REVISTA DA **ABRALIN**

APRESENTAÇÃO

Novos caminhos para a documentação e o estudo de línguas dos povos originários: apresentação

Filomena SANDALO
Universidade Estadual de Campinas (UNICAMP)

OPEN ACCESS

COMO CITAR
Sandalo, F. (2026). Novos caminhos para a documentação e o estudo de línguas dos povos originários: apresentação. *Revista da Abralin*, v. 24, n. 2, p. 1-8, 2025.

> *...academic linguists (me among them) continue documenting minoritized languages. They continue to believe in their "mission", building archives and celebrating discoveries of structures, lexicons, verbal arts, even when they feel some kind of malaise."*
>
> Bruna Franchetto, 2026, p. 16

Introdução

O objetivo deste dossiê é o de abrir caminhos para novas metodologias de registro e de documentação de dados coletados em trabalho de campo linguístico, com benefícios mútuos para as populações indígenas e para os linguistas.

Este dossiê apresenta, assim, várias propostas diferentes de documentação das línguas da América do Sul, especialmente do Brasil, com base em estudos linguísticos. Esperamos poder colaborar através de nosso conhecimento científico com a manutenção das línguas, com novas pesquisas e com a revitalização de línguas cujos falantes, por motivos adversos, foram levados a uma situação de substituição de suas línguas originárias e que atualmente buscam alternativas para sua retomada ou revitalização.

A inspiração para este dossiê teve como origem dois eventos:

i) o simpósio temático 14 do congresso da ABRALIN de 2023 - O *papel da linguística na produção de conhecimentos para fortalecer e vitalizar as línguas originárias* (https://www.congresso2023.abralin.org/simposio/view?ID_SIMPOSIO=86), e

ii) o projeto temático e-Science FAPESP, desenvolvido pela autora deste artigo em colaboração com linguistas, pesquisadores indígenas e cientistas da computação, *Corpora anotados digitais de línguas indígenas brasileiras com traduções automáticas* (DACILAT, Fapesp Processo 22/09158-5).

As atividades nesses eventos mostraram uma necessidade de organizar uma publicação que pudesse apresentar em uma única publicação diferentes esforços, no Brasil, de documentação e preservação de suas línguas nativas, fortemente ameaçadas atualmente. Vários artigos que compõem este volume são resultantes do simpósio da ABRALIN, de fato.

Neste artigo inicial, apresentamos os trabalhos incluídos neste dossiê e, em seguida, apresentamos também o projeto DACILAT, mencionado acima, e seus resultados já obtidos, já que este projeto é também uma proposta de documentação das línguas do Brasil.

1. Os artigos deste volume temático

No primeiro artigo, Bruna Franchetto, da Universidade Federal do Rio de Janeiro, apresenta um panorama da vitalidade das línguas indígenas no Brasil e dos esforços para documentá-las, um pré-requisito para sua preservação ou revitalização. A autora apresenta ainda uma reflexão crítica crucial sobre o papel do linguista diante dos povos originários e suas línguas.

O segundo artigo apresenta uma discussão sobre a natureza dos dados documentados. É uma tradição na literatura separar (i) dados coletados através de textos (dados naturais ou naturalísticos) de (ii) dados coletados através de sentenças elicitadas pelo/s pesquisador/es (dados elicitados). Kristina Balykova, Universidade do Texas/Austin, EUA, discute esta classificação e defende que a divisão entre espontâneo e elicitado é simplista e deveria ser abolida.

No terceiro artigo, Eva-Maria Roessler, da Universidade do Minho, Portugal, e Pikygi Torales, pesquisador indígena e falante nativo da língua aché, falada no Paraguai, problematizam um passo importante para o início da documentação linguística: qual ortografia usar. As línguas indígenas em geral contam com diferentes ortografias em paralelo e uma uniformização ortográfica é um passo para um corpus sistemático. Essa sistematização precisa ser feita, entretanto, de modo a não fomentar uma língua padrão dominante. O caso da língua aché é particularmente interessante porque até recentemente ela era considerada um dialeto do guarani e, portanto, a ortografia do guarani era a empregada. Atualmente o aché é reconhecido como uma língua autônoma.

A partir deste ponto, trazemos artigos específicos com propostas e experiências de documentação linguística e cultural, através da coleta de vários tipos de materiais linguísticos.

Angela Chagas (Universidade Federal do Pará), Eduardo Vasconcelos (Universidade Federal do Amapá) e Fernando De Carvalho (Universidade Federal do Rio de Janeiro) apresentam uma proposta que inclui o uso de ferramentas, hoje nas mãos dos pesquisadores dedicados à documentação, como ELAN (um software gratuito do *Max Planck Institute*) e FLEX (um software de criação de dicionários do SIL Technology), para a documentação da língua Wai Wai. O projeto apresentado registra entrevistas com professores indígenas, pessoas contando suas trajetórias de vida ou ensinando como fazer artefatos tradicionais. A compilação, preservação e disponibilização desse material pode ajudar futuramente os próprios Wai Wai e outros pesquisadores a desenvolverem diversos tipos de materiais e

atividades que contribuam para que a língua continue sendo falada pelos membros dessa comunidade indígena e não seja substituída pelo português, língua dominante em todo o Brasil.

O quinto artigo apresenta quatro projetos distintos realizados na Universidade de São Paulo (USP), todos multidisciplinares, envolvendo linguística, tecnologia digital e a aplicação de inteligência artificial (I.A.) para apoiar a documentação e o fortalecimento de línguas indígenas brasileiras, por meio de ferramentas tecnológicas, como assistentes de escrita (corretores ortográficos, completadores de palavras e sentenças e tradutores) e bancos de dados linguísticos. O uso da inteligência artificial ainda é extremamente desafiador para as línguas minorizadas, o que exige criatividade por parte dos especialistas em tecnologia e I.A. ao adaptarem os modelos de linguagem pensados para línguas com grandes recursos como o inglês, alemão e português para trabalharem com quantias muito baixas de dados. Trata-se, portanto, de um projeto bastante inovador que dialoga e se contrapõe ao projeto DACILAT descrito mais adiante nesta introdução.

Recebemos também várias experiências de elaboração de dicionários. Dicionários são bastante importantes no processo de revitalização linguística e são requisitados pelos próprios indígenas no atual processo de levante indígena.

Assim, o sexto artigo apresenta uma metologia inovativa de documentação lexical ao discutir a elaboração de dicionários multimídias de línguas indígenas. Especificamente, o artigo de uma equipe interdisciplinar do Museu Paraense Emílio Goeldi e da Universidade do Novo México, EUA, apresenta uma metodologia de elaboração de dicionários multimídias digitais, como uma tecnologia social desenvolvida para apoiar a aprendizagem e revitalização de línguas indígenas. O artigo também discute o papel destes dicionários na revitalização de línguas originárias do Brasil, a partir de dados de oito línguas.

A sétima contribuição é a de Carmen Añn Brasolin, doutoranda em Antropologia Social na Universidade Estadual de Campinas, e trata da documentação audiovisual da língua karara a partir de entrevistas com sua última falante fluente. O artigo narra como a etnografia se aliou à documentação linguística para elaborar um estudo inédito sobre o povo Karara e identificar esta língua como parte da família Aruak. A documentação emergencial do karara foi financiada pelo *Endangered Languages Documentation Programme*, mencionado no primeiro artigo deste dossiê. São evidentes os desafios da documentação e análise de línguas extremamente ameaçadas.

O oitavo artigo apresenta um relato de experiência inserido no contexto dos movimentos ou iniciativas de retomada de fragmentos de línguas ancestrais, dado o contexto de um apagamento brutal que sofreram, em particular, na região nordeste do país. Trata-se de experiências que precisam ser registradas e analisadas, dada a sua difusão e incremento na atualidade, sua importância histórica e política, e, no âmbito da linguística, sua contribuição para a reconstrução diacrônica possível de línguas silenciadas pelos séculos de colonização violenta no Brasil. Deste modo trazemos aqui o trabalho de Rosani Maciel Calado, professora indígena e doutoranda na Universidade Federal Rural de Pernambuco, e sua equipe, descrevem o trabalho de elaboração de um dicionário ilustrado de uma língua do estado de Pernambuco que já não é falada por seu povo e qual resta um léxico: xukuru do ororubá.

Finalmente, da Universidade Estadual de Feira de Santana, recebemos um artigo que se situa dentro da área de filologia. O estudo de Bruna Trindade Carneiro, Alícia Duhá Dose e Zenaide Novais

Carneiro trata de uma língua que quase desapareceu: a Língua Geral Amazônica, criada no período colonial a partir do contato entre povos indígenas e missionários. Essa língua foi usada durante séculos na Amazônia, mas foi proibida pelo governo português no século XVIII. As pesquisadoras analisaram um antigo manuscrito escrito nessa língua e guardado na Universidade de Coimbra, em Portugal. Seguindo os métodos filológicos, mediu-se letras, comparando caligrafias e usando ferramentas digitais para descobrir quem de fato escreveu o manuscrito. Ao digitalizar e divulgar o manuscrito, a pesquisa transforma um registro colonial em patrimônio vivo, mostrando como a memória e a tecnologia podem andar juntas para devolver às comunidades amazônicas a voz escrita de sua própria história.

Deste modo apresentamos uma coletânea que traz experiências de regiões diferentes, com propostas e reflexões de pesquisadores de universidades variadas, do Brasil e do exterior, ao se depararem com o estudo e documentação de línguas sul-americanas de diversos níveis de vitalidade.

Esta apresentação termina com a apresentação de uma proposta original da Universidade Estadual de Campinas (UNICAMP) para a documentação linguística que pretende dialogar com os métodos de documentação textual e dicionários apresentados nos capítulos deste dossiê temático.

2. O projeto DACILAT

O projeto DACILAT se baseia em uma concepção de construção de corpus com base na elaboração de corpora de narrativas e dicionários interligados computacionalmente, incluindo ainda tradução automática. Este projeto usa de ferramenta computacional inovativa e completamente desenvolvida no Brasil, na Universidade Estadual de Campinas. Diferentemente das correntes principais de abordagens computacionais que partem de uma enorme quantidade de dados (*big data*), assumimos uma posição oposta que contesta a obrigatoriedade de 'big data' para pesquisa de línguas com poucos materiais disponíveis. Para línguas em perigo de extinção, em particular, grandes corpora não podem ser constituídos, e isso por si só as excluiria desse tipo de projetos. Defendemos que a anotação acrescida aos textos e um Parser de regras é o que torna corpora relevantes para pesquisas e aplicações para fins educativos e sociais.

Em suma, nosso trabalho é baseado em elaboração de corpora anotados automaticamente por um Parser sintático (analisador sintático) construído pelo projeto que garanta uma documentação textual acrescida de informações gramaticais que abram a possibilidade de novas pesquisas sobre as línguas documentadas e também materiais para ensino das línguas e de suas gramáticas nas escolas indígenas. Trabalhamos na elaboração de um Parser de regras, facilmente adaptável para qualquer língua, baseado na gramática gerativa chomskyana e na linguagem *Corpus Search* (https://corpussearch.sourceforge.net). O projeto entrega (i) corpora anotados do Kadiwéu e do Nheengatu na Plataforma Tycho Brahe (https://www.tycho.iel.unicamp.br/dacilat) que contemple edição de textos, Parser sintático, sistemas de buscas lexicais e sintáticas, e dicionários, e (ii) traduções automáticas.

Além de linguistas e cientistas da computação, este projeto conta fundamentalmente com professores indígenas na equipe e com a colaboração de falantes nativos em campo. A Plataforma digital e online Tycho Brahe é fácil de ser aprendida e ser usada em escolas indígenas ou por alunos de graduação indígenas ou não.

A plataforma é um repositório capaz de ser uma biblioteca que analisa e traduz materiais de quaisquer tipologias linguísticas dado que temos linguistas e falantes da língua trabalhando nas adaptações necessárias.

Neste momento, o corpus da língua kadiwéu está desenvolvido e é nele que vamos nos concentrar em exemplos ao apresentar as funcionalidades de nossa plataforma. Na Plataforma Tycho Brahe, o kadiwéu conta atualmente com 40 narrativas anotadas automaticamente para classes de palavras (POS), morfemas e estrutura sintática (atualmente cerca de 10000 palavras). Além disso, a língua conta, na mesma plataforma, com um dicionário de atualmente cerca de 7000 entradas. Abaixo estão os links específicos para o corpus e o dicionário Kadiwéu:

- Corpus Kadiwéu: https://www.tycho.iel.unicamp.br/viewer/C12Elemento
- Dicionário Kadiwéu: https://www.tycho.iel.unicamp.br/lexicon/C12

Todos nossos dados (com transcrições e áudios) e anotações estão disponíveis na Plataforma Tycho Brahe. Nada no corpus, até o momento, está fechado, e pode ser visitado online e usado para outras pesquisas ou para educação indígena livremente. Obviamente as comunidades indígenas que estão colaborando com o projeto podem pedir para fechar determinado assunto e isso será feito.

Resta ainda mencionar que os tamanhos dos corpora e dicionários são variáveis, pois estão sempre sendo ampliados. Qualquer nova narrativa pode ser alimentada na Plataforma através de uma ferramenta que chamamos de Edictor, e esta narrativa será automaticamente anotada e, por sua vez, alimentará o dicionário com cada nova palavra que ocorrer. Enfim, é um trabalho de documentação que não se esgota e cujas informações gramaticais alimentam novos trabalhos sobre as línguas. Além disso, nossos corpora, por serem formado por narrativas originárias, podem alimentar o estudo etnológico/antropológico das línguas documentadas e permitir comparações linguísticas e mitológicas na região do Chaco e/ou Amazônica.

3. A Plataforma Tycho Brahe

A *Plataforma Tycho Brahe* é o nosso conjunto de ferramentas **totalmente online**, e é pioneira em sua aplicação para as línguas originárias da América do Sul. A sua interface baseada na web permite a

disseminação imediata para as comunidades indígenas e para outros pesquisadores de linguística ou antropologia que queiram trabalhar com dados de línguas indígenas, além de permitir a criação de corpora anotados gramaticalmente significativamente maiores do que era possível anteriormente.

A Plataforma Tycho Brahe foi desenvolvida pelo cientista da computação e aluno de doutorado em Linguística Luiz Henrique Lima Veronesi, sob supervisão de Charlotte Galves, pesquisadora principal do projeto. Antes de sitiar línguas indígenas, a plataforma abrigou corpora do português brasileiro e europeu.

Abaixo apresentamos nossas ferramentas que constituem a Plataforma Tycho Brahe, as abertas para o público e as fechadas.

3.1. As ferramentas abertas ao público

A Plataforma Tycho Brahe conta com ferramentas computacionais de documentação linguística abertas para o público em geral interessado em dados anotados de línguas indígenas ou não.

3.1.1. O Visualizador (https://www.tycho.iel.unicamp.br/viewer)

A ferramenta de visualização permite que qualquer pessoa usando a internet possa visitar e baixar dados de quaisquer de nossos corpora. A Plataforma Tycho (https://www.tycho.iel.unicamp.br/home) é uma Biblioteca Digital para muitas línguas, não só originárias do Brasil.

Como mencionado, neste projeto, concentramos nossos esforços na língua indígena kadiwéu, falada no Mato Grosso do Sul, e no nheengatu, língua possivelmente mista usada como língua franca por vários povos no norte do Brasil. Mas temos corpora iniciados por alunos indígenas na UNICAMP de tukano, sateré-mawé.

As narrativas do kadiwéu são de vários gêneros: narrativas mitológicas, narrativas histórias e cantos (choro Kadiwéu). O choro é um importante gênero ritualístico da arte ameríndia brasileira. O choro Kadiwéu são cantos contendo poemas na língua, sendo um gênero que irá trazer um desafio especial para a tradução.

É importante mencionar que a anotação dos corpora segue a necessidade de seus criadores e ou falantes e é totalmente parametrizada. Nosso Parser, que será apresentado mais adiante, aceita diversas anotações de corpora, desde que contenham etiquetas POS (Part-of-Speech), que também podem ser parametrizadas para cada língua. Por exemplo, a anotação do corpus Kadiwéu conta com etiquetas POS incluindo algumas etiquetas específicas para línguas com sintaxe de verbos aplicativos (*e.g.* VBAPL) e para nomes em estruturas genitivas (*e.g.* N$), etiquetas morfológicas e traduções de palavras e morfemas, além de árvores sintáticas e áudios. Segue o link de uma sentença para ilustração das anotações automáticas feitas pelo Parser: https://www.tycho.iel.unicamp.br/viewer/sentence/27ce2ac5-24ed-11e5-b0c8-94de80bbbd4a.

3.1.2. A Ferramenta de Buscas (https://www.tycho.iel.unicamp.br/search)

As buscas nos corpora são também abertas para quaisquer visitantes fomentando novas pesquisas sobre as línguas. É possível fazer buscas de palavras ou por etiquetas sintáticas (POS) ou por estruturas sintáticas através do uso da linguagem Corpus Search, já que o corpus é anotado sintaticamente (ver árvore sintática no link acima).

O trabalho com o corpus kadiwéu vem evidenciando novas estruturas da língua e as buscas permitem encontrar todas suas ocorrências. E assim novos artigos sobre a gramática desta língua tem sido publicados.

3.1.3. Dicionários (https://www.tycho.iel.unicamp.br/lexicon)

Geramos ainda dicionários das línguas dos corpora automaticamente. Os dicionários são cruciais para escolas indígenas. E são também visitados livremente. Os dicionários podem também ser encontrado na entrada da página da Plataforma.

O dicionário Kadiwéu conta com palavras como entradas e estas palavras são anotadas com etiquetas POS e informações gramaticais. Verbos e nomes contam ainda com seus paradigmas de conjugação. Os materiais que alimentam os dicionários são as palavras das próprias narrativas, mas também dados coletados nas comunidades indígenas e em publicações sobre as línguas. As fontes dos dados são discriminadas dentro do dicionário.

3.1.4. A Ferramenta *Syntree* (https://www.tycho.iel.unicamp.br/syntrees)

Esta ferramenta permite ao visitante elaborar automaticamente análises sintáticas de quaisquer sentenças nos corpora ou de sentenças criadas pelo falante, por exemplo um professor indígena, ajudando na educação escolar indígena. O professor indígena pode criar uma frase em frente aos alunos e o analisador sintático (Parser) fará sua análise, que poderá ser discutida e comparada com estruturas do português. Além da visualização da estrutura gerada, é possível fazer o download de imagens (árvores sintáticas) das sentenças.

3.2. A Plataforma Tycho Brahe: Ferramentas Fechadas (área reservada na página da plataforma)

A Plataforma tem ainda ferramentas fechadas desenvolvida pela equipe do projeto.

3.2.1. O Parser

O Parser é uma de nossas ferramentas principais e original para línguas indígenas: além de fazer traduções de palavras e anotações automáticas palavra por palavra e morfema por morfema, ele elabora automaticamente as análises sintáticas das sentenças nos corpora, base para a nossa proposta de tradução automática e de educação linguística. Nosso Parser é um Parser de regras e toma como modelo o Parser elaborado por Beatrice Santorini na Universidade da Pensilvânia para o francês antigo e adaptado ao espanhol e ao português por Catarina Magro do Centro de Linguística da Universidade de Lisboa. É baseado na linguagem Corpus Search e está sendo adaptado para o kadiwéu e para o nheengatu por Filomena Sandalo e Charlotte Galves.

Nosso Parser é baseado em uma concepção de gramática universal, portanto é facilmente adaptado para novas línguas, bastando acrescentar regras particulares e desativar regras não ativas na língua em questão. Para a adaptação de novos Parsers para outras línguas, o trabalho de um linguista especialista na língua é fundamental, bem como de falantes nativos das línguas.

3.2.2. O Edictor: edição de textos e tradução automática pelo Parser

O Edictor é nossa ferramenta de edição manual, que permite a entrada de narrativas e correção manual de qualquer erro gerado pelo Parser. A anotação é automática, mas erros são gerados eventualmente e corrigidos manualmente no Edictor.

4. Considerações Finais

O presente dossiê traz propostas e experiências inovadoras, para além do projeto DACILAT apresentado acima, bem como reflexões sobre processos e materiais que devam ser armazenados em corpora linguísticos. Salientamos neste volume o papel e a importância da linguística para a documentação das línguas originárias das Américas.

De modo geral, todos os trabalhos defendem que a documentação linguística deve colaborar em:

a) disseminação de conhecimento que possam rapidamente ser compartilhados com as comunidades indígenas;

b) preservação e vitalização de línguas;

c) incentivo do protagonismo de pesquisadores indígenas no estudo sistemático e na divulgação de suas próprias línguas;

d) melhoria da educação bilíngue e do acesso de estudantes indígenas ao ensino superior, seja nas licenciaturas indígenas ou em cursos de graduação e pós-graduação, considerando, inclusive, o aumento de vestibulares específicos em várias universidades brasileiras.