

Estado da publicação: O preprint não foi publicado em outro meio.

Lexicografia e linguística histórica das línguas Tupi: situação atual e próximos passos

Sebastian Drude, Lorrain Tyson dos Santos Araújo, Joshua Birchall, Saulo Rodrigo Teixeira Brito, Natalia Chousou-Polydouri, Cristina Martins Fargetti, Helder Perri Ferreira, Thomas Finbow, Ana Vilacy Galúcio, Lev Michael, Denny Moore, Andrey Nikulin, Luís H. Araújo Oliveira, Luciana Storto

<https://doi.org/10.1590/SciELOPreprints.18401>

Submetido em: 2026-10-13

Postado em: 2026-10-14 (versão 1)

(AAAA-MM-DD)

Lexicografia e linguística histórica das línguas Tupi: situação atual e próximos passos

Lexicography and historical linguistics of the Tupí languages: current situation and next steps

Sebastian Drude¹, Lorrar Tyson dos Santos Araújo¹⁰, Joshua Birchall²,
Saulo Rodrigo Teixeira Brito¹, Natalia Chousou-Polydouri³, Cristina Martins Fargetti⁴,
Helder Perri Ferreira¹, Thomas Finbow⁵, Ana Vilacy Galúcio¹, Lev Michael⁶,
Denny Moore¹, Andrey Nikulin^{7,9}, Luís H. Araújo Oliveira⁸, Luciana Storto⁵

¹ Museu Paraense Emílio Goeldi (MPEG), Belém, Brasil

² University of New Mexico, Albuquerque, EUA

³ Institute for Mediterranean Studies – FORTH, Rethymno, Grécia

⁴ Universidade Estadual Paulista (UNESP), Araraquara, Brasil

⁵ Universidade de São Paulo (USP), São Paulo, Brasil

⁶ University of California, Berkeley, EUA

⁷ Universidade Federal de Goiás (UFG), Goiânia, Brasil

⁸ Universidade Federal do Pará (UFPA), Belém, Brasil

⁹ Universidade do Estado de Mato Grosso (UNEMAT), Cáceres, Brasil

¹⁰ Pesquisador independente

ORCID

Lorrar Tyson dos Santos Araújo (<https://orcid.org/0000-0002-8196-5669>);

Joshua Birchall (<https://orcid.org/0000-0002-6110-6904>);

Saulo Rodrigo Teixeira Brito (<https://orcid.org/0009-0004-6425-9641>);

Natalia Chousou-Polydouri (<https://orcid.org/0000-0002-5693-975X>);

Sebastian Drude (<https://orcid.org/0000-0002-2970-7996>);

Cristina Martins Fargetti (<https://orcid.org/0000-0001-8999-8601>);

Helder Perri Ferreira (<https://orcid.org/0009-0007-0510-6890>);

Thomas Finbow (<https://orcid.org/0000-0003-4803-6603>);

Ana Vilacy Galúcio (<https://orcid.org/0000-0003-0168-1904>);

Lev Michael (<https://orcid.org/0000-0002-9612-5996>);

Denny Moore (<https://orcid.org/0000-0001-6249-1317>);

Andrey Nikulin (<https://orcid.org/0000-0003-2237-564X>);

Luís H. Araújo Oliveira (<https://orcid.org/0000-0002-0634-7908>);

Luciana Storto (<https://orcid.org/0000-0003-3012-6862>)

Autor de correspondência: Sebastian Drude (Sebastian_Drude@museu-goeldi.br)

Preprint — versão a ser depositada no SciELO Preprints; ainda não revisada por pares. Licença CC-BY.

Resumo. Este artigo apresenta um panorama do estado atual da lexicografia das línguas da família Tupi e das perspectivas de avanço, tanto na descrição lexicográfica individual quanto na pesquisa histórico-comparativa. Baseia-se nos resultados de uma oficina sobre dicionários e bancos de dados lexicais de línguas Tupi, realizada no Museu Paraense Emílio Goeldi (Belém) em março de 2023, que reuniu dezenove pesquisadores de diversas instituições. Constata-se que há hoje mais pesquisadores trabalhando sobre mais línguas Tupi do que em qualquer momento anterior, com ferramentas e padrões digitais mais maduros. Identificam-se os principais desafios — da padronização de conceitos e de campos de bancos de dados à escolha da forma de citação — e os encaminhamentos propostos pelo grupo. Uso de IA: as gravações da oficina foram transcritas e resumidas com ferramentas de IA, que também auxiliaram na revisão do texto; esses resumos serviram de base às seções sobre os projetos e sobre os desafios; a IA não forneceu citações nem referências, e os autores respondem integralmente pelo texto.

Abstract. This article presents an overview of the current state of lexicography for the languages of the Tupí family and of prospects for the coming years, both in individual lexicographic description and in historical-comparative research. It is based on the results of a workshop on dictionaries and lexical databases of Tupí languages, held at the Museu Paraense Emílio Goeldi (Belém, Brazil) in March 2023, which brought together nineteen researchers from several institutions. The overview shows that more researchers are now working on more Tupí languages than ever before, with more mature digital tools and standards. The main challenges are identified, from the standardization of basic concepts and database fields to theoretical lexicographic questions such as the choice of citation form, together with the courses of action proposed by the group. AI use: the workshop recordings were transcribed and summarized with AI tools, which also assisted in revising the text; these summaries served as the basis for the sections on individual projects and on challenges; AI supplied no citations or references, and the authors take full responsibility for the text.

Palavras-chave: lexicografia; línguas Tupi; linguística histórica comparativa; bancos de dados lexicais; línguas indígenas brasileiras; documentação linguística.

Keywords: lexicography; Tupí languages; comparative historical linguistics; lexical databases; Brazilian Indigenous languages; language documentation.

Sumário

1	Contextualização	4
1.1	A família Tupi e o que sabemos sobre sua história	4
1.2	O panorama lexicográfico: avanços e desafios	6
1.3	Ferramentas e iniciativas disponíveis	7
1.4	A Oficina de março de 2023	8
1.5	Procedimento e uso de ferramentas de IA	8
2	Projetos lexicográficos e comparativos sobre línguas Tupi.....	9
2.1	A lexicografia do Aweti.....	9
2.2	Terminologia etnográfica e lexicografia do Juruna	9
2.3	O Projeto Comparativo Tupi-Guarani de Berkeley	10
2.4	Línguas Mondé: desafios com propriedades suprasegmentais.....	10
2.5	Wayoró: lexicografia em situação de urgência extrema	10
2.6	Nheengatu: riqueza documental e heterogeneidade.....	11
2.7	Lexedata: infraestrutura para dados lexicais comparativos em CLDF	11
2.8	Dicionário Etimológico Tupi	12
2.9	Portal Japiim: plataforma para a publicação de dicionários multimídia.....	13
2.10	Documentação linguística e lexicografia do Karitiana	13
2.11	Terminologia de parentesco nas línguas Tupi	14
2.12	Dicionários multimídia para línguas indígenas.....	15
2.13	<i>Corpus</i> comparativo Nhandewa–Mbyá	15
3	Desafios e próximos passos	16
3.1	Padronização de conceitos e um primeiro produto coletivo	16
3.2	Ferramentas e formatos: em direção a um consenso técnico	17
3.3	Questões lexicográficas transversais.....	17
3.4	Comunicação, colaboração e participação indígena	18
3.5	Perspectivas para a linguística histórico-comparativa	19
4	Considerações finais.....	19
	Contribuição dos autores	20
	Financiamento.....	20
	Conflito de interesses.....	20
	Ética e consentimento	20
	Disponibilidade de dados	21
	Referências	21

1 CONTEXTUALIZAÇÃO

1.1 A família Tupi e o que sabemos sobre sua história

O Brasil é, com suas mais de 150 línguas indígenas pertencentes a pelo menos 16 famílias distintas — além de cerca de oito línguas isoladas —, um dos maiores ‘hotspots’ de diversidade linguística do mundo (Anderson, 2011; Forkel & Hammarström, 2014ff), comparável apenas à Papua Nova Guiné e a algumas regiões da África Subsaariana. Porém, uma grande parte dessa diversidade ainda permanece insuficientemente documentada e descrita. Esta situação tem melhorado visivelmente desde o início do século, impulsionado pelo desenvolvimento da linguística documental como subdisciplina (Himmelmann, 1998) com seu foco na documentação e revitalização de línguas ameaçadas (Moore et al., 2008). Nesse contexto, surgiram numerosas experiências de documentação linguística no sentido moderno: coleções representativas de registros de fala em áudio e vídeo, anotados com texto e acessíveis em acervos. Há também cada vez mais descrições gramaticais de diferentes aspectos das línguas, não raro como dissertações e teses universitárias.

Dentro desse panorama, a família Tupi,¹ uma das maiores da América do Sul, é um caso exemplar, estando entre as mais estudadas, especialmente nas terras baixas. Porém, ainda há muitas questões em aberto. A classificação interna da família distingue dez subfamílias com diferentes propostas para seu agrupamento em um número variável — de duas a sete — de unidades mais abrangentes (e.g., Galúcio et al., 2015; Nikulin & Carvalho, 2022; Rodrigues & Cabral, 2012). A subfamília mais extensa é de longe o Tupi-Guarani, que reúne aproximadamente quarenta línguas, algumas com diversas variedades — sendo às vezes comparada às línguas românicas em termos de grau de diferenciação interna — com ampla distribuição nas terras baixas tropicais da América do Sul.

As situações de vitalidade variam enormemente: o Guarani Paraguaio, cooficial no Paraguai com mais de cinco milhões de falantes, e outras variedades Guarani com (dezenas de) milhares de falantes, e o Nheengatu amazônico, cooficial em São Gabriel da Cachoeira (AM) com entre 6.000 e 8.000 falantes nativos, contrastam com línguas que contam apenas com um ou alguns pouquíssimos falantes fluentes remanescentes, como é o caso do Wayoró (Ajuru) ou do Aruá, ambos em Rondônia. Essa diversidade sociolinguística determina perfis de urgência e de tipo de trabalho lexicográfico muito distintos de um caso a outro.

Os estudos diacrônicos da família Tupi são ao mesmo tempo avançados e ainda inconclusivos. O ramo ‘Maweti-Guarani’, que agrupa o Tupi-Guarani com as línguas Aweti e Sateré-Mawé, é provavelmente o agrupamento (acima das dez subfamílias tradicionais) mais bem estabelecido da família (cf. Corrêa da Silva, 2010; Drude, 2006; Meira & Drude, 2015; Rodrigues, 1985; Rodrigues & Dietrich, 1997, entre outros). Muitos também concordariam que Maweti-Guarani e as famílias Mundurukú e Juruna fazem parte de um grande grupo ‘Oriental’ (Galúcio et al., 2015; Nikulin & Carvalho, 2022; Rodrigues & Cabral, 2012), mas a configuração interna deste ainda não está clara. Entre as

¹ Segue-se aqui a prática americanista de grafar nomes de famílias de línguas e línguas individuais com inicial maiúscula, estendida aqui também a família Bantu por coerência interna.

famílias ‘ocidentais’ (Mondé, Tupari, Arikém, Puruborá e Ramarama), situadas em Rondônia e regiões adjacentes, vale o mesmo, com exceção da aparentemente incontestada relação entre as famílias (ou melhor, línguas) Puruborá e Ramarama (Karo) (Galúcio & Gabas Jr., 2002, 2014; Rodrigues & Cabral, 2012). As evidências para formar unidades maiores, como, por exemplo, uma possível junção de Arikém e Tupari (‘Tuparikém’, Nikulin & Carvalho, 2022), não são conclusivas.

Em termos de reconstruções, existem estudos para as famílias Mondé (Moore, 2014), Tupari (Galúcio & Nogueira, 2012, 2018; Moore & Galúcio, 1994; Nikulin & Andrade, 2020), Mundurukú (Picanço, 2003, 2010, 2019) e Juruna (Carvalho, 2019; Fargetti & Rodrigues, 2008, 2021), além de propostas de reconstrução do Proto-Tupi-Guarani (Carvalho, 2022, 2023; Jensen, 1989; Mello, 2000; Schleicher, 1998), do Maweti-Guarani (Corrêa da Silva, 2010; Meira & Drude, 2015), do Ramarama–Puruborá (Lobato & Galúcio, 2023) e do Proto-Tupi como um todo (Nikulin & Carvalho, 2022; Rodrigues, 2005, 2007; Rodrigues & Dietrich, 1997). Mas esses trabalhos, em sua maioria, partem de dados escassos: em muitos casos, a reconstrução de segmentos individuais de proto-línguas apoia-se em apenas alguns poucos conjuntos de cognatos, ficando aquém dos requisitos metodológicos mínimos da linguística histórica (Campbell, 2013).

De modo geral (cf. Nikulin & Carvalho, 2019), após um início promissor em torno do ano 2000, a pesquisa histórico-comparativa — sobre a família, e em geral no Brasil — ficou muito aquém de seu potencial. Nos últimos anos, os métodos filogenéticos computacionais (MFCs) passaram a integrar o instrumental dos linguistas históricos, sendo aplicados a numerosas famílias (Michael & Chousou-Polydouri, 2019), inclusive a Tupi (Galúcio et al., 2015; Gerardi & Reichert, 2021; Michael et al., 2015). Como argumentam Michael & Chousou-Polydouri (2019), os MFCs e o Método Comparativo (MC) são complementares, com pontos fortes diferentes.² Uma vez que MFCs tomam julgamentos de cognância como insumo básico, sua precisão depende diretamente da qualidade dos dados disponíveis. Bancos de dados lexicais comparativos ricos e interoperáveis são, portanto, cruciais tanto para a lexicografia em si quanto para o trabalho histórico. Idealmente, o Método Comparativo e os MFCs são aplicados de forma articulada, fortalecendo-se mutuamente: reconstruções sólidas melhoram os julgamentos de cognância que alimentam as análises filogenéticas, e os resultados filogenéticos oferecem e fundamentam hipóteses de subagrupamento a serem testadas pelo MC.

² O Método Comparativo se destaca na reconstrução e no estabelecimento de cognância por correspondências regulares, mas não dispõe de procedimento explícito para avaliar subagrupamentos diante de evidências conflitantes, nem fornece cronologias absolutas. A filogenética, por sua vez, emprega modelos matemáticos explícitos para tratar do subagrupamento, mesmo sob informação complexa ou conflitante, e pode estimar tempos de divergência. Ambas abordagens são objeto de debate metodológico ativo — no caso do MC, quanto ao estatuto epistemológico das reconstruções e aos limites da comparação de longa distância; no caso dos MFCs, quanto à sensibilidade dos resultados à qualidade dos julgamentos de cognância e à modelagem de empréstimos (cf., para o contexto sul-americano, Ramirez, 2020 cap. 2; Greenhill et al., 2020 dão uma boa introdução atual sobre os MFCs).

1.2 O panorama lexicográfico: avanços e desafios

O cenário da lexicografia das línguas indígenas brasileiras melhorou significativamente nas últimas décadas. Um levantamento baseado no portal especializado *lexicografiaindigena* (Lopes, 2024) indica que o número de trabalhos que contêm informações lexicográficas publicados ao longo do presente século — incluindo listas de palavras, vocabulários, glossários e dicionários mais abrangentes — supera a produção acumulada das sete décadas anteriores. Esse crescimento quantitativo reflete tanto o aumento no número de pesquisadores quanto o surgimento de novas ferramentas e plataformas de publicação.

No entanto, em diversos lugares do mundo, uma parte substancial dos dados lexicais coletados permanece sob custódia privada, inacessível à comunidade científica e às próprias comunidades de falantes (Seyfeddinipur et al., 2019). Os pesquisadores de campo coletam dados lexicais — hoje em dia em sua maioria em formatos digitais —, mas esses recursos raramente chegam a dicionários publicados de qualidade, e quando publicados, usualmente não são interoperáveis com outros projetos. No Brasil, mesmo a iniciativa ‘ProDocLin’ — que incluiu explicitamente uma fase voltada para o desenvolvimento de dicionários digitais — não alterou fundamentalmente esse cenário.

Dentro da família Tupi, a concentração temática das publicações existentes agrava o problema. A grande maioria dos dicionários publicados — estimados em torno de 60 a 70 ao todo — é dedicada ao Guaraní ou ao Tupinambá e ao Nheengatu.³ Para as línguas das demais subfamílias, isto é, fora do Tupi-Guaraní, há apenas quatro dicionários publicados (Alves, 2004; Crofts & others, 1977; Landin, 1983; Pacheco Ribeiro, 2010)⁴. A plataforma *lexicografiaindigena* confirma isto: dos aproximadamente 240 trabalhos ali identificados como relevantes para línguas da família Tupi e publicados desde 2000, ~150 são dedicados a variedades do Guaraní ou Tupinambá/Nheengatu, ~55 a outras línguas Tupi-Guaraní, e somente 35 tratam de línguas não-Tupi-Guaraní.

O problema é, no entanto, não apenas de quantidade, mas de cobertura e qualidade. O dicionário do Mundurukú, por exemplo, não contempla tons, ainda que o Mundurukú — assim como outras línguas Tupi — seja tonal. E também há desafios de acessibilidade e interoperabilidade: cada pesquisa acumulou ao longo das décadas materiais lexicais em formatos heterogêneos — cadernos de campo, planilhas de configurações diversas, bancos de dados no Toolbox/Shoebox (Wimbish & Buseman, 1987) e, mais recentemente, no FieldWorks Language Explorer (FLEX) (SIL International, 2014). Esses acervos não foram concebidos tendo em mente a interoperabilidade entre projetos, nem a comparação sistemática entre línguas, e muitas vezes tampouco a publicação para uso comunitário.

A massa crítica de dados e de pesquisadores necessária para um salto qualitativo no campo existe; o que falta é coordenação, padrões comuns e uma infraestrutura técnica adequada.

³ Essa estimativa é o resultado de uma [busca no banco de dados bibliográficos do Glottolog](#), eliminando entradas duplicadas e textos sobre dicionários.

⁴ Mas veja seções 2.9 e 2.12 neste trabalho.

1.3 Ferramentas e iniciativas disponíveis

Nos últimos anos, o panorama das ferramentas disponíveis para o trabalho lexicográfico em línguas indígenas mudou significativamente, criando condições técnicas antes inexistentes para superar a lacuna identificada acima. O software *FLE*x, desenvolvido pelo SIL International, apesar de diversos problemas, está sendo aplicado e recomendado para a gestão de dados lexicais em línguas indígenas no Brasil (Beier & Michael, 2022). Sua estrutura relacional robusta, a integração com a interlinearização de textos e a exportação em múltiplos formatos — incluindo o *LIFT* (Lexicon Interchange FormAT) — tornaram-no um ponto de partida possível para novos projetos.

Para o compartilhamento e a publicação de dados comparativos, o padrão *CLDF* (Cross-Linguistic Data Formats; Forkel et al., 2018) emergiu como a solução mais promissora. O CLDF organiza os dados em arquivos CSV interligados — tabelas de línguas, formas, conceitos, conjuntos de cognatos, entre outros — com metadados em JSON e referências bibliográficas em formato BibTeX, garantindo compatibilidade com repositórios acadêmicos e com as principais ferramentas de análise filogenética computacional (LingPy, Edictor, MrBayes, BEAST2). O *Lexedata* (Kaiping et al., 2022) funciona especificamente como ponte entre o trabalho cotidiano em planilhas e o armazenamento e a publicação em CLDF, sem exigir competências avançadas de programação.

Para a publicação de dicionários interativos com dados multimídia, destinados ao uso comunitário, serão discutidas neste artigo duas soluções técnicas que foram desenvolvidas no Brasil e aplicadas a várias línguas indígenas⁵: (1) O Portal Japiim, do atual Museu Nacional dos Povos Indígenas (FUNAI), gera, a partir de um único banco de dados, produtos em três formatos: interface web em HTML com busca cruzada entre dicionários, versão impressa em PDF (via LaTeX) e aplicativo Android para uso offline — essencial para comunidades em zonas remotas sem acesso regular à internet. (2) os Dicionários Multimídia para Línguas Indígenas, desenvolvidos através de uma parceria entre o Museu Goeldi, a Universidade de New Mexico e as comunidades indígenas, dentro do conceito de Tecnologia Social, oferecem uma solução semelhante que consiste em compilar dados linguísticos e arquivos de mídia em um formato que permite fácil acesso e compartilhamento em diversos formatos (HTML, PDF e aplicativos para uso offline), com baixo custo técnico e de infraestrutura (Brito et al., 2025; Galúcio et al., 2026).⁶

Quanto a bancos de dados publicados, iniciativas como o TuLeD (Tupian Lexical Database; Gerardi et al., 2021) e o banco de terminologia de parentesco MPEG-KIN (Birchall et al., 2022; integrado ao KinBank global: Passmore et al., 2023) constituem infraestruturas já consolidadas que demonstram a viabilidade e a utilidade de dados lexicais abertos e interoperáveis para a família Tupi.

⁵ Uma terceira plataforma conhecida é a Plataforma Dicweb Indígena - Dicionários das línguas indígenas da Amazônia, sendo desenvolvida pela Universidade Federal de Rondônia, disponível no endereço eletrônico <https://dicweb.app/>.

⁶ Usando esta tecnologia, já estão disponíveis dicionários de algumas centenas de entradas das línguas Kanoé, Oro Win, Puruborá, Sakurabiat, Salamã e Wanjam; outros estão sendo preparados.

1.4 A Oficina de março de 2023

Para inventariar a situação atual, identificar oportunidades de colaboração e definir próximos passos coletivos, S. Drude organizou nos dias 9 e 10 de março de 2023, no Museu P. E. Goeldi (Belém), uma Oficina sobre Dicionários e Bancos de Dados Lexicais de Línguas Tupi. O evento reuniu dezenove pesquisadores de instituições brasileiras e internacionais — MPEG, USP (Departamento de Linguística e FFLCH), UNESP, UFG, UFPA, o então Museu do Índio/FUNAI, UC Berkeley, Universidade de Bristol, Universidade do Novo México e Universidade de Zurique —, cada qual apresentando o estado de seu trabalho lexicográfico ou comparativo sobre uma ou mais línguas da família. A oficina foi organizada no âmbito de um projeto financiado pela Deutsche Forschungsgemeinschaft (DFG) voltado à documentação lexicográfica do Aweti e à construção de infraestrutura de dados para a família Tupi, mas foi concebida desde o início para ir além desse projeto específico e articular uma comunidade mais ampla de pesquisadores.

O programa de dois dias teve o primeiro dia focado em um panorama do trabalho individual de cada pesquisador e o segundo em tecnologias (bancos de dados, formatos, ferramentas) e na pesquisa comparativa, culminando em uma discussão coletiva sobre próximos passos. As seções que seguem sintetizam as contribuições individuais e os debates produzidos na oficina, assim como o progresso registrado desde então, identificando tanto os avanços verificados em cada projeto quanto os obstáculos e as questões em aberto que o grupo reconheceu como prioritários para o campo.

1.5 Procedimento e uso de ferramentas de IA

Para a elaboração deste artigo foram usadas ferramentas de inteligência artificial (IA) em três etapas. Primeiro, as gravações em vídeo e áudio das apresentações e discussões da oficina ocorrido em 2023 foram transcritas automaticamente, por reconhecimento automático de fala (ASR), com o WhisperX 3.8.6 (modelo Whisper ‘turbo’ via faster-whisper 1.2.1, com separação de falantes pelo pyannote.audio 4.0.4); os falantes foram identificados manualmente a partir dos áudios. Em seguida, a partir dessas transcrições e dos slides de cada apresentação, foram geradas com o modelo de linguagem Claude (Anthropic) [Sonnet 4.6 e Opus 4.6, março e abril 2026] resumos de cada projeto apresentado e dos resultados das discussões; esses resumos serviram de apoio à redação das seções sobre os projetos individuais e sobre os desafios e próximos passos, que foram conferidos e corrigidos pelos autores. Por fim, o texto final foi verificado quanto a erros e inconsistências (grafia de nomes, convenções de citação, aspas, etc.) por uma ferramenta de IA (Claude Sonnet 5.5 e Fable 5, em setembro) e, em alguns casos, esta auxiliou na formulação (por exemplo, na abreviação de um parágrafo). O texto foi redigido pelos autores; a IA não contribuiu com nenhuma citação ou referência, e os autores assumem integralmente a responsabilidade pelo conteúdo e por todas as fontes citadas.

2 PROJETOS LEXICOGRÁFICOS E COMPARATIVOS SOBRE LÍNGUAS TUPI

2.1 A lexicografia do Aweti

O Aweti, língua da família Maweti-Guarani falada no Alto Xingu (Mato Grosso), é o objeto do projeto lexicográfico de S. Drude mencionado acima. A pesquisa lexicográfica sobre o Aweti remonta ao final dos anos 1990, quando os primeiros dados foram coletados com o programa Shoebox (SIL). A trajetória tecnológica posterior — de Shoebox para Toolbox e deste para o FLEx, com perspectivas de exportação para CLDF — ilustra os desafios metodológicos que praticamente todos os projetos de longa duração sobre línguas indígenas enfrentam: ortografia instável nas fases iniciais, estruturação e organização de campos em constante revisão, e a necessidade de articular dados textuais com dados lexicais em ambientes de trabalho que se transformam ao longo do tempo.

O banco de dados no FLEx acumulou 1.400 entradas para temas simples e complexos e 75 para afixos e clíticos, com uma estrutura de campos que inclui forma subjacente, forma ortográfica de citação, categorias morfológicas e sintáticas, glosas e definições em português e inglês, informações enciclopédicas, paradigmas de flexão e metadados de fonte e data. A terminologia de parentesco Aweti, por sua vez, está organizada em um banco separado, em formato Toolbox, com colunas para relação, protótipo, sexo do falante e sexo do referente — estrutura que antecipa e alimenta diretamente os projetos comparativos de parentesco na família Tupi. A migração para o CLDF é uma ação em andamento dentro do projeto Heisenberg/DFG.

2.2 Terminologia etnográfica e lexicografia do Juruna

O trabalho lexicográfico sobre o Juruna (ou ‘Yudjá’, da família Juruna), povo do médio Xingu, é desenvolvido por Cristina Fargetti (UNESP Araraquara/LINBRA) e se apoia numa reflexão teórico-metodológica de longa data sobre os Estudos do Léxico de línguas indígenas. Sua contribuição central é a abordagem que denomina ‘Terminologia Etnográfica’ (2018): qualquer estudo terminológico/terminográfico sério sobre uma língua indígena deve partir de pesquisa etnográfica genuína — da compreensão das categorias nativas tal como os próprios falantes as organizam —, em vez de impor classificações externas (taxonômicas científicas ou da cultura ocidental).

O argumento tem consequências práticas diretas para a lexicografia comparativa, uma vez que os estudos terminológicos também alimentam propostas lexicográficas (dicionários gerais de língua). Se a terminologia para fauna e flora precisa ser documentada a partir da classificação nativa, o *tertium comparationis* necessário para a comparação entre línguas não pode ser simplesmente o nome científico da espécie, pois este pode não corresponder às unidades de categorização de nenhuma das línguas comparadas. Fargetti vem publicando estudos lexicais sobre o Juruna em campos semânticos variados — cultura material, astronomia, numerais, aves, culinária, borboletas —, acumulando uma documentação que é ao mesmo tempo rigorosamente linguística e etnograficamente fundamentada (Fargetti, 2021, 2026).

2.3 O Projeto Comparativo Tupi-Guarani de Berkeley

O projeto de comparação lexical e reconstrução filogenética do Tupi-Guarani conduzido, entre outros, por Zachary O'Hagan (UC Berkeley) e Natalia Chousou-Polydouri (Universidade de Zurique), teve início em Berkeley por volta de 2010, a partir de uma investigação sobre o Omagua e o Kokama. O projeto cresceu para uma comparação sistemática de 38 variedades da família, abarcando hoje cerca de 540 conceitos primários, aproximadamente 20.000 formas e cerca de 5.000 conjuntos de cognatos. A análise filogenética resultou em uma nova proposta de classificação interna do Tupi-Guarani (Michael et al., 2015), com uma hipótese subsequente de localizar o berço do Proto-TG na bacia do baixo Xingu (O'Hagan et al., 2019).

A trajetória tecnológica — de planilhas offline para Google Planilhas com múltiplos colaboradores, e, por fim, ao plano atual de converter o conjunto de dados para o formato CLDF e continuar seu desenvolvimento usando o Lexedata — ilustra as dificuldades do trabalho colaborativo de longo prazo sem financiamento central nem pessoa dedicada exclusivamente a ele. Entre as lições aprendidas pela equipe ao longo desse percurso estão: coletar metadados mais completos desde o início (traduções originais, número de página das fontes), documentar as decisões sobre cognância no momento em que são tomadas, e cultivar colaborações com especialistas de cada língua desde as fases iniciais.

2.4 Línguas Mondé: desafios com propriedades suprasegmentais

A documentação das línguas da família Mondé é conduzida há décadas por Denny Moore (pesquisador emérito do Museu Goeldi) e abrange o Suruí/Paiter (~1.300 falantes), Cinta-Larga (~2.000), Gavião (~800), Zoró (~800), o Aruá (~4) e o Salamã (com apenas dois semi-falantes idosos). O principal desafio linguístico das línguas Mondé é o sistema tonal e o prolongamento vocálico, cuja existência não foi reconhecida pelas primeiras pesquisas linguísticas (por missionários e outros), gerando décadas de materiais escritos de forma inadequada. O sistema tonal do Gavião é, possivelmente, o mais complexo de toda a família Tupi.

A situação atual combina oportunidade com obstáculos. Os grupos Mondé têm populações relativamente grandes, membros das comunidades familiarizados com recursos digitais e lideranças articuladas — condições favoráveis para um projeto lexicográfico participativo. Entretanto, a necessidade de primeiro resolver a fonologia de forma satisfatória, negociar ortografias com as comunidades e coletar dados lexicais de forma sistemática significa que o caminho até um dicionário publicável ainda é longo. Moore (2012, 2014) indicou, por outro lado, que as correspondências tonais e segmentais entre as línguas da família são passíveis de reconstrução, e que o Proto-Mondé deve ser acessível à reconstrução completa — mas apenas após a elucidação das análises fonológicas individuais.

2.5 Wayoró: lexicografia em situação de urgência extrema

O Wayoró (Ajuru), da família Tupari, falado na Terra Indígena Rio Guaporé (Rondônia), é documentado por Fernanda Nogueira (ILC-UFPA) em uma situação sociolinguística de urgência extrema: no momento da oficina havia apenas uma falante fluente

(Dona Paulina Macurapi), com cerca de dez semi-falantes, e nenhuma criança na TI adquirindo a língua indígena como primeira língua.

O banco de dados no FLEx, iniciado em 2011 sem objetivo explicitamente lexicográfico, acumulava em 2023 722 entradas organizadas em 13 campos semânticos definidos em diálogo com o cacique da comunidade. O projeto elaborou dois produtos: um vocabulário bilíngue em PDF, publicado deliberadamente mesmo com imperfeições — por entender que o material disponível já pode ajudar a comunidade —, e um aplicativo (APK) para Android com pesquisa por palavra e reprodução de arquivos de áudio. Essa experiência ilustra um dilema recorrente na lexicografia de línguas ameaçadas: por um lado, a tensão entre o ideal de completude e rigor científico e, por outro, a urgência de disponibilizar material antes que a última geração de falantes se torne inacessível.

2.6 Nheengatu: riqueza documental e heterogeneidade

O Nheengatu (Língua Geral Amazônica), da família Tupi-Guarani, conta com cerca de 6.000 a 8.000 falantes nativos na bacia do Rio Negro (AM), é cooficial em São Gabriel da Cachoeira desde 2002 e foi oficializado no estado do Amazonas em julho de 2023; seu panorama lexicográfico é analisado por Thomas Finbow (FFLCH/USP).

O Nheengatu resulta de um contínuo linguístico cuja documentação escrita na Amazônia começa com os registros dos capuchinhos franceses no Maranhão (d'Abbeville, 1614; d'Evreux, 1615), sendo o *Compêndio de Bettendorff* (1687/1800) a primeira obra maior produzida para a região; pelo vínculo com a língua brasílica descrita por Anchieta (1595) e por outros jesuítas (Araújo, 1618; Figueira, 1621), essa tradição recua ao século XVI — o que faz desse contínuo o de mais longa tradição documental entre as línguas indígenas brasileiras (Finbow, 2025; Rodrigues, 1986). Para o século XVIII, o levantamento de Ávila (2021) reúne cerca de treze textos da língua geral amazônica, seguidos de uma série de obras lexicográficas e gramaticais nos séculos XIX e XX.

Paradoxalmente, essa relativa abundância documental constitui um desafio específico. Sem padronização entre as múltiplas normas ortográficas em uso hoje (cf. D'Angelis, 2023), e ainda mais variedades usadas no passado, especialmente no século XIX, é difícil comparar dados entre fontes e oferecer um recurso lexicográfico coerente. O dicionário de Ávila (2021), elaborado no FLEx, com aproximadamente 700 páginas, constitui o produto mais substancial do período recente. Mas o trabalho tem suas limitações, como a ausência de áudio de falantes nativos e a proliferação de formas introduzidas por autores sem competência nativa. Do ponto de vista diacrônico, o Nheengatu oferece uma janela privilegiada para o estudo da mudança lexical de uma língua Tupi-Guarani ao longo de quatro séculos, e a elaboração de um *corpus* histórico anotado do Tupi Antigo figura entre os planos de Finbow, que prevêem também sanar algumas das deficiências apontadas em Ávila (2021).

2.7 Lexedata: infraestrutura para dados lexicais comparativos em CLDF

O Lexedata é um conjunto de ferramentas de código aberto, escritas em Python, para editar e manter conjuntos de dados lexicais no padrão CLDF, desenvolvido por Ge-reon Kaiping, Melvin Steiger e Natalia Chousou-Polydouri (Universidade de Zurique)

(Kaiping et al., 2022). O problema que o Lexedata tenta resolver é que o CLDF é excelente para publicar e arquivar dados comparativos, mas não foi projetado para o trabalho cotidiano de coleta, edição e curadoria. Na prática, os pesquisadores trabalham em planilhas Excel ou programas especializados; as ferramentas de análise filogenética usam formatos dedicados; e a conversão entre esses ambientes exige habilidades de programação que a maior parte dos linguistas não possui.

O Lexedata opera como interface de linha de comando entre o CLDF e diferentes contextos de uso, com importadores e exportadores para xlsx, TSV (compatível com LingPy e Edictor) e formatos filogenéticos (nexus/XML), além de ferramentas de edição em lote e geração de relatórios de cobertura. Um caso de uso exemplar é a migração do banco Berkeley TG — treze anos de planilhas Google — para o CLDF (cf. seção 2.3), tarefa que sem o Lexedata teria sido inviável. A ferramenta está sendo também aplicada a projetos sobre as famílias Aruak e Bantu, e sua adoção pelos projetos da família Tupi é uma das perspectivas concretas que emergiram da oficina.

2.8 Dicionário Etimológico Tupi

O Dicionário Etimológico Tupi, elaborado por Andrey Nikulin (IFSIT/UFG), foi iniciado em 2022 com o objetivo de documentar e analisar o léxico da família em perspectiva diacrônica. Com 373 entradas no momento da submissão deste artigo, o dicionário cobre conjuntos de cognatos com distribuição em mais de uma subfamília, representados nas ortografias nativas das línguas documentadas sem padronização artificial. A obra é elaborada em LaTeX — escolha incomum para lexicografia, justificada pela excelência e flexibilidade tipográfica e pela possibilidade de manter entradas incompletas comentadas na fonte sem que apareçam na versão publicada.

Os princípios editoriais são inspirados no *Uralic Etymological Dictionary* (Aikio, 2020): referência completa para cada dado, reconstrução própria do Proto-Tupi e das protolínguas intermediárias (não mera compilação de reconstruções alheias), comentários sistemáticos sobre irregularidades identificadas, e citação das formas provenientes de reconstruções publicadas com a devida referência. Tal formato já foi utilizado pelo autor no dicionário etimológico que integra sua fonologia histórica das línguas Mataguaias (Nikulin & Carol, 2024).

A classificação interna adotada por Nikulin distingue cinco ramos: Kepkiriwat, Tuparikém, Rama-Puru, Mondé⁷ e Oriental (incluindo Maweti-Guarani, Mundurukú e Juruna).⁸ Prevê-se a publicação do Dicionário em formato de livro, bem como — em futuro mais distante — a elaboração de um dicionário etimológico digital no âmbito da série EvoLex, idealizada por Alexandre François e Mathieu Dehouck (CNRS—LaTTiCe); a conversão para o formato CLDF por S. Drude está sendo finalizada. O projeto é potenci-

⁷ Incluindo notas sobre possíveis cognatos em Arara do Rio Branco, uma língua escassamente documentada que o autor acredita estar relacionada à família Mondé.

⁸ Todas as línguas não-Tupi-Guarani são incluídas; a subfamília Tupi-Guarani é representada pelas línguas Kawaiwete, Parintintin, Siriono, Guarasugwe, Gwarayu, Tapiete, Guarani Paraguaio, Mbyá, Tupi Antigo, Kamaiurá, Apyãwa, Suruí-Aikewara, Tenetehara, Ka'apor e Emérillon-Teko.

almente colaborativo: Nikulin convida os especialistas a contribuir com entradas nas línguas de sua especialidade para as próximas edições. O Dicionário inclui, ainda, capítulos que tratam do subagrupamento, evolução fonológica e aspectos da morfossintaxe diacrônica das línguas Tupi.

2.9 Portal Japiim: plataforma para a publicação de dicionários multimídia

O Portal Japiim é uma plataforma desenvolvida no âmbito do Programa de Documentação de Línguas Indígenas (ProDocLin) para a publicação e a disponibilização de dicionários multimídia de línguas indígenas; ele é desenvolvido por Helder Ferreira (hoje no Museu Goeldi, em 2023 no Museu do Índio, atual Museu Nacional dos Povos Indígenas/FUNAI). A partir de um único banco de dados relacional com cerca de 90 tabelas (metade para *backup*), o Japiim gera três tipos de produto: interface web em HTML (com busca cruzada entre dicionários e mapa de distribuição geográfica das línguas), versão impressa em PDF via LaTeX, e aplicativo Android para uso offline. A inserção de dados no sistema pode ser feita via tabelas Excel com interface gráfica ou por importação direta de FLEx via arquivo no formato LIFT (cf. SIL LSDev, 2015/2026).

A filosofia de dados do Japiim privilegia formatos abertos (CSV, TSV, TXT) como garantia de sustentabilidade a longo prazo: as plataformas mudam, mas os dados em texto puro resistem à obsolescência tecnológica. Um aspecto técnico de particular interesse para o trabalho colaborativo é a funcionalidade de compartilhamento de imagens entre dicionários da plataforma — que permite usar uma fotografia disponível em um dicionário como ilustração em outro enquanto uma imagem mais adequada não estiver disponível. A questão das imagens culturalmente apropriadas é apontada por Ferreira como um dos maiores gargalos práticos na produção de dicionários ilustrados para línguas indígenas.

Desenvolvimentos recentes no código da plataforma ampliam consideravelmente suas capacidades de interoperabilidade. Foi implementada a exportação nos formatos CLDF (*Cross-Linguistic Data Formats*), tanto no módulo de lista de palavras quanto no de dicionário, facilitando a integração com ferramentas e repositórios consolidados da linguística comparativa. Paralelamente, foi introduzida uma ferramenta de mapeamento semi-automático das entradas aos quase 4.000 conceitos do Concepticon, sujeita à confirmação do autor do dicionário, preservando o seu controle editorial. O diferencial do Portal Japiim é operar inteiramente a partir das glosas e definições em português, o que o torna particularmente útil para línguas indígenas do Brasil. Em conjunto, a possibilidade de importar dados do FLEx, mapear entradas pelo Concepticon e exportar a base nos formatos CLDF posiciona o Japiim como ferramenta estratégica para estudos históricos e léxico-comparativos.

2.10 Documentação linguística e lexicografia do Karitiana

O Karitiana, única língua viva da família Arikém (Tupi), falada na Terra Indígena Karitiana próxima a Porto Velho (Rondônia), por uma comunidade de aproximadamente 400 falantes nativos, é objeto de décadas de pesquisa de Luciana Storto (Departamento de Linguística, USP). Há hoje seis aldeias diferentes, algumas fora da TI, e muitos falantes se mudaram para núcleos urbanos, o que tem fragmentado a comunidade de fala e

enfraquecido a língua, já que 60% das crianças que nascem em áreas urbanas não falam a língua indígena. Do ponto de vista descritivo e teórico, o Karitiana é uma das línguas indígenas brasileiras mais intensivamente estudadas — especialmente quanto à sintaxe —, e a documentação do léxico e da cultura sempre acompanhou a pesquisa gramatical.

Um léxico com cerca de 2.000 entradas, iniciado no Shoebox e migrado para o Toolbox e depois para o FLEx, reúne transcrição fonética, ortográfica, tradução em português e inglês, categoria gramatical e exemplos de uso. A necessidade de converter esse banco de dados em um produto lexicográfico publicável para a comunidade — com imagens, áudio e organização por campos semânticos adequados — foi apontada como a ação mais urgente do projeto. Entre 2024 a 2026, Storto orientou a mestrande Shuanny G. Ramires na elaboração de um dicionário destinado à escolas indígenas, aplicando a metodologia do MPEG; a obra conta com aproximadamente 750 itens lexicais e inclui áudio, transcrição fonética e ortográfica além de campos semânticos. A ortografia Karitiana, definida em um projeto colaborativo com a comunidade entre 1993 e 1997, tem sido ensinada por professores indígenas nas escolas através do material preparado por Storto, mas não se pode dizer que o uso efetivo tenha seguido o padrão estabelecido no material didático. Diferenças dialetais e idioletais têm aparecido no uso da língua escrita, fenômeno que merece estudo linguístico próprio; oficinas de escrita seriam necessárias para padronizar o uso da ortografia por parte dos falantes em todas as aldeias.

2.11 Terminologia de parentesco nas línguas Tupi

A terminologia de parentesco nas línguas Tupi é objeto de um programa de pesquisa conduzido por Luís H. Araújo Oliveira (UFPA/MPEG) e Joshua Birchall (University of New Mexico/MPEG), desenvolvido no âmbito de projetos financiados pela British Academy (2016–2018) e pelo ERC VARIKIN (Variation in Language, Cognition and Social Norms Regarding Family; 2015–2021).⁹ O resultado mais expressivo é o banco de dados MPEG-KIN, com dados de terminologia de parentesco padronizados em 100 relações de referência para 31 línguas Tupi, estruturados em CLDF e integrados ao KinBank global (1.229 línguas). A análise histórica resultou em uma reconstrução de 35 termos de parentesco do Proto-Tupi-Guarani (Carvalho & Birchall, 2022), identificando um sistema de tipo fusão bifurcada (*bifurcate merging*) como configuração da protolíngua, o que apontaria para um sistema do tipo Iroquês, talvez com traços de Dravidiano, na tipologia de Murdock (1970).

O equilíbrio metodológico entre as perspectivas *êmica* (a categorização nativa tal como os falantes a estruturam) e *ética* (um conjunto padronizado de relações de referência externo à língua, útil para comparação tipológica) é um ponto central deste programa.

⁹ Neste contexto, cabe lembrar que Felzke e Moore (2019) apresentaram um estudo diacrônico da terminologia de parentesco dos grupos Mondé; e Nogueira e colegas (2019) apresentaram um estudo semelhante para os grupos Tupari.

2.12 Dicionários multimídia para línguas indígenas

Os dicionários multimídia para línguas indígenas (DMLI) são desenvolvidos por Ana Vilacy Galúcio, Saulo Brito e Joshua Birchall e colaboradores, em uma parceria entre as comunidades indígenas, o Museu Goeldi e a Universidade do Novo México. O desenvolvimento dos DMLI iniciou-se em 2019, como uma resposta à demanda específica de algumas comunidades indígenas do estado de Rondônia (Brasil) por ferramentas que contribuíssem para a aprendizagem das suas línguas tradicionais e incluíssem a possibilidade de escutar os sons da língua. Essa demanda suscitou o desenvolvimento de um produto desenvolvido dentro do conceito de uma tecnologia social replicável, de uso livre e fácil acesso, que pode ser usada pelas comunidades em larga escala e sem necessidade de acesso à internet (Galúcio et al., 2026). A ferramenta foi inicialmente voltada para línguas em situação de risco extremo, como as línguas Puruborá, Kanoé, Wayoró, Sakurabiat, Oro Win e Moré. Para essas comunidades, o dicionário multimídia é concebido como instrumento de potencial reversão da quebra de transmissão intergeracional, e não apenas como registro acadêmico. O Acervo de Línguas Indígenas do MPEG (ALIM), com registros audiovisuais de 83 línguas, mais de 1.500 horas de áudio e mais de 500 horas de vídeo, é uma das fontes de dados para esses projetos.

A metodologia de elaboração dos DMLI utiliza uma estrutura de bancos de dados lexicais e de arquivos multimídia e alguns recursos computacionais para o processamento das informações, o que permite gerar dicionários digitais a partir de arquivos CSV padronizados e preenchidos com dados lexicais e um conjunto organizado de arquivos de mídia, os quais, por sua vez, podem ser convertidos para os formatos HTML e PDF (Brito et al., 2025). Os produtos HTML gerados podem conter arquivos de mídia embutidos ou vinculados externamente. A utilização da mídia embutida permite o acesso ao conteúdo sem necessidade de conexão à internet. Um diferencial dos DMLI é exatamente a possibilidade de uso offline, através de aplicativos Android, pois considera-se essencial garantir o uso por comunidades remotas, onde o acesso à internet ainda é muito restrito. A produção de aplicativos Android com o HTML do dicionário integrado segue em desenvolvimento, visando automatizar a criação de arquivos APK que atendam aos requisitos técnicos necessários para serem disponibilizados em plataformas como a Google Play Store a partir da versão HTML do dicionário.

A solução é deliberadamente acessível: requer apenas software livre e uma planilha organizada, e já foi aplicada com sucesso no desenvolvimento de sete dicionários, publicados na Plataforma de Dicionários Multimídia, disponível no portal do Museu Goeldi¹⁰, que também oferece orientação a usuários interessados em montar dicionários com essa metodologia. A perspectiva de colaboração com o Portal Japiim para as etapas seguintes é considerada um caminho natural para ampliar o alcance do programa.

2.13 *Corpus* comparativo Nhandewa–Mbyá

Cabe aqui registrar um projeto mais recente, conduzido por T. Finbow e colaboradores na USP, de constituição rápida de um *corpus* lexical para o Nhandewa (Tupi-Guarani), variedade diaspórica falada em São Paulo e no Paraná (a grafia com ⟨w⟩ distingue-

¹⁰ <https://dicionarios.museu-goeldi.br/>

a da variedade do Paraguai e do Mato Grosso do Sul). Embora o Nhandewa tenha sido a primeira variedade guarani do Brasil a ser tópico de algum estudo etnográfico e linguístico em tempos modernos (Nimuendajú, 1914/1987), a atenção dos linguistas voltou-se depois sobretudo para o Mbyá, e falta hoje material lexical e gramatical em português acessível a não falantes. Essa lacuna é crítica no contexto de uma assessoria em curso a duas comunidades de Mongaguá (SP), onde a transmissão intergeracional está interrompida há três gerações e a retomada da língua depende de professores monolíngues em português (para o projeto de revitalização, ver Finbow et al., 2026).

A solução metodológica adotada é a transliteração sistemática de uma obra lexicográfica robusta da variedade mais próxima: o dicionário Mbyá de Dooley (2006), com 1.806 entradas, numerosas sentenças exemplares traduzidas para o português e uma descrição gramatical detalhada. O fundamento é a observação de que o Nhandewa e o Mbyá do litoral paulista compartilham semelhanças fonológicas que não são fruto de contato (Ivo, 2018), ainda que o contato intenso entre os dois povos na segunda metade do século XX possa ter reforçado paralelos estruturais. Diferenças estruturais impedem a transferência direta em muitos casos, mas a proximidade permite identificar raízes e construções básicas comuns sem percorrer o ciclo longo de levantamento de *corpus* do zero, análise e só então produção de material.

Todas as entradas de Dooley já foram transliteradas e verificadas, e uma parcela significativa das sentenças exemplares foi incluída no âmbito de dois projetos de Iniciação Científica na USP. A fase seguinte é a correção do *corpus* com falantes nativos de Nhandewa, a quem se apresentam as formas transliteradas — ou sua tradução em português — para avaliação, identificando assim as divergências entre as duas variedades. O produto esperado é duplo: para as comunidades, uma obra de consulta que serve de modelo para materiais pedagógicos e de base empírica para discutir a norma ortográfica, ainda lacunar nas cartilhas; para a linguística, um *corpus* comparativo alinhado das duas variedades guaranis mais faladas no Brasil.

3 DESAFIOS E PRÓXIMOS PASSOS

A oficina de março de 2023 não se limitou à apresentação de projetos em andamento: produziu também um diagnóstico coletivo dos principais obstáculos a superar e das ações mais urgentes a empreender. As sessões de discussão — especialmente a grande discussão coletiva ao final do primeiro dia — revelaram um grupo com visão convergente sobre os objetivos comuns e com propostas concretas sobre como avançar. Os pontos que seguem organizam os temas identificados, com indicação das soluções em curso ou propostas.

3.1 Padronização de conceitos e um primeiro produto coletivo

O ponto de partida identificado pelo grupo como mais urgente e prontamente realizável é a adoção de uma lista básica padronizada de conceitos — na ordem de 500 itens —, assumindo os participantes o compromisso coletivo de preencher essa lista para cada uma de suas línguas, em formato compatível (FLEX ou CLDF), até uma data acordada. Essa lista seria o primeiro produto coletivo dos pesquisadores reunidos na oficina. A lista

de 400 conceitos já em uso no MPEG foi mencionada como candidata natural, embora a escolha final da lista — e de como ela se relaciona com padrões internacionais como o Concepticon (List et al., 2016) — precise ser discutida e acordada. Uma consideração será o léxico cultural, para possibilitar a reconstrução de elementos da protocultura Tupi, como foi feito por Rodrigues (2010).

Esse primeiro produto coletivo serviria de base tanto para a pesquisa histórico-comparativa — com a terminologia de parentesco já se perfilando como campo semântico piloto para a reconstrução, por conta dos dados existentes e dos passos já dados — quanto para a identificação de lacunas e assimetrias entre os projetos, e daria ao grupo uma plataforma comum sobre a qual construir colaborações futuras campo a campo.

3.2 Ferramentas e formatos: em direção a um consenso técnico

A oficina avançou em direção a um consenso sobre as ferramentas tecnológicas recomendadas para o grupo: FLEEx como ferramenta de trabalho cotidiano para a lexicografia de línguas individuais, CLDF como padrão de publicação e interoperabilidade para trabalho comparativo, e Lexedata como ferramenta para editar e harmonizar os diferentes conjuntos de dados. Falta ainda uma ferramenta para a mediação entre esses dois ambientes, sendo o Japiim um bom candidato. Esse consenso representa um avanço significativo em relação ao cenário anterior de total dispersão tecnológica.

Questões técnicas importantes, porém, permanecem abertas. A integração de áudio e vídeo nas entradas do FLEEx continua sendo um problema sério — mencionado em ao menos três dos projetos aqui relatados, sem que uma solução acordada tenha surgido. A conversão de bases de dados existentes em outros formatos (Toolbox, Excel, LaTeX) para FLEEx e CLDF requer tanto ferramentas específicas quanto esforço manual considerável. Atualmente, o Lexedata apoia somente a conversão entre tabelas de diferentes formatos para o CLDF. E a padronização dos campos internos do FLEEx entre projetos — condição para que os bancos individuais sejam comparáveis sem retrabalho — ainda está por ser negociada coletivamente. O projeto de criar um mecanismo para converter entre o formato LaTeX do Dicionário Etimológico de Nikulin e o CLDF foi identificado como uma ação técnica prioritária, e tem sido implementado por S. Drude e outros no MPEG desde o encontro.

3.3 Questões lexicográficas transversais

Alguns problemas de natureza teórico-lexicográfica atravessam praticamente todos os projetos da família Tupi sem solução de validade geral.

O primeiro é a **forma de citação** (o lema das entradas) nas línguas com morfologia em que verbos e muitos nomes (e.g. os nomes relacionais; neste contexto, ver Salanova & Nikulin, 2025) ocorrem obrigatoriamente com algum prefixo pessoal ou ‘relacional’ ou então uma forma particular do tema com um início específico. A discussão na oficina — que opôs defensores do lema não-flexionado (mais abstrato, mas potencialmente irreconhecível pelos falantes), da forma flexionada canônica (se ela existir na língua e for reconhecida, podendo, porém, inflar certas letras do índice alfabético) e da solução de múltiplas colunas no banco de dados — chegou à conclusão provisória de que não há solução universal. A forma de citação deve ser determinada língua por língua, em diálogo

com professores, pesquisadores e ativistas indígenas, que são os interlocutores mais qualificados para decidir sobre a usabilidade de um dicionário em cada comunidade.

O segundo problema é a **representação das formas nos dados comparativos**: transcrição fonológica (que permite avaliação científica das correspondências, mas exige análise prévia e pode tornar-se obsoleta quando a análise muda) ou ortografia prática (que respeita os sistemas já estabelecidos pelas comunidades, em muitos casos é mais estável, mas dificulta a verificação das correspondências por quem não domina todas as convenções). A solução de incluir uma coluna adicional com a transcrição fonológica ao lado da ortografia parece a mais equilibrada, mas requer trabalho adicional considerável — e, para algumas línguas com poucos dados, a transcrição seria uma inferência, não uma observação direta.

Um terceiro tema de interesse transversal é o **banco de imagens compartilhado**. A dificuldade de obter imagens culturalmente adequadas foi mencionada em ao menos quatro dos projetos aqui relatados: imagens de busca genérica são frequentemente inapropriadas para contextos indígenas específicos, e as fotografias precisam ser feitas diretamente com as comunidades. O Portal Japiim já oferece funcionalidade de compartilhamento de imagens entre dicionários, assim como a plataforma de dicionários multimídia para línguas indígenas desenvolvida no MPEG. A proposta de incentivar o uso de desenhos feitos pelos próprios falantes merece atenção especial, tanto pela qualidade reconhecida do grafismo indígena em muitas comunidades da família Tupi quanto pelo valor da representatividade da autoria comunitária nos produtos lexicográficos.

3.4 Comunicação, colaboração e participação indígena

O grupo reconheceu que, para além dos projetos individuais, são necessários mecanismos de comunicação e coordenação contínua. As propostas concretas formuladas na oficina incluem: a criação de uma lista de e-mail para troca de experiências e dúvidas; a realização de oficinas temáticas por campo semântico (flora, fauna, termos de paisagem, manufaturas), que podem ser realizadas online, com cada pesquisador apresentando seus dados sobre um campo específico; e a publicação coletiva dos resultados desta e de futuras oficinas.

Um ponto que emergiu com força das discussões foi a necessidade de que professores e acadêmicos indígenas sejam parceiros ativos — e não apenas consultores periféricos — nas decisões lexicográficas fundamentais: forma de citação, organização de campos semânticos, estrutura dos verbetes. Em várias comunidades de fala de línguas Tupi há já professores e pesquisadores indígenas com formação universitária (como os Yudjá na UFG, ou pesquisadores e professores Karitiana, Makurap, Aruá, Tupari, Suruí/Paiter, Guaratira-Sakurabiat e de outras etnias de Rondônia) que têm tanto o conhecimento linguístico quanto o apoio comunitário para contribuir efetivamente para essas decisões. A inclusão ativa dessas vozes é não apenas um imperativo ético, mas uma condição prática para que os dicionários produzidos tenham uso real nas comunidades a que se destinam.

3.5 Perspectivas para a linguística histórico-comparativa

A oficina revelou um campo de pesquisa histórico-comparativa em condições favoráveis de crescimento. O grupo Berkeley acumulou treze anos de dados sobre o Tupi-Guarani. O Dicionário Etimológico Tupi de Nikulin já inclui muitas entradas com cognatos interfamiliares. O MPEG-KIN oferece uma base de parentesco em CLDF para 31 línguas. A reconstrução do Proto-Tupi-Guarani avançou significativamente, e a reconstrução das famílias individuais — Arikém, Mondé, Tupari, Mundurukú — e unidades mais abrangentes — Maweti-Guarani, Rama-Puru — está em progresso.

Os próximos passos na linguística diacrônica da família dependem, porém, de avanços metodológicos que a oficina identificou: a integração dos dados de ortografia e de transcrição fonológica em um único banco interoperável; a migração dos principais bancos para CLDF como formato comum; e a extensão da comparação lexical sistemática para além do Tupi-Guarani, incluindo os ramos mais divergentes da família. A questão específica das línguas tonais — subfamílias Mondé, Mundurukú e, possivelmente, Ramarama, Juruna e Tupari — constitui um desafio que merece um fórum especializado, dado que o tom opera em três níveis distintos de problemas, ainda não resolvidos: análise fonológica adequada, representação ortográfica e representação em formatos comparativos interfamiliares.

O campo semântico do parentesco emergiu na oficina como o mais maduro para pesquisa comparativa na família: há dados sistematizados, pesquisadores com experiência, metodologia testada e uma reconstrução do proto-TG já publicada como ponto de partida. Os planos de continuidade incluem a reconstrução do parentesco para os Proto-Tupari, Proto-Mondé e eventualmente para o Proto-Tupi como um todo.

O Nheengatu, com sua documentação histórica contínua de quatro séculos, oferece oportunidades únicas para estudos de mudança lexical dentro do Tupi-Guarani; a elaboração de um *corpus* histórico anotado que permita rastrear essas mudanças de forma sistemática é um objetivo de médio prazo com alta viabilidade. A possível integração dos dados arqueológicos, genéticos e/ou históricos com os bancos de dados sincrônicos existentes — incluindo o Dicionário Etimológico de Nikulin e o banco Berkeley TG — completaria uma infraestrutura de dados capaz de possibilitar análises filogenéticas com ancoragem temporal.

4 CONSIDERAÇÕES FINAIS

A Oficina sobre Dicionários e Bancos de Dados Lexicais de Línguas Tupi de março de 2023 revelou um campo em momento de transição: a passagem de uma fase de projetos individuais heterogêneos para uma fase com um grande potencial para uma colaboração coordenada, com padrões técnicos mais maduros e uma comunidade de pesquisadores com objetivos comuns em crescente articulação. A família Tupi oferece, para essa transição, condições excepcionalmente favoráveis: um número relativamente grande de pesquisadores ativos, acervos de dados já substanciais em várias línguas, ferramentas digitais maduras e padrões abertos de interoperabilidade.

Os desafios não são insuperáveis. Padronizar uma lista mínima de conceitos, acordar quais campos básicos (‘colunas em uma tabela’) devem ser incluídos como obrigatórios e quais como recomendados no FLEx ou em outros formatos, migrar os principais bancos de dados para o CLDF, criar um canal de comunicação permanente entre pesquisadores e — talvez o mais importante — garantir que as comunidades e os acadêmicos indígenas sejam parceiros ativos do trabalho lexicográfico: esses são passos concretos e realizáveis no horizonte imediato. A meta mais ambiciosa — ter dicionários publicados para as principais línguas da família, bancos de dados comparativos interoperáveis e reconstruções avançadas em um horizonte de cinco a dez anos — está ao alcance desse grupo de pesquisadores, desde que os próximos passos identificados na oficina sejam efetivamente implementados e que novos financiamentos permitam sustentar a continuidade do trabalho.

CONTRIBUIÇÃO DOS AUTORES

Taxonomia CRediT. Sebastian Drude: Administração de projeto, Aquisição de financiamento, Conceituação, Investigação, Escrita – rascunho original, Escrita – revisão e edição. Thomas Finbow: Conceituação, Investigação, Escrita – rascunho original, Escrita – revisão e edição. Lorrain Tyson dos Santos Araújo, Joshua Birchall, Saulo Rodrigo Teixeira Brito, Natalia Chousou-Polydouri, Cristina Martins Fargetti, Helder Perri Ferreira, Ana Vilacy Galúcio, Lev Michael, Denny Moore, Andrey Nikulin, Luís H. Araújo Oliveira e Luciana Storto: Conceituação, Investigação, Escrita – revisão e edição.

FINANCIAMENTO

O trabalho de Sebastian Drude contou com o apoio da Deutsche Forschungsgemeinschaft (DFG, Fundação Alemã de Pesquisa), Programa Heisenberg, processo DR 857/3-1. A oficina contou com apoio desta bolsa e do Museu P. E. Goeldi. Ana Vilacy Galúcio contou com o apoio da Bolsa de Produtividade do CNPq, processos 304766/2022-4 e 314726/2026-8. Saulo Brito conta com o apoio da bolsa MPEG/CNPq (DTI-B, processo 380059/2026-6). Luciana Storto recebe apoio do CNPq pela bolsa PQ2: processo 312613/2023-7. Os demais autores não receberam financiamento específico para participar da oficina de 2023 ou para elaborar este texto.

CONFLITO DE INTERESSES

Os autores declaram não haver conflito de interesses.

ÉTICA E CONSENTIMENTO

Não houve pesquisa com seres humanos na elaboração deste artigo. A oficina de março de 2023 foi gravada com a concordância dos participantes, em sua maioria coautores deste artigo; os demais são citados no texto.

DISPONIBILIDADE DE DADOS

Este artigo é uma revisão: as obras, bancos de dados e ferramentas discutidos estão referenciados no texto e na lista de referências, com identificador persistente (DOI), quando houver. O único material não publicado usado na redação deste artigo são as gravações, transcrições e resumos da oficina de março de 2023 (ver seção 1.5), que não são disponibilizados publicamente, pois a concordância dos participantes se referiu à gravação da oficina, e não à sua publicação; podem ser solicitados ao autor de correspondência, com a anuência dos participantes envolvidos.

REFERÊNCIAS

- Aikio, A. (2020). *URALIC ETYMOLOGICAL DICTIONARY (draft version of entries A-Ć)*. https://www.academia.edu/41659514/URALIC_ETYMOLOGICAL_DICTIONARY_draft_version_of_entries_A_%C4%86
- Alves, P. M. (2004). *O léxico do Tupari: Proposta de um dicionário bilíngüe* [Tese de doutorado, Universidade Estadual Paulista “Júlio de Mesquita”].
- Anchieta, J. de. (1595). *Arte de Gramática da língua mais usada na costa do Brasil* (Trabalho original publicado em 1595; edição de 1946). Ed. Anchieta.
- Anderson, G. (2011). Language Hotspots: What (applied) linguistics and education should do about language endangerment in the twenty-first century. *Language and Education*, 25, 273–289. <https://doi.org/10.1080/09500782.2011.577218>
- Araújo, A. de. (1618). *Catecismo na Lingoa Brasilica*. <http://archive.org/details/catecismo-na-lingoa-brasilica-1618>
- Ávila, M. T. (2021). *Proposta de dicionário nheengatu-português*. USP. <http://archive.org/details/avila-nheengatu>
- Beier, C., & Michael, L. (2022). Managing Lexicography Data: A Practical, Principled Approach Using FLEx (FieldWorks Language Explorer). Em A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Org.), *The Open Handbook of Linguistic Data Management* (p. 301–314). The MIT Press. <https://doi.org/10.7551/mitpress/12200.003.0029>
- Bettendorff, J. F. (1800). *Compendio da Doutrina Christã na Lingua Portugueza, e Brasilica* (2º ed.). Offic. de Simão Thaddeo Ferreira. (Trabalho original publicado 1687, Imprensa Miguel Deslandes)
- Birchall, J., Araújo [Oliveira], L. H., Passmore, S., & Jordan, F. M. (2022). *MPEGKin: A database of kinship terminology from the Tupian and Cariban language families* (Versão v1.0) [Conjunto de dados]. Zenodo. <https://doi.org/10.5281/ZENODO.6474213>
- Brito, S., Birchall, J., & Galúcio, A. V. (2025). *MPEG-Linguistica/CSV2RMD: CSV2RMD 1.0* (Versão v1.0.0) [Programa de computador]. Zenodo. <https://doi.org/10.5281/ZENODO.14894549>
- Campbell, L. (2013). *Historical linguistics: An introduction* (3rd ed.). Edinburgh University Press.
- Carvalho, F. O. de. (2019). Revisitando o Proto-Jurúna: A reconstrução da série de oclusivas orais. Em E. dos S. Oliveira, E. A. Vasconcelos, & R. D. Sanches (Org.), *Estudos linguísticos na Amazônia* (p. 215–236). Pontes.

- Carvalho, F. O. de. (2022). On the Guaranian evidence for two Proto-Tupi-Guarani affricates. *Journal of Language Relationship*, 20, 81.
- Carvalho, F. O. de. (2023). Proto-Tupi-Guarani had no palatalized velar stop. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 18(1), e20220013. <https://doi.org/10.1590/2178-2547-bgoeldi-2022-0013>
- Carvalho, F. O. de, & Birchall, J. (2022). A comparative reconstruction of Proto-Tupi-Guarani kinship terminology. *LIAMES: Línguas Indígenas Americanas*, 22, e022001. <https://doi.org/10.20396/liames.v22i00.8666489>
- Corrêa da Silva, B. C. (2010). *Mawé/Awetí/Tupí-Guaraní: Relações Linguísticas e Implicações Históricas* [Tese de doutorado, Universidade de Brasília].
- Crofts, M. & others. (1977). *Dicionário bilingüe em português e mundurukú*. Fundação Nacional do Índio.
- d'Abbeville, C. (1614). *Histoire de la Mission des Peres Capucins en l'Isle de Maragnan et terres circonuoisines*. L'Imprimerie de François Huby.
- d'Evreux, Y. (1615). *Voyage au Nord du Brésil* (F. Obermeier, Org.). <http://archive.org/details/voyageaunorddubresil1615>
- D'Angelis, W. da R. (2023). A língua Nheengatu e suas ortografias: Questões técnicas e de política linguística. *LIAMES: Línguas Indígenas Americanas*, 23, e023004–e023004. (Brazil; Contemporary). <https://doi.org/10.20396/liames.v23i00.8670910>
- Dooley, R. (2006). *LÉXICO GUARANI, DIALETO MBYÁ com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística*. Summer Institute of Linguistics.
- Drude, S. (2006). On the position of the Awetí language in the Tupí family. Em W. Dietrich & H. Symeonidis (Org.), *Guarani y “Maweti-Tupi-Guarani”*. *Estudios históricos y descriptivos sobre una familia lingüística de America del Sur* (p. 11–45). LIT Verlag.
- Fargetti, C. M. (2018). Estudios del léxico de lenguas indígenas: ¿terminología? In M. G. González, M.-D. Sánchez-Palomino, & I. V. Mateos (Eds.), *Terminología: A necessidade da colaboración* (pp. 343–368). Vervuert Verlagsgesellschaft. <https://doi.org/10.31819/9783964568090-019>
- Fargetti, C. M. (2021). Terminologia da cultura material Juruna. Letraria. <https://www.letraria.net/terminologia-da-cultura-material-juruna/>
- Fargetti, C. M. (2026). Terminologia Etnográfica: Uma abordagem da música em línguas minoritárias e/ou minorizadas. Em M. Meirelles, *Terminologia na América-Latina: Passado, presente e futuro/Terminologia en Latino América: Pasado, presente y futuro* (p. 477–498). CirKula. <https://online.fliphtml5.com/kfqsfbq/#p=477>
- Fargetti, C. M., & Rodrigues, C. L. R. (2008). Consoantes do Xipaya e do Juruna—Uma comparação em busca do proto-sistema. *Alfa*, 52(2), 535–563.
- Fargetti, C. M., & Rodrigues, C. L. R. (2021). Vogais do Juruna e do Xipaya – uma comparação. *MOARA – Revista Eletrônica do Programa de Pós-Graduação em Letras* ISSN: 0104-0944, (58), 186. <https://doi.org/10.18542/moara.v0i58.10855>

- Felzke, L. F., & Moore, D. (2019). Terminologias de parentesco dos grupos da família linguística Mondé. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 14(1), 15–32. <https://doi.org/10.1590/1981.81222019000100003>
- Figueira, L. (1621). *Arte de Grammatica da Lingua Brasilica*. Manoel da Silua.
- Finbow, T. (2025). A Sociophilological Account of the Formation and Evolution of the Term Língua Geral, with Emphasis on Amazonia. *Cadernos de Linguística*, 6(1), 806. <https://doi.org/10.25189/2675-4916.2025.v6.n1.id806>
- Finbow, T., Storto, L. R., Autuori, J. D., & Geiger, M. (2026). Processos e Produtos do Projeto “Processamento Ético de Línguas Indígenas” do Centro de Inteligência Artificial da USP (2022-2025). *Revista da ABRALIN*, 89–115. <https://doi.org/10.25189/rabralin.v24i2.2387>
- Forkel, R., & Hammarström, H. (2014ff). *Glottolog 4.4—GlottoScope*. Glottolog. <https://glottolog.org/langdoc/status>
- Forkel, R., List, J.-M., Greenhill, S. J., Rzymiski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1), 180205. <https://doi.org/10.1038/sdata.2018.205>
- Galúcio, A. V., Birchall, J., Rocha, I., Brito, S. R. T., Soares, M. A. R., Marcon, I. N., Rodrigues Junior, D. D. C., Santos, E., Rocha, V. S., Nogueira, A. F. D. S., Pinheiro, F. H. F., Solano, J. K. S., & Silva, B. P. (2026). Dicionários multimídia: Uma tecnologia social para auxiliar a aprendizagem e revitalização de línguas minorizadas. *Revista da ABRALIN*, 116–144. <https://doi.org/10.25189/rabralin.v24i2.2384>
- Galúcio, A. V., & Gabas Jr., N. (2002). Evidências para o agrupamento genético Karo-Puroborá, Tronco Tupí. *Proceedings of the 17º Encontro nacional da ANPOLL*. The 17º Encontro nacional da ANPOLL.
- Galúcio, A. V., & Gabas Jr., N. (2014). *Revisiting the hypothesis of a Ramarama-Purubora branch of the Tupian family*. Amazônicas V, Belém, PA.
- Galúcio, A. V., Meira, S., Birchall, J., Moore, D., Gabas Jr., N., Drude, S., Storto, L., Picanço, G., & Rodrigues, C. R. (2015). Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim Do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10(2), 229–274. <https://doi.org/10.1590/1981-81222015000200004>
- Galúcio, A. V., & Nogueira, A. F. de S. (2012). Comparative study of the Tupari branch of the Tupi family: Contributions to understanding its historical development and internal classification. *Memorias del V Congreso de Idiomas Indígenas de Latinoamérica, 6–8 de octubre de 2011, Universidad de Texas*. https://etnolinguistica.wdfiles.com/local--files/artigo%3Agalucio-nogueira-2011/galucio_nogueira_2011_comparative.pdf
- Galúcio, A. V., & Nogueira, A. F. de S. (2018). From object nominalization to object focus: The innovative A-alignment in the Tuparian languages (Tupian family). *Journal of Historical Linguistics*, 8(1), 95–127. <https://doi.org/10.1075/jhl.16025.gal>
- Gerardi, F. F., & Reichert, S. (2021). The Tupí-Guaraní language family: A phylogenetic classification. *Diachronica*, 38(2), 151–188. <https://doi.org/10.1075/dia.18032.fer>

- Gerardi, F. F., Reichert, S., & Aragon, C. C. (2021). TuLeD (Tupían lexical database): Introducing a database of a South American language family. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-020-09521-5>
- Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). Bayesian Phylolinguistics. Em R. D. Janda, B. D. Joseph, & B. S. Vance (Org.), *The Handbook of Historical Linguistics* (1º ed., p. 226–253). Wiley. <https://doi.org/10.1002/9781118732168.ch11>
- Himmelman, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36(1), 161–195.
- Ivo, I. P. (2018). *Características fonéticas e fonologia do Guarani no Brasil* [Doutora em Linguística, Universidade Estadual de Campinas (UNICAMP)]. <https://doi.org/10.47749/T/UNICAMP.2018.1090892>
- Jensen, C. J. S. (1989). *O Desenvolvimento Histórico da Língua Wayampi*. Ed. UNICAMP.
- Kaiping, G. A., Steiger, M. S., & Chousou-Polydouri, N. (2022). Lexedata: A toolbox to edit CLDF lexical datasets. *Journal of Open Source Software*, 7(72), 4140. <https://doi.org/10.21105/joss.04140>
- Landin, D. (1983). *Dicionário e léxico karitiâna/português*. SIL. <http://www.sil.org/americas/brasil/PUBLICNS/DICTGRAM/KTDic.pdf>
- List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A Resource for the Linking of Concept Lists. Em N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Org.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 2393–2400). European Language Resources Association (ELRA). <https://aclanthology.org/L16-1379/>
- Lobato, F., & Galúcio, A. V. (2023). Primeiros passos na reconstrução do ramo Ramarama–Puruborá da família Tupí. Em B. F. s Lima-Padovani, J. Andrade, L. M. S. Rodrigues, S. Facundes, M. da C. V. Pereira, & P. do N. Costa (Org.), *Estudos sobre Línguas e Culturas Amazônicas* (V. 1, p. 111–123). Grupo de Estudos sobre Línguas e Culturas Indígenas da Amazônia.
- Lopes, J. D. (2024). *Catálogo de Materiais Lexicográficos das Línguas Indígenas do Brasil*. Jorge Domingues Lopes. <https://www.lexicografiaindigena.com.br/>
- Meira, S., & Drude, S. (2015). A summary reconstruction of Proto-Maweti-Guarani segmental phonology. *Boletim Do Museu Paraense Emílio Goeldi: Ciências Humanas*, 10(2), 275–296. <https://doi.org/10.1590/1981-81222015000200005>
- Mello, A. A. S. (2000). *Estudo histórico da família lingüística Tupí-Guaraní: Aspectos fonológicos e lexicais* [Tese de doutorado, Universidade Federal de Santa Catarina].
- Michael, L., & Chousou-Polydouri, N. (2019). Computational phylogenetics and the classification of South American languages. *Language and Linguistics Compass*, 13(12), e12358. <https://doi.org/10.1111/lnc3.12358>
- Michael, L., Chousou-Polydouri, N., Bartolomei, K., Donnelly, E., Meira, S., Wauters, V., & O'Hagan, Z. (2015). A Bayesian Phylogenetic Classification of Tupí-Guaraní. *LIAMES: Línguas Indígenas Americanas*, 15(2), 193. <https://doi.org/10.20396/liames.v15i2.8642301>
- Moore, D. (2012, julho 16). Tone-length correspondences within the Mondé branch of the Tupi family. Em *Symposion Historical variation and variation by contact*

- among the Tupi languages*. 54 Congresso Internacional de Americanistas, Vienna.
- Moore, D. (2014, maio 28). *Aspects of the Proto-Mondé sound inventory*. Línguas Amazônicas V, Belém.
- Moore, D., & Galúcio, A. V. (1994). Reconstruction of Proto-Tupari consonants and vowels. Em *Proceedings of the meeting of the Society for the Study of the Indigenous Languages of the Americas and the Hokan-Penutian workshop* (V. 8, p. 119–137). University of California.
- Moore, D., Galúcio, A. V., & Gabas Jr., N. (2008). O Desafio de Documentar e Preservar as Línguas Amazônicas. *Scientific American Brasil, Edição Especial 3: Amazônia: A Floresta e o Futuro*, 36–43.
- Murdock, G. P. (1970). Kin Term Patterns and Their Distribution. *Ethnology*, 9(2), 165–208. <https://doi.org/10.2307/3772782>
- Nikulin, A., & Andrade, R. (2020). The rise and fall of approximants in the Tuparian languages. *Journal of Language Relationship*, 18(3–4), 284–319. <https://doi.org/10.31826/jlr-2020-183-412>
- Nikulin, A., & Carol, J. (2024). *Historical phonology of Mataguayan*. Language Science Press. <https://doi.org/10.5281/zenodo.13907413>
- Nikulin, A., & Carvalho, F. O. de. (2019). Estudos Diacrônicos de Línguas Indígenas Brasileiras: Um Panorama. *Macabéa - Revista Eletrônica do Netlli*, 8(2), 255–305. <https://doi.org/10.47295/mren.v8i2.1910>
- Nikulin, A., & Carvalho, F. O. de. (2022). A revised reconstruction of the Proto-Tupian vowel system. *Boletim Do Museu Paraense Emílio Goeldi. Ciências Humanas*, 17(2), e20210035. <https://doi.org/10.1590/2178-2547-BGOELDI-2021-0035>
- Nimuendajú, C. (1987). *As lendas da criação e destruição do mundo como fundamentos da religião dos Apapocúva-Guaraní* (C. Emmerich & E. B. V. de Castro, Trans.). Editora HUCITEC; Editora da Universidade de São Paulo. (Trabalho original publicado 1914, Zeitschrift für Ethnologie)
- Nogueira, A. F. S., Galúcio, A. V., Soares-Pinto, N., & Singerman, A. R. (2019). Termos de parentesco nas línguas Tuparí (família Tupí). *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 14(1), 33–64. <https://doi.org/10.1590/1981.81222019000100004>
- O'Hagan, Z., Chousou-Polydouri, N., & Michael, L. (2019). Phylogenetic classification supports a Northeastern Amazonian Proto-Tupí-Guaraní Homeland. *LIAMES: Línguas Indígenas Americanas*, 19, e019018. <https://doi.org/10.20396/liames.v19i0.8655791>
- Pacheco Ribeiro, M. de J. (2010). *Dicionário sateré-mawé/português*. Guajará-Mirim: Universidade Federal de Rondônia.
- Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., Birchall, J., Bown, C., Calladine, J., Deb, A., Diederer, A., Metsäranta, N. P., Araújo [Oliveira], L. H., Schembri, R., Hickey-Hall, J., Honkola, T., Mitchell, A., Poole, L., Rácz, P. M., ... Jordan, F. M. (2023). Kinbank: A global database of kinship terminology. *PLOS ONE*, 18(5), e0283218. <https://doi.org/10.1371/journal.pone.0283218>
- Picanço, G. L. (2003). *Language Change. A comparison of Mundurukú and Karuaya*.

- Picanço, G. L. (2010). A reconstruction of nasal harmony in proto-mundurukú (tupi). *International Journal of American Linguistics*, 76(4), 411–438. <https://doi.org/10.1086/656313>
- Picanço, G. L. (2019). *A fonologia diacrônica do Proto-Mundurukú (Tupí)* (1a. edição). Appris Editora.
- Ramirez, H. (2020). *Enciclopédia das línguas Arawak: Acrescida de seis novas línguas e dois bancos de dados*. Editora CRV.
- Rodrigues, A. D. (1985). Relações internas na família lingüística Tupí-Guaraní. *Revista de Antropologia*, 27/28, 33–53.
- Rodrigues, A. D. (1986). *Línguas Brasileiras. Para o Conhecimento das Línguas Indígenas*. Ed. Loyola.
- Rodrigues, A. D. (2005). As vogais orais do Proto-Tupí (UNB, Trad.). Em A. D. Rodrigues & A. S. A. C. Cabral (Org.), *Novos estudos sobre línguas indígenas* (p. 35–46). Editora da UNB.
- Rodrigues, A. D. (2007). As consoantes do Proto-Tupí. Em A. S. A. C. Cabral & A. D. Rodrigues (Org.), *Línguas e Culturas Tupí* (p. 167–203). Curt Nimuendajú.
- Rodrigues, A. D. (2010). Linguistic Reconstruction of Elements of Prehistoric Tupi Culture. Em E. B. Carlin & S. Van De Kerke (Org.), *Linguistics and Archaeology in the Americas* (p. 1–10). BRILL. https://doi.org/10.1163/9789047427087_002
- Rodrigues, A. D., & Cabral, A. S. A. C. (2012). Tupían. Em L. Campbell & V. M. Grondona (Org.), *The indigenous languages of South America: A comprehensive guide* (p. 495–574). Mouton de Gruyter. <https://doi.org/10.1515/9783110258035.495>
- Rodrigues, A. D., & Dietrich, W. (1997). On the linguistic relationship between Mawé and Tupí-Guaraní. *Diachronica*, 14(2), 265–304.
- Salanova, A. P., & Nikulin, A. (2025). A Typology of Relationalizing and Absolutizing Morphology in Lowland South American Languages. *International Journal of American Linguistics*, 91(1), 43–96. <https://doi.org/10.1086/732880>
- Schleicher, C. O. (1998). *Comparative and Internal Reconstruction of the Tupi-Guarani Language family* [Tese de doutorado, University of Wisconsin–Madison].
- Seyfeddinipur, M., Drude, S., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B., Cruz, H., Epps, P. L., Ferreira, V., Galúcio, A. V., Hellwig, B., Hinte, O., Holton, G., Jung, D., Kasinskaite Buddeberg, I., Krifka, M., Kung, S., Monroig, M., Neba, A. N., ... Woodbury, T. (2019). Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation*, 13, 545–563.
- SIL International. (2014). *FLEX: FieldWorks Language Explorer™ - Dictionary Creation Software* [Programa de computador]. <https://software.sil.org/fieldworks/>
- SIL LSDev. (2026). *LIFT-standard* [C#]. <https://github.com/sillsdev/lift-standard> (Trabalho original publicado 2015)
- Wimbish, J., & Buseman, A. (1987). *Field Linguist's Toolbox* (Versão 1.6.4) [Windows]. SIL International. <https://software.sil.org/toolbox/>

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença [Creative Commons CC-BY](#).
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.