buenos	

Fonte: Elaboração dos autores.

Qual seria o significado da ascensão repentina da palavra "antigo" no *corpus* da RBH: antigo está sendo usado como adjetivo ou trata-se de um crescimento nos estudos sobre a antiguidade? Enquanto a RHHJ viu ascender rapidamente as palavras "debate" e "gênero", a RBH viu ascender "área" e "universidade". Em termos de debate público contemporâneo, o que isso significa? Como as questões mais relevantes no debate público atual chegam à produção dessas revistas? Estaria a RBH fazendo uma defesa da própria história enquanto a RHHJ se insere de forma mais contundente em outros temas? O objetivo de estudos semântico-temporais com este não é responder a essas questões, mas levantá-las para que possam ser respondidas pelo campo ou auxiliá-lo a gerar outras questões derivadas delas. Um exemplo de como um estudo neste formato pode ser útil a outros historiadores está no uso feito por Carmem Zeli de Vargas Gil no artigo "A sala de aula nas publicações da *Revista História Hoje* (2015-2017)" publicado entre as páginas 84 e 108 desta mesma edição.

REFERÊNCIAS

ANDERSEN, H.; BARKER, P.; CHEN, X. *The Cognitive Structure of Scientific Revolutions*. Cambridge: Cambridge University Press, 2006.

CARDOSO, O. The social flow of historical narratives and its many names. *Esboços*: histórias em contextos globais, v. 26 n. 43, p. 573-596, set.-dez. 2019.

CHRISTENSEN, Garret; FREESE, Jeremy; MIGUEL, Edward. *Transparent and Reproducible Social Science Research*: How to Do Open Science. Oakland: University of California Press, 2019.

DEELMAN, E. et alii. The future of scientific workflows. *The International Journal of High Performance Computing Applications*, v. 32 n. 1, p. 159-175, 2017.

BEZJAK, S. et alii. *Manual de Formação em Ciência Aberta*. 2018. Disponível em: https://book.fosteropenscience.eu/pt/book.pdf. https://doi.org/10.5281/zenodo.1212496. Acesso em: 30 mai. 2021.

IOANNIDIS, J. Why Most Published Research Findings Are False. *PLoS Med* v. 2, n. 8: e124, 2005. https://doi.org/10.1371/journal.pmed.0020124.

KANWAL, J.; SMITH, K.; CULBERTSON, J.; KIRBY, S. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, Amsterdam, v. 165, p. 45-52, 2017.

KOSCHTIAL, Claudia; KÖHLER, Thomas; FELDEN, Carsten (orgs.). *e-Science*:

Open, Social and Virtual Technology for Research Collaboration. Cham: Springer, 2021.

KUHN, T. S. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press, 1996.

MARCUM, J. A. From Paradigm to Disciplinary Matrix and Exemplar. In: KINDI, V.; ARABATZIS, T. *Kuhn's The structure of scientific revolutions revisited*. New York: Routledge, 2012.

MORENO-SÁNCHEZ, I.; FONT-CLOS, F.; CORRAL, Á. Large-Scale Analysis of Zipf's Law in English Texts, *PLoS ONE*, San Francisco/Cambridge, v. 11, n. 1, 2016. https://doi.org/10.1371/journal.pone.0147073.

OECD (Organisation for Economic Co-operation and Development). Making Open Science a Reality. (OECD Science, Technology and Industry Policy Papers, 25). Paris: OECD Publishing, 2015.

PONTIKA, Nancy; KNOTH, Petr; CANCELLIERI, Matteo; PEARCE, Samuel. Fostering Open Science to Research using a Taxonomy and an eLearning Portal. In: *iKnow*: 15th International Conference on Knowledge Technologies and Data Driven Business, p. 21-22, out. 2015, Graz, Austria. https://doi.org/10.1145/2809563.2809571.

SILVA, E. A.; SILVA, J. M. Ofício, Engenho e Arte: Inspiração e Técnica na Análise de Dados Qualitativos. *Revista Latino-americana de Geografia e Gênero*, Ponta Grossa, v. 7, n. 1, p. 132-154, 2016.

WILLIAMS, J. R.; BAGROW, J. P.; REAGAN, A. J.; ALAJAJIAN, S. E.; DANFORTH, C. M.; DODDS, P. S. Zipf's law is a consequence of coherent language production. *arXiv*, Ithaca. https://arxiv.org/abs/1601.07969v2, 2016.

ZIPF, G. K. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley, 1949.

Artigo submetido em 30 de maio de 2021.