Aryon: um aplicativo Shiny para documentação e análise de línguas indígenas brasileiras

Mateus Zaparoli [1], **Katiuska Rowe**[2], **Magnun Rochel Madruga**[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

[2]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{mateuszaparoli7,katyfigueiredo0111}@gmail.com, magnun@ufmg.br

Abstract. *In order to gather scientific efforts for the documentation of Brazilian languages, this work presents partial results of the development of an R Shiny web application for documentation and exploratory analysis of indigenous languages. The objective is to concentrate lexicographic descriptions of different languages by different years and researchers in a single platform. It aims at making the platform a reliable source for querying quantitative information about different language levels. In the current version, the application provides vocabularies of two languages, namely: a)* ***Krenak****, a language spoken in the East of Minas Gerais, belonging to the Botocudo family of the Macro-Jê trunk; and b)* ***Terena****, a language of the Aruak family spoken in Mato Grosso do Sul and in the west of the State of São Paulo. The vocabularies available in the application are Bruno Rudolph [Rudolph 1909], for Krenak, and Denise Silva [Silva 2013], for Terena language.*

Resumo. *No sentido de reunir esforços científicos para documentação das línguas brasileiras, este trabalho apresenta resultados parciais do desenvolvimento de uma aplicação web R Shiny para a documentação e análise exploratória de línguas indígenas. O objetivo é concentrar em uma plataforma descrições lexicográficas de diversas línguas por diferentes épocas e pesquisadores. Além disso, objetiva-se também tornar a plataforma uma fonte confiável para consulta de informações quantitativas sobre os diferentes níveis linguísticos. Na versão atual, a aplicação disponibiliza vocabulários de duas línguas, a saber: a)* ***Krenak****, língua falada no leste de Minas Gerais, pertencente à família Botocudo do tronco Macro-Jê; e b)* ***Terena****, língua da família Aruak falada no Mato Grosso do Sul e no oeste do Estado de São Paulo. Os vocabulários disponíveis no aplicativo são os de Bruno Rudolph [Rudolph 1909], para o Krenak, e o de Denise Silva [Silva 2013], para língua Terena.*

1. Introdução

A Organização das Nações Unidas (ONU) e UNESCO afirmam que pelo menos 50% das línguas faladas de hoje estarão extintas ou seriamente ameaçadas até 2100. Por outro lado, as previsões mais pessimistas – e talvez realistas – afirmam que 90-95% serão extintas ou estarão seriamente ameaçadas até o final do século XXI. Nesse contexto, as

línguas indígenas brasileiras são parte das línguas que apresentam um alto grau de perigo de extinção nas próximas décadas, como consequência da colonização do Brasil e da expansão do português para a formação da identidade nacional. Conforme a Unesco, a humanidade pode resultar em apenas 300-600 línguas orais não ameaçadas até o final deste século [Moseley 2010], o que representaria somente algo entre 4% e 8% da diversidade linguística verificada atualmente.

O Brasil possui uma enorme diversidade linguística e étnica em seu território. As línguas indígenas brasileiras pertencem a diferentes grupos linguísticos, que se subdividem em importantes famílias relacionadas ou a ramos de línguas, tais como: Macro-Jê, Tupi, Caribe, Pano, Tucano, Aruaque, Katukina, Maku, Nambikwara, Chapacura, Yanomami, Mura-Pirahã e Guaicuru. Os troncos linguísticos Macro-Jê e Tupi são exclusivos do Brasil, enquanto as demais famílias de línguas estendem-se a países vizinhos.

Acredita-se que o Brasil tenha aproximadamente 150 línguas atualmente faladas, embora não haja concordância quanto ao número exato. Em um trabalho minucioso sobre essa questão, [D'Angelis 2019] afirma que, de forma otimista, pode-se dizer que o Brasil apresenta atualmente em torno de 100 línguas. O cenário é preocupante, sobretudo no se refere à documentação e revitalização dessas línguas ainda vivas. Conforme [Moseley 2010], 13% têm uma descrição gramatical parca da língua, 38% possuem uma descrição avançada, 29% têm descrição científica e 19% delas apresentam descrição científica insignificante. Embora o trabalho de descrição tenha aumentado no Brasil em função dos esforços das universidades brasileiras e outras organizações internacionais, as documentações sobre as línguas são esparsas, de difícil acesso em bibliotecas do Brasil ou mesmo do mundo. Os materiais, sobretudo os lexicográficos, são raros ou carecem de digitalização e tratamento adequado.

No sentido de reunir esforços científicos para documentação das línguas brasileiras, este trabalho apresenta resultados parciais do desenvolvimento de uma aplicação web R Shiny para a documentação e análise exploratória de línguas indígenas. O objetivo é concentrar em uma plataforma descrições lexicográficas de diversas línguas por diferentes épocas e pesquisadores. Além disso, objetiva-se também tornar a plataforma uma fonte confiável para consulta de informações quantitativas sobre os diferentes níveis linguísticos. Sobre o projeto, o nome **Aryon** homenageia o grande pesquisador brasileiro *Aryon Dall'Igna Rodrigues*, pioneiro nos estudos das línguas indígenas da América do Sul. Seu livro *Línguas brasileiras: para o conhecimento das línguas indígenas* [Rodrigues 1994] é considerado um dos cem livros do século, portanto, um clássico do pensamento científico brasileiro.

2. Aryon: uma aplicação web R Shiny

O **Projeto Aryon** é uma aplicação web implementada na linguagem R que utiliza recursos do pacote Shiny (cf. [R Core Team 2021] [Chang et al. 2023]), cujo intuito é oferecer às comunidades indígenas e científicas uma ferramenta voltada à consulta de vocabulários e à visualização de gráficos e estatísticas sobre estruturas que perpassam os níveis fonético, fonológico, morfológico, sintático e semântico das línguas brasileiras.

Na versão atual, a aplicação disponibiliza vocabulários de duas línguas, a saber: a) **Krenak**, língua falada no lete de Minas Gerais, pertencente à família Botocudo do tronco Macro-Jê; e b) **Terena**, língua da família Aruak falada no Mato Grosso do Sul e no

oeste do Estado de São Paulo. Os vocabulários disponíveis no aplicativo são os de Bruno Rudolph [Rudolph 1909], para o Krenak, e o de Denise Silva [Silva 2013], para língua Terena.

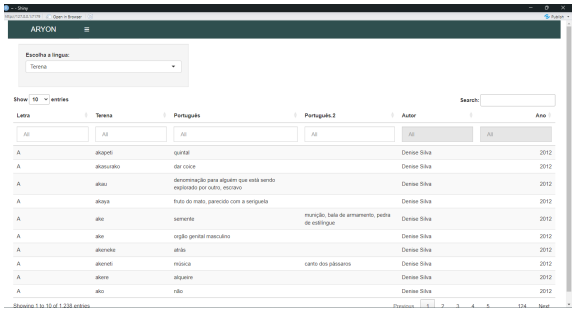


Figura 1. Apresentação da Estrutura de Vocabulários no Aryon

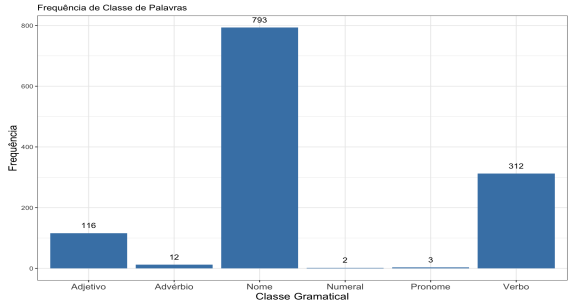
A Figura 1 mostra uma das abas do aplicativo em que se disponibilizam os vocabulários. Nessa aba, o usuário pode consultar pela palavra na língua indígena, português ou em outra língua em que ela tenha sido originalmente descrita. No caso do Krenak, por exemplo, esta língua foi descrita alemão, língua materna do autor. A tradução para o português foi dada pela equipe do projeto. É possível, ainda, buscar por autor, ano ou família linguística. Na seção a seguir, apresentamos um exemplo de análise exploratória possível de ser realizada no aplicativo.

3. Explorando o léxico Terena (Aruak)

Através da disponibilização do vocabulário, faz-se uma anotação linguística da categoria gramatical das palavras disponíveis. Com essa informação, pode-se explorar o funcionamento do léxico, sua composição em categorias gramaticais, o tamanho de cada classe e o número de palavras das classes aberta e fechada.

Na Figura 2, apresentamos a frequência das classes de palavras da língua Terena, conforme análise gramatical disponibilizada em [Silva 2013].

Figura 2. Frequência de categoria gramatical (Terena)



O vocabulário de Silva disponibiliza centenas de frases na língua e suas respectivas traduções em português. Com isso, é possível analisar, de forma imediata, a morfologia e sintaxe do Terena. O léxico da autora disponibilizado no **Aryon** conta com 312 entradas e 1245 ocorrências, numa razão tipo/ocorrência de 0.25.

4. Considerações Finais

O trabalho apresentado tem como objetivo criar uma ferramenta computacional que seja capaz de congregar informações confiáveis para pesquisas em Linguística, mas sobretudo uma ferramente capaz de ser utilizada pelas comunidades indígenas brasileiras em suas escolas, computadores e celulares. Como está em fase inicial, o projeto está em fase de aprimoramento da interface de consulta pelos usuários.

Referências

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R*. R package version 1.7.4.9002.

D'Angelis, W. d. R. (2019). Línguas indígenas no brasil: quantas eram, quantas são, quantas serão. *Revitalização de línguas indígenas: o que é*, pages 13–26.

Moseley, C. (2010). *Atlas of the World's Languages in Danger*. Unesco.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, A. D. (1994). *Línguas brasileiras: para o conhecimento das línguas indígenas*, volume 11. Edições Loyola.

Rudolph, B. (1909). *Wörterbuch der Botokudensprache*. Thaden.

Silva, D. (2013). Estudo léxicográfico da língua terena: proposta de um dicionário terena-português.