LINGUASAGEM

O PAPEL DA INTELIGÊNCIA ARTIFICIAL NA REVITALIZAÇÃO DE LÍNGUAS EM EXTINÇÃO POR MEIO DO PROCESSAMENTO DE LINGUAGEM NATURAL

Valéria Vieira dos Santos[1]
Joceli Catarina Stassi-Sé[2]

RESUMO
Este artigo investiga o papel do Processamento de Linguagem Natural (PLN) e da Inteligência Artificial (IA) na preservação e revitalização de línguas em extinção. Com a previsão de que quase metade das cerca de 7.000 línguas do mundo pode desaparecer até o final do século (Crystal, 2000; Unesco, 2003), tecnologias como reconhecimento automático de fala (ASR), tradução automática e corpora digitais surgem como ferramentas relevantes, ainda que limitadas. Projetos como o Rosetta Project, o Masakhane e iniciativas latino-americanas demonstram que avanços recentes, incluindo aprendizado transferido, modelos de grandes linguagens (LLMs) e técnicas de anotação semiautomática, podem ampliar o alcance dessas tecnologias. No entanto, a aplicação de PLN em línguas minoritárias enfrenta contradições importantes: a exigência de grandes volumes de dados pode desviar esforços de documentação tradicional (livros, dicionários, materiais pedagógicos), e a introdução de novas tecnologias pode favorecer a substituição linguística em vez da preservação. A partir da análise de experiências recentes e da literatura mais atual (Pinhanez *et al.*, 2024), o artigo discute como a integração entre inovação tecnológica, participação comunitária e práticas linguísticas tradicionais é condição indispensável para que a IA contribua para a diversidade linguística global.

PALAVRAS-CHAVE: Processamento de Linguagem Natural; Inteligência Artificial; Revitalização de Línguas; Línguas em Extinção; Documentação Linguística.

ABSTRACT
This article investigates the role of Natural Language Processing (NLP) and Artificial Intelligence (AI) in the preservation and revitalization of endangered languages. With forecasts suggesting that nearly half of the world's approximately 7,000 languages may disappear by the end of the century (Crystal, 2000; Unesco, 2003), technologies such as automatic speech recognition (ASR), machine translation, and digital corpora emerge as relevant, though limited, tools. Projects such as the Rosetta Project, Masakhane, and Latin American initiatives demonstrate that recent advances, including transfer learning, large language models (LLMs), and semi-automatic annotation techniques, can expand the reach of these technologies. However, the application of NLP to minority languages faces important contradictions: the demand for large datasets can divert efforts away from traditional documentation (books, dictionaries, pedagogical materials), and the introduction of new technologies may foster language substitution rather than

[1] Doutoranda pelo Programa de Pós-Graduação em Linguística pela Universidade Federal de São Carlos (UFSCar). Atua na área de Linguística e Processamento de Linguagem Natural, com foco em tecnologias, IA e hesitação. E-mail: valeriavsantos93@gmail.com.
[2] Professora Associada do Departamento de Metodologia de Ensino (DME) e docente permanente do Programa de Pós-Graduação em Linguística (PPGL) da Universidade Federal de São Carlos (UFSCar). Possui Doutorado em Estudos Linguísticos pela UNESP. Desenvolve pesquisas nas áreas de Gramática Discursivo-Funcional, Linguística Aplicada e formação docente. E-mail: jocelistassise@ufscar.br.

preservation. Drawing on recent experiences and the latest literature (Pinhanez *et al.*, 2024), the article argues that the integration of technological innovation, community participation, and traditional linguistic practices is an indispensable condition for AI to contribute to global linguistic diversity.

KEYWORDS: Natural Language Processing; Artificial Intelligence; Language Revitalization; Endangered Languages; Language Documentation.

Introdução

A crise global de extinção de línguas é uma grande ameaça à diversidade cultural e linguística do mundo. Com estimativas sugerindo que quase metade das cerca de 7.000 línguas faladas hoje pode desaparecer até o final deste século (Krauss, 1992; Unesco, 2003), a preservação de línguas ameaçadas tornou-se uma tarefa emergencial.

O desaparecimento de uma língua implica a perda não só de um meio de comunicação, mas também de tradições, conhecimento ancestral e de uma visão de mundo única, que são parte integrante da identidade das comunidades que as falam.

Nesse contexto, as tecnologias de Inteligência Artificial (IA) e, em particular, o Processamento de Linguagem Natural (PLN), emergem como ferramentas promissoras para a preservação e revitalização dessas línguas. O PLN é um subcampo da IA que trata da interação entre computadores e a linguagem humana, facilitando tarefas como tradução automática, reconhecimento de fala e análise de texto. Porém, apesar de seu potencial, essas tecnologias ainda enfrentam limitações significativas, sobretudo no caso de línguas minoritárias com poucos registros disponíveis.

No campo da revitalização linguística, o PLN oferece soluções tecnológicas que ajudam a documentar e preservar línguas ameaçadas de extinção, muitas das quais possuem poucos registros escritos ou orais. Porém, um dos maiores desafios é a criação de corpora digitais, grandes bases de dados linguísticos que alimentam modelos de IA. Sem esses dados, os modelos permanecem limitados; ao mesmo tempo, reunir corpora extensos pode desviar recursos de práticas tradicionais de documentação, criando um dilema que precisa ser considerado.

Este artigo tem como objetivo investigar como a IA e o PLN podem ser usados na preservação e revitalização de línguas ameaçadas. Para isso, pretende-se:

[1] Explorar o papel das tecnologias de IA e PLN na documentação e preservação de línguas em risco de extinção, com foco no uso de sistemas de reconhecimento de fala e tradução automática.
[2] Discutir as limitações e desafios enfrentados na aplicação dessas tecnologias em línguas com poucos recursos, como a falta de dados linguísticos digitalizados.
[3] Analisar casos de sucesso de iniciativas de preservação que fazem uso de IA, destacando como essas soluções têm ajudado a revitalizar línguas indígenas e minoritárias.
[4] Propor novas direções de pesquisa e desenvolvimento para superar os desafios técnicos e éticos no uso de IA na revitalização linguística.

A crise das línguas em extinção representa um dos maiores desafios contemporâneos à diversidade cultural e linguística global. Segundo a UNESCO, mais de 50% das cerca de 7.000 línguas faladas hoje correm risco de desaparecer até o final deste século. Essa perda linguística não se limita à comunicação, mas afeta profundamente a identidade cultural, o patrimônio imaterial e o conhecimento ancestral das comunidades. As línguas são veículos fundamentais para transmitir saberes tradicionais, práticas espirituais e modos de vida que se perdem quando uma língua morre.

As causas dessa crise são variadas, incluindo pressões econômicas, políticas e sociais. A globalização, a expansão de línguas majoritárias como o inglês e o espanhol, e políticas de assimilação cultural têm contribuído para o enfraquecimento das línguas minoritárias. Além disso, o deslocamento populacional e a migração para áreas urbanas resultam na perda de fluência entre gerações mais jovens, um fenômeno conhecido como substituição linguística. Assim, a extinção de uma língua não apenas reduz a diversidade linguística, mas também implica a perda de formas únicas de interpretar o mundo e de transmitir um vasto repertório de conhecimentos tradicionais.

> Cerca de 97% da população mundial fala apenas 4% das línguas do mundo; e, inversamente, cerca de 96% das línguas são faladas por apenas 3% da população mundial. A maior parte da diversidade linguística do planeta, então, está sob a tutela de um número muito pequeno de pessoas (Bernard, 1996, p. 142, tradução nossa[3]).

[3] "About 97% of the world's people speak about 4% of the world's languages; conversely, about 96% of the world's languages are spoken by about 3% of the world's people. Most of the world's language heterogeneity, then, is under the stewardship of a very small number of people" (Bernard, 1996, p. 142).

> Mesmo línguas com muitos milhares de falantes já não estão sendo adquiridas por crianças; pelo menos 50% das mais de seis mil línguas do mundo estão perdendo falantes. Estima-se que, em várias regiões do globo, cerca de 90% dessas línguas possam ser substituídas por idiomas dominantes até o final do século XXI (Bernard, 1996, p. 142, tradução nossa[4]).

Essa crise tem motivado linguistas e organizações internacionais a desenvolver estratégias para preservar e revitalizar línguas ameaçadas. Tecnologias como o PLN e a IA são ferramentas importantes nesse processo, permitindo documentar e preservar línguas que muitas vezes têm poucos registros (Bird; Chiang, 2012).

Contudo, a criação de corpora digitais robustos permanece um grande desafio, dada a natureza predominantemente oral de muitas dessas línguas, exigindo inovações na coleta, curadoria e análise de dados.

A importância da preservação de línguas para as comunidades

A preservação de uma língua é essencial para a comunidade que a fala, pois está profundamente ligada à identidade cultural, à coesão social e à transmissão intergeracional de conhecimento (Crystal, 2000). As línguas não são apenas ferramentas de comunicação, mas também repositórios de saberes tradicionais e valores culturais. Quando uma língua desaparece, perde-se a capacidade de se expressar em um determinado idioma e um vasto conjunto de saberes e práticas que dificilmente pode ser transmitido em sua totalidade por outras línguas.

A preservação linguística é, portanto, fundamental para a autodeterminação das comunidades. Línguas minoritárias são usadas cotidianamente para expressar histórias, mitos e conhecimentos ecológicos - como o uso sustentável de recursos naturais, crenças e sabedoria popular - transmitidos de geração a geração. Comunidades que mantêm suas línguas vivas estão mais aptas a resistir à assimilação cultural e a defender seus direitos. No contexto de povos indígenas e minorias, preservar a língua é uma forma de preservar a autonomia e a dignidade da comunidade.

Outro aspecto importante é a transmissão de conhecimento. Em culturas orais, a língua é o meio através do qual práticas e saberes são passados. Quando uma língua

[4] "Even languages with many thousands of speakers are no longer being acquired by children; at least 50% of the world's more than six thousand languages are losing speakers. We estimate that, in most world regions, about 90% of the languages may be replaced by dominant languages by the end of the 21st century" (Bernard, 1996, p. 142).

desaparece, as práticas sustentáveis relacionadas à biodiversidade e aos modos de vida também se perdem, o que pode ter consequências para a própria sustentabilidade dessas comunidades e para a preservação de registros históricos que dependem dessa transmissão.

No caso das comunidades indígenas no Brasil, a preservação das línguas tem um valor simbólico ainda maior. Como observado no artigo *Preservação das Línguas Indígenas e Direito à Memória: O Caso dos Kokama* (2019), a língua é vista como um elemento essencial para a manutenção do direito à memória e para a preservação do patrimônio cultural de uma comunidade. Quando uma língua desaparece, há uma ruptura na transmissão de saberes e práticas culturais que muitas vezes não podem ser traduzidos ou substituídos por outras línguas.

Além disso, a preservação de línguas indígenas está diretamente relacionada ao fortalecimento da identidade cultural das comunidades. Ao proteger e promover o uso de suas línguas maternas, as comunidades indígenas podem reafirmar sua existência e resistir às pressões de assimilação cultural. O exemplo do povo Kokama, que iniciou um processo de revitalização linguística para preservar sua memória e identidade, é um testemunho de como a língua pode ser um pilar de resistência e orgulho cultural.

A preservação linguística também contribui para a diversidade cultural e intelectual da humanidade. Cada língua traz consigo uma perspectiva sobre o mundo, incluindo conhecimentos ecológicos, medicinais e sociais que são inestimáveis para o futuro da humanidade. Por isso, o direito à preservação linguística e à memória coletiva deve ser garantido como parte das políticas públicas, especialmente em países como o Brasil, onde a diversidade cultural é uma das maiores riquezas nacionais.

A extinção de línguas no Brasil e a revitalização

A extinção de línguas é um fenômeno global, mas no Brasil, onde há uma enorme diversidade linguística, a situação é particularmente alarmante. De acordo com Queren Souza de Castro e Selmo Azevedo Apontes (2020), a extinção de línguas indígenas no país é motivada por uma série de fatores, incluindo desastres naturais, genocídios e pressões sociolinguísticas vindas de línguas oficiais, como o português. Essas línguas, além de frequentemente desvalorizadas pelas próprias comunidades que as falam, enfrentam o risco de desaparecimento acelerado. Para mitigar esse cenário, os autores propõem uma abordagem que vincula a revitalização linguística à revitalização cultural,

reconhecendo que a preservação das línguas também depende da manutenção e fortalecimento das práticas culturais das comunidades falantes.

Em uma linha complementar, Marcus Maia (2006) ressalta a importância de programas educacionais bilíngues e interculturais como estratégias centrais na preservação dessas línguas. Ele defende a criação de dicionários enciclopédicos que não apenas registrem o vocabulário, mas também incluam o conhecimento cultural inerente a essas línguas, como mitos, canções e práticas tradicionais. Segundo Maia (2006), o envolvimento ativo das comunidades no desenvolvimento de materiais didáticos é crucial para garantir o sucesso dos projetos de revitalização, uma vez que as próprias comunidades são os principais guardiões de seus idiomas e tradições.

Essa preocupação com a extinção de línguas no Brasil é reforçada por Rodrigues (2013), que observa que cerca de 75% das línguas indígenas desapareceram desde 1500 e que o Brasil enfrenta o risco de perder até um terço de suas línguas nativas nos próximos 15 anos, caso medidas urgentes não sejam tomadas. Segundo o autor, entre 45 e 60 idiomas ameríndios devem desaparecer até 2030, o que destaca a necessidade de que universidades e centros de pesquisa brasileiros implementem ações concretas para promover a documentação, revitalização e fortalecimento dessas línguas e de seu patrimônio cultural.

Ademais, estudos recentes apontam que o Brasil corre sério risco de perder, no prazo de 15 anos, um terço de suas línguas nativas em razão de muitas contarem não muito mais que uma dezena de falantes. Segundo estimativas, devem ser extintas até 2030, entre 45 a 60 idiomas ameríndios, situação que sinaliza a importância de que as universidades e centros de investigações brasileiros incentivam ações concretas e permanentes para promover a documentação, a descrição, a revitalização e a reconstrução da história filogenética das línguas indígenas sobreviventes, pois esta é uma tarefa de caráter urgente, urgentíssimo (Rodrigues, 2013).

O que é Processamento de Linguagem Natural (PLN) e suas aplicações?

O Processamento de Linguagem Natural (PLN), também conhecido como *Natural Language Processing* (NLP), é um campo interdisciplinar que une computação, linguística e inteligência artificial (IA), com o objetivo de facilitar a interação entre máquinas e linguagem humana. Segundo Jurafsky e Martin (2023), o PLN busca

desenvolver sistemas capazes de compreender, interpretar, gerar e responder à linguagem natural de forma útil e significativa.

Para isso, a área combina avanços em linguística computacional, aprendizado de máquina e ciência de dados, possibilitando a realização de tarefas como análise de sentimentos, tradução automática, respostas a perguntas e reconhecimento de fala. Mais do que *compreender* a linguagem em sentido humano, esses sistemas operam a partir de padrões estatísticos e representações numéricas, produzindo resultados úteis em tarefas específicas.

Tokenização é o processo de segmentar um fluxo de texto em unidades menores, *tokens*, que podem ser palavras, subpalavras (como em BPE/*SentencePiece*) ou caracteres. Trata-se de um passo fundamental de pré-processamento, que não é uma aplicação em si, mas uma técnica que alimenta outras tarefas do PLN, como análise morfológica (identificação de variantes e afixos), análise sintática (relações entre palavras na sentença) e análise semântica (interpretação de sentidos no contexto). A segmentação de sentenças é uma tarefa relacionada, mas distinta da tokenização.

A importância do PLN não reside em tarefas isoladas como a resolução de ambiguidades, mas em sua capacidade de viabilizar aplicações como tradução automática, resposta a perguntas, geração de texto e reconhecimento de fala. Para atingir esses objetivos, o PLN precisa enfrentar desafios centrais da linguagem humana, entre eles a ambiguidade, a variabilidade dialetal e a dependência do contexto.

O Processamento de Linguagem Natural (PLN) envolve um conjunto de componentes que operam em diferentes níveis linguísticos, permitindo que sistemas computacionais analisem, compreendam e gerem linguagem humana de forma automatizada. Esses componentes são organizados em tarefas interdependentes, que vão desde o reconhecimento de unidades mínimas de significado até a interpretação de contextos mais amplos.

Muitas vezes, em materiais introdutórios, o PLN é explicado a partir de uma divisão em níveis (léxico, morfologia, sintaxe, semântica, pragmática), inspirada na análise linguística. Essa organização funciona como recurso pedagógico para compreender a complexidade da linguagem, mas não reflete a forma como os sistemas modernos de PLN efetivamente operam. Os sistemas atuais, em especial os baseados em aprendizado profundo, não seguem etapas sequenciais, mas integram múltiplos tipos de informação simultaneamente. Por exemplo, modelos neurais de tradução ou reconhecimento de fala aprendem representações distribuídas que já capturam aspectos

morfológicos, sintáticos e semânticos de maneira conjunta. Ainda assim, a noção de *níveis linguísticos* continua útil para discutir os desafios específicos enfrentados por essas tecnologias, como o tratamento de morfologia complexa em línguas indígenas brasileiras.

Um dos elementos iniciais e mais recorrentes ao longo do PLN é a tokenização, que consiste na segmentação do texto em unidades menores, os tokens, fundamentais para que qualquer sistema de PLN possa operar adequadamente. Ela serve como ponto de partida para diversas outras tarefas, como classificação textual, tradução automática ou reconhecimento de entidades nomeadas.

Portanto, em vez de etapas fixas e sucessivas, o PLN deve ser entendido como um conjunto integrado de técnicas que trabalham em paralelo, combinando diferentes camadas analíticas. Em conjunto, essas técnicas visam simular a capacidade humana de compreender e produzir linguagem, sendo cada uma delas indispensável para o desenvolvimento de aplicações robustas e sensíveis às nuances do uso linguístico.

Aplicações do Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) possui um conjunto diversificado de aplicações práticas que, embora muitas vezes invisíveis ao usuário final, já fazem parte do cotidiano de milhões de pessoas em escala global. Essas aplicações abrangem desde a comunicação interpessoal mediada por tecnologia até processos analíticos e automatizados em diferentes setores da sociedade.

Uma das aplicações mais conhecidas é a tradução automática, presente em ferramentas como *Google Translate* e *DeepL*, que utilizam redes neurais profundas e técnicas de aprendizado a partir de corpora bilíngues para converter textos entre diferentes idiomas. Apesar de seus avanços, esses sistemas ainda apresentam limitações importantes em contextos com poucos dados disponíveis, como no caso das línguas indígenas e ameaçadas, onde corpora bilíngues são escassos ou inexistentes. Com avanços constantes, essas ferramentas tornaram-se cada vez mais eficazes ao lidar com expressões idiomáticas, contextos específicos e diferentes registros de linguagem, mas sua eficácia continua diretamente vinculada à disponibilidade e à qualidade dos dados de treinamento.

Outro campo amplamente difundido é o do *reconhecimento automático de fala* (ASR, na sigla em inglês). Assistentes virtuais como *Siri*, *Google Assistant* e *Alexa* utilizam modelos estatísticos e redes neurais para converter sinais acústicos em representações numéricas, que são então transcritas em texto. Esse processo não depende

de conhecimento fonético explícito: os modelos aprendem, a partir de grandes corpora de áudio transcritos, a mapear diretamente sinais acústicos para sequências linguísticas. Essa tecnologia permite que usuários interajam com dispositivos por meio de comandos de voz e que tarefas automatizadas sejam executadas a partir da fala, embora sua eficácia ainda dependa fortemente da disponibilidade de grandes volumes de dados anotados, o que limita sua aplicação em línguas minoritárias.

A análise de sentimento também representa uma aplicação relevante, especialmente para empresas e setores que lidam com grandes volumes de *feedback* textual. Através do PLN, algoritmos são capazes de identificar e classificar sentimentos expressos em comentários, postagens em redes sociais ou avaliações de produtos, possibilitando uma compreensão mais profunda da percepção pública sobre marcas, serviços e eventos. No entanto, fatores como ironia, sarcasmo e contextos culturais específicos continuam sendo desafios importantes para essa tecnologia.

Outra aplicação importante é o resumo automático, no qual sistemas de PLN são empregados para condensar textos extensos, identificando suas informações mais relevantes. Essa tecnologia é especialmente útil para profissionais que precisam lidar com grandes volumes de informação, como jornalistas, pesquisadores e gestores. Ainda assim, resumos gerados automaticamente podem omitir nuances ou apresentar vieses, exigindo validação humana em contextos críticos.

O **s**istema de pesquisa e recuperação de informação é outro exemplo marcante da presença do PLN na vida cotidiana. Ferramentas como o Google utilizam modelos de linguagem para interpretar as intenções dos usuários, indo além da simples correspondência de palavras-chave, e oferecendo resultados que consideram o contexto e os significados implícitos nas buscas.

No campo da *geração de linguagem natural* (NLG), sistemas de PLN são capazes de produzir texto de forma automatizada a partir de diferentes tipos de dados, incluindo dados estruturados (como tabelas e bancos de dados) e não estruturados (como corpora textuais e multimodais). Isso inclui desde relatórios gerados a partir de planilhas até textos produzidos por modelos treinados em grandes conjuntos de dados não estruturados. Exemplos de aplicação são assistentes virtuais e *chatbots*, que geram respostas a consultas de usuários, bem como sistemas para criação de relatórios automatizados e conteúdos jornalísticos. Essa funcionalidade amplia a eficiência operacional e garante consistência na apresentação de informações.

Os *chatbots* e assistentes virtuais são aplicações centrais no uso contemporâneo do PLN. Sistemas como o *ChatGPT*, entre outros, são treinados para compreender perguntas, interpretar intenções comunicativas e gerar respostas coerentes e apropriadas. Eles estão presentes em atendimentos automatizados, sistemas de suporte técnico e plataformas educacionais, entre outras áreas, revolucionando a forma como interagimos com a tecnologia e tornando a comunicação com máquinas mais fluida e natural.

Essas diversas aplicações demonstram o potencial do PLN não apenas como um recurso técnico, mas como uma ponte entre linguagem humana e processamento computacional, fundamental para o avanço da inteligência artificial em contextos sociais e profissionais. Ainda assim, é importante destacar que sua eficácia está sempre condicionada à disponibilidade e qualidade dos dados, além da necessidade de adaptação a diferentes contextos culturais e linguísticos.

Desafios do PLN

Apesar dos avanços significativos no Processamento de Linguagem Natural (PLN), muitos desafios permanecem devido à complexidade inerente à linguagem humana. Um dos problemas centrais é a ambiguidade linguística, que se refere à multiplicidade de significados que uma palavra pode ter dependendo do contexto. Um exemplo clássico é a palavra ‘banco’, que pode significar tanto uma instituição financeira quanto um assento. Resolver essa ambiguidade, um processo conhecido como desambiguação lexical, é uma das tarefas mais complexas, pois requer que os sistemas identifiquem o significado correto com base no contexto em que a palavra é utilizada. Esse desafio se agrava no contexto da preservação e revitalização de línguas em extinção, já que muitas dessas línguas possuem palavras e expressões com significados profundamente enraizados em tradições culturais locais, tornando a desambiguação ainda mais difícil.

A variabilidade linguística é um desafio universal para o PLN. Todas as línguas apresentam diferenças de sotaque, dialeto e estilo, o que torna difícil desenvolver sistemas que funcionem igualmente bem em todos os contextos. No caso das línguas ameaçadas, o problema não é que variem mais do que as majoritárias, mas que geralmente não existem *corpora* suficientes para representar essa diversidade interna. Assim, a falta de dados adequados impede que tecnologias de PLN capturem plenamente a riqueza cultural

e histórica dessas línguas, o que constitui um obstáculo adicional à documentação e revitalização.

A compreensão do contexto e a coerência também são um desafio, já que são cruciais para a geração de linguagem natural e para a tradução. Para que os modelos de PLN sejam eficazes, eles precisam não apenas entender palavras isoladas, mas também interpretar o contexto mais amplo em que essas palavras são utilizadas, mantendo a coerência ao longo de um texto ou uma conversa. Em línguas ameaçadas, muitas das quais são predominantemente orais e dependem fortemente do contexto cultural para atribuir significado às palavras, essa tarefa se torna ainda mais complexa. O desafio não está em fenômenos exclusivos dessas línguas, mas na escassez de dados contextuais suficientes para treinar modelos capazes de refletir adequadamente essas nuances.

Um dos maiores obstáculos para o uso de PLN na revitalização de línguas ameaçadas é a escassez de recursos digitais. A maioria das pesquisas concentra-se em línguas amplamente faladas, como inglês e espanhol, que possuem vastos corpora disponíveis. Já as línguas minoritárias raramente contam com dados suficientes para treinar modelos robustos. Essa limitação levanta um dilema importante: o esforço para reunir dados em formato digital, com curadoria de qualidade e em volume suficiente, pode acabar desviando linguistas, antropólogos e professores de tarefas imediatamente mais produtivas, como a elaboração de livros, dicionários e materiais pedagógicos que já têm impacto direto na comunidade. Essa contradição precisa ser considerada em qualquer projeto de PLN para revitalização linguística: é necessário equilibrar a construção de corpora digitais com práticas de documentação tradicionais, de modo que uma atividade complemente, e não substitua, a outra.

Outro desafio é a interpretação de nuances emocionais, ironias e expressões culturais simbólicas, fenômenos presentes em qualquer língua. Esses aspectos, comuns tanto em línguas majoritárias quanto minoritárias, tornam-se ainda mais difíceis de modelar quando há escassez de dados anotados e contextualizados, condição frequente em línguas ameaçadas.

Ferramentas de inteligência artificial (IA) amplamente conhecidas no campo do PLN, como *Google Translate*, *DeepL* e *OpenAI GPT*, têm desempenhado um papel importante no avanço das tecnologias linguísticas. O *Google Translate*, por exemplo, utiliza redes neurais profundas para traduzir mais de 100 idiomas, permitindo não só a tradução de texto, mas também de fala e imagens. Essa ferramenta tem sido fundamental

na disseminação de informações entre diferentes idiomas, mas ainda enfrenta desafios quando se trata de línguas menos representadas ou com poucos recursos disponíveis.

DeepL, conhecido por sua qualidade superior em traduções automáticas, especialmente em textos técnicos e científicos, também se destaca por utilizar modelos de redes neurais avançados que fornecem traduções mais naturais e precisas. No entanto, seu foco em línguas amplamente faladas limita seu impacto no contexto da revitalização de línguas ameaçadas.

Por sua vez, o *OpenAI GPT* é um dos modelos de linguagem mais atuais disponíveis, capaz de realizar tarefas que incluem geração de texto, tradução automática e simulação de diálogos. O GPT tem sido amplamente utilizado em diversas aplicações de PLN, incluindo *chatbots*, assistentes virtuais e geração automática de conteúdo. Ainda assim, esses modelos exigem ajustes finos (*fine-tuning*) ou estratégias como aprendizado transferido para se tornarem realmente úteis em cenários de baixa disponibilidade de dados, como ocorre com línguas indígenas e minoritárias.

A preservação e revitalização de línguas em extinção requerem, portanto, o desenvolvimento de novas estratégias e adaptações das tecnologias atuais de PLN, de modo que possam lidar com as complexidades linguísticas e culturais dessas línguas.

Como essas ferramentas podem ajudar na revitalização de línguas ameaçadas?

As ferramentas de Processamento de Linguagem Natural (PLN), apoiadas por Inteligência Artificial (IA), podem desempenhar um papel relevante na preservação e revitalização de línguas ameaçadas, fornecendo meios tecnológicos para documentar, transcrever e disseminar essas línguas. Uma das formas mais exploradas nesse contexto é a documentação e transcrição. Ferramentas de reconhecimento automático de fala (ASR) podem ser utilizadas para auxiliar na transcrição de línguas ameaçadas, embora sua eficácia dependa de corpora de áudio suficientemente grandes e diversificados, algo que geralmente não existe para essas línguas. Ainda assim, quando combinadas a métodos de coleta manual e à participação das comunidades, essas tecnologias podem facilitar a criação de corpora linguísticos úteis para o registro de tradições orais e para a preservação do conhecimento cultural das comunidades falantes.

Os sistemas de tradução automática também podem ser adaptados para trabalhar com línguas minoritárias e ameaçadas, criando modelos personalizados que permitam a tradução entre essas línguas e línguas majoritárias. No entanto, essa adaptação enfrenta

sérias limitações: a tradução neural moderna depende de grandes volumes de dados paralelos, raros nesse contexto. Uma alternativa é recorrer a técnicas como aprendizado transferido, dicionários bilíngues e corpora pequenos, que permitem ao menos resultados iniciais. Quando bem implementados, esses modelos podem apoiar o aprendizado bilíngue, tornando as línguas ameaçadas mais acessíveis tanto para falantes nativos quanto para aprendizes, e incentivando a transmissão intergeracional do idioma.

Outro uso do PLN é a análise de textos tradicionais. Ferramentas de análise semântica podem ser empregadas para estudar e interpretar textos antigos, auxiliando linguistas a recuperar e compreender documentos históricos cruciais para a cultura e identidade de uma comunidade. Ainda que essas tarefas possam ser parcialmente automatizadas, elas continuam dependendo fortemente do trabalho manual de linguistas e falantes nativos, já que algoritmos sozinhos não dão conta de reconstruir sentidos culturais específicos. Além disso, tais recursos podem apoiar a formalização de gramáticas e a criação de dicionários bilíngues e monolíngues, que permanecem como instrumentos fundamentais no processo de revitalização linguística.

Ao adaptar essas tecnologias para atender às especificidades das línguas ameaçadas, é possível documentá-las de forma mais sistemática. Contudo, é preciso reconhecer a contradição: a ênfase excessiva em modelos de PLN pode desviar tempo e energia da produção de materiais pedagógicos tradicionais; como livros, gramáticas e dicionários, que são de utilidade imediata para as comunidades. O desafio está em encontrar um equilíbrio, em que a tecnologia complemente, e não substitua, as práticas tradicionais de documentação. A utilização de ferramentas de PLN pode, de fato, ampliar o acesso das novas gerações ao idioma, mas somente se articulada a uma estratégia mais ampla que combine inovação tecnológica, práticas linguísticas tradicionais e participação comunitária.

Materiais didáticos digitais e dicionários na revitalização

Os materiais didáticos digitais e os dicionários interativos têm desempenhado um papel significativo em projetos de revitalização linguística, especialmente quando combinados com plataformas tecnológicas de fácil acesso. Esses recursos possibilitam não apenas o registro lexical e gramatical das línguas, mas também promovem o engajamento da comunidade falante com o idioma em contextos educacionais e cotidianos. Como destacam Grenoble e Whalen (2006), a produção de materiais

adaptados à realidade sociocultural das comunidades é fundamental para o sucesso das iniciativas de revitalização.

Além disso, o uso de dicionários digitais bilíngues ou multimodais (com imagens, áudio e vídeo) facilita o aprendizado intergeracional e pode ser incorporado a aplicativos móveis e ambientes virtuais de aprendizagem. Tais ferramentas não apenas documentam a língua, mas incentivam seu uso ativo, contribuindo para a reconstrução de práticas comunicativas em risco de desaparecimento (Bird; Simons, 2003).

No caso da língua Apurinã, falada no sudoeste do Amazonas, foram desenvolvidos materiais como livros de alfabetização e dicionários bilíngues (Apurinã-Português). Esses materiais permitem que a língua seja ensinada em contextos formais, como escolas, e informais, como em casa, proporcionando um meio acessível para o aprendizado da gramática, vocabulário e das narrativas tradicionais da comunidade. Esses recursos ajudam a manter viva a ligação entre a língua e a identidade cultural dos falantes. A língua é um repositório de sabedoria ancestral, e os materiais didáticos digitais permitem que as gerações futuras tenham acesso não apenas ao idioma em si, mas também ao vasto conhecimento cultural, como histórias, mitos e práticas sociais transmitidas pela língua.

Os dicionários bilíngues, em particular, desempenham o seu papel no processo de revitalização. Eles documentam e organizam o vocabulário da língua ameaçada de forma sistemática, garantindo que ele não se perca com o tempo. Além disso, os dicionários atuam como ferramentas de ensino e referência, permitindo que os falantes nativos e os aprendizes possam consultar palavras, seus significados e contextos de uso. Isso facilita o ensino da língua, tanto para falantes nativos que desejam reforçar seus conhecimentos quanto para novos aprendizes que estão se familiarizando com o idioma. Os dicionários e materiais didáticos digitais possuem ainda uma importância simbólica significativa: eles representam um esforço ativo de registro e formalização da língua, muitas vezes em contextos em que não há ortografia padronizada. O desenvolvimento desses recursos costuma ser feito em colaboração entre linguistas e membros da comunidade, assegurando que o registro seja autêntico e culturalmente respeitoso.

É importante diferenciar, contudo, ferramentas de documentação linguística de sistemas de PLN. Plataformas como o *FieldWorks Language Explorer* (FLEx) e o *EUDICO Linguistic Annotator* (ELAN) não são sistemas de PLN, mas programas de anotação e análise que permitem criar glossários, dicionários e transcrições de gravações orais. O FLEx é amplamente usado para análise morfológica e registro lexical, enquanto

o ELAN facilita a transcrição e anotação de materiais audiovisuais, essenciais para línguas predominantemente orais. Esses recursos, embora distintos do PLN, fornecem dados estruturados que podem futuramente alimentar aplicações computacionais mais avançadas.

Assim, materiais didáticos digitais, dicionários e ferramentas de anotação não constituem PLN em si, mas cumprem uma função estratégica na preservação cultural e linguística. Eles garantem que as línguas minoritárias continuem sendo faladas e transmitidas entre gerações, ao mesmo tempo em que preparam terreno para o desenvolvimento de tecnologias de PLN mais sofisticadas no futuro.

IA e revitalização de línguas em extinção

A Inteligência Artificial tem se mostrado uma aliada promissora nos esforços de revitalização de línguas em extinção, especialmente por meio de técnicas como aprendizado transferido, redes neurais profundas e modelagem linguística baseada em poucos dados (Anastasopoulos, 2019; Hauenstein *et al.*, 2022). Em cenários de baixa disponibilidade de recursos, modelos pré-treinados podem ser adaptados para línguas minoritárias, permitindo avanços iniciais na geração de vocabulário, padrões morfológicos e estruturas sintáticas básicas.

Projetos como o *Masakhane*, voltado para línguas africanas, demonstram que a IA pode ser usada de forma colaborativa e eticamente orientada para construir tradutores automáticos, corpora paralelos e ferramentas educacionais (Adams *et al.*, 2021). Como aponta Bird (2020), a efetividade tecnológica só se concretiza quando ocorre em diálogo estreito com as comunidades falantes, garantindo sensibilidade cultural e respeito às formas de uso da língua.

Ferramentas de ASR (reconhecimento automático de fala) baseadas em IA podem transcrever fala em texto, o que é especialmente útil em línguas de tradição predominantemente oral. Esses sistemas utilizam redes neurais treinadas em grandes corpora de áudio para correlacionar sinais acústicos com representações linguísticas. Embora promissora, a adaptação para línguas minoritárias enfrenta sérios limites: mesmo com técnicas de aprendizado transferido, a escassez de gravações confiáveis compromete a precisão dos modelos.

De modo semelhante, ferramentas de tradução automática como *Google Translate* e *DeepL* foram desenvolvidas para línguas majoritárias e dependem de vastos corpora

bilíngues. A adaptação para línguas ameaçadas requer métodos alternativos, como dicionários bilíngues ou textos paralelos, mas sua utilidade prática ainda é restrita. Há, inclusive, o risco de que o esforço de coleta de dados digitais desvie recursos de iniciativas mais imediatas, como a produção de materiais didáticos e dicionários pedagógicos, de impacto comprovado nas comunidades.

Além da tradução, a IA pode contribuir na análise de textos tradicionais e na criação de gramáticas e dicionários digitais. Ferramentas de análise semântica e morfológica podem auxiliar na sistematização de estruturas gramaticais, mesmo em línguas sem longa tradição escrita. No entanto, tais aplicações dependem de dados de qualidade e, sem eles, correm o risco de gerar recursos artificiais pouco úteis para as comunidades.

Tecnologias emergentes, como o aprendizado profundo, os *modelos de grandes linguagens* (LLMs) e o aprendizado transferido, ampliam as possibilidades de superação desses limites. Combinadas à participação ativa das comunidades falantes e à colaboração entre linguistas e desenvolvedores, elas podem produzir soluções mais eficazes e personalizadas. Ainda assim, é fundamental reconhecer que o sucesso dessas iniciativas depende de equilibrar inovação com práticas tradicionais de documentação, e de avaliar criticamente os riscos de que novas tecnologias incentivem a substituição linguística em vez da preservação.

Assim, o PLN e a IA não devem ser vistos apenas como soluções tecnológicas, mas como aliados potenciais no esforço global de preservação. Seu impacto positivo dependerá de uma aplicação ética, contextualizada e integrada a métodos de documentação comunitária, garantindo que a tecnologia fortaleça, e não fragilize, a diversidade linguística.

Projeto Rosetta

O *Projeto Rosetta*, uma iniciativa colaborativa liderada pela *National Geographic* em parceria com a *Long Now Foundation*, é amplamente considerado um dos esforços mais ambiciosos e inovadores no campo da preservação linguística. Inspirado pela original Pedra de Roseta, que permitiu aos estudiosos decifrar hieróglifos egípcios ao comparar textos em diferentes línguas, o projeto busca criar uma versão moderna desse artefato, com o objetivo de preservar as línguas do mundo para as gerações futuras.

Segundo informações divulgadas pela própria fundação (Long Now Foundation, 2024), o acervo reúne dados linguísticos básicos de mais de 1.500 línguas em risco de extinção, incluindo alfabetos, dicionários, textos traduzidos e elementos culturais. A proposta é preservar esse conteúdo tanto digital quanto fisicamente, de forma que possa ser utilizado por futuras gerações em iniciativas de revitalização linguística.

Uma das inovações do Projeto Rosetta é a criação do *Rosetta Disk*, um disco de níquel com aproximadamente o tamanho de um CD, projetado para resistir à passagem do tempo por milhares de anos. O disco contém informações linguísticas gravadas em alta resolução, utilizando tecnologias de micro gravação, que permite armazenar até 13.000 páginas de texto. Essas informações podem ser lidas com a ajuda de um microscópio, garantindo que mesmo em um futuro distante, as línguas registradas não sejam completamente perdidas, mesmo que seus falantes tenham desaparecido.

A estrutura do Rosetta Disk inclui amostras de centenas de línguas ameaçadas, organizadas de maneira a fornecer um panorama das estruturas linguísticas de cada idioma. O conteúdo linguístico inclui vocabulários essenciais, expressões idiomáticas, textos tradicionais e gramaticais básicas. A robustez do disco físico, no entanto, é complementada por uma base de dados digital, onde o material linguístico é armazenado e atualizado continuamente por linguistas e membros das comunidades falantes.

Um dos pilares do Projeto Rosetta é a colaboração ativa com comunidades indígenas e falantes de línguas minoritárias. Essa abordagem participativa garante que a documentação não seja apenas técnica, mas culturalmente sensível, incluindo narrativas, canções e rituais, além da língua em si.Esse engajamento promove também o empoderamento comunitário, já que os falantes contribuem diretamente para a construção dos arquivos linguísticos.

Do ponto de vista tecnológico, iniciativas como o Projeto Rosetta utilizam PLN e IA para organizar grandes volumes de dados linguísticos. Contudo, esse projeto também enfrenta limitações severas: muitas línguas ameaçadas têm poucos registros escritos ou orais, e sua natureza predominantemente oral dificulta a criação de sistemas de escrita padronizados. Isso limita o potencial imediato de aplicação de técnicas avançadas de IA, reforçando a necessidade de combinar esforços tecnológicos com práticas tradicionais de documentação.

Esse dilema evidencia uma contradição central: sem corpora digitais, o PLN é pouco eficaz; mas o esforço para criar corpora em escala pode desviar tempo e recursos de atividades de impacto direto, como a produção de livros, dicionários e materiais

didáticos para uso imediato pelas comunidades. Nesse sentido, o Projeto Rosetta deve ser visto menos como substituto das práticas tradicionais e mais como um repositório complementar de longo prazo.

Outro desafio é garantir que o material documentado seja acessível às futuras gerações, especialmente em comunidades com pouco acesso à tecnologia. Embora o Rosetta Disk tenha sido projetado para durar milhares de anos, sua utilidade depende da capacidade futura de ler e interpretar os registros. Assim, seu valor é mais simbólico e de arquivo, funcionando como um *seguro* cultural, do que como ferramenta prática de revitalização no presente.

Em conclusão, o Projeto Rosetta representa um marco na documentação linguística, mas sua contribuição imediata à revitalização é limitada. Ao aliar inovação tecnológica com colaboração comunitária, o projeto cria condições para que o conhecimento cultural e linguístico seja preservado em longo prazo, sem substituir a necessidade urgente de práticas pedagógicas e materiais acessíveis que mantenham as línguas vivas no cotidiano das comunidades.

Projetos de tradução automática para línguas indígenas e ameaçadas

Em colaboração com comunidades indígenas no México, linguistas e cientistas de dados têm utilizado ferramentas de Processamento de Linguagem Natural (PLN) para desenvolver sistemas de tradução automática voltados a línguas como o Mixteco, o Nahuatl e o Quechua. Iniciativas como o *AmericasNLP* — workshop focado no avanço do PLN para línguas indígenas da América Latina — vêm contribuindo ativamente para a construção de modelos de tradução automática e reconhecimento de fala, em contextos com poucos recursos linguísticos disponíveis (Mager *et al.*, 2021).

Esses sistemas auxiliam na criação de materiais bilíngues, dicionários digitais e conteúdos educacionais, promovendo a preservação do conhecimento linguístico e cultural das comunidades. No entanto, a efetividade dessas ferramentas depende da coleta de corpora suficientes e da curadoria cuidadosa dos dados, tarefas que muitas vezes competem com esforços mais imediatos, como a elaboração de livros e materiais pedagógicos tradicionais. A tradução automática, quando adaptada com a participação ativa dos falantes, torna-se uma ferramenta eficaz tanto para a documentação quanto para a revitalização das línguas ameaçadas (Oncevay *et al*., 2020).

Ferramentas de PLN para criação de dicionários e materiais didáticos digitais

O Processamento de Linguagem Natural (PLN) tem sido apontado como uma ferramenta importante para apoiar a criação de dicionários digitais e materiais didáticos voltados a línguas ameaçadas. Esses recursos são especialmente relevantes para revitalizar idiomas com escassa documentação formal. No entanto, é importante distinguir entre ferramentas de PLN propriamente ditas e softwares de anotação ou documentação linguística. Entre as ferramentas mais utilizadas estão o *FLEx* (FieldWorks Language Explorer), que oferece funcionalidades para análise morfossintática, criação de glossários e elaboração de dicionários, e o *ELAN* (EUDICO Linguistic Annotator), voltado à transcrição e anotação de dados audiovisuais em pesquisas de campo (Bouda *et al.*, 2012).

Essas tecnologias não constituem PLN em sentido estrito, mas fornecem dados estruturados que podem ser utilizados posteriormente em aplicações de PLN. São, ainda assim, amplamente recomendadas por especialistas em documentação linguística por permitirem a coleta e a organização sistemática de dados linguísticos e culturais, essenciais para o desenvolvimento de materiais pedagógicos em contextos de educação bilíngue e intergeracional (Bird; Simons, 2003). Seu uso colabora diretamente com a continuidade cultural e com o fortalecimento das identidades linguísticas das comunidades falantes.

Desafios e oportunidades do uso de PLN e Ia na revitalização

O Processamento de Linguagem Natural (PLN) tem se mostrado uma ferramenta poderosa na preservação de línguas ameaçadas, oferecendo soluções tecnológicas inovadoras para documentar e revitalizar idiomas que estão à beira da extinção. Apesar desse potencial, sua aplicação enfrenta obstáculos significativos, tanto técnicos quanto culturais, que limitam a eficácia dos esforços de preservação. Esses desafios estão diretamente relacionados à natureza complexa da linguagem humana, especialmente no que diz respeito a línguas minoritárias que apresentam características únicas, muitas vezes profundamente enraizadas em contextos culturais específicos. Paralelamente, as oportunidades trazidas pela Inteligência Artificial (IA), especialmente com o avanço de técnicas como o aprendizado transferido e os modelos de grandes linguagens (LLMs), indicam caminhos promissores, embora ainda incipientes.

Outro risco que precisa ser reconhecido é o de substituição linguística. Diversas experiências históricas mostram que a chegada de novas tecnologias não fortaleceu línguas minoritárias, mas promoveu a adoção de línguas majoritárias associadas a essas tecnologias. Para que o uso de PLN e IA realmente favoreça a preservação, é indispensável que as ferramentas sejam implementadas em diálogo com as comunidades, de modo a fortalecer o valor social da língua local. Caso contrário, há o perigo de que a tecnologia, em vez de contribuir para a revitalização, acelere processos de aculturação e substituição linguística.

Um obstáculo técnico igualmente relevante é a complexidade semântica e cultural de muitas línguas ameaçadas. Esses idiomas frequentemente carregam expressões idiomáticas, conceitos culturais e categorias cognitivas que não encontram equivalentes diretos em línguas majoritárias. Os sistemas de IA têm dificuldades em capturar tais nuances, uma vez que técnicas como a tradução automática tendem a padronizar e generalizar significados, resultando em interpretações imprecisas ou culturalmente inadequadas. Termos relacionados a práticas espirituais, conhecimentos ecológicos tradicionais ou formas específicas de organização social são altamente contextuais e, portanto, difíceis de serem preservados com fidelidade por meio de ferramentas digitais.

A Inteligência Artificial oferece, sim, oportunidades - como o aprendizado transferido, os modelos multilingues e as técnicas de anotação semiautomática - mas seu real impacto depende diretamente da colaboração entre linguistas, desenvolvedores e comunidades. É a articulação entre conhecimento técnico e sensibilidade cultural que pode garantir que essas ferramentas reforcem o uso cotidiano das línguas, em vez de ameaçá-las.

Entre as contribuições mais recentes nesse campo, destaca-se o trabalho de Pinhanez *et al*. (2024), que propõe uma análise detalhada de como tecnologias de IA podem ser aplicadas a línguas minoritárias em contextos de baixa disponibilidade de dados. O estudo discute, de forma prática, os limites técnicos e os ganhos possíveis, reforçando a ideia de que soluções automatizadas não substituem, mas complementam os esforços tradicionais de documentação linguística.

Uma das principais contribuições de Pinhanez *et al*. (2024) é a ênfase no uso de modelos adaptativos que exploram técnicas de aprendizado transferido e treinamento leve para lidar com a escassez de corpora. O artigo demonstra que, mesmo com recursos limitados, é possível desenvolver protótipos de sistemas de tradução ou reconhecimento de fala, desde que haja integração entre métodos estatísticos e conhecimento linguístico

pré-existente. Essa abordagem ajuda a mitigar o dilema apontado em diversos trabalhos - o de exigir grandes volumes de dados para resultados de qualidade - sem ignorar a realidade das línguas ameaçadas.

Além do aspecto técnico, Pinhanez *et al*. (2024) também destacam a dimensão social do problema. Os autores argumentam que a simples criação de ferramentas não garante preservação, podendo inclusive gerar efeitos contrários se não houver participação comunitária e preocupação com o impacto cultural. Essa visão converge com a necessidade, apontada neste artigo, de que linguistas, desenvolvedores e comunidades trabalhem em conjunto, garantindo que a tecnologia fortaleça a língua e sua identidade cultural, em vez de contribuir para a sua substituição.

Outro avanço promissor é o uso de técnicas de anotação semiautomática, que aceleram o processo de transcrição e anotação de dados linguísticos. Essas ferramentas permitem que linguistas ou membros das comunidades falantes revisem e ajustem automaticamente transcrições geradas por IA, economizando tempo e garantindo maior precisão na criação de corpora linguísticos. Ferramentas como o *FieldWorks Language Explorer* (FLEx) e o *ELAN* já oferecem funcionalidades que facilitam a anotação e a documentação de línguas, mas o futuro dessas tecnologias depende da integração mais avançada de IA para melhorar a eficiência e a precisão.

O caminho para o futuro da IA na preservação de línguas ameaçadas exige uma combinação de inovação tecnológica e sensibilidade cultural. Por um lado, é necessário investir em novas técnicas de PLN capazes de lidar com a diversidade e a complexidade das línguas minoritárias, especialmente aquelas de tradição oral predominante. Por outro lado, a participação ativa das comunidades é crucial para garantir que os modelos e sistemas desenvolvidos reflitam suas necessidades e valores culturais.

Com o avanço contínuo das tecnologias de *deep learning*, os modelos de IA estão se tornando cada vez mais capazes de capturar a riqueza semântica e cultural das línguas, adaptando-se às suas especificidades. Ao mesmo tempo, o desenvolvimento de tecnologias multimodais que integram áudio, texto e até mesmo vídeo abre novas possibilidades para a preservação de tradições orais e visuais, ampliando o escopo das iniciativas de documentação e revitalização linguística.

Apesar das limitações ainda presentes, o futuro da inteligência artificial no campo da revitalização linguística oferece oportunidades promissoras. A incorporação de tecnologias emergentes, como o *aprendizado profundo* (*deep learning*) e os modelos de linguagem de larga escala (*Large Language Models* - LLMs), tem potencial para

transformar significativamente esse cenário, sobretudo quando associadas a processos colaborativos entre linguistas, tecnólogos e comunidades falantes.

O avanço das redes neurais e das técnicas de aprendizado profundo tem possibilitado que modelos de IA se tornem mais adaptáveis a línguas com poucos recursos. Entre essas abordagens, destaca-se o *aprendizado transferido* (*transfer learning*), que permite a reutilização de conhecimentos adquiridos em línguas majoritárias para aplicação em línguas minoritárias (Zhang *et al*., 2021). Modelos como o *GPT*, desenvolvido pela *OpenAI*, demonstram que arquiteturas de larga escala podem ser ajustadas, por meio de *fine-tuning* ou *prompt engineering*, para atender a contextos linguísticos menos representados (Brown *et al.*, 2020; Liu *et al.*, 2023), ampliando as possibilidades de inclusão dessas línguas em sistemas de Processamento de Linguagem Natural.

O sucesso dessas inovações depende diretamente da colaboração entre linguistas, desenvolvedores e comunidades. A criação de soluções voltadas especificamente para línguas ameaçadas requer uma abordagem interdisciplinar, na qual os linguistas fornecem não apenas dados, mas também orientações metodológicas capazes de assegurar que a tecnologia respeite as estruturas e nuances da língua. A articulação entre conhecimento técnico e sensibilidade cultural é fundamental para o desenvolvimento de ferramentas eficazes e apropriadas, capazes de contribuir não só para a documentação, mas também para a reativação do uso cotidiano da língua.

Outro elemento central nesse processo é o papel das próprias comunidades. Os falantes detêm um saber vivo e complexo que não pode ser reduzido apenas a bases de dados. Como destacam Bird (2020) e Dwyer (2006), a participação ativa dessas comunidades no design, coleta de dados e aplicação das tecnologias é indispensável para garantir que os resultados reflitam suas realidades socioculturais e reforcem sua autonomia, evitando que a tecnologia se torne mais uma forma de apagamento cultural.

Ferramentas como o *FieldWorks Language Explorer* (FLEx) e o *ELAN* já têm se destacado por incentivar o engajamento comunitário no processo de documentação linguística. No entanto, como apontam Zaidan *et al.* (2021), o futuro da revitalização dependerá da ampliação de iniciativas que priorizem abordagens participativas, integrando tecnologia e tradição em prol da diversidade linguística e da preservação de identidades culturais.

Conclusão

O artigo mostrou que, embora o PLN e a IA ofereçam contribuições relevantes para a preservação e revitalização de línguas ameaçadas, seu uso exige cautela e reflexão crítica. Entre as potencialidades, destacam-se ferramentas de transcrição automática, tradução personalizada, criação de dicionários digitais e desenvolvimento de materiais didáticos interativos. No entanto, desafios persistem: a escassez de corpora, a variabilidade interna das línguas, a dificuldade em captar nuances culturais e o risco de que a tecnologia, em vez de apoiar, acabe por acelerar a substituição linguística.

As contradições discutidas, como a dependência de grandes volumes de dados e os possíveis efeitos culturais adversos, evidenciam que o PLN não pode ser tratado como solução única, mas sim como complemento às práticas tradicionais de documentação linguística. O caminho mais promissor está na cooperação interdisciplinar e participativa: linguistas, desenvolvedores e comunidades precisam atuar de forma conjunta para produzir tecnologias culturalmente sensíveis, socialmente úteis e alinhadas às necessidades locais.

Publicações recentes, como Pinhanez *et al.* (2024), reforçam essa visão ao indicar que as soluções de IA terão impacto positivo apenas se forem desenvolvidas em diálogo com os falantes e articuladas a políticas linguísticas de longo prazo. Assim, mais do que instrumentos técnicos, o PLN e a IA devem ser compreendidos como aliados potenciais na luta contra a perda da diversidade linguística, desde que implementados de forma crítica, ética e centrada nas comunidades.

REFERÊNCIAS

ADAMS, O.; *et al*. **Masakhane:** machine translation for Africa, by Africans. 2021. Disponível em: https://www.masakhane.io. Acesso em: 02 abr. 2025.

ANASTASOPOULOS, A. **Low-resource multilingual NLP:** from transfer learning to endangered languages. 2019.

BERNARD, H. R. Language death: the great extinction. *In:* GRENOBLE, L. A.; WHALEY, L. J. (orgs.). **Endangered Languages:** Language Loss and Community Response. Cambridge: Cambridge University Press, 1996. p. 142.

BIRD, S. Decolonising speech and language technology. *In:* **Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona: ICCL, 2020.

BIRD, S.; CHIANG, D. Machine translation for language preservation. *In:* **Proceedings of the Conference on Computational Linguistics (COLING)**. Mumbai: ACL, 2012.

BIRD, S.; SIMONS, G. Seven dimensions of portability for language documentation and description. **Language**, v. 79, n. 3, p. 557–582, 2003.

BOUDA, P.; *et al*. **ELAN and FLEx as tools for linguistic documentation**. 2012.

BROWN, T.; *et al*. Language Models are Few-Shot Learners. *In:* **Advances in Neural Information Processing Systems (NeurIPS)**. 2020.

CASTRO, Q. S.; APONTES, S. A. **A extinção de línguas indígenas no Brasil:** causas e perspectivas de revitalização. 2020.

CRYSTAL, D. **Language Death**. Cambridge: Cambridge University Press, 2000.

DWYER, A. M. Ethics and practicalities of cooperative fieldwork and analysis. *In:* GIPPERT, J.; HIMMELMANN, N.; MOSEL, U. (eds.). **Essentials of Language Documentation**. Berlin: Mouton de Gruyter, 2006.

GRENOBLE, L. A.; WHALEN, D. H. **Saving Languages:** An Introduction to Language Revitalization. Cambridge: Cambridge University Press, 2006.

HAUENSTEIN, A.; *et al*. **Neural approaches for low-resource speech recognition**. 2022.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3. ed. Draft, 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 02 abr. 2025.

KRAUSS, Michael E. The world's languages in crisis. **Language**, Washington, D. C., v. 68, n. 1, p. 4-10, 1992.

LIU, P.; *et al*. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. **ACM Computing Surveys**, v. 55, n. 9, 2023.

LONG NOW FOUNDATION. **The Rosetta Project**. 2024. Disponível em: https://rosettaproject.org. Acesso em: 02 abr. 2025.

MAIA, M. **Programas de Educação Bilíngue e Intercultural para a Preservação de Línguas Indígenas**. 2006.

MAGER, M.; *et al*. AmericasNLP: A New Workshop on NLP for Indigenous Languages of the Americas. *In:* **Proceedings of NAACL**. 2021.

MOTA, A. F.; SAMPAIO, J. M. N. Preservação das línguas indígenas e direito à memória: O caso dos Kokama. **Revista Brasileira de Linguística Antropológica**, Brasília, v. 11, n. 2, p. 238-261, 2019. Disponível em: https://periodicos.ufmg.br/index.php/relin/article/view/61404. Acesso em: 02 abr. 2025.

ONCEVAY, A.; *et al*. Machine Translation for Indigenous Languages of the Americas: the AmericasNLP Shared Task. *In:* **Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas**. 2020.

PINHANEZ, C.; *et al*. **Artificial Intelligence for Low-resource Languages:** opportunities and challenges. 2024.

RODRIGUES, A. D. Línguas indígenas: 500 anos de contato. **Revista Brasileira de Linguística Antropológica**, v. 5, n. 2, p. 15–33, 2013.

UNESCO. **Language Vitality and Endangerment**. Paris: UNESCO, 2003.

ZAIDAN, O.; *et al*. **Participatory Approaches in Language Technology for Endangered Languages**. 2021.

ZHANG, Z.; *et al*. **Cross-lingual Transfer Learning for Low-resource NLP**. Computational Linguistics, 2021.

Como referenciar este artigo:

SANTOS, Valéria Vieira dos; STASSI-SÉ, Joceli Catarina. O papel da inteligência artificial na revitalização de línguas em extinção por meio do processamento de linguagem natural. **revista Linguasagem**, São Carlos, v.49, n.1, p. 152-176, 2025.

Submetido em: 04/04/2025
Aprovado em: 27/08/2025