

LINGUÍSTICA COMPUTACIONAL E LÍNGUAS INDÍGENAS BRASILEIRAS

COMPUTATIONAL LINGUISTICS AND BRAZILIAN INDIGENOUS LANGUAGES

Arthur B. Scandelari¹
Dione M. Gomes²

RESUMO: Apresentamos os desafios da Linguística Computacional no processamento das línguas indígenas brasileiras, abordando os avanços recentes, as principais dificuldades e as perspectivas para o uso das tecnologias disponíveis na preservação e revitalização dessas línguas. Para tanto, fazemos uma breve introdução à área da Linguística Computacional e uma revisão, baseada em Mager et al. (2018), das ferramentas mais promissoras para a pesquisa. As línguas indígenas brasileiras representam um conjunto linguístico diversificado, com mais de 250 línguas, faladas por diferentes povos. Essa diversidade desafia as abordagens convencionais da Linguística Computacional, que não têm sido capazes de lidar com as particularidades dessas línguas, o que pode ser visto no descompasso entre as soluções tecnológicas e suas aplicações às línguas indígenas. A falta de investimento e de interesse social e político nas pesquisas de línguas nativas se somam à realidade linguística. A tradição oral e a ausência de sistemas ortográficos, bem como estruturas gramaticais distintas daquelas das línguas mais faladas, dificultam a criação de bases de dados necessárias para o treinamento de modelos computacionais e exigem inovações metodológicas que considerem suas especificidades. Além disso, a adoção de visões e conceitos eurocêntricos agravam os desafios. Mediante esta exposição, esperamos contribuir para a compreensão da situação atual entre a Linguística Computacional e as línguas indígenas brasileiras e reforçar a necessidade de aproximação das pesquisas em tecnologias e línguas nativas.

Palavras-chave: Processamento de Língua Natural (PLN). Línguas indígenas brasileiras. Diversidade linguística. Inteligência Artificial. Tipologia linguística.

ABSTRACT: We present the challenges of Computational Linguistics in processing Brazilian Indigenous languages, addressing recent advancements, the main difficulties, and the perspectives on using available technologies in the preservation and revitalization of these languages. To this end, we provide a brief introduction to the field of Computational Linguistics and a review, based on Mager et al. (2018), of the most promising tools for research. Brazilian Indigenous languages represent a diverse linguistic set, with more than 250 languages spoken by different peoples. This diversity challenges conventional approaches in Computational Linguistics, which have struggled to address the particularities of these languages, as seen in the disconnect between technological solutions and their application to Indigenous languages. The lack of investment and social and political interest in research on native languages adds to the linguistic reality. The oral tradition and the absence of orthographic systems, as well as grammatical structures distinct from those of the more widely spoken languages, complicate the creation of databases necessary for training computational models and demand methodological innovations that consider their specificities. Furthermore, the adoption of Eurocentric views and concepts exacerbates the challenges. Through this exposition, we aim to contribute to the understanding of the current situation between Computational Linguistics and Brazilian Indigenous languages and emphasize the need for closer integration between research on technologies and native languages.

Keywords: Natural Language Processing (NLP). Brazilian Indigenous Languages. Linguistic Diversity. Artificial Intelligence. Linguistic Typology.

¹ Doutor em Linguística pelo Programa de Pós-Graduação em Linguística da Universidade de Brasília. scandelari@gmail.com

² Professor Titular do Departamento de Linguística, Português e Línguas Clássicas da Universidade de Brasília. dione98@gmail.com

INTRODUÇÃO

Com o avanço da tecnologia e o uso crescente de sistemas de Processamento de Língua Natural (PLN), que constituem um subcampo da Inteligência Artificial, a Linguística Computacional tem se mostrado promissora na coleta de dados e análise linguística. Seus recursos são particularmente atraentes em um país como o Brasil, afinal, nossa nação possui uma das maiores diversidades linguísticas do planeta, o que representa um patrimônio cultural valioso e uma fonte de conhecimento linguístico e antropológico. Contudo, muitas línguas indígenas estão em processo de desaparecimento, e seu estudo e preservação enfrentam desafios significativos.

Nesse contexto, a busca pela aplicação de ferramentas computacionais surge na esperança de que as tecnologias atuais possam contribuir para a preservação, a revitalização e a valorização das línguas indígenas. A ampliação das bases de dados, por meio da automação da coleta de informações, o trabalho com quantidades expressivas de línguas e, eventualmente, até a automação das análises linguísticas são usos esperados das ferramentas disponíveis³.

Muitos pesquisadores de línguas indígenas buscam utilizar tais recursos em seus trabalhos, procurando instrumentos que sejam relativamente fáceis de usar e de testar, sem necessidade de programação, e que os auxiliem na extração ou na identificação automática de informações. Todavia, essas ferramentas nem sempre atendem a suas demandas. Diante da dificuldade de aplicar os recursos mais propalados da área – como sistemas de tradução automática (Bing, Google Tradutor) e modelos de linguagem de grande escala (*Large Language Models*, LLMs), a exemplo do ChatGPT e do Gemini –, pesquisadores recorrem a mecanismos específicos, voltados a línguas indígenas. Os resultados, porém, tendem a ser limitados, apresentando falhas frequentes no processamento de construções gramaticais complexas e na precisão lexical. Isso ocorre porque pesquisas com diferentes variáveis e voltadas às línguas indígenas demandariam a articulação de múltiplas tarefas da Linguística Computacional, como o reconhecimento automático de fala para línguas de tradição oral, a análise morfológica para línguas

³ Enfatizamos que iniciativas de documentação, digitalização e “virtualização” de línguas indígenas devem, necessariamente, envolver a participação ativa das comunidades falantes em todas as etapas do processo, desde a coleta até a disponibilização dos dados. Em conformidade com princípios internacionais, como o direito à Consulta Livre, Prévia e Informada (CLPI), previsto na Convenção nº 169 da Organização Internacional do Trabalho (OIT), tais projetos devem ser conduzidos somente com o consentimento explícito dessas comunidades, respeitando eventuais restrições quanto ao acesso, uso e circulação de seus conhecimentos linguísticos. É essencial, ainda, que tais iniciativas resultem em benefícios concretos e culturalmente relevantes para os próprios povos envolvidos, evitando práticas extrativistas ou a apropriação indevida de seus patrimônios linguísticos. Para mais informações: <https://www.gov.br/funai/pt-br/arquivos/conteudo/ascom/2014/doc/11-nov/convencaooit.pdf>

polissintéticas e a criação de *treebanks*, as quais ainda não estão disponíveis para usuários que não possuem especialização técnica para operar em *softwares* específicos.

O objetivo principal deste artigo é apresentar os desafios da Linguística Computacional no processamento de línguas indígenas brasileiras, fornecendo uma síntese das ferramentas disponíveis atualmente e discutindo as limitações das abordagens convencionais diante da diversidade tipológica dessas línguas. Essa discussão se justifica pela necessidade de estratégias tecnológicas que auxiliem na preservação e revitalização de um patrimônio de mais de 250 línguas, muitas em risco de desaparecimento, e pela urgência de superar visões que marginalizam as chamadas “línguas de baixo recurso”.

Para alcançar esse propósito, o texto está organizado da seguinte forma: primeiramente, realizamos uma breve introdução à área da Linguística Computacional. Em seguida, descrevemos o contexto linguístico na América e no Brasil e os impactos da escassez de dados digitais. Depois, apresentamos uma revisão das ferramentas computacionais com potencial prático para as línguas da região. Por fim, analisamos os desafios específicos do processamento de dados e encerramos com as considerações finais.

Esclarecemos que o levantamento das aplicações da Linguística Computacional nas línguas indígenas americanas, assim como a lista das ferramentas mais usadas na área, se baseia principalmente em Mager *et al.* (2018)⁴.

LINGÜÍSTICA COMPUTACIONAL

O século XX marca o desenvolvimento da Linguística Computacional, com o surgimento dos primeiros computadores e as tentativas de usá-los para lidar com desafios linguísticos. Após a Segunda Guerra Mundial, pesquisadores começaram a explorar aplicações civis para os cálculos e a computação recém-desenvolvidos. O final dos anos 1940 e o início dos anos 1950 foram marcantes nesse sentido, com destaque para os avanços possibilitados pelo Memorando de Weaver (1949) e pelo Teste de Turing (1950), voltados a um dos temas mais desafiadores da área, a tradução automática.

Nas décadas de 1980 e 1990, com o avanço da computação e a difusão da Internet, a Linguística Computacional se expandiu, e foram desenvolvidos sistemas mais sofisticados de análise de texto, geração de linguagem e interação entre humanos e máquinas. Atualmente, testemunhamos a evolução da Inteligência Artificial, do aprendizado de máquinas e do processamento de *big data* (‘grandes volumes

⁴ Agradecemos ao professor Thiago Pires (UnB), que gentilmente fez a seleção de referências bibliográficas e esclareceu nossas dúvidas a respeito de processos computacionais específicos

de dados⁵). As redes neurais⁵ e a aprendizagem profunda (*deep learning*) têm se destacado nas tarefas linguísticas⁶.

Entre as aplicações modernas mais conhecidas, podemos citar as perguntas e respostas (DialogFlow, Watson), os motores de busca (Google, Yahoo!), os assistentes de língua natural (Alexa, Siri), os sistemas de tradução (Bing, Google Tradutor) e os modelos de linguagem de grande escala (LLMs) (ChatGPT, Gemini).

Para desenvolver tais sistemas, capazes de entender, interpretar e até gerar a linguagem humana de forma automatizada, a Linguística Computacional combina conhecimentos de diferentes áreas, como Linguística, Ciência da Computação e Inteligência Artificial. Além disso, precisa extrair informações linguísticas de grandes volumes de dados textuais, com base em diferentes técnicas, a exemplo de algoritmos de processamento de texto, modelos probabilísticos e os já mencionados, aprendizado de máquina e redes neurais.

Esclarecemos que, dependendo do seu foco em um ou outro campo do conhecimento, a Linguística Computacional é tratada por outros nomes, com algumas diferenças entre eles. A chamada “Linguística Computacional” é mais teórica e tem ênfase na Linguística, usando conceitos teóricos e códigos matemáticos para representar e formalizar suas ações. O Processamento de Língua Natural (PLN) se concentra em engenharia e ferramentas que forneçam soluções práticas. O Tratamento Automático de Línguas Naturais é mais abrangente e contempla as duas primeiras. Constatamos, no entanto, que a literatura nem sempre distingue as terminologias.

A seguir, expomos o contexto linguístico na América, com destaque para o Brasil, as ferramentas computacionais disponíveis, alguns exemplos de uso em línguas indígenas brasileiras e os desafios para a disseminação dessas tecnologias na pesquisa nacional.

LÍNGUAS INDÍGENAS AMERICANAS E RECURSOS COMPUTACIONAIS

A diversidade linguística é uma característica notável do continente americano. Os fenômenos linguísticos encontrados nas cerca de novecentas línguas indígenas da região, divididas em dezenas de famílias, são distintos daqueles comumente estudados na Linguística Computacional. Suas ferramentas se aplicam a línguas amplamente pesquisadas, com tradição escrita e quantidades imensas de dados

⁵ Redes neurais são modelos computacionais de inteligência artificial inspirados no funcionamento do cérebro humano. São capazes de aprender padrões complexos a partir de grandes conjuntos de dados. Além disso, dispõem de sistemas adaptativos, que possibilitam aos computadores se adequarem a novas situações (Jurafsky; Martin, 2023).

⁶ A aprendizagem profunda é um tipo de rede neural que envolve múltiplas camadas interconectadas, o que permite o processamento de dados novos e o aprendizado de funções para as quais não foi programada (Jurafsky; Martin, 2023).

disponíveis digitalmente. A situação é diferente quando tentamos utilizá-las em línguas menos conhecidas, sem sistema ortográfico e sem investimentos suficientes em pesquisa.

Em consequência, existem poucas ferramentas destinadas a essas línguas. As tarefas mais pesquisadas nas américas são a tradução automática e a análise morfológica, mas ainda produzem resultados insuficientes. De acordo com Mager *et al.* (2018, p. 56, tradução nossa⁷),

A grande diversidade dessas línguas apresenta desafios científicos interessantes, p.ex., adaptação de abordagens bem estabelecidas, criação de novos métodos, coleta de novos dados. Enfrentar esses desafios contribui para construir modelos computacionais mais gerais de linguagem e para obter uma visão mais profunda na compreensão da linguagem humana.

Alguns dos desafios recorrentes encontrados nas línguas indígenas brasileiras são: morfologia rica com alto grau de síntese morfológica (línguas polissintéticas), ausência de ortografia ou de normalização ortográfica, variação dialetal, tradição oral, falta de programas de alfabetização e educação para os falantes. Além disso, poucos dados disponíveis e escassez de programas de pré-processamento linguístico aumentam o desafio (Mager *et al.*, 2018, p. 56).

Pelo fato de manterem forte tradição oral e serem, normalmente, ágrafas, as línguas indígenas brasileiras têm baixa presença na Internet e nos textos digitalizados. Apesar de jovens indígenas usarem as redes sociais nas comunidades e divulgarem conteúdos e manifestações artísticas, essa tendência ainda é incipiente, e o volume de dados continua escasso para as necessidades dos modelos computacionais (OEI, 2024; Teixeira de Sousa, 2020).

Há, no entanto, outro fator que precisa ser registrado quanto à falta de projetos para as línguas indígenas, o qual tem relação com o conceito de “línguas de baixo recurso” (Cieri *et al.*, 2016, tradução nossa⁸). A definição do que é ser uma língua baixa em recursos e os critérios de sua apuração são um tema controverso e debatido. Ela pode incluir línguas que dispõem de poucos dados disponíveis para treinamento de modelos de PLN, como vimos, mas sobretudo aquelas que possuem poucos dados digitais em relação à projeção internacional da língua.

Em outras palavras, esse conceito acaba funcionando como uma medida relativa. Como a comparação costuma ser feita com o Inglês, que tem uma presença digital altíssima, há um viés para rotular a maioria das línguas como de baixo recurso (Hämäläinen, 2021). Nesse sentido, até o português pode ser considerado baixo em recursos, pois a quantidade de falantes e o alcance que tem no mundo

⁷ No original: The great diversity of these languages poses interesting scientific challenges, e.g., adapting well established approaches, creation of new methods, collecting new data. Tackling these challenges contributes to building more general computational models of language, and to get a deeper insight into human language understanding (Mager *et al.*, 2018, p. 56).

⁸ No original: [...] low resource languages [...].

não condiz com sua relativa escassez de dados digitais (vide a inexistência de um *corpus* anotado semanticamente).

O problema da adoção desse conceito é que ele acaba influenciando quais línguas receberão investimento e quais não. O debate envolve autores que defendem abertamente a inutilidade de investir nas línguas que receberam o rótulo de “baixo recurso”, definido pelo critério exposto, reduzindo ainda mais o interesse científico, político e social nessas línguas (Cieri *et al.*, 2016; Hämäläinen, 2021). Para uma discussão sobre a viabilidade de criar *corpora* para línguas com pouca presença digital, conferir Teixeira de Sousa (2020), que apresenta os desafios oriundos da escassez e heterogeneidade dos dados, mas defende a necessidade de metodologias flexíveis e de valorização da representatividade para a constituição desses materiais.

FERRAMENTAS COMPUTACIONAIS

Entre as ferramentas com potencial prático nas línguas indígenas, Mager *et al.* (2018) citam o *corpus* paralelo e os dados morfológicos anotados. Aquele se refere ao conjunto de textos em duas ou mais línguas que são alinhados de forma correspondente, havendo um equivalente direto na outra língua que possibilita comparar as versões. Os textos paralelos foram comuns no período colonial da América do Sul, pela ação de missionários europeus que traduziam a bíblia para línguas indígenas, e outros textos supostamente educativos, a fim de aculturar e evangelizar os povos originários. O inconveniente no uso de documentos antigos, todavia, é a tendência à reprodução do léxico, dos termos e dos valores registrados nesses textos. Vemos isso, atualmente, em modelos de linguagem de grande escala baseados em Inteligência Artificial Generativa (IAG), como ChatGPT e Gemini, que propagam discursos preconceituosos de suas bases de dados ou que incorporam palavras de uma língua indígena como se fossem aplicáveis a outras línguas, levando a uma predominância de termos e expressões provenientes das famílias linguísticas mais conhecidas, em detrimento das demais⁹. Não obstante, o *corpus* paralelo é importante para criar e treinar sistemas de Tradução Automática Estatística (SMT), os quais traduzem textos com base na tradução mais provável, calculada a partir do *corpus* disponível.

Os dados morfológicos anotados (ou categorizados) são importantes para ferramentas de etiquetagem morfossintática, que leem textos e categorizam gramaticalmente seus elementos, e para analisadores morfológicos, que extraem automaticamente as formas morfológicas de uma palavra. Os dados morfológicos anotados são usados para treinar e avaliar modelos de linguagem e sistemas de

⁹ Agradecemos a um(a) revisor(a) anônimo(a) pelas reflexões sobre o impacto da Inteligência Artificial na disseminação de palavras de origem guaraní ou tupí no contexto latino-americano.

PLN. Uma proposta de anotação morfossintática no Kadiwéu (Guaikurú) pode ser conferida em Sandalo e Galves (2023), que problematizam a transposição automática, para línguas indígenas, de esquemas originalmente desenvolvidos para línguas indo-europeias, e fornecem uma proposta de anotação categorial, morfológica e sintática que respeita as especificidades estruturais da língua analisada.

Outro recurso útil é o *treebank*, um banco de dados que parte do *corpus* previamente anotado com estruturas sintáticas, revisado por um linguista, e identifica automaticamente a sintaxe e a semântica de um texto. É empregado tanto para desenvolver quanto para avaliar modelos de análise sintática. A partir dele, é possível treinar modelos e desenvolver gramáticas formais. Embora seja raro encontrar *treebanks* aplicados às línguas indígenas americanas, verificamos que foram desenvolvidos *treebanks* para algumas línguas brasileiras, como Apurinã (Arawak) (Rueter *et al.*, 2021), Mbyá Guaraní (Tupí-Guaraní) (Thomas, 2019) e Nheengatú (Tupí-Guaraní) (Alencar, 2023; 2024).

Em relação à documentação e à revitalização de línguas indígenas, processos essenciais para preservar a diversidade linguística e cultural da humanidade, existem ferramentas que podem potencializá-las, como o reconhecimento automático da fala, os mecanismos de alinhamento forçado e a já citada anotação morfológica automática.

O reconhecimento automático da fala (*automatic speech recognition*) é uma tecnologia que capta a voz humana, reconhece, decodifica e converte em texto escrito. É utilizado com frequência na criação de legendas e nas transcrições. Essas e outras aplicações têm se beneficiado dos avanços no aprendizado de máquinas (*machine learning*), por meio do qual os computadores identificam padrões e fazem previsões, de modo a realizar tarefas para as quais não foram programados.

Os mecanismos de alinhamento forçado (*forced alignment tools*) recebem sons e transcrições e os ordenam temporalmente entre si, alinhando a emissão do fonema ao respectivo texto. São empregados para transcrição de áudio e legendagem automática de vídeos.

Mager *et al.* (2018) afirmam que a Wikipédia costuma ser uma fonte útil para montar bancos de dados e etiquetadores de línguas faladas nas Américas, já que o *site* possui páginas dedicadas a essas línguas. Contudo, os autores lembram que, mesmo quando há ferramentas disponíveis, a falta de padronização ortográfica e a dificuldade de processar caracteres específicos acaba restringindo o uso dos recursos digitais. A disponibilização das listas de palavras mais frequentes em uma língua, incluindo suas flexões, pode ajudar a sanar alguns desses problemas, além de ser útil na criação de corretores ortográficos.

Em relação às análises morfológicas, a Linguística Computacional tem se concentrado na lematização, na estemização ('redução ao radical'), na segmentação e na *reinflection* ('recolocação de flexão'). A lematização é o processo que visa à identificação automática da forma de citação das palavras, o lema, e depende do reconhecimento de todo o contexto para obter um resultado confiável. A

estemização busca reconhecer o radical das palavras, de modo a unir palavras com radicais semelhantes. Ambas as técnicas se baseiam em algoritmos e reduzem a variação morfológica no texto, pelo fato de agrupar palavras, facilitando a compreensão do sentido geral.

A segmentação é usada para separar automaticamente os morfemas e depende do conhecimento da estrutura da língua, sendo importante no pré-processamento de texto. A *reinflection* é usada para recolocar a flexão nas palavras que passaram por lematização estemização, servindo para prever como será a flexão morfológica, por meio do processamento do lema em conjunto com as regras gramaticais.

Apesar da variedade de recursos existentes direcionados à morfologia, Mager *et al.* (2018, p. 58-59) alertam que eles não estão adaptados às particularidades das línguas indígenas, cuja complexidade morfológica exige mais do que apenas remover sufixos para obter um radical. Por exemplo, uma única palavra da língua Mundurukú (Mundurukú) representa uma sentença completa em outras línguas:

- (1) *o'=su-bu-dakat-∅*
 3a.SING.SUJ=3a.SING.OBJ-dedo-cortar-PERF
 'Ele cortou o dedo (da criança).' (Gomes, 2024, comunicação pessoal)¹⁰

Outra ferramenta de uso potencial para as línguas indígenas brasileiras é o processamento de língua baseado em regras (*rule-based*), que verifica o conteúdo textual a partir de um conjunto de critérios pré-definidos. Esse conteúdo pode envolver a correção gramatical, a análise morfológica, a etiquetagem, entre outros. A título de exemplo, Mager *et al.* (2018) citam um programa que pode ser útil para algumas línguas indígenas, devido aos materiais disponíveis de séculos passados. Segundo Mager *et al.* (2018, p. 60, tradução nossa¹¹),

Para as línguas uto-astecas, existe uma ferramenta computacional chamada *chachalaca* que realiza análises morfológicas (Thouvenot, 2011) do Náuatle. Esse é um *software* baseado em regras focado no Náuatle Clássico, sendo capaz de gerar mais de um candidato de análise morfológica por palavra. Ele se baseia em gramáticas que descrevem a maioria das construções de palavras do século XVI.

A seguir, listamos mais algumas ferramentas que têm sido testadas nas línguas indígenas das Américas e que podem vir a beneficiar a pesquisa linguística em nosso continente:

¹⁰ As seguintes abreviaturas e símbolos são usados na glosa: OBJ: objeto; PERF: aspecto perfectivo; SING: singular; SUJ: sujeito; 3a: terceira pessoa ativa/processual; =: fronteira de clítico.

¹¹ No original: *For Uto-Aztecan languages, there exists a computational tool called "chachalaca" that performs morphological analysis (Thouvenot, 2011) of Nahuatl. This is a rule-based software focused on Classical Nahuatl, it is able to generate more than one morphological analysis candidate per word. It is based on grammars that describe most of the 16th-century-word constructions.*

a) Aprendizagem semissupervisionada (*semi-supervised*): parte de dados etiquetados e não etiquetados, sendo um método híbrido de treinamento computacional para compreender a língua, rotular, fazer classificações e previsões. É a mais comum, pelo fato de combinar dados classificados e não classificados e agilizar a anotação manual de muitos dados;

b) Aprendizagem não supervisionada (*unsupervised*): parte de dados não etiquetados, de modo que não faz previsões, mas é capaz de identificar padrões, rotular, agrupar dados e oferecer análises a partir de grande quantidade de dados. Esse recurso opera pela representação vetorial do espaço (matemática vetorial) e necessita de uma quantidade muito extensa de dados para funcionar corretamente. Quando há tantas informações, é capaz de identificar os contextos e minimizar significativamente os erros. Entretanto, como essa ferramenta se baseia em informações não categorizadas previamente, se não houver toda essa disponibilidade de dados, pode gerar resultados incoerentes, motivo pelo qual também requer supervisão humana;

c) Autômato de estado finito (*finite state approach*): é composto por um conjunto finito de estados e uma série de transições entre esses estados, que são ativadas por entradas específicas. Assim, pode atuar em áreas distintas, como definição de estado, definição de transições, processamento de texto, reconhecimento de padrões, entre outros. Na definição de estado, cada estado representa uma condição específica do processamento. Por exemplo, um estado pode representar a identificação de um substantivo em uma sentença, enquanto outro expressaria a identificação de um verbo. Na definição de transições, as transições entre os estados são determinadas pelas entradas, as quais podem ser palavras individuais ou sequências de caracteres. A entrada de uma palavra, por exemplo, pode fazer o autômato transitar de um estado que representa o sujeito para outro que representa o predicado;

d) Transdutor de estado finito (*finite-state transducer*): além de realizar as tarefas do autômato de estado finito, também pode fazer transformações ou associações entre sequências de entrada e sequências de saída. Neste sentido, ele produz uma sequência de saída à medida que avança. Essa capacidade é útil para tarefas como a tradução automática, em que o programa recebe uma sentença em uma língua e produz a tradução correspondente em outra, na saída. O transdutor também é capaz de gerar textos.

Mager *et al.* (2018, p. 60, tradução nossa¹²) observam que o transdutor de estado finito é empregado com frequência para realizar análises morfológicas:

A maioria dos métodos que encontramos que lidam com morfologia são baseados em FST [Transdutor de Estado Finito]. Contudo, as línguas indígenas do continente são muito diversas e seria caro construir tais analisadores com conhecimento especializado para cada língua, além do fato de os analisadores precisarem ser constantemente atualizados para lidar com a mudança linguística.

Em nosso ponto de vista, além do desafio de lidar com os interesses envolvidos no conceito de “línguas de baixo recurso”, essa parece ser uma das principais dificuldades envolvendo as línguas indígenas. Faltam investimentos para construir programas com conhecimento especializado sobre cada língua, incluindo o custo futuro de atualização constante.

Mager *et al.* (2018, p. 60, tradução nossa¹³) também destacam a tradução automática nos estudos sobre línguas indígenas, como uma ferramenta de uso potencial pelas comunidades:

[a] Tradução Automática é uma tarefa natural para línguas indígenas, pois pode fornecer uma oportunidade de comunicação com línguas mais populares. O desenvolvimento de sistemas de TA [tradução automática] para línguas indígenas seguiu as tendências na área, desde as abordagens baseadas em regras até as estatísticas e neurais.

A tradução de conteúdos importantes para as comunidades, como textos sagrados, narrativas históricas, documentos legais e materiais educacionais e literários, pode se beneficiar desse tipo de ferramenta. Os autores observam, todavia, que traduções de línguas com poucos dados disponíveis ainda são um desafio na área. Os sistemas atuais costumam apresentar defeitos e falhar em traduzir construções complexas.

Outros produtos que podem ser potencializados pela Linguística Computacional, e que são resultados comuns do trabalho entre linguistas e comunidades indígenas, são os dicionários, os glossários e os materiais didáticos, sendo todos relevantes no apoio e na preservação das línguas e culturas indígenas.

Mager *et al.* (2018) observam que a coleta automática de dados é uma ferramenta capaz de amenizar a carência de dados e facilitar a criação de grandes repositórios de textos traduzidos. Há,

¹² No original: Most of the methods that we found that deal with morphology are based on FST. However, the indigenous languages of the continent are far too diverse, it would be expensive to build such analyzers with expert knowledge for each language, besides the fact that the analyzers need to be constantly updated to cope with language change (Creutz and Lagus, 2005).

¹³ No original: Machine Translation is a natural task for indigenous languages, since it might provide a communication window with more popular languages. The development of MT systems for indigenous languages have follow the trends in the field, from rule-based, to statistical and neural based approaches.

entretanto, a necessidade de investimento na área para adaptar os sistemas e lidar adequadamente com a realidade das línguas indígenas. Lembramos que teses, dissertações e artigos costumam ter textos analisados e traduzidos, mas muitos não estão digitalizados e precisam ser consultados um a um. O desafio da Linguística Computacional acaba sendo não só a escassez de dados, mas de tempo e pessoas para coletar e tratar dados que já foram produzidos academicamente.

Apesar disso, o emprego de modelos computacionais apresenta tendência crescente. Segundo Mager *et al.* (2018, p. 63), o estudo das línguas indígenas tem aumentado a partir de 2013, porém essas pesquisas se concentram nas línguas da América do Norte. Os autores constatarem que abordagens morfológicas e de modelagem de subpalavras, em que as palavras são segmentadas em partes menores para facilitar a análise, tendem a produzir resultados positivos para línguas com morfologia complexa.

Por fim, chamamos atenção para o fato de que algumas abordagens computacionais estão fundamentadas em teorias linguísticas. Isso significa que, embora estudemos os autômatos como algo dado, existe um princípio norteador por trás de seus processos, o qual é fornecido por uma gramática teoricamente orientada. No caso do autômato de estado finito, por exemplo, essa orientação é chomskyana. Assim, cada tipo de autômato terá relação com um tipo de gramática, a depender de sua base teórica. Um autômato pode ser sensível ao contexto, por exemplo, se a teoria em que se fundamenta incluir questões funcionais entre suas variáveis de análise.

CONSIDERAÇÕES FINAIS

Neste artigo, apresentamos os desafios da Linguística Computacional no processamento de línguas indígenas, as ferramentas computacionais mais empregadas e suas limitações para a pesquisa sobre línguas indígenas brasileiras. A diversidade dessas línguas, que apresentam especificidades gramaticais, requer a adaptação dos modelos computacionais existentes, possivelmente com abordagens inovadoras, para que as funcionalidades da Linguística Computacional se tornem uma realidade também entre elas. As soluções dependem não só do interesse científico, mas também político e social, voltado ao reconhecimento e à valorização dos povos nativos.

Para se aprofundar na Linguística Computacional, as obras de referência são Jurafsky e Martin (2023), Bird, Klein e Loper (2019) e Manning e Schütze (1999). Para leituras adicionais sobre temas que podem auxiliar as pesquisas em línguas indígenas, sugerimos os seguintes artigos: Brunato *et al.* (2020), Li *et al.* (2020), Virk *et al.* (2021), Kargaran *et al.* (2023) e Linders e Louwerse (2023). Os links de todos esses estudos estão disponíveis nas referências bibliográficas.

Por fim, Mager *et al.* (2018) disponibilizam *online* uma lista atualizada¹⁴ com os avanços do PLN em línguas indígenas das Américas, reunindo pesquisas, recursos linguísticos e ferramentas técnicas com o objetivo de mapear e organizar os trabalhos sobre PLN no continente, além de incentivar pesquisadores a ampliar seus esforços na criação de tecnologias para essas línguas.

REFERÊNCIAS

- ALENCAR, L. F. Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework. Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL). *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL). SBC. *Anais* [...] 25 set. 2023. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25445>. Acesso em: 28 jun. 2024.
- ALENCAR, L. F. A Universal Dependencies Treebank for Nheengatu. *In*: CONFERÊNCIA INTERNACIONAL DE PROCESSAMENTO COMPUTACIONAL DO PORTUGUÊS, 16., 2024, Santiago de Compostela. *Anais* [...] Santiago de Compostela: Associação de Linguística Computacional (ACL), 2024. v. 2. Disponível em: <https://aclanthology.org/2024.propor-2.0.pdf>. Acesso em: 28 jun. 2024.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python*. Analyzing Text with the Natural Language Toolkit. 2019. Disponível em: <https://www.nltk.org/book/>. Acesso em: 15 jul. 2024.
- BRUNATO, D.; CIMINO, A.; DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. Profiling-UD: a Tool for Linguistic Profiling of Texts. *In*: CONFERÊNCIA SOBRE RECURSOS E AVALIAÇÃO DE LÍNGUA, 12., 2020. *Anais* [...], 2020, p. 7145-7151. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.883.pdf>. Acesso em: 15 jul. 2024.
- CIERI, C.; MAXWELL, M.; STRASSEL, S.; TRACEY, J. Selection Criteria for Low Resource Language Programs. *In*: CONFERÊNCIA INTERNACIONAL SOBRE RECURSOS E AVALIAÇÃO DE LÍNGUA, 10., 2016, Portorož. *Anais* [...]. Portorož: Associação Europeia de Recursos Linguísticos (ELRA), 2016. p. 4543-4549. Disponível em: <https://aclanthology.org/L16-1720/>. Acesso em: 15 jul. 2024.
- HÄMÄLÄINEN, M. Endangered Languages are not Low-Resourced. *Preprints*, 2021. Disponível em: <https://doi.org/10.20944/preprints202104.0113.v1>. Acesso em: 15 jul. 2024.
- JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3 ed. 2023. Disponível em: <https://web.stanford.edu/~jrafsky/slp3/>. Acesso em: 15 jul. 2024.
- KARGARAN, A. H.; IMANI, A.; YVON, F.; SCHÜTZE, H. GlotLID: Language Identification for Low-Resource Languages. *Findings of the Association for Computational Linguistics*, EMNLP, p. 6155-6218, dez. 2023. Disponível em: <https://aclanthology.org/2023.findings-emnlp.410.pdf>. Acesso em: 15 jul. 2024.
- LI, XiuHong; LI, Zhe; SHENG, J.; SLAMU, W. Low-Resource Text Classification via Cross-lingual Language Model Fine-tuning. *In*: CONFERÊNCIA NACIONAL DA CHINA SOBRE LINGÜÍSTICA COMPUTACIONAL, 19.,

¹⁴ Disponível em: <https://github.com/pywirrarika/naki>

2020. *Anais [...]*, 2020. p. 994-1005. Disponível em: <https://aclanthology.org/2020.ccl-1.92.pdf>. Acesso em: 17 jul. 2024.

LINDERS, G. M.; LOUWERSE, M. M. *Lingualyzer*. A computational linguistic tool for multilingual and multidimensional text analysis. *Behav Res*, 2023. Disponível em: <https://doi.org/10.3758/s13428-023-02284-1>. Acesso em: 17 jul. 2024.

MAGER, M.; GUTIERREZ-VASQUES, X.; SIERRA, G.; MEZA, I. Challenges of language technologies for the indigenous languages of the Americas. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 20-26, p. 55-69. 2018. Disponível em: <https://aclanthology.org/C18-1006/>. Acesso em: 10 jul. 2024.

MANNING, C.; SCHÜTZE, H. *Foundations of statistical Natural language Processing*. Cambridge: MIT Press, 1999. Disponível em: <http://nlp.stanford.edu/fsnlp/>. Acesso em: 16 jul. 2024.

ORGANIZAÇÃO DE ESTADOS IBERO-AMERICANOS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA (OEI). *Línguas indígenas no mundo digital*. inventário de recursos e lacunas. OEI, 2024. Disponível em: <https://oei.int/pt/escritorios/secretaria-geral/publicacoes/linguas-indigenas-no-mundo-digital-inventario-de-recursos-e-lacunas/>. Acesso em: 28 jun. 2024.

RUETER, J.; FREITAS, M. F. P.; FACUNDES, S. S.; HÄMÄLÄINEN, M.; PARTANEN, N. Apurinã Universal Dependencies Treebank. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. *Anais... /n. PROCEEDINGS OF THE FIRST WORKSHOP ON NATURAL LANGUAGE PROCESSING FOR INDIGENOUS LANGUAGES OF THE AMERICAS*. Online: Association for Computational Linguistics, 2021. Disponível em: <https://www.aclweb.org/anthology/2021.americasnlp-1.4>. Acesso em: 30 nov. 2023.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: o problema de Anchieta. *Cadernos de Estudos Linguísticos*, v. 65, p. e023007-e023007, 2023.

TEIXEIRA DE SOUSA, L. Sobre a constituição de corpora para línguas com poucos recursos. *Revista Linguística*, v. 16, n. 1, p. 43-61, 30 abr. 2020.

THOMAS, G. Universal Dependencies for Mbyá Guaraní. */n. WORKSHOP ON UNIVERSAL DEPENDENCIES*, 3., 2019, Paris. *Anais [...]*. Paris: Associação de Linguística Computacional (ACL), 2019. Disponível em: <https://aclanthology.org/W19-8008>. Acesso em: 10 jun. 2024.

VIRK, S. M.; FOSTER, D.; MUHAMMAD, A. S.; SALEEM, R. A Deep Learning System for Automatic Extraction of Typological Linguistic Information from Descriptive Grammars. *Proceedings of Recent Advances in Natural Language Processing*, p. 1480-1489, 2021. Disponível em: <https://aclanthology.org/2021.ranlp-1.166.pdf>. Acesso em: 10 jun. 2024.