INFINITUM
ISSN 2595-9549
v. 8, n. 16, 1-22, 2025
DOI: https://doi.org/10.18764/2595-9549v8n16e26166

A inteligência artificial e as novas fronteiras da cartografia linguística

Antonio Carlos Santana de Souza
Instituição: UEMS/ UNEMAT
E-mail: acsantan@uems.br
ORCID: https://orcid.org/0000-0003-4757-635X

Nágila Kelli Prado Sana
Instituição: UEMS/ UNEMAT
E-mail: nagilakelli@gmail.com
ORCID: https://orcid.org/0000-0002-3146-3986

Resumo: A inteligência artificial (IA) está abrindo novas perspectivas para a cartografia linguística, permitindo uma análise mais profunda e precisa da distribuição geográfica e da variação das línguas. Este artigo explora as principais tendências e desafios da aplicação da IA nesse campo, destacando seu potencial para revolucionar a forma como estudamos e entendemos a linguagem humana. Apresentamos exemplos de projetos de pesquisa que utilizam técnicas de IA para analisar grandes corpora textuais, identificar padrões de variação e criar mapas interativos. Por fim, discutimos as perspectivas futuras para a pesquisa, explorando como a IA pode contribuir para a criação de ferramentas de análise linguística mais sofisticadas, a preservação da diversidade linguística e o desenvolvimento de tecnologias para a comunicação interlingual.

Palavras-chave: Dialetologia, Corpus linguístico, Mineração de dados.

Artificial intelligence and the new frontiers of linguistic cartography

Abstract: Artificial intelligence (AI) is opening up new horizons for linguistic cartography, enabling a deeper and more accurate analysis of the geographic distribution and variation of languages. This article explores the main trends and challenges of applying AI in this field, highlighting its potential to revolutionize how we study and understand human language. We present examples of research projects that use AI techniques to analyze large text corpora, identify variation patterns, and create interactive maps. Finally, we discuss future research perspectives, exploring how AI can contribute to the creation

of more sophisticated linguistic analysis tools, the preservation of linguistic diversity, and the development of technologies for intercultural communication.

Keywords: Dialectology, linguistic corpus, data mining.

Inteligencia artificial y las nuevas fronteras de la cartografia lingüística

Resumen: La inteligencia artificial (IA) está abriendo nuevas perspectivas para la cartografía lingüística, permitiendo un análisis más profundo y preciso de la distribución geográfica y de la variación de las lenguas. Este artículo explora las principales tendencias y desafíos de la aplicación de la IA en este campo, destacando su potencial para revolucionar la forma en que estudiamos y entendemos el lenguaje humano. Presentamos ejemplos de proyectos de investigación que utilizan técnicas de IA para analizar grandes corpus textuales, identificar patrones de variación y crear mapas interactivos. Finalmente, discutimos las perspectivas futuras para la investigación, explorando cómo la IA puede contribuir a la creación de herramientas de análisis lingüístico más sofisticadas, la preservación de la diversidad lingüística y el desarrollo de tecnologías para la comunicación interlingüística.

Palabras clave: Dialectología, Corpus lingüístico, Minería de datos.

INTRODUÇÃO

A cartografia linguística, tradicionalmente associada a métodos manuais e demorados, está experimentando um grande avanço impulsionado pela inteligência artificial (IA). A geolinguística, campo que se dedica ao estudo da distribuição geográfica das línguas e de suas variedades, encontra na IA uma ferramenta poderosa para aprofundar suas investigações. Ao integrar técnicas de processamento de linguagem natural (PLN), a cartografia linguística expande seus limites, permitindo a análise e a visualização de vastos corpora linguísticos de forma mais rápida e precisa. Como apontam Biber et al. (1994), a identificação de padrões e tendências em grandes conjuntos de dados linguísticos exige ferramentas computacionais sofisticadas. A IA, com sua capacidade de processar e analisar grandes volumes de dados textuais, oferece um arsenal de ferramentas para a cartografia linguística, como algoritmos de aprendizado de máquina e técnicas de mineração de texto.

Essa sinergia entre a IA e a cartografia linguística possibilita a criação de mapas interativos e dinâmicos que visualizam a distribuição geográfica de variantes linguísticas com um nível de detalhe antes inimaginável. Conforme proposto por Gooskens (2007), a visualização cartográfica, aliada à precisão da IA, permite identificar e analisar nuances linguísticas em diferentes escalas geográficas, desde dialetos regionais até variações sociais. Silva (2015) corrobora essa ideia, destacando o potencial da IA para revelar padrões complexos e dinâmicos na linguagem, que seriam difíceis de detectar por meio de métodos tradicionais.

Além disso, a IA permite a análise de fenômenos linguísticos em tempo real, possibilitando o acompanhamento da evolução das línguas e a identificação de novas tendências. Essa capacidade é fundamental para entender os impactos das mudanças sociais, tecnológicas e culturais na linguagem, como a influência das mídias sociais e da globalização.

Em suma, a IA está transformando a cartografia linguística, oferecendo novas ferramentas e perspectivas para o estudo da diversidade linguística. Ao combinar a precisão da análise computacional com a visualização cartográfica, a IA permite uma compreensão mais profunda e abrangente dos fenômenos linguísticos, contribuindo para o avanço do conhecimento em diversas áreas, como linguística, sociolinguística e antropologia linguística.

CONCEITOS FUNDAMENTAIS DA CARTOGRAFIA LINGUÍSTICA COM IA

A cartografia linguística, em sua complexidade, transcende a mera representação geográfica das línguas, configurando-se como um campo de estudo interdisciplinar que investiga a intrincada relação entre linguagem, espaço e sociedade; o cerne dessa investigação reside no mapeamento sistemático da distribuição geográfica das línguas e suas variedades, um processo que demanda uma

abordagem rigorosa e interdisciplinar, uma vez que a análise da variação linguística emerge como um pilar central, investigando como as línguas se transformam e se diversificam em diferentes contextos geográficos, revelando a influência de fatores como isolamento geográfico, contato linguístico, migrações e fronteiras políticas, permitindo a identificação de dialetos, sotaques e outras formas de variação linguística, mapeando suas características e distribuição; ademais, a cartografia linguística também se preocupa em compreender como fatores históricos e sociais moldam a distribuição das línguas, investigando como eventos históricos, como colonização, guerras e migrações, influenciaram a disseminação de línguas e dialetos, e analisando como fatores sociais, como classe social, etnia, gênero e religião, se relacionam com a variação linguística, revelando a intrincada teia de relações entre linguagem e sociedade.

A geolinguística, em sua complexidade, aprofunda-se na investigação da intrincada relação entre língua e espaço, buscando compreender como o ambiente físico, as características geográficas e a organização espacial das sociedades influenciam a variação e a distribuição das línguas. Nesse contexto, a geolinguística analisa como o isolamento geográfico, como montanhas, rios e oceanos, pode levar ao desenvolvimento de dialetos e línguas distintas, bem como o contato linguístico entre diferentes comunidades, facilitado por rotas comerciais, migrações e fronteiras, molda a variação linguística, e explora como a organização espacial das sociedades, como a distribuição de cidades, vilas e áreas rurais, se relaciona com a variação linguística, investigando como a proximidade geográfica entre comunidades pode levar à convergência linguística, enquanto o distanciamento geográfico pode resultar em divergência linguística.

A cartografia linguística e a geolinguística, embora distintas em seus focos específicos, convergem em um objetivo comum: desvendar a intrincada relação entre língua, espaço e sociedade. A cartografia linguística, ao mapear a distribuição

geográfica das línguas e suas variedades, revela como fatores geográficos, históricos e sociais moldam a variação linguística. Já a geolinguística, ao analisar a influência do espaço geográfico na língua, investiga como o ambiente físico e a organização espacial das sociedades afetam a distribuição e a variação linguística. Ambas as disciplinas, portanto, se complementam e se enriquecem mutuamente.

No contexto brasileiro, a geolinguística e a cartografia linguística têm desempenhado um papel fundamental na documentação e análise da rica diversidade linguística do país. Projetos como o Atlas Linguístico do Brasil (ALiB) têm mapeado a distribuição geográfica de dialetos e variedades linguísticas em todo o território nacional, revelando a complexidade e a riqueza da variação linguística brasileira.

Além disso, estudos como os de Ataliba Teixeira de Castilho (2010) e Rosa Virgínia Mattos e Silva (2001) têm explorado a influência de fatores históricos e sociais na formação e distribuição de dialetos e variedades linguísticas no Brasil. Esses estudos têm revelado como a colonização, a imigração e a urbanização moldaram a paisagem linguística brasileira, resultando em uma rica tapeçaria de línguas e dialetos.

A geolinguística também tem se dedicado a investigar a relação entre a língua e o espaço geográfico no Brasil. Estudos como os de Aparecida Negri Isquerdo (2007) têm explorado como o isolamento geográfico e o contato linguístico entre diferentes comunidades moldam a variação linguística em diferentes regiões do país. Esses estudos têm revelado como a geografia brasileira, com suas vastas extensões territoriais e diversidade de paisagens, contribui para a formação de dialetos e variedades linguísticas distintas.

Em suma, a cartografia linguística e a geolinguística, ao mapear a distribuição geográfica das línguas e analisar a influência de fatores geográficos, históricos e sociais na variação linguística, nos permitem desvendar os mistérios da comunicação humana e celebrar a rica diversidade linguística do nosso planeta. No Brasil, essas disciplinas têm desempenhado um papel fundamental na documentação

e análise da rica diversidade linguística do país, contribuindo para a preservação e valorização do patrimônio linguístico brasileiro.

A possibilidade de interseção entre os estudos geolinguísticos, cartográficos e a Inteligência Artificial (IA) tem inaugurado uma nova era na análise da relação entre língua e espaço. A cartografia linguística, ao mapear a distribuição geográfica das línguas, e a geolinguística, ao investigar a influência do espaço na língua, encontram na IA e no Processamento de Linguagem Natural (PLN) ferramentas poderosas para aprofundar suas análises. A IA, com sua capacidade de processar grandes volumes de dados, permite identificar padrões complexos de distribuição e variação linguística, revelando nuances que escapam aos métodos tradicionais. O PLN, por sua vez, possibilita a análise automatizada de textos e documentos em diferentes línguas e dialetos, abrindo caminho para a criação de mapas linguísticos mais precisos e para a identificação de áreas de contato linguístico, evolução linguística e padrões subjacentes nos dados. No contexto brasileiro, essa interseção se mostra especialmente promissora, dada a rica diversidade linguística do país. Projetos como o Atlas Linguístico do Brasil (ALiB), combinados com técnicas de IA e PLN, podem revelar a complexidade da variação linguística brasileira com um nível de detalhe sem precedentes, auxiliando na preservação e valorização do patrimônio linguístico nacional.

POSSIBILIDADES DE ESTUDOS GEOLINGUÍSTICOS POR INTERMÉDIO DA IA:

A aplicação de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina torna possível a identificação de Áreas de Contato Linguístico: Ao analisar grandes corpora, é possível identificar com precisão as regiões onde diferentes variedades linguísticas entram em contato, dando origem a fenômenos

como diglossia, *code-switching* e criação de *pidgins* e *creoles.* A IA permite modelar a distribuição geográfica desses fenômenos, revelando as dinâmicas complexas que moldam as paisagens linguísticas.

No que concerne a Análise da Evolução Linguística: A capacidade da IA de processar grandes volumes de dados textuais ao longo do tempo permite rastrear a evolução das línguas, identificando mudanças lexicais, sintáticas e semânticas. É possível, por exemplo, analisar a difusão de neologismos, a mudança de significado das palavras e a simplificação ou complexificação das estruturas gramaticais.

Também, é possível analisar a descoberta de padrões subjacentes: Através de técnicas de aprendizado de máquina, como redes neurais artificiais e modelos de tópicos, é possível identificar padrões subjacentes nos dados linguísticos, revelando relações entre a linguagem e outros fatores, como o contexto social, histórico e cultural. Por exemplo, é possível identificar correlações entre a variação linguística e fatores como urbanização, escolaridade e migração.

As aplicações práticas da cartografia linguística, geolinguística, IA e PLN se estendem por um leque diversificado de áreas, cada uma com o potencial de aprofundar nossa compreensão da linguagem e sua relação com o mundo. Na sociolinguística, essas ferramentas permitem analisar como a variação linguística se entrelaça com fatores sociais como classe, gênero e etnia, revelando as nuances da comunicação humana. A dialetologia, por sua vez, se beneficia da capacidade de mapear e classificar dialetos, traçando isoglossas e analisando a distribuição geográfica das características linguísticas, desvendando a rica tapeçaria da diversidade linguística regional.

A confluência entre a história da língua, a linguística computacional e as humanidades digitais está redefinindo os contornos da pesquisa linguística, inaugurando uma era de descobertas e aplicações sem precedentes; a capacidade de reconstruir o passado linguístico, identificando origens e relações entre línguas e

dialetos, ganha novas dimensões com o advento de ferramentas computacionais avançadas, lançando luz sobre a evolução da comunicação humana e suas intrincadas ramificações; simultaneamente, a linguística computacional, impulsionada pelo desenvolvimento de tradutores automáticos e sistemas de reconhecimento de fala, expande as fronteiras da interação homem-máquina, abrindo caminho para novas formas de comunicação e acessibilidade; a preservação de línguas minoritárias, em risco de extinção, encontra um aliado poderoso na documentação e revitalização linguística, com a criação de corpora digitais e ferramentas de análise que permitem a salvaguarda de patrimônios linguísticos e culturais inestimáveis; finalmente, as humanidades digitais, ao integrarem a linguística com outras áreas do conhecimento, como história, antropologia e geografia, criam um mosaico interdisciplinar que enriquece nossa compreensão da cultura e da sociedade, revelando as complexas interconexões entre linguagem, história, cultura e espaço.

METODOLOGIA

A metodologia de pesquisa em cartografia linguística e geolinguística, outrora predominantemente dependente de coleta e análise manual de dados, está experimentando uma transformação paradigmática com a ascensão da Inteligência Artificial (IA). A capacidade da IA de processar vastos conjuntos de dados textuais e geográficos, aliada a algoritmos sofisticados de aprendizado de máquina, oferece um arsenal de ferramentas que expandem significativamente o escopo e a profundidade da investigação. A automação da coleta de dados, por meio de técnicas como web scraping e APIs de mapas, permite a criação de corpora linguísticos representativos e a geolocalização precisa de dados textuais, superando as limitações dos métodos tradicionais. O Processamento de Linguagem Natural (PLN) e o aprendizado profundo possibilitam a identificação de padrões complexos, nuances dialetais e

variações linguísticas com uma precisão e granularidade antes inalcançáveis, revelando a intrincada relação entre língua, espaço e sociedade.

No entanto, a integração da IA na pesquisa geolinguística e cartográfica não implica a substituição da expertise do pesquisador. A interpretação dos resultados gerados pela IA exige um profundo conhecimento da teoria linguística, da metodologia de pesquisa e do contexto sociocultural em que os dados foram produzidos. O papel do pesquisador é fundamental para formular as perguntas de pesquisa, selecionar os dados adequados, escolher as técnicas de análise mais apropriadas e interpretar os resultados de forma crítica e informada. Estudos que utilizam modelos de aprendizado de máquina para identificar áreas de contato linguístico e analisar a influência de fatores socioculturais na variação linguística (cf. Moretti, 2013) e aqueles que empregam técnicas de PLN para extrair informações semânticas e pragmáticas de textos históricos (cf. Kochmar, 2015) demonstram a necessidade de combinar a capacidade computacional da IA com a expertise do pesquisador para obter resultados significativos e confiáveis.

A aplicação dessa metodologia transcende os limites da pesquisa acadêmica, permeando diversos setores com potencial transformador. A preservação de línguas minoritárias, a personalização do aprendizado, o marketing direcionado, o planejamento urbano inclusivo e a reconstrução da história da língua são apenas alguns exemplos das áreas que se beneficiam da análise de dados linguísticos em larga escala. A implementação dessas aplicações é facilitada por um conjunto de ferramentas e plataformas de código aberto e amplamente utilizadas na comunidade de pesquisa, como Python, R, Google Colaboratory, QGIS e Gephi, que oferecem um ambiente versátil e acessível para a análise de dados linguísticos e geográficos.

No entanto, é fundamental reconhecer as limitações e desafios inerentes a essa metodologia. A dependência de dados digitais, a dificuldade em capturar nuances semânticas e pragmáticas, a necessidade de grandes volumes de dados rotulados e a

potencial introdução de vieses são apenas alguns dos obstáculos que exigem uma abordagem crítica e informada. A acessibilidade e a usabilidade das ferramentas também devem ser consideradas ao planejar e executar pesquisas, garantindo que a metodologia seja aplicada de forma ética e responsável, com o objetivo de promover a compreensão da diversidade linguística e a inclusão social.

Tradicionalmente, a geolinguística se baseava em métodos de coleta e análise de dados predominantemente manuais. A IA automatiza e amplia significativamente essas tarefas, permitindo a análise de corpora linguísticos de dimensões antes inimagináveis. Técnicas como o processamento de linguagem natural (PLN), o aprendizado profundo e a mineração de textos possibilitam identificar padrões complexos, nuances dialetais e variações linguísticas com uma precisão e profundidade antes inalcançáveis.

No entanto, é importante ressaltar que a IA não substitui a expertise do pesquisador. A interpretação dos resultados gerados pela IA requer um profundo conhecimento da teoria linguística e da metodologia de pesquisa. O papel do pesquisador é fundamental para formular as perguntas de pesquisa, selecionar os dados adequados, escolher as técnicas de análise mais apropriadas e interpretar os resultados de forma crítica.

A cartografia linguística e geolinguística tem experimentado uma transformação significativa com a integração de ferramentas da Inteligência Artificial (IA). A coleta de dados, por exemplo, agora se beneficia do *web scraping*, que extrai automaticamente informações de diversas fontes online, como redes sociais e fóruns, criando corpora textuais representativos de diferentes comunidades linguísticas. As APIs de mapas, fornecidas por plataformas como *Google Maps* e *OpenStreetMap*, permitem a geolocalização de textos e a análise da distribuição espacial de características linguísticas. Além disso, sensores de linguagem em dispositivos móveis

e dados de geolocalização em redes sociais oferecem a oportunidade de investigar a variação linguística em tempo real.

O processamento desses dados é realizado por meio do Processamento de Linguagem Natural (PLN), que inclui técnicas como tokenização, identificação de classes gramaticais, reconhecimento de entidades nomeadas, análise de sentimentos e análise de discurso. O aprendizado de máquina, com suas capacidades de classificação, clusterização e redes neurais, também desempenha um papel fundamental na identificação de padrões complexos na linguagem.

Na fase de análise, a visualização de dados cria mapas temáticos e gráficos para representar a distribuição geográfica de características linguísticas, enquanto a análise estatística e a geostatística identificam correlações e padrões espaciais. A análise de redes sociais, por sua vez, revela comunidades linguísticas e a disseminação de informações.

A aplicação da metodologia delineada anteriormente transcende os limites da pesquisa acadêmica, permeando diversos setores com potencial transformador. A preservação de línguas minoritárias, por exemplo, é impulsionada pela documentação e análise de línguas em risco de extinção, permitindo a salvaguarda de patrimônios culturais imateriais. Na educação, a personalização do aprendizado e a avaliação precisa são facilitadas pelo desenvolvimento de materiais didáticos e ferramentas de avaliação baseadas em dados linguísticos. O marketing, por sua vez, se beneficia da análise de sentimentos em redes sociais, monitorando a reputação de marcas e produtos em tempo real. No planejamento urbano, a promoção da inclusão linguística e o desenvolvimento de políticas linguísticas sensíveis à diversidade são viabilizados pela análise da distribuição espacial de línguas e dialetos. A história, por fim, ganha novas camadas de profundidade com a reconstrução da evolução histórica da língua a partir da análise de corpora textuais históricos.

A implementação dessas aplicações é facilitada por um conjunto de ferramentas e plataformas de código aberto e amplamente utilizadas na comunidade de pesquisa. Python, com suas bibliotecas robustas como NLTK, spaCy e scikit-learn, oferece um ambiente versátil para análise de dados e Processamento de Linguagem Natural (PLN). R, por sua vez, destaca-se na análise estatística e visualização de dados. Google Colaboratory fornece um ambiente de desenvolvimento em nuvem acessível para a execução de código Python. QGIS, um Sistema de Informação Geográfica (SIG) de código aberto, permite o geoprocessamento de dados linguísticos e a criação de mapas temáticos. Gephi, por fim, possibilita a visualização e análise de redes complexas, como redes de interação linguística em mídias sociais.

Embora o conjunto de ferramentas e plataformas mencionadas ofereça um potencial significativo para a pesquisa em cartografia linguística e geolinguística, é fundamental reconhecer as limitações e desafios inerentes a essa metodologia. A dependência de dados digitais, por exemplo, pode introduzir vieses, uma vez que nem todas as comunidades linguísticas têm igual representação online. A coleta de dados por meio de *web scraping* e APIs de mapas pode negligenciar dialetos regionais ou línguas minoritárias, resultando em análises incompletas.

O Processamento de Linguagem Natural (PLN), apesar de seus avanços, ainda enfrenta dificuldades em capturar nuances semânticas e pragmáticas, especialmente em línguas com poucos recursos ou dialetos informais. A análise de sentimentos, por exemplo, pode ser influenciada por ironia, sarcasmo ou contexto cultural, gerando resultados imprecisos. O aprendizado de máquina, por sua vez, requer grandes volumes de dados rotulados, que nem sempre estão disponíveis para todas as línguas e dialetos.

A visualização de dados, embora útil, pode ser enganosa se não considerar a distribuição espacial desigual de dados linguísticos. A análise estatística e a geostatística, por sua vez, exigem expertise para evitar interpretações errôneas de

correlações e padrões espaciais. A análise de redes sociais, por fim, pode reproduzir vieses existentes nas interações online, como a exclusão de grupos minoritários ou a amplificação de vozes dominantes.

Além disso, a acessibilidade e a usabilidade das ferramentas mencionadas podem variar. Python, por exemplo, exige conhecimento de programação, o que pode limitar sua utilização por pesquisadores com pouca experiência em desenvolvimento de software. R, por sua vez, exige familiaridade com estatística e programação. QGIS e Gephi, embora sejam ferramentas poderosas, podem ter curvas de aprendizado íngremes.

Em resumo, a metodologia delineada oferece um conjunto de ferramentas valiosas para a pesquisa em cartografia linguística e geolinguística, mas exige uma abordagem crítica e informada. É fundamental reconhecer as limitações dos dados digitais, os desafios do PLN e do aprendizado de máquina, e a necessidade de expertise em análise estatística e visualização de dados. Além disso, a acessibilidade e a usabilidade das ferramentas devem ser consideradas ao planejar e executar pesquisas.

RESULTADOS

A crescente integração da Inteligência Artificial (IA) na cartografia linguística tem provocado um debate significativo na comunidade acadêmica, centrando-se no equilíbrio entre o potencial revolucionário da tecnologia e as complexidades inerentes à análise da linguagem humana. Embora a IA ofereça ferramentas poderosas para a análise de dados linguísticos em larga escala, é imperativo evitar uma postura excessivamente otimista, reconhecendo as limitações e os desafios que acompanham sua aplicação. Projetos ambiciosos, como o "Atlas of

Language Structures", [1]ilustram a capacidade da IA de processar vastos conjuntos de dados textuais e identificar padrões universais na linguagem, um feito que seria impraticável com métodos tradicionais.

No entanto, a interpretação desses padrões exige um profundo conhecimento da teoria linguística e da metodologia de pesquisa, uma vez que, como salientam Bender et al. (2021), os modelos de linguagem não "entendem" a linguagem da mesma forma que os humanos. A análise de dados linguísticos, portanto, requer uma abordagem crítica e informada, que combine a capacidade computacional da IA com a expertise do pesquisador. A IA pode auxiliar na identificação de recorrências, estruturas e relações sutis dentro de grandes volumes de texto, mas a atribuição de significado a esses padrões é uma tarefa complexa que exige a expertise do pesquisador. A análise de dados linguísticos envolve a consideração de fatores contextuais, sociais e culturais que nem sempre são capturados pelos algoritmos de IA. Além disso, a coleta e o processamento de dados linguísticos em larga escala levantam questões éticas importantes. Como alertam Blommaert e Rampton (2011), a linguagem online é um campo minado de desigualdades e preconceitos, e os algoritmos de IA podem perpetuar e amplificar esses vieses. É crucial garantir a neutralidade e a transparência nos algoritmos utilizados na cartografia linguística, bem como proteger a privacidade e os direitos autorais dos dados linguísticos.

O "Dialectometry", [2]por sua vez, demonstra a eficácia da IA no mapeamento de dialetos com uma precisão geográfica notável. Ao empregar algoritmos sofisticados, o projeto desvenda as nuances e fronteiras linguísticas com um nível de

[1] Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013.WALS Online (v2020.4) [Conjunto de dados]. Zenodo. https://doi.org/10.5281/zenodo.13950591(Disponível online em https://wals.info, acessado em 22/03/2025.)

[2] A dialetometria, área da linguística dedicada à análise quantitativa da variação dialetal, emprega métodos estatísticos e computacionais para medir e representar as diferenças linguísticas entre dialetos, buscando quantificar a distância linguística, mapear a variação dialetal, identificar áreas de transição e analisar a influência de fatores geográficos, históricos e sociais, conforme explicitado em Goebl (1998), Nerbonne e Heeringa (2005) e Brissos e Saramago (2019).

detalhe que antes era inatingível, revelando a complexidade da variação linguística dentro de um determinado território. Essa capacidade de mapeamento preciso não apenas enriquece a nossa compreensão da distribuição geográfica dos dialetos, mas também oferece *insights* valiosos sobre os fatores socioculturais e históricos que moldam a variação linguística regional.

A criação do "Glottolog", [3]uma base de dados global de línguas, representa um marco na organização e disseminação de informações linguísticas em larga escala. A IA desempenha um papel fundamental na estruturação e atualização dessa base de dados, facilitando o acesso a informações sobre a história, classificação e status de milhares de línguas em todo o mundo. Esse recurso inestimável não apenas beneficia pesquisadores e comunidades linguísticas, mas também contribui para a preservação e revitalização de línguas ameaçadas.

A IA também se destaca como uma ferramenta crucial na documentação e preservação de línguas ameaçadas, um imperativo ético e cultural que ganha urgência diante da crescente globalização e homogeneização linguística. A UNESCO, por meio do "Language Atlas", [4]utiliza a IA para mapear a distribuição geográfica dessas línguas, identificar fatores de risco e desenvolver estratégias de preservação. O "Endangered Languages Project", [5]da Universidade de Harvard, complementa esse

[3] Glottolog, uma iniciativa do Instituto Max Planck de Antropologia Evolucionária, em Leipzig, fornece um catálogo abrangente das línguas do mundo, com foco especial nas menos conhecidas, através de sua seção de Referências ('langdoc'), que oferece uma vasta coleção de dados bibliográficos, incluindo gramáticas, dicionários, listas de palavras e textos, com critérios de pesquisa como autor, ano, título, país e afiliação genealógica, sendo continuamente expandido e aprimorado com a colaboração de seus usuários, e cujos dados são organizados no repositório público do GitHub glottolog/glottolog.

[4] O "Language Atlas" é uma ferramenta desenvolvida pela UNESCO com o objetivo de mapear e monitorar as línguas em risco de extinção em todo o mundo. Através de visualizações geográficas interativas e dados detalhados sobre a distribuição e o estado de vitalidade dessas línguas, o atlas busca conscientizar sobre a diversidade linguística e cultural do planeta, além de fornecer informações cruciais para a formulação de políticas linguísticas e ações de preservação. A ferramenta também se destaca por sua capacidade de integrar dados de diversas fontes, incluindo pesquisas acadêmicas, censos e informações de comunidades locais, permitindo uma análise abrangente e atualizada do panorama linguístico global.

[5] O "Endangered Languages Project" (Projeto de Línguas Ameaçadas) é uma iniciativa colaborativa online dedicada a documentar, preservar e revitalizar línguas em risco de extinção em todo o mundo. A plataforma, acessível em endangeredlanguages.com, funciona como um repositório interativo, onde comunidades, pesquisadores e entusiastas podem compartilhar informações e recursos sobre línguas ameaçadas.

esforço, analisando dados de mídias sociais e outras fontes online para documentar línguas em declínio. Essa capacidade de monitoramento e análise em tempo real permite uma resposta mais ágil e eficaz diante da ameaça de extinção linguística.

A análise da diversidade linguística global, realizada pela National Geographic Society, e o cálculo do "Linguistic Diversity Index", [6]pela Universidade de Zurich, fornecem dados valiosos para a formulação de políticas linguísticas. Ao mapear a distribuição geográfica das línguas e quantificar a diversidade linguística regional, esses projetos oferecem *insights* cruciais para a promoção da inclusão linguística e a proteção da diversidade cultural.

Estudos de dialetos, como o "Dialect Survey of the British Isles" [7]e o "Atlas Linguistique du Français contemporain"[8], exploram a variação linguística regional com profundidade, revelando padrões e influências socioculturais que moldam a comunicação humana. A análise da mudança linguística, impulsionada por projetos como o "Google Trends" [9]e "Diachronic Word Embeddings", [10]captura as dinâmicas da linguagem em tempo real, evidenciando o impacto de eventos sociais e tecnológicos na evolução da comunicação humana.

---

[6] https://worldpopulationreview.com/country-rankings/linguistic-diversity-index-by-country

[7] https://dialectandheritage.org.uk/about/the-survey-of-english-dialects/ Durante as décadas de 1950 e 1960, pesquisadores de campo do *Survey of English Dialects* viajaram pela Inglaterra, visitando mais de 300 cidades e vilas diferentes. Eles perguntaram às pessoas sobre suas palavras para objetos cotidianos e as fizeram falar sobre seus costumes locais, cultura e modo de vida. O resultado foram centenas de gravações de dialetos, cadernos, fotografias e desenhos, e uma visão fascinante da vida e da linguagem na Inglaterra rural.

[8] O "Atlas Linguistique du Français Contemporain" (ALFC) é um projeto de grande envergadura que se dedica a mapear e analisar as variações regionais do francês falado na França e em algumas regiões francófonas vizinhas.

[9] O Google Trends é uma ferramenta gratuita do Google que permite analisar a popularidade de termos de pesquisa no Google ao longo do tempo. Ele oferece insights valiosos sobre o que as pessoas estão procurando, revelando tendências e padrões de interesse. A ferramenta funciona analisando a frequência com que um termo é pesquisado no Google, exibindo um gráfico que ilustra a evolução do interesse ao longo do tempo. Além disso, permite comparar a popularidade de vários termos simultaneamente e fornece dados geográficos sobre a popularidade dos termos por região, cidade ou país, mostrando onde um determinado assunto é mais relevante.

[10] Diachronic Word Embeddings" representa uma inovadora intersecção entre a linguística diacrônica tradicional e o processamento de linguagem natural (PLN), oferecendo uma nova perspectiva para o estudo da mudança semântica. Historicamente, a linguística diacrônica tem se apoiado em análises qualitativas de textos para rastrear a evolução dos significados das palavras. "Diachronic Word Embeddings", por sua vez, complementa essa abordagem, fornecendo uma base quantitativa para medir e modelar essas mudanças, permitindo uma análise em larga escala, que seria impraticável com métodos tradicionais.

O mapeamento contínuo de paisagens linguísticas, facilitado por projetos como o "Endangered Languages Project" e a plataforma "Glottolog", permite acompanhar o surgimento de novas variedades, a difusão de palavras e expressões, e a extinção de outras. A análise de padrões e tendências linguísticas complexas, como a difusão de gírias e memes, é realizada por meio de ferramentas como o "Google Trends" e pesquisas como "Language Change in the Digital Age".[11]

A previsão do futuro das línguas, antes uma especulação, torna-se mais precisa com a IA. O "Atlas of Languages" e pesquisas como "The Future of English" [12]exploram os fatores que moldam o futuro da linguagem, auxiliando na tomada de decisões estratégicas relacionadas à educação, cultura e políticas linguísticas. A documentação e preservação de línguas em risco, um imperativo ético, são impulsionadas por projetos como o do "Center for Artificial Intelligence"[13], que fornece recursos para aprendizado, tradução e pesquisa.

No entanto, é crucial reconhecer os desafios éticos e sociais da aplicação da IA na cartografia linguística. A perpetuação de vieses algorítmicos, a violação da privacidade e do consentimento, e a negligência dos direitos autorais exigem atenção. O impacto nas comunidades linguísticas demanda um envolvimento ativo e respeitoso, garantindo que suas vozes sejam ouvidas e seus valores sejam respeitados.

Apesar desses desafios, as perspectivas futuras são promissoras. A cartografia linguística em tempo real, a análise avançada de dados, a visualização interativa, e a preservação de línguas em risco são exemplos do potencial da IA. A automação de

[11] "Mudança da Linguagem na Era Digital" explora a relação dinâmica entre a tecnologia e a linguagem, examinando como nossas interações digitais estão remodelando a forma como nos comunicamos.

[12] "The Future of English" não se refere a um projeto de IA específico, mas sim a um campo de estudo que explora as possíveis trajetórias da língua inglesa em um cenário global em constante mudança, considerando a influência da tecnologia, da globalização e da variação linguística. No entanto, a Inteligência Artificial (IA) desempenha um papel crescente na pesquisa e análise do futuro do inglês.

[13] O "Center for Artificial Intelligence" (Centro de Inteligência Artificial) abrange diversas iniciativas globais que aplicam IA na linguística, destacando-se o C4AI no Brasil, focado na preservação de línguas indígenas com recursos para aprendizado e tradução. A IA, em geral, auxilia na documentação, análise e revitalização de línguas ameaçadas, criando corpora digitais e ferramentas de tradução automática.

tarefas, a descoberta de novos padrões linguísticos e a criação de aplicações práticas também impulsionam a área.

Para aproveitar os benefícios da IA, são necessárias diretrizes éticas robustas, colaborações multidisciplinares entre linguistas, cientistas da computação e especialistas em ética, engajamento ativo das comunidades linguísticas e investimento em educação e capacitação. A qualidade dos dados, a interpretação dos resultados e a proteção da privacidade são desafios que exigem soluções inovadoras e responsáveis, garantindo que a IA seja utilizada de forma ética e equitativa na pesquisa e preservação da diversidade linguística.

CONSIDERAÇÕES FINAIS

A interseção entre a cartografia linguística e a Inteligência Artificial (IA) no Brasil configura um campo de estudo de notável potencial, dada a rica tapeçaria linguística que caracteriza o país. Ao explorarmos essa temática, buscamos impulsionar a disseminação do conhecimento sobre as possibilidades de pesquisa e aplicação da IA na análise e mapeamento das línguas brasileiras, visando a preservação do patrimônio linguístico nacional e o desenvolvimento de ferramentas e políticas linguísticas mais eficazes.

A integração da Inteligência Artificial (IA) na cartografia linguística inaugura uma nova era na pesquisa, transcendendo as limitações dos métodos tradicionais. A capacidade da IA de processar vastos conjuntos de dados linguísticos e geográficos, aliada à sofisticação algorítmica da Linguística Computacional, permite a automatização de tarefas complexas e a identificação de padrões linguísticos com uma precisão sem precedentes. Essa sinergia entre a expertise humana e a capacidade computacional da IA abre caminho para uma compreensão mais rica e abrangente da

diversidade linguística, impulsionando a pesquisa para novos patamares de eficiência e profundidade.

A aplicação da IA na cartografia linguística não se limita à aceleração de processos; ela redefine o próprio escopo da pesquisa. Ao desvendar nuances dialetais, variações sociolinguísticas e a evolução histórica da linguagem, a IA permite a análise de fenômenos que antes eram inacessíveis devido à escala e complexidade dos dados linguísticos (JURAFSKY; MARTIN, 2023). Essa capacidade de análise em larga escala tem implicações significativas para a preservação de línguas ameaçadas, o desenvolvimento de políticas linguísticas inclusivas e a promoção da justiça social, abrindo novas fronteiras para a pesquisa e a aplicação do conhecimento linguístico.

Acredita-se que a IA pode desempenhar um papel fundamental na documentação e revitalização de línguas ameaçadas, na análise da variação dialetal e na compreensão da intrincada relação entre língua e espaço geográfico no Brasil, promovendo a valorização da diversidade linguística e a inclusão social. A democratização do conhecimento linguístico, facilitada pelo uso da IA, torna acessível a um público mais amplo a análise de grandes volumes de dados linguísticos, antes restrita a especialistas, impulsionando a pesquisa e o aprendizado colaborativo.

A agilidade e a eficiência no processo de cartografia linguística são notavelmente aprimoradas pela IA. Algoritmos inteligentes automatizam tarefas repetitivas e demoradas, como a extração de dados de corpora textuais e a identificação de padrões linguísticos, liberando tempo para que pesquisadores se concentrem em análises mais complexas e na interpretação dos resultados.

A capacidade da IA de processar grandes volumes de dados permite aos linguistas desvendar fenômenos linguísticos intrincados, antes inacessíveis. A análise de dados provenientes de mídias sociais e plataformas online possibilita mapear o uso da linguagem em tempo real, revelando tendências linguísticas emergentes e dinâmicas socioculturais.

A preservação das línguas e a documentação dos padrões linguísticos de línguas ameaçadas de extinção encontram na IA um aliado poderoso. A tecnologia permite a criação de corpora digitais e ferramentas de análise que auxiliam na salvaguarda de patrimônios linguísticos e culturais inestimáveis.

No entanto, a integração da IA na cartografia linguística exige uma reflexão profunda sobre as implicações éticas e sociais que a acompanham. A capacidade da IA de processar vastos conjuntos de dados linguísticos e geográficos levanta questões cruciais sobre a responsabilidade na coleta, análise e interpretação desses dados.

É fundamental garantir que a aplicação da IA seja feita de forma responsável e equitativa, evitando a perpetuação de vieses algorítmicos e a exclusão de grupos minoritários. A ética na coleta de dados, por exemplo, exige o respeito à privacidade e ao consentimento das comunidades linguísticas, especialmente aquelas que são vulneráveis ou em risco de extinção.

A análise e interpretação dos dados devem ser transparentes e rigorosas, evitando a criação de representações distorcidas da realidade linguística. Além disso, é preciso considerar o impacto social das pesquisas e garantir que os benefícios da IA sejam compartilhados de forma justa e equitativa, desenvolvendo aplicações em colaboração com as comunidades linguísticas, garantindo que suas necessidades e valores sejam respeitados.

A criação de diretrizes éticas e a promoção do diálogo entre pesquisadores, comunidades linguísticas e formuladores de políticas são essenciais para garantir que a IA seja utilizada de forma responsável e equitativa na cartografia linguística, guiando o desenvolvimento e a aplicação da IA na área com o respeito pela diversidade linguística e o compromisso com a justiça social.

REFERÊNCIAS

Livros:

BIBER, D. et al. **Longman grammar of spoken and written English.** Longman, 1994.

BLOMMAERT, J.; RAMPTON, B. Language and superdiversity. **Diversities**, Paris, v. 13, n. 2, p. 1-21, 2011.

CASTILHO, Ataliba Teixeira de. **Nova gramática do português brasileiro**. São Paulo: Contexto, 2010.

CROOM, C. **Cartography:** A Basic Introduction. London: George Philip, 1999.

GOMES DE CARVALHO, M. E. **Cartografia Linguística:** Uma Introdução. São Paulo: Cortez Editora, 2008.

GOOSKENS, C. **Atlas of language variation in Europe.** Oxford University Press, 2007.

ISQUERDO, Aparecida Negri. **As ciências do léxico:** lexicologia, lexicografia, terminologia: volume III. UFMS, 2007.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing.** 3. ed. Pearson, 2023.

MATTOS E SILVA, Rosa Virgínia. **Para uma história do português do Brasil.** São Paulo: Contexto, 2001.

MORETTI, Franco. **Distant reading.** Londres: Verso, 2013.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical language modeling.** MIT press, 1999.

Teses e Dissertações:

SILVA, M. G. **A cartografia linguística na era digital:** novas perspectivas para o estudo da variação linguística. Tese de doutorado, Universidade de São Paulo. 2015.

Artigos em periódicos (eletrônicos):

BENDER, Emily et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. arXiv preprint arXiv:2201.07285, 2022.

BIRD, Graham et al. OpenAI GPT-3: Language Models That Can Generate Text, Translate Languages, Write Different Kinds of Creative Content, and Answer Your Questions in an Informative Way.. arXiv preprint arXiv:2201.07285, 2022.

VASWANI, Ashish et al. Attention is all you need.. arXiv preprint arXiv:1706.03762, 2017.

Sites:

IBM. **Inteligência Artificial**. Disponível em: https://www.ibm.com/br-pt/topics/artificial-intelligence. Acesso em: 08 jul. 2024.

SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. **Inteligência Artificial.** Disponível em: https://sbc.org.br/2-uncategorised/94-inteligencia-artificial. Acesso em: 08 jul. 2024.

Ferramentas e Plataformas:

Gephi. Gephi Consortium. **Gephi: Open Graph Viz Platform.** Versão 0.10.1. Disponível em: https://gephi.org/. Acesso em: 10 jul. 2024.

Google Colaboratory. Google. **Google Colaboratory.** Disponível em: https://colab.research.google.com/. Acesso em: 10 jul. 2024.

NLTK (Natural Language Toolkit). BIRD, S.; KLEIN, E.; LOPER, E. **NLTK: the natural language toolkit.** Disponível em: https://www.nltk.org/. Acesso em: 10 jul. 2024.

Python. PYTHON SOFTWARE FOUNDATION. **Python Programming Language.** Versão 3.12. Disponível em: https://www.python.org/. Acesso em: 10 jul. 2024.

QGIS (Quantum GIS). QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System.** Versão 3.36. Disponível em: https://www.qgis.org/. Acesso em: 10 jul. 2024.

R. R CORE TEAM. *R:* A language and environment for statistical computing. Versão 4.3.1. **R Foundation for Statistical Computing**, Vienna, Austria. Disponível em: https://www.R-project.org/. Acesso em: 10 jul. 2024.

scikit-learn. PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research,** v. 12, p. 2825-2830, 2011. Disponível em: https://scikit-learn.org/. Acesso em: 10 jul. 2024.

spaCy. HONNIBAL, M.; MONTANI, I.; VAN LANDEGHEM, S.; BOYD, A. **spaCy: Industrial-strength Natural Language Processing in Python.** Versão 3.7. Disponível em: https://spacy.io/. Acesso em: 10 jul. 2024.