DOI: 10.5433/1981-8920.2025v30n2p188

COMUNICAÇÃO NO CONTEXTO DAS HUMANIDADES DIGITAIS: FERRAMENTAS GRATUITAS OPERADAS POR COMPUTADOR COMO COMPLEMENTO ÀS METODOLOGIAS TRADICIONAIS

COMMUNICATIONS IN THE CONTEXT OF DIGITAL HUMANITIES: FREE COMPUTER-OPERATED TOOLS AS A COMPLEMENT TO TRADITIONAL METHODOLOGIES

Cristiano Magrini Rodrigues [a]
Rejane de Oliveira Pozobon[b]

RESUMO
Objetivo: Apresentar uma proposta de metodologia para tratamento de dados em Comunicação que une a Análise Discursiva Argumentativa à Connected Comcept Analysis por meio de softwares para computador sob a perspectiva da Linguística de Corpus. **Metodologia**: O protocolo, desenvolvido em uma tese de doutorado, apresenta a possibilidade de se expandir a análise do discurso a um conjunto textual volumoso sem abandonar a característica interpretativa da AD. O esquema metodológico está inserido no contexto das Humanidades Digitais, abordagem que une a força da computação às pesquisas das áreas Sociais e Humanas. **Conclusões**: O procedimento descrito aponta um caminho gratuito e replicável para pesquisadores com necessidade de trabalhar com grande quantidade de textos, mostrando diretrizes e detalhando funcionalidades em contextos de intensa produção de conteúdo.

Descritores: Humanidades Digitais. Metodologia da pesquisa. Tratamento de dados. Ferramentas gratuitas.

1 INTRODUÇÃO

As metodologias para tratamento de dados na Comunicação são tão diversas quanto os tipos de pesquisas e as possibilidades de uso dentro do

[a] Doutor em Comunicação pela Universidade Federal de Santa Maria (UFSM), Santa Maria, Brasil. E-mail: cristiano.magrinirodrigues@gmail.com

[b] Doutora em Ciências da Comunicação pela Universidade do Vale do Rio dos Sinos. Docente do Programa de Pós-graduação em Comunicação da Universidade Federal de Santa Maria, Santa Maria, Brasil. E-mail: rejanepozobon@gmail.com

campo. Neste texto, é apresentada uma dessas possibilidades. Trata-se de uma metodologia desenvolvida para uma tese de doutorado (Rodrigues, 2023) que une a Análise Discursiva Argumentativa (Amossy, 2018a; 2018b; Charaudeau, 2015, Orlandi, 2015; Pêcheux, 1990) à *Connected Comcept Analysis* por meio de *softwares* que trabalham sob a perspectiva da Linguística de Corpus e demonstra a capacidade de se expandirem as inferências da análise do discurso francesa a um conjunto textual volumoso. A principal vantagem desse pacote metodológico está na possibilidade de se trabalhar com grande quantidade de textos sem renunciar à capacidade interpretativa da análise discursiva. A metodologia se insere no escopo do que defendem as Humanidades Digitais (Alves, 2016), abordagem que vincula práticas consolidadas das Ciências Sociais e Humanas à computação e que é defendida por um grupo de pesquisadores como uma forma de ampliar as possibilidades das metodologias tradicionais com o uso de ferramentas operadas por computador.

Nas próximas páginas, discorre-se, sobretudo, acerca dos softwares de linguística de corpus sob dois pontos, que se traduzem nos objetivos do texto: o primeiro, já mencionado, é apresentar uma forma de permitir que as características da análise discursiva sejam aplicadas a uma amostra exponencialmente extensa, algo inviável quando se trata da AD tradicional; o segundo, demonstrar que essa análise pode ser realizada por qualquer pesquisador interessado ao apresentar um esquema metodológico gratuito e replicável – inclusive para quem deseja trabalhar com textos mesmo que sem adentrar na esfera discursiva.

2 HUMANIDADES DIGITAIS: A FORÇA DA COMPUTAÇÃO ALIADA ÀS PESQUISAS DAS ÁREAS SOCIAIS E HUMANAS

A abordagem deste artigo se inclui na perspectiva das Humanidades Digitais (HD). Elas são compreendidas como uma comunidade de práticas que reúne pesquisadores de diferentes campos de estudo. Isso significa que não se trata de algo delimitado, que venha a constituir um campo de saber específico. Pelo contrário, trata-se de uma abordagem multidisciplinar na qual campos diferentes de estudo encontram afinidades. Essas afinidades dizem respeito,

sobretudo, ao uso das metodologias digitais e à possibilidade de compartilhamento de ferramentas em projetos distintos.

Apesar de as discussões sobre as HD propriamente ditas já ocorrerem há pelo menos duas décadas, o fato de elas serem tratadas mais como comunidade e menos como disciplina ou campo acadêmico mantém difusa a possibilidade de consolidar uma definição. A classificação das Humanidades Digitais como comunidade é resgatada por Alves (2016) – signatário do texto fundador da Associação das Humanidades Digitais, entidade que reúne pesquisadores lusófonos interessados no tema – dá-se a partir de uma das literaturas fundadoras da área: a obra *Companion to Digital Humanities*, de 2004 e relançada em 2016. O livro marca o início do uso da expressão Humanidades Digitais em larga escala e defende, sob a ideia de comunidade, que a riqueza das HD está na pluralidade (Schreibman; Siemens; Unsworth, 2016). Ademais, a literatura recente dedicada a trabalhar criticamente elementos da Ciência de Dados é taxativa: vivemos em um momento histórico no qual podemos extrair informações dos dados e manipulá-las como nunca foi possível (Moreira; Carvalho; Horváth, 2019). Assim, Humanidades Digitais são atividades acadêmicas e científicas que têm como questão primordial a união entre computação e Humanidades (Schreibman, Siemens, Unsworth, 2016).

Em se tratando de comunidade, a questão linguística é um fator determinante, pois a linguagem é base do discurso e todo discurso é carregado de sentidos. Nas Humanidades, “sentidos” são elementos-chave: início e fim. Nessa via, as abordagens sobre Comunicação podem compartilhar as metodologias. A metodologia demonstrada aqui aborda a possibilidade de análises textuais a partir de *softwares*.

Juruá (2021) reforça a importância do estudo das linguagens com o apoio computacional. Ora, a produção de sentidos está justamente incrustada no universo da Cultura e da Comunicação e, por isso, acredita-se que qualquer tentativa de avançar nas Humanidades Digitais carrega, em algum nível, a força que a computação proporciona para um processamento linguístico robusto que contribui para as interpretações do cotidiano. Trata-se de trabalhar com uma abordagem em que as inteligências artificial e humana se complementam.

Ambos, máquinas e humanos, trabalham voltados ao mesmo universo semântico, cada um com suas particularidades e limitações.

Käsper e Maurer (2020) dizem que as técnicas baseadas em uso de máquinas para análises de textos remontam aos anos 1960 e, desde então, o uso sistemático de computadores vem sendo aprimorado para estudos nas áreas das Humanidades. Possibilidades mais complexas de trabalho, como procedimentos de lematização e categorização, por exemplo, só foram possíveis com a evolução da computação na década de 1980. Na contemporaneidade, a inteligência artificial já permite análises lexicais robustas a partir de estatísticas trabalhadas por computador que vêm originar intepretações de perfis sociopsicológicos e históricos, entre outros.

Um exemplo é a *Connected Comcept Analysis* (CCA). Com ela, é possível apresentar uma análise que une elementos quantitativos e qualitativos dos textos. Nichols *et al.* (2021) explicam que os programas de computador identificam narrativas e mostram a relação entre palavras, conceitos e recursos. Segundo Lindgen (2016), a técnica também permite o processamento de grande quantidade de material textual e até mesmo considera a sensibilidade do discurso. Outras informações possíveis incluem correlações e co-ocorrências formadas pelo conjunto lexical dos textos.

A *Connected Comcept Analysis* deriva em algum grau da *Network Text Analysis* (NTA). A NTA é uma técnica que engloba abordagens como Representações Gráficas, Pontos de Destaque, Análise de Mapas, Análise de Co-ocorrência de Palavras, Análise de Redes de Palavras e Análise de Ressonância Central[1] (Lindgren, 2016). Há bastante semelhança nesses procedimentos com a abordagem pela CCA. A diferença está na forma como os conceitos são identificados e trabalhados no decorrer da análise.

Lindgren (2016) diz que a NTA pode ser integrada à Análise do Discurso e a outras abordagens analíticas, podendo estas serem trianguladas com os resultados. Por sua vez, a CCA tem mais afinidade com as Humanidades Digitais porque faz uma codificação qualitativa para identificar os temas; na NTA, isso é

1 Em inglês, esses termos são conhecidos como Knowledge Graphing, Popping, Map Analytics, Co-Word Analysis, Word Network Analysis e Centering Resonance Analytics.

automatizado. Assim, nota-se na CCA que o aspecto sociocultural é presente em todo o processo da pesquisa, enquanto na NTA, os conceitos teóricos e filosóficos acabam mais restritos à leitura final dos dados e dos gráficos. Ou seja, a CCA busca a integração, não a triangulação, dando destaque ao processo interpretativo em todas as fases.

A CCA mescla análise comparativa constante, teoria do discurso e análise de texto em rede. Conforme Lindgren (2016), esse foi um método concebido especialmente para levar a sensibilidade do discurso para o *big data*, possibilitando que fossem analisadas publicações massivas da internet provenientes de redes sociais digitais, blogues e fóruns sem a necessidade de selecionar apenas uma parte dos dados. Como ele mesmo explica, o método supre a necessidade de ser uma análise que integre as sensibilidades semióticas e semânticas às análises de redes – exigência do caráter socialmente conectado dos meios de comunicação *on-line*. Em síntese, a CCA converge para a Análise do Discurso. Dadas essas características, orientam-se as próximas linhas pensando numa abordagem voltada à linguística.

Para Orlandi (2015), a comunicação é produtora de sentidos e esses sentidos são expressos pela palavra. Percebe-se a importância dessa união justamente porque a AD considera que as narrativas são estruturadas a partir de contextos histórico, político e social. Ter a tecnologia como aliada para desbravar esses sentidos de maneira ampla ainda é bastante controverso para os mais puristas, porém, a experiência tem demonstrado resultados profícuos. Com o uso dos programas adequados, pode-se aceder a grafos e métricas (Nichols *et al*., 2021). Isso é importante para saber como um determinado tema é abordado no *corpus* analisado.

Sobre a CCA, Lindgren (2016) explica que se trata de um método baseado em codificação comparativa constante: a partir da construção de conceitos e de análise de redes, as palavras ou frases são conectadas. Entre os resultados, estão grafos e redes dos principais conceitos interligados aos textos analisados. São esses *outputs* que amparam a análise discursiva.

Biber (1993) aponta que os textos em um determinado idioma são linguisticamente semelhantes, isto é, há um padrão lógico e não aleatório que é

seguido pelos falantes nativos. Yuan, Feng e Danowski (2013) indicam que programas como WORDij permitem análises de redes semânticas e revelam digitalmente a estrutura de significados de um texto. Isso representa de maneira indireta a estrutura cognitiva do seu criador justamente porque há um sentido estrutural por trás da escrita. Sardinha (2000) complementa, ao dizer que a linguagem humana analisada dessa forma corresponde a um sistema probabilístico naturalizado pelo escritor. Em resumo, as análises consideram a linguagem em uso.

A calibragem e olhar humanos são necessários para que as investigações não se percam em um universo de dados. Daí a necessidade de se construir espaços conceituais que retomam a literatura específica do campo em estudo para direcionar e dar sentido às informações, conectando conceitos aos dados e analisando sentidos (Juruá, 2021) de modo metodologicamente coerente. Dessa forma, a observação a partir de contextos socioculturais na CCA coloca essas análises como possibilidades de compreensão da sociedade e demonstra bem o funcionamento do princípio das Humanidades Digitais, somando o dinâmico processamento computacional às teorias e estudos das ciências Sociais e Humanas.

3 CONCEITOS BÁSICOS NA CONSTRUÇÃO DO CORPUS NO CONTEXTO DA CONNECTED COMCEPT ANALYSIS

A *Connected Comcept Analysis* é um procedimento que permite análises do discurso frente à massiva oferta de dados disponíveis. Vincula-se à Análise de Redes e é uma técnica que visa suprir a necessidade de observação de produções textuais cada vez mais volumosas. Pode ser aplicada a qualquer tipologia textual, como comentários nas redes sociais ou matérias jornalísticas, por exemplo. Como comentado anteriormente, não há necessidade de se excluir grande quantidade de informação para selecionar pequenas amostras. Adotando tal procedimento, a análise se aproxima da chamada leitura à distância (*distant reading*), contraposta à leitura atenta (*close reading*), essa mais comum à análise discursiva clássica (Lindgren, 2016). Ou seja, o analista não se debruça exaustivamente sobre todos os textos, delegando à máquina o processo de

leitura de todo o *corpus*.

Tomemos por exemplo uma análise dos acontecimentos de 8 de janeiro de 2023 a partir de uma *hashtag* na rede social X (anteriormente Twitter). É humanamente inviável realizar uma leitura atenta do assunto, mesmo que em um recorte de poucas horas. O volume de informação publicado naquele dia e nos imediatamente posteriores é demasiado extenso para que a atenção humana dê conta. Da mesma forma, é inexequível uma leitura atenta das reportagens sobre esse mesmo acontecimento, pois foram muitas e nos mais variados veículos. Obviamente, o refinamento da pesquisa dita os parâmetros, os recortes e os limites, no entanto, uma análise mais robusta para quantidades numerosas de textos exige a força da computação.

A CCA permite uma leitura que se alterna entre a redução quantitativa (condensa muitos textos) e a validação qualitativa (necessita da interpretação do analista). O aspecto sociocultural é levado em consideração em todo o processo e, dessa forma, a CCA relaciona análise comparativa, teoria do discurso e análise de textos em rede de um modo que integra a computação às Humanidades.

Quanto à coleta, Recuero (2017) explica que ela pode ser feita de modo qualitativo ou de modo quantitativo (pela datificação). O modo qualitativo envolve entrevistas, questionários e outras ferramentas tradicionais das pesquisas em Comunicação que devem ser digitalizadas em texto para, então, serem processadas pelos *softwares*. Diferentemente, o modo quantitativo realiza a coleta diretamente nas redes ou nas plataformas.

Sobre as coletas quantitativas, a pesquisadora explica que elas geralmente focam em bases de dados preexistentes, como são as redes sociais digitais. As formas mais comuns incluem a extração automática de dados da internet por meio de *softwares* (*web scapping*), as API (*Application Programming Interface*), fornecidas pelas próprias plataformas *on-line*, o monitoramento de mídias sociais, o rastreio de *cookies*, as pesquisas *on-line* e análises de comportamento dos usuários em aplicativos, programas ou serviços *on demand*, por exemplo. Alguns propiciam análises mais sociológicas, como o *scrapping* e as API; outros estão mais voltados às pesquisas de mercado, como o

rastreamento de *cookies* e análise de comportamento em aplicativos e páginas. Nesses casos, a coleta é feita automaticamente por meio de *crawlers*. Muitos deles estão disponíveis publicamente e a maioria é desenhada focando um banco de dados específico. NodeXL, NetVizz e Gephi são exemplos. A depender de qual base de dados será realizada a coleta, pode ser necessária autorização para acesso à API, como é o caso para Facebook e X. No entanto, a depender da forma como a análise for realizada, o pesquisador pode elaborar seu próprio banco de dados a fim de facilitar o trabalho e refinar a análise. No caso a ser apresentado logo adiante, os textos foram coletados e salvos em arquivos de texto, visando a compatibilidade com os programas e mais precisão no processamento das informações por eles.

Para as pesquisas mais sociológicas, foco da CCA, as etapas do processo após seleção do *corpus* são tokenização, seleção, conceitualização, conexão e visualização (Lindgren, 2016). Vejamos brevemente a definição desses conceitos antes da apresentação da metodologia.

A tokenização é quantitativa e objetiva. Ela divide o texto em unidades menores (palavras) e isso facilita a análise. Ela prepara os textos para modelagens e análises pelos programas usados em Linguísticas de *Corpus*, por exemplo. A seleção é a primeira parte qualitativa da análise. O uso de uma lista de palavras (*stoplist*) ajuda a realizar a limpeza do *corpus* e refina os *outputs*. A conceitualização permite que sejam identificados e definidos conceitos sensibilizadores a partir das palavras restantes após a etapa anterior. Hipoteticamente, em uma pesquisa sobre comentários na rede social X sobre desarmamento, a palavra “vida” pode levar a diferentes entendimentos dos usuários sobre o tema a partir de forma como substantivo está ligado a outros termos (co-ocorrência). Isso leva à etapa de conexão, quando a parte quantitativa (co-ocorrência) é analisada qualitativamente para descrever o conteúdo dos textos e aos conceitos apresentados na etapa anterior. Aplicações e softwares como Voyant-Tools, AntConc, WORDij e IRaMuTeQ auxiliam no processamento de dados nas etapas de tokenização, seleção e conceitualização. Além das interpretações que podem ser obtidas diretamente a partir das suas ferramentas nativas, é possível exportar os dados para

processamento em programas de análise de redes como Gephi ou NodeXL, resultando em visualizações intuitivas e personalizáveis e, assim, tem-se acesso a dados interpretáveis.

Para a visualização, a análise de redes é uma opção que oferece retorno interessante se pensarmos no valor que as representações gráficas das redes agregam. Venturini, Munk e Jacomy (2018) afirmam que eles permitem situar um ponto em relação a outro e indicam o percurso e a proximidade entre eles. Essa é a principal estratégia de leitura que se faz presente na apresentação dos *layouts* oferecidos pelos algoritmos dos programas de análise de redes. Também é possível trabalhar com algoritmos que estabelecem o equilíbrio na disposição dos nós, realizando um balanceamento que reduz o cruzamento de linhas e facilita a legibilidade dos grafos.

Outro recurso proporcionado é o de densidade visual, que facilita a leitura dos grafos, tornando-os mais intuitivos. Os nós se reúnem em grupos, os clusters; os buracos estruturais se apresentam como zonas esparsas; os nós centrais, hubs, ficam em posições intermediárias; entre as regiões, notam-se pontes, ou seja, arestas que ligam os pontos. A linguagem é lógica e a CCA também. Por isso, a apresentação sob o aspecto de redes também serve para apresentar sínteses daquilo que é tratado em outros tipos de textos, como reportagens e transcrições de entrevistas, por exemplo.

4 PROCEDIMENTO METODOLÓGICO: USO DE *SOFTWARES* GRATUITOS DE PROCESSAMENTO LINGUÍSTICO

Comecemos pelo básico: a coleta dos dados. A pesquisa que originou este artigo teve como matéria-prima textos completos publicados na seção de editoriais de jornais de grande circulação no Brasil. Observaram-se as edições impressas e, posteriormente, foram buscadas as versões digitais desses textos publicadas nas páginas dos veículos na internet. A primeira decisão no momento da formação do banco de dados era definir qual a melhor forma para coleta e armazenamento dessas informações. Em casos como esse, a extração pode ser feita principalmente de duas formas: salvando a página inteira da internet ou separando os textos e salvando no formato .txt (codificação UTF-8). No primeiro

caso, a vantagem é a rapidez no salvamento. Não há necessidade de copiar e colar texto por texto, o que pode economizar muito tempo quando se trata de um material vasto. Porém, dentro das limitações dos programas utilizados, a desvantagem tem consequências. Trata-se da necessidade de limpar o texto posteriormente, pois as páginas completas apresentam termos que geram ruídos porque contêm vocábulos que não fazem parte do conteúdo a ser analisado e essas palavras interferem no resultado. Além disso, em algum momento, será preciso converter as páginas salvas em .html ou .pdf para arquivos compatíveis com os *softwares*. Já no segundo caso, a vantagem é de se preparar o *corpus* desde o início, tornando-o compatível para rodar na maioria dos programas, pois os arquivos em .txt são largamente aceitos. A desvantagem esbarra na demora da coleta, pois acaba sendo preciso copiar, colar e salvar texto por texto.

A título de ilustração, têm-se as imagens 1 e 2. Elas demonstram a leitura inicial nos dois tipos de coleta a partir do Voyant Tools[2], aplicação disponível gratuitamente *on-line* e com recursos bastante intuitivos. Pode-se notar a acuidade dos *outputs* na imagem 2, enquanto na imagem 1 os termos de interesse do texto ficam em segundo plano, uma vez que outras expressões da página preponderam e desviam o foco.

Definido e coletado o *corpus*, cabe passar para a manipulação dos dados. Dentro da perspectiva teórica que se apresenta aqui, faz-se necessário, primeiro, garantir o cumprimento das necessidades técnicas para *rodar* o banco de dados nos *softwares*. Para essa adequação, recomenda-se que o texto completo seja copiado das páginas e colado em um documento do programa Notepad++, codificação binária (Unicode) UTF-8 (8-BIT *Unicod Transformation Format*). Cria-se um arquivo de texto (.txt) para cada unidade textual e se nomeia o documento de modo que seja fácil identificá-lo e mantê-lo em ordem. Sugere-se nomenclatura com uma palavra ou letra e a data de publicação. Por exemplo, o esquema “Xaaaammdd”. Nesse caso “X” identifica a fonte do texto. Por sua vez,

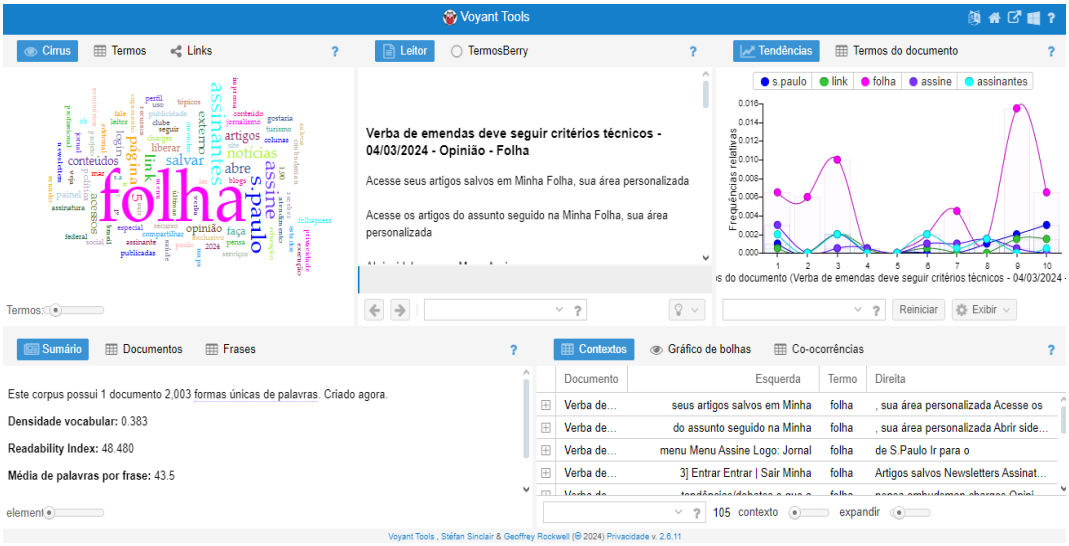
2 O Voyant Tools é um leitor de textos disponível on-line. Ele foi desenvolvido para auxiliar estudantes e pesquisadores interessados nas Humanidade Digitais na leitura e interpretação de textos. A aplicação oferece diversos recursos, alguns bastante semelhantes aos demais softwares recomendados neste capítulo. É uma excelente porta de entrada para exploração do corpus, no entanto, há algumas limitações e certa instabilidade que podem causar empecilhos em pesquisas extensas.

o formato ano-mês-dia permite a apresentação cronológica ordenada no sistema Windows e nos *softwares* de análise. Assim, um texto da Folha de S. Paulo pode ser nomeado como F20230803, por exemplo. Esse procedimento inicial de salvamento facilita a manipulação dos textos pelos programas, excluindo qualquer tipo de formatação ou *hiperlinks*, bem como organiza cronologicamente o *corpus* para as buscas. Reitera-se a vantagem da estratégia de coleta texto a texto para constituir uma base de dados o mais limpa possível. Além disso, esse trabalho mais demorado evita futura necessidade de conversão de formatos de arquivo para leitura nos programas e permite eliminar outras ocorrências que possam interferir no *corpus*. Nas páginas dos jornais, por exemplo, é comum uma chamada com *hiperlink* para outras matérias entre um parágrafo e outro. Também, o fato de que o salvamento das páginas em outros formatos, como .pdf ou .html, frequentemente incluem informações que não são de interesse, como comentários, índices, listas, legendas de imagens, *tags*, entre outras comuns às páginas na *web*.

A montagem do banco de dados aqui defendida ainda permite que ele possa ser rodado nos principais *softwares* gratuitos e possibilita a inclusão dos textos em bancos de dados facilmente convertidos para outros formatos, específicos para as análises que necessitam de processamento em .xls (Excel) ou .txt (Notepad++).

Com o uso do AntConc, programa gratuito para análise de documentos escritos que se baseia na Linguística de *Corpus* e que ajuda em pesquisas que acionam a Análise do Discurso (Käsper, Maurer, 2020; Gonçalves, 2016), pode-se organizar e processar com facilidade grande volume de textos. Nele, a primeira etapa após o carregamento dos arquivos compreende o recurso da *stoplist*, uma lista de palavras que é inserida a fim de filtrar o *corpus*. Essa lista contém as *stopwords*, palavras que o *software* deve passar a ignorar porque são muito frequentes na maioria dos textos em um idioma, mas que não agregam significado às análises.

Figura 1 – *Layout* da aplicação *Voyant Tools* – processamento da página inteira



Fonte: *Voyant Tools* e autores (2024)

Figura 2 – *Layout* da aplicação *Voyant Tools* – processamento apenas do texto de interesse

Fonte: *Voyant Tools* e autores (2024)

Alguns exemplos de *stopwords* são os artigos, as preposições e as conjunções. Esses grupos de palavras geralmente são desconsiderados nas atividades de mineração de textos. Um aspecto importante é que as *stoplists* são editáveis, o que permite partir de uma lista genérica facilmente encontrada na internet para a formação de outras, com ajustes personalizados pelo analista de acordo com o *corpus* em análise. Essa etapa pode ser orientada a partir de Lindgren (2016). Dessa forma, calibra-se a *stoplist* e, posteriormente, os arquivos já inseridos são novamente processados para que o resultado dos *outputs*

contenha as palavras mais relevantes para o *corpus*. Ainda no AntConc é possível verificar se os textos apresentam as palavras-chave convenientes ao problema de pesquisa. A ferramenta *Concordance Plot* indica quanta vezes os termos aparecem e quais as suas exatas posições nos arquivos. Funcionalidades como essa garantem que os textos analisados realmente tratam das temáticas de interesse da pesquisa.

Na etapa seguinte, a exploração do *corpus* acontece pelo *software* WORDij. Trata-se, na verdade, de um conjunto de programas de ciência de dados que processa a linguagem natural. Nele, há possibilidade de trabalhar com publicações de mídias sociais, notícias, entrevistas, discursos, e-mails, sites, grupos focais ou quaisquer outras formas cuja manipulação em formato textual possa ser viabilizada e transformada em redes semânticas. Em comparação com uma simples categorização das mensagens, a perspectiva de rede garante a visualização da relação entre as palavras nos textos. O WORDij transforma os textos em um conjunto de nós e pares de palavras conectados e executa as suas operações a partir da ideia de redes entre elementos semânticos codificados por processos ou por outras unidades sociais:

> As pessoas que encadeiam palavras de maneira semelhante são semelhantes umas às outras em comportamento de fala/ação. Se duas pessoas falam ou escrevem da mesma forma, por causa dessa semelhança de codificação semântica, elas podem ser semelhantes em outros níveis. Assumimos que a linguagem reflete a percepção e o comportamento. As pessoas que falam mais com as outras tendem a se comportar de forma mais semelhante umas com as outras, dadas as circunstâncias contextuais semelhantes. Isso provavelmente ocorre porque elas percebem seus ambientes e suas escolhas de comportamento dentro deles de maneira semelhante (WORDij, c2025, local. 1, tradução nossa)[3].

Os dados processados pelo WORDij são convertidos em diferentes formatos. Os *outputs* permitem a manipulação das informações em outros programas e no próprio WORDij. Entre as ferramentas que o *software* oferece,

[3] No original, em inglês: “People who string words together in a similar manner are similar to one another in speech/act behaviors. If two people talk alike or write alike, because of this semantic encoding similarity they may be similar on other levels. We assume that language reflects perception and behavior. The people who talk more like each other are likely to behave more similarity to one another given similar contextual circumstances. This is probably because they perceive their environments and their choices for behavior within them more similarly.” (WORDij, c2025, local. 1). Disponível em: https://www.wordij.net/about.html. Acesso em: 07 ago. 2022.

duas se destacam pela funcionalidade: *WordLink* – lê os arquivos em .txt elaborados anteriormente no AntConc e os converte para os formatos necessários para análises no próprio programa ou fora dele – e *Utilities* – apresenta o recurso *Proper Nouns*, um identificador de substantivos próprios e de suas conexões com o restante do texto.

No caso dessa última ferramenta, nota-se grande utilidade devido à oportunidade de ajustar termos compostos para que eles sejam lidos como uma única palavra. O funcionamento é bastante próximo ao processo de lematização, mas mais preciso. Em resumo, o recurso *Proper Nouns* combina dois ou mais termos e eles passam a serem lidos como uma única palavra. Para isso, bastam comandos simples em um arquivo .txt, semelhante ao executado na *stoplist*. O programa identifica automaticamente no texto as palavras grafadas em maiúsculo, característica dos substantivos próprios na Lingua Portuguesa. Para melhor resultado, unem-se todos os textos em um único arquivo .txt a ser inserido na seção *Input text file* da aba *Utilites*. Depois, é só criar um arquivo para cada *output* na mesma pasta. Em um desses arquivo, ele identificará os substantivos próprios e em outro, oferecerá uma sugestão para substituição. Ambos são editáveis.

Posteriormente, basta adicionar o arquivo com as substituições nas opções avançadas na caixa *String Replace File* e rodar novamente a análise. O programa salvará uma série de formatos que podem ser rodados em diferentes programas, entre eles, o Gephi. Por exemplo, as palavras “Albert” e “Einstein” configuram, inicialmente, duas entradas diferentes na lista obtida no AntConc. Sabe-se que juntas, elas se referem a uma só pessoa, o célebre físico teórico alemão Albert Einstein. A funcionalidade em questão agrega os dois termos para que eles apareçam únicos sempre que estiverem juntos em um texto. Basta ajustar uma lista no Bloco de Notas (por exemplo, definir “entrada->saída”, da seguinte forma: *Albert Einstein->Albert_*Einsten). Porém, pessoas diferentes podem ter nome ou sobrenome em comum. Um caso recorrente na imprensa brasileira e que ilustra bem a situação é o do ex-presidente Jair Bolsonaro e dos seus quatro filhos envolvidos na política. Com o auxílio dessa funcionalidade, a ferramenta identifica os substantivos próprios no texto e oferece sugestões de

apresentação dos *outputs*. É uma forma de aprimorar os dados, convertendo diversas palavras em um único termo e permitindo respostas e visualizações mais precisas. Isso resulta em refinamento e apresentação mais precisa dos nós e arestas que sintetizam o texto. O material pode ser exportado com o recurso *NodeTric*, que é salvo em um formato de arquivo .net preparado para ser lido como rede semântica e que resulta em grafos com destaque para a palavras e suas ligações no texto, a exemplo da imagem 3, cujos dados foram processados no WORDij e elaborados graficamente no Gephi[4].

A etapa final conta com o auxílio de mais um programa, o IRaMuTeQ. Esse *software* é bastante útil para análises de conteúdo, lexicometria e análises discursivas. Uma das funcionalidades mais utilizadas no IRaMuTeQ é a Classificação Hierárquica Descendente (CHD). Trata-se de uma ferramenta que auxilia na identificação dos diversos temas abordados no *corpus*. Ela apresenta classes dentro de um conjunto de textos ou segmentos de textos e indica como os termos se agrupam na narrativa. Como consequência disso, também permite que se visualize os subtemas abordados.

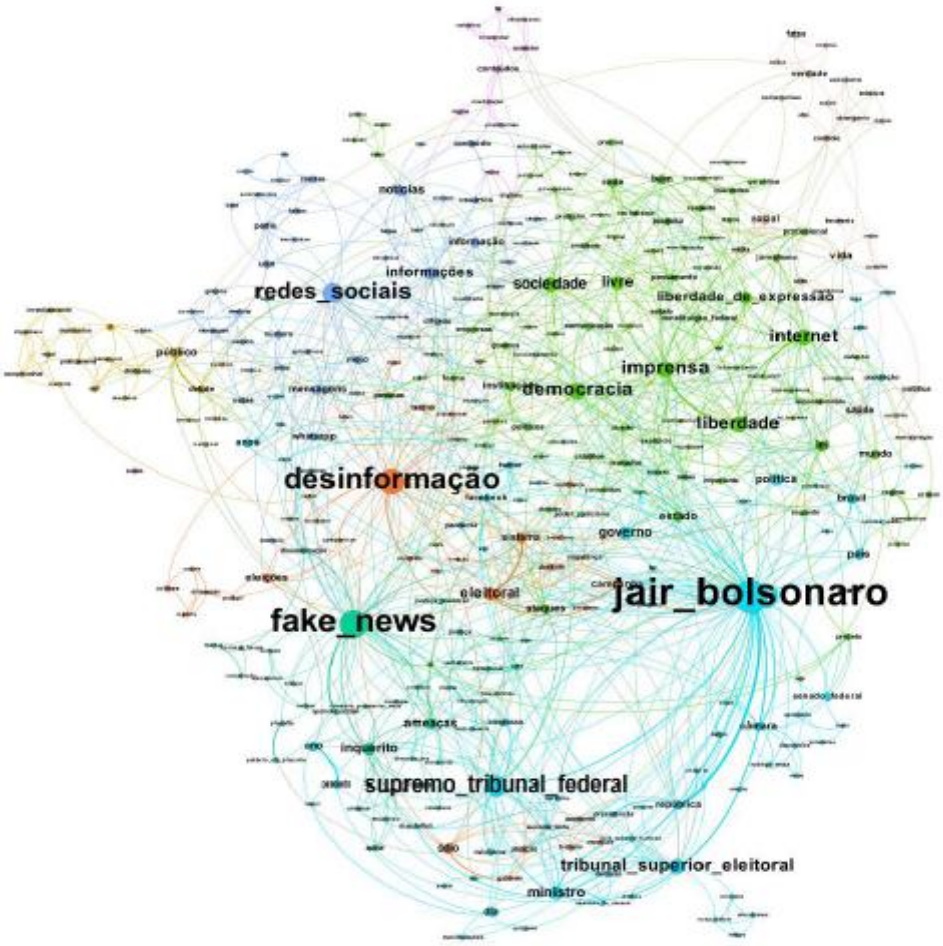
As características do IRaMuTeQ exigem que o pesquisador se baseie em três níveis de divisão do seu banco de dados, seguindo a gramática do *software*. São eles o *corpus*, isto é, todo o material da coleta; os textos, separados para serem analisados no programa porque fazem parte da temática da pesquisa; e os segmentos de textos, na prática, a unidade básica do programa a partir do fracionamento dos textos completos, forma pela qual são realizados os cálculos estatísticos no *software*.

O tamanho do texto influencia na forma como os algoritmos do programa fazem os cálculos. Por isso, períodos muito curtos nem sempre permitem que o IRaMuTeQ realize as operações necessárias para a segmentação e a comparação dos trechos. Ou seja, quanto maior a quantidade de material a ser analisada, melhores serão os resultados apresentados. Do outro modo, análises de textos muito curtos inviabilizam alguns tipos de análises. Observa-se que a pouca variedade de formas existentes em um texto curto dificulta ou até mesmo

[4] No Gephi, é preciso adequar o arquivo para melhor visualização, aplicando filtros e estabelecendo o melhor algoritmo de distribuição.

impossibilita que o programa rode alguns dos seus algoritmos.

Figura 3 – Grafo do jornal *O Estado de S. Paulo*



Fonte: Elaborado pelos autores (2023).

Retomando a funcionalidades do IRaMuTeQ, a Classificação Hierárquica Descendente e a Análise Fatorial de Correspondência (AFC) são recursos que retornam ao analista gráficos, dendogramas e filogramas a partir do cruzamento entre vocabulário e classes de palavras e entregam representações visuais (Salviati, 2017)[5]. No entanto, a construção da base de dados para ser rodada no IRaMuTeQ exige uma preparação. Os procedimentos de formatação dos textos já realizados anteriormente facilitam agora, porque o texto ainda passa por uma nova limpeza a fim de que seja lido corretamente no programa. É importante notar que há uma série de caracteres que o IRaMuTeQ não lê: aspas, apóstrofo,

[5] Consulte em Salviati (2017) os detalhes acerca do vocabulário do IRaMuTeQ.

cifrão, recuos de parágrafos, margens e tabulações. Salvati (2017) afirma que também é preponderante adequar a escrita para a forma de texto corrido, isto é, sem mudança de linha; evitar o uso de palavras compostas, unindo-as pelo sinal gráfico sublinhado (*underline*); padronizar as siglas e os substantivos próprios para a mesma grafia, entre outras questões.

Merece atenção o fato de que os algoritmos do IRaMuTeQ necessitam de uma determinada codificação para que os dados sejam processados adequadamente. O manual disponível na página oficial do programa sugere que os caracteres estejam codificados em UTF-8 ou em SP1252 (Salviati, 2017).

Cria-se uma tabela no Excel com os identificadores de cada objeto de estudo – um jornal, por exemplo –, data, texto e ID (cada texto receberá um número para ser lido como único pelo programa. Vale ter em mente que em alguns casos, pode ser necessária ainda a aplicação de uma macro no Notepad++, garantido a subtração dos elementos gráficos inadequados. Em seguida, depois do tratamento, é fundamental a criação de um cabeçalho na tabela do Excel. Ele permite a separação dos textos em linhas de comandos. O cabeçalho determinará as entradas e as demais informações na base de dados. Aqui, a nomenclatura dos textos com as datas pode ser útil. Conforme o tamanho do *corpus* e o que se busca nele, a data pode apontar para informações relevantes, encaminhando a análise para um resultado mais acurado, por exemplo, identificação de padrões e outras características cujo recorte temporal ajudar a explicar determinadas tendências discursivas. É preponderante garantir que as informações estejam formatadas corretamente para a composição do cabeçalho. Ele precisa ter um identificador único para cada entrada (ID), além da demais informações que serão lidas pelo IRaMuTeQ.

Formatadas as informações e elaborada a linha de comando, a etapa seguinte corresponde à criação de uma tabela dinâmica contendo a linha de comando e o texto já preparado. Essa tabela é exportada para o Notepad++ para ser salva em arquivo .txt. E devido às informações do cabeçalho, pode-se explorar os textos conforme datas desejadas, realizando a seleção no próprio IRaMuTeQ. Explicar o passo a passo dessa etapa rende, por si só, um tutorial próprio. No entanto, há diversos guias na internet que detalham o passo a passo

da criação das tabelas dinâmicas no *software*.

Assim, os textos preparados são inseridos no IRaMuTeQ, que faz um cálculo da extensão de cada um, divide-os conforme o tamanho e as palavras lematizadas e oferece dados dispostos em classes de palavras semelhantes entre si e diferentes das demais (Ramos, Lima, Amaral-Rosa, 2018).

Infelizmente, a limitação de um artigo inviabiliza a descrição detalhada desses procedimentos para o IRaMuTeQ, bem como para o ajuste das visualizações no Gephi. Para tal, o analista iniciante precisará empreender algum esforço na busca de executar a tarefa. Felizmente, também para esses casos, há tutoriais disponibilizados voluntariamente por cientistas de dados na internet. A leitura da metodologia da tese (Rodrigues, 2023) referenciada abaixo pode complementar as informações abordadas neste texto e abrir novos horizontes para quem busca explorar mais a fundo o *corpus* textual em uma pesquisa.

5 CONSIDERAÇÕES FINAIS

O presente artigo apresentou uma proposta de análise aos pesquisadores em Comunicação desejam trabalhar com grande volume de textos. Longe de ser um passo a passo detalhado, o foco se direciona a apontar possibilidades para pós-graduandos e pesquisadores que sofrem com a escassez de recursos no desenvolvimento das suas pesquisas, dado que alguns programas que facilitam esses procedimentos cobram assinaturas. Essas, mesmo com descontos, são muitas vezes inacessíveis.

As orientações aqui expostas não eximem que o interessado busque explicações para além do presente material. Pelo contrário, é preponderante complementar o estudo a partir dos manuais dos programas e de outras formas. Apesar disso, nosso texto indica uma sequência lógica que pode ser seguida ao traçar algumas diretrizes que abreviam o caminho ao indicar um percurso testado. Já não é preciso partir do zero, permitindo que os analistas, principalmente os iniciantes, ganhem tempo para explorar as funcionalidades que mais se adequam aos seus problemas de pesquisa.

Para além, insere a metodologia dentro de uma justificativa teórica coerente com as discussões realizadas pelos defensores das Humanidades

Digitais, que propõem a integração das facilidades computacionais às necessidades de descrições e respostas mais robustas detalhadas em um universo de rápida e massiva produção de conteúdo.

REFERÊNCIAS

ALVES, Daniel. As Humanidades Digitais como uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português. **Ler História**, Lisboa, n. 69, p. 91–103, dez. 2016. Disponível em: https://journals.openedition.org/lerhistoria/2496. Acesso em: 08 jul. 2025.

AMOSSY, Ruth. **A Argumentação no Discurso**. São Paulo: Contexto, 2018a.

AMOSSY, Ruth. O ethos na intersecção das disciplinas: retórica, pragmática, sociologia dos campos. *In*: AMOSSY, Ruth. **A Argumentação no Discurso**. São Paulo: Contexto, 2018b.

BIBER, Douglas. Representativeness in Corpus Design. **Literary and Linguistic Computing**, [*S.l.*], v. 8, n.4, p. 243-257, out. 1993. DOI: http://dx.doi.org/10.1093/llc/8.4.243. Disponível em: https://academic.oup.com/dsh/article-abstract/8/4/243/928942.

CHARAUDEAU, Patrick. **Discurso político**. São Paulo: Contexto, 2015.

GONÇALVES, Julia de Souza Borba. TUTORIAL ANTCONC – software para a realização de análises qualitativas. **Documentos LANTRI**. Franca, n.1, set. 2016. Disponível em: https://www.lantri.org/documentoslantri. Acesso em: 08 jul. 2025.

JURUÁ, Mayra. A pesquisa em ciências humanas, sociais aplicadas, linguística, letras e artes (CHSSALLA), algoritmos, educação e cultura: elementos para as humanidades digitais brasileiras. *In*: COSTA, Leonardo; ROCHA, Renata (orgs.). **Cultura e Ciência de Dados**. Salvador: EduFBA, 2021.

KÄSPER, Marge; MAURER, Liina. Starting Points in French Discourse Analysis Lexicometry to Study Political Tweets. **Digital Humanities in the Nordic & Baltic Countries**. Tartu: University of Tartu, College of Foreing Languages and Cultures, p.379-387, 2020. Disponível em: https://ceur-ws.org/Vol-2612/poster2.pdf. Acesso em 08 jul. 2025.

LINDGREN, Simon. Introducing Connected Concept Analysis: a network approach to big text datasets. **Text&Talk**, Alemanha, v. 36, n. 3, p. 341–362, 2016. DOI: https://doi.org/10.1515/text-2016-0016. Disponível em: https://www.degruyter.com/document/doi/10.1515/text-2016-0016/html. Acesso em: 08 jul. 2025.

MOREIRA, João Mendes; CARVALHO, André Carlos Ponce de Leon Ferreira, HORVÁTH, Thomáš. **A general Introduction to Data Analytics**. Hoboken, NJ: John Wiley & Sons, 2019.

NICHOLS, Bruno; KLEINA, Nilton; MARIOTTO, Djiovanni J. F.; SAMPAIO, Rafael. CPI do Circo ou CPI do Fim do Mundo? A guerra de narrativas sobre a Comissão Parlamentar de Inquérito da COVID-19 no YouTube. *In:* ENCONTRO ANUAL DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM CIÊNCIAS SOCIAIS., 45, 2021, [*S.l.*], **Anais** [...], [*S.l.*]: AMPOCS, 2021. p. 1-32. Disponível em: https://biblioteca.sophia.com.br/terminal/9666/VisualizadorPdf?codigoArquivo=590&tipoMidia=0. Acesso em: 08 jul. 2025.

ORLANDI, Eni Puccinelli. **Análise de discurso**: princípios & procedimentos. Campinas: Pontes Editores, 12ª ed., 2015.

PÊCHEUX, Michel. **O discurso**: estrutura ou acontecimento. Campinas: Pontes, 1990.

RAMOS, Maurivan Güntzel; LIMA, Valderez Marina do Rosário; AMARAL-ROSA, Marcelo Prado. Contribuições do software IRAMUTEQ para a Análise Textual Discursiva. In: CONGRESSO IBERO-AMERICANO EM INVESTIGAÇÃO QUALITATIVA. 7. 2018. Fortaleza. **Anais** [...] Fortaleza: Unifor, 2018, p. 505-514. Disponível em: https://repositorio.pucrs.br/dspace/bitstream/10923/14665/2/Contribuicoes_do_software_IRAMUTEQ_para_a_Analise_Textual_Discursiva.pdf. Acesso em: 08 jul. 2025.

RECUERO, Raquel. **Introdução à análise de redes sociais**. Salvador: Edufba, 2017.

RODRIGUES, Cristiano Magrini. **Campo jornalístico na pós-verdade:** estratégias argumentativas de um jornalismo político desafiado pela desinformação. 2023. 283f. Tese (Doutorado em Comunicação) – Universidade Federal de Santa Maria, Santa Maria, 2023. Disponível em: https://repositorio.ufsm.br/handle/1/29667. Acesso em: 08 jul. 2025.

SALVIATI, Maria Elisabeth. **Manual do Aplicativo Iramuteq** (versão 0.7 Alpha 2 e R 3.2.3). Planaltina, Embrapa Cerrados, 2017. Disponível em: http://www.iramuteq.org/documentation/fichiers/manual-do-aplicativo-iramuteq-par-maria-elisabeth-salviati. Acesso em: 08 jul. 2025.

SARDINHA, Tony Beber. Linguística de Corpus: histórico e problemática. **Delta**: documentação de estudos em linguística teórica e aplicada. São Paulo. v. 16, n. 2, p. 323-367, 2000, Disponível em: https://www.scielo.br/j/delta/a/vGknQkZQGsGYbrQfKmTZY4s/?format=pdf&lang=pt. Acesso em 08 jul. 2025.

SCHREIBMAN, Susan; SIEMENS, Ray; UNSWORTH, John. **A new companion to Digital Humanities**. West Sussex: John Wiley & Sons, 2016.

VENTURINI, Tommaso; MUNK, Anders; JACOMY, Mathieu. Ator-rede versus Análise de Redes versus Redes Digitais: falamos das mesmas redes? **Galáxia**, São Paulo, n. 38, p. 5-27, maio/ago., 2018. Disponível em: https://revistas .pucsp.br/index.php/galaxia/article/view/36645. Acesso em 08 jul. 2025.

WORDIJ (Semantic Network Tools). **About: Conceptualizing Semantic Networks**. c2025.Disponível em: https://www.wordij.net/about.html. Acesso em: 08 jul. 2025.

YUAN, Elaine; FENG, Miao; DANOWSKI, James. “Privacy” in Semantic Networks on Chinese Social Media: The Case of Sina Weibo. **Journal of Communication**, Oxford, v. 63, n. 6, p. 1011-1031, 2013. Disponível em: https://www.researchgate.net/publication/259552367_Privacy_in_Semantic_Networks_on_Chinese_Social_Media_The_Case_of_Sina_Weibo. Acesso em 08 jul. 2025.

COMMUNICATIONS IN THE CONTEXT OF DIGITAL HUMANITIES: FREE COMPUTER-OPERATED TOOLS AS A COMPLEMENT TO TRADITIONAL METHODOLOGIES

ABSTRACT
Objective: The paper presents a methodology for data processing in Communications research that combines Argumentative Discursive Analysis with Connected Concept Analysis through computer software from the perspective of Corpus Linguistics. **Methodology**: Developed during a doctoral degree, the protocol offers the possibility of expanding discursive analysis to a voluminous textual set without abandoning the interpretative characteristic of DA. The methodological scheme is situated within the context of Digital Humanities, an approach that combines the power of computational analysis with research in the Social Sciences and Humanities. **Conclusions**: The described procedure offers a free and replicable path for researchers who need to work with large amounts of text, providing guidelines and detailing functionalities in contexts of intense content production.

Descriptors: Digital Humanities. Research methodology. Data processing. Free tools.

LA COMUNICACIÓN EN EL CONTEXTO DE LAS HUMANIDADES DIGITALES: HERRAMIENTAS INFORMÁTICAS GRATUITAS COMO COMPLEMENTO A LAS METODOLOGÍAS TRADICIONALES

RESUMEN
Objetivo: El texto presenta una propuesta de metodología para el procesamiento de datos en Comunicación que combina el Análisis Discursivo Argumentativo con el Connected Comcept Analysis a través de software informático desde la perspectiva de la Lingüística de Corpus. **Metodología**: El protocolo, desarrollado en una tesis doctoral, presenta la posibilidad de ampliar el análisis del discurso a un conjunto textual voluminoso sin abandonar la característica interpretativa del AD. El esquema metodológico se inserta en el contexto de las Humanidades Digitales, enfoque que combina el poder de la informática con la investigación en las áreas de las Ciencias Sociales y Humanidades. **Conclusiones**: El procedimiento descrito apunta a un camino gratuito y replicable para investigadores que necesitan trabajar con una gran cantidad de textos, mostrando pautas y detallando funcionalidades en contextos de intensa producción de contenidos.

Descriptores: Humanidades Digitales. Metodología de investigación. Procesamiento de datos. Herramientas libres.

Recebido em: 27.07.2024
Aceito em: 26.06.2025