I Speak Kanoe: Engineering a Dataset for an Endangered Isolated Language

Gabrielly A. Gomes[1] **, Iago S. Aragão**[1] **, Martony D. Silva**[1] **, Patricia M. L. de Drumond**[1]

[1]Centro de Educação Aberta e à Distância (CEAD) – Universidade Federal do Piauí (UFPI)
Campus Universitário Ministro Petrônio Portella – Ininga, Teresina – PI, 64049-550

{gabrielly.gomes, iago.aragao, martony.silva, patricia.medyna}@ufpi.edu.br[1]

Abstract. *This paper present Eu Falo Kanoê, a relational dataset for documentation, study, and revitalization of the Kanoê language, a critically endangered linguistic isolate from southern Rondônia (Brazil). Consolidating 18 documentary sources (1992–2025) into a normalized PostgreSQL schema with five tables, it contains 1,684 lexical entries, 1,937 IPA pronunciation records (90.8% coverage), 2,134 sense annotations, and 612 contextualized sentences, all traceable via foreign keys. Curation followed FAIR and CARE principles for indigenous data governance. Released under CC BY-NC-SA, it provides a replicable foundation for NLP in low-resource settings and linguistic research on language isolates.*

Resumo. *Este artigo apresenta Eu Falo Kanoê, um conjunto de dados relacional para documentação, estudo e revitalização da língua Kanoê, isolada e em perigo crítico de extinção no sul de Rondônia (Brasil). Consolidando 18 fontes documentais (1992–2025) em um esquema PostgreSQL normalizado com cinco tabelas, o recurso contém 1.684 entradas lexicais, 1.937 registros de pronúncia em IPA (90,8% de cobertura), 2.134 anotações de sentido e 612 sentenças contextualizadas, todas rastreáveis por chaves estrangeiras. A curadoria seguiu as diretrizes FAIR e CARE para governança de dados indígenas. Disponibilizado sob licença CC BY-NC-SA, oferece base replicável para PLN em cenários de baixo recurso e pesquisa linguística sobre línguas isoladas.*

1. Introduction

Linguistic diversity is among humanity's most fragile cultural assets. According to recent estimates, one language disappears approximately every two weeks, and more than half of the roughly 7,000 languages spoken worldwide are projected to vanish by the end of the century [UNESCO 2024]. Language loss is not merely a linguistic phenomenon: it entails the erosion of cosmologies, traditional knowledge systems, and collective memories that constitute indigenous identities.

In Brazil, 305 ethnic groups coexist and approximately 274 indigenous languages are spoken [FUNAI 2022]. Among these is Kanoê (also known as Kapixána), spoken in southern Rondônia near the Bolivian border. Kanoê is classified as a linguistic isolate, with no proven genetic affiliation to any other South American language family [Bacelar 1992, Bacelar 2004], which makes it particularly valuable for typological and comparative studies. This singularity, however, contrasts with extreme vulnerability: as

of 2012, only three monolingual speakers of Kanoê were documented, all members of a single family in the Omerê Indigenous Territory [Bacelar 2012].

The preservation challenge is compounded by the fragmentation of existing records. Decades of linguistic, pedagogical, and ethnographic work have produced a substantial body of documentation, but this material remains scattered across theses, pedagogical booklets, and institutional reports, most of them available only as scanned PDF files with inconsistent orthographic conventions. This dispersion hinders both systematic linguistic research and the development of digital tools capable of supporting revitalization efforts. In the broader context of natural language processing (NLP), Kanoê exemplifies the class of *extremely low-resource* languages for which no structured corpus, lexicon, or benchmark has been made publicly available [Joshi et al. 2020, Hedderich et al. 2021].

This paper presents *I Speak Kanoê*, a relational dataset engineered to address this gap. The resource unifies 18 documentary sources into a normalized schema of five tables, implemented in PostgreSQL and released publicly on Hugging Face. The contribution is not the production of new linguistic primary data, which remains the domain of field researchers in collaboration with the Kanoê community, but the systematic *data engineering* work required to transform heterogeneous legacy materials into a structured, queryable, verifiable, and reusable resource. The dataset preserves provenance for every entry through foreign keys linking lexemes, senses, and sentences to their primary bibliographic sources, enabling both computational applications and philological traceability.

The remainder of the paper is organized as follows. Section 2 surveys related work on datasets for indigenous and low-resource languages. Section 3 describes the data acquisition, curation, and normalization pipeline. Section 4 details the schema, quantitative characteristics, and data dictionary. Section 5 discusses potential research applications and known limitations. Section 6 concludes and outlines directions for future work.

2. Related Work

The documentation of endangered languages in structured form has become an active research area at the intersection of computational linguistics, digital humanities, and linguistic anthropology. [Mager et al. 2018] survey the state of language technology for indigenous languages of the Americas, identifying data scarcity, orthographic variation, and absence of benchmarks as the primary obstacles to computational research. Their review highlights that most indigenous languages of the continent lack even minimal lexical resources in machine-readable formats, a situation that remains largely unchanged for Brazilian languages.

More recently, [Cavalin et al. 2023] investigated native language identification for seven Brazilian indigenous languages using bible-based training data, reporting that accuracy drops sharply when models are evaluated on non-biblical texts, particularly for languages with unstable orthographic conventions. This work, together with the broader research program documented in [Pinhanez et al. 2024], underscores the need for curated, provenance-aware datasets to complement opportunistic web-mined corpora. The AMERICASNLP shared tasks [Mager et al. 2021] further demonstrate that even modest amounts of high-quality parallel data can yield meaningful improvements in machine translation for indigenous languages, provided that the data are carefully curated and properly attributed.

From a data governance perspective, [Bird 2020] argues for an explicitly decolonial approach to speech and language technology, cautioning against treating indigenous knowledge as raw material for the taking. This concern is operationalized in the CARE (Collective benefit, Authority to control, Responsibility, and Ethics) Principles for Indigenous Data Governance [Carroll et al. 2020], which complement the FAIR (Findable, Accessible, Interoperable, and Reusable) Principles [Wilkinson et al. 2016] by emphasizing collective benefit, authority to control, responsibility, and ethics. The present work adopts both frameworks as design constraints, favoring transparent provenance, restrictive licensing, and non-commercial reuse.

Regarding Kanoê specifically, the foundational linguistic descriptions were produced by [Bacelar 1992] and [Bacelar 2004], which remain the most comprehensive phonological and grammatical accounts of the language. More recent initiatives include the multimedia Kanoê-Portuguese dictionary [Marcon et al. 2023], hosted at the Museu Paraense Emílio Goeldi, and the study of conceptual metaphors by [Unternbäumen et al. 2025]. To the best of our knowledge, no publicly available structured relational dataset for Kanoê exists prior to the present work; existing resources are distributed as isolated PDFs or as web interfaces that do not expose the underlying data for programmatic access or bulk download.

3. Methodology

The construction of *I Speak Kanoê* followed a four-stage pipeline: (i) source acquisition and inventory, (ii) manual extraction and transcription, (iii) normalization and cross-source consolidation, and (iv) relational modeling and loading into a managed database. Each stage was designed to preserve provenance and to enable reproducibility.

3.1. Source Acquisition and Inventory

Eighteen documentary sources were selected for inclusion, covering the period 1992–2025. The corpus of sources comprises one doctoral thesis [Bacelar 2004], one master's dissertation [Bacelar 1992], four peer-reviewed articles (including [Bacelar and Silva Junior 2003, Bacelar and Pereira 2009, Unternbäumen et al. 2025]), six pedagogical primers produced under the PRODOCLIN program [Bacelar 2010, Bacelar 2013, Bacelar 2014b, Bacelar 2014a], and six ethnographic and technical reports [Bacelar and Canoê 2011, Bacelar et al. 2012, Bacelar et al. 2011, Bacelar 2012, Bacelar 2015]. All sources were obtained from public repositories, including the FUNAI digital library and the PRODOCLIN/Museu do Índio portal, and each was registered in the `bibliografia` table with its full metadata (title, authors, year, type, file reference, and annotations).

Selection criteria required that a source (i) contain primary linguistic material on Kanoê, (ii) be attributed to identifiable authors, and (iii) be accessible for consultation, either online or through deposit libraries. Secondary works that merely cite Kanoê were excluded from the inventory but remain available for future inclusion.

An extensive literature review was conducted to identify all available materials. The concentration of sources on a single author (Bacelar) reflects the historical reality of Kanoê studies, as his fieldwork constitutes the vast majority of the primary linguistic documentation available for this critically endangered language.

3.2. Manual Extraction and Transcription

Because most sources exist only as scanned PDFs of variable quality, automated text extraction proved unreliable, particularly for phonetic transcriptions involving IPA symbols and diacritics. All lexical entries, sense descriptions, and example sentences were therefore extracted manually by trained annotators, with systematic cross-checking against the original page images. For each extracted entry, the following information was recorded when available: the Kanoê form in its source orthography, the IPA transcription, the Portuguese gloss or definition, cultural or pragmatic notes, and the bibliographic source identifier.

A representative example illustrates the density of the extracted information. The entry *päräpärä* [pæɾæpæɾæ] "pineapple" is recorded with a cultural note indicating that the fruit is cultivated in family plots (*roças*), and is cross-referenced to the pedagogical primer in which it first appears. Every such entry is accompanied by a numeric foreign key pointing to the source, ensuring that downstream users can retrieve the exact page of origin when needed.

3.3. Normalization and Cross-Source Consolidation

Legacy sources exhibit substantial orthographic variation: the same lexeme may appear spelled differently across a thesis from 1992, a primer from 2010, and a report from 2015. Three classes of inconsistency were identified and treated systematically: (i) variation in the representation of nasal vowels (e.g., *ã* vs. *an*), (ii) alternation between the digraphs *k*/*c*/*qu* and *w*/*u* in consonant clusters, and (iii) inconsistent marking of stress and tone on polysyllabic forms. Normalization followed the conventions established in [Bacelar 1992, Bacelar 2004], which are the most detailed phonological descriptions of the language, while the original spelling of each source was preserved in the `grafia` field of the `pronuncia` table to allow reconstruction of the source-specific form.

Variant forms and polysemy were handled without collapsing distinctions. When a single `palavra` is attested with multiple pronunciations, each variant was stored as a separate `pronuncia` row linked to the same lexeme: 251 lexemes have two or more pronunciation variants, with a maximum of three variants per lexeme. Similarly, 252 lexemes are polysemous and have two or more `significado` rows, with a maximum of four distinct senses per lexeme.

Cross-source consolidation was performed by manual reconciliation. When two sources disagreed on a gloss or transcription, both readings were retained as separate `significado` or `pronuncia` records, each with its own bibliographic link, leaving downstream users free to adjudicate. This policy favors transparency over apparent consistency and preserves the philological record of the language as it has actually been documented.

3.4. Relational Modeling and Validation

The consolidated data were loaded into a managed PostgreSQL instance hosted on Supabase. The choice of a relational backend, rather than a denormalized tabular file or a document store, was motivated by three requirements of the project: (i) referential integrity between lexemes, senses, pronunciations, sentences, and sources must be enforced

at the database level; (ii) queries across multiple annotation layers (for example, "all polysemous nouns attested in pedagogical primers") must be expressible in standard SQL; and (iii) future extensions (audio pointers, morphological segmentation, syntactic annotation) must be integrable without restructuring the core schema.

Validation was performed along two axes. *Technical validation* consisted of automated checks for referential integrity, orphan foreign keys, duplicate primary keys, and schema-level constraints; the full validation script is released alongside the dataset. *Linguistic validation* consisted of manual cross-source review of a stratified sample of 10% of the entries, focusing on coherence between the `grafia`, `ipa`, and `traducao_primaria` fields. Discrepancies flagged during linguistic validation were either resolved by consulting the primary sources or recorded as open issues in the project's public tracker.

Ethical considerations followed the spirit of the CARE Principles [Carroll et al. 2020]. The dataset was built exclusively from publicly available documentary sources that were themselves produced under consent protocols established with the Kanoê community since the 1990s [Bacelar 2012]. The license adopted (CC BY-NC-SA 4.0) prohibits commercial reuse and requires derivative works to preserve the same terms, aligning with the non-commercial orientation of the original materials.

4. Dataset Description

This section details the architecture and contents of the I Speak Kanoê dataset. It first presents the relational schema and data dictionary, followed by a quantitative summary of the extracted records. Finally, it addresses the handling of missing data and the design choices made to ensure philological transparency.

4.1. Storage, Schema, and Data Dictionary

The dataset is organized as five interrelated tables connected through primary and foreign keys. The schema separates the *lexeme* (an abstract word identity), the *pronunciation* (one or more phonetic realizations of that lexeme, including variant forms), the *sense* (one or more lexical meanings, each potentially from a distinct source), the *sentence* (a contextualized example of use), and the *bibliography* (the documentary source from which each piece of information was drawn). This separation reflects standard lexicographic practice and supports one-to-many relationships at multiple levels: a lexeme may have many pronunciations, many senses, and many example sentences.

As shown in Figure 1, the five tables follow PostgreSQL conventions (`int8` for 64-bit integers, `text` for variable-length strings, and `jsonb` for binary JSON payloads). The `palavra` table holds canonical lexemes with their part-of-speech tag and an extensible `tags_nosql` field; `pronuncia` stores one or more phonetic realizations per lexeme, including the source orthography (`grafia`), the IPA transcription (`ipa`), the form type (standard, variant, allomorph), and a `jsonb` field reserved for morphological segmentation. The `significado` table captures one or more senses per lexeme, each with a primary Portuguese gloss, an optional extended definition, cultural notes, a semantic-domain tag, and a foreign key to the source in `bibliografia`. The `frase` table records example sentences in Kanoê and Portuguese, with contextual annotations and foreign keys to both the associated sense and lexeme. Finally, `bibliografia`

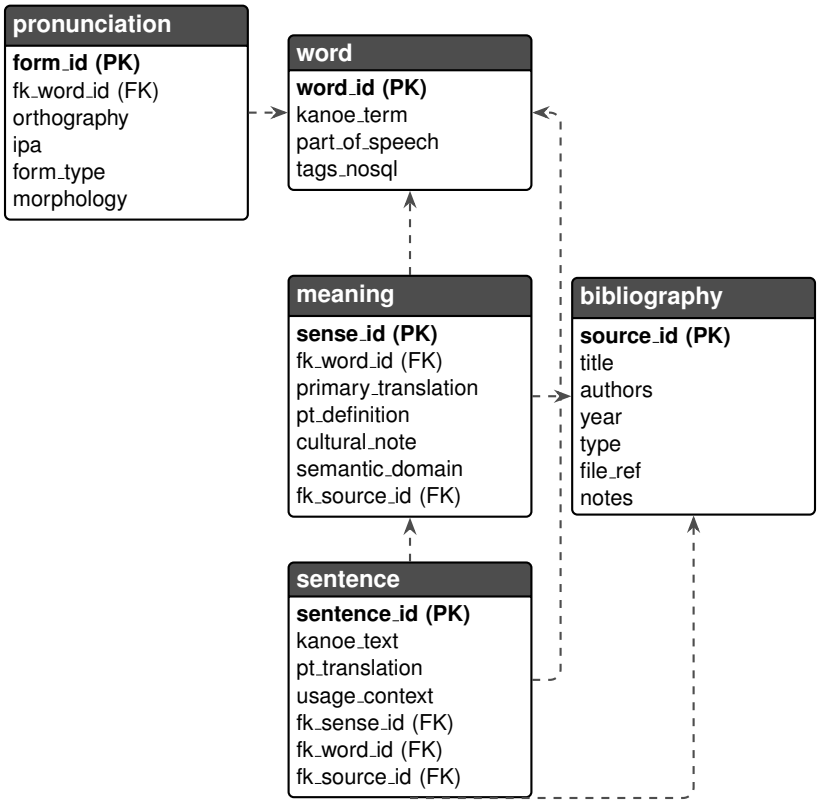

Figure 1. Entity-relationship diagram of the I Speak Kanoê dataset. Dashed arrows denote foreign-key references. Primary keys are shown in bold.

stores source metadata: title, authors, year, type (thesis, article, primer, report, other), file reference, and free-text annotations. All foreign keys enforce referential integrity at the database level.

Listing 1 illustrates how the schema supports typical lexicographic queries. The query retrieves, for a given Kanoê lexeme, its canonical form, IPA transcription, Portuguese gloss, and the bibliographic source from which the gloss was taken.

The output of Listing 2 is not an automatic resolution but a ranked list of philological conflicts in which each divergent IPA form is paired with its source. Curators adjudicate the conflict by consulting the original documents, and the decision is recorded either by adding a `tipo_forma` distinction (for example, marking one variant as "dialectal") or by preserving both readings with their respective bibliographic links, following

```
SELECT p.termo_kanoe, pr.ipa, s.traducao_primaria,
       b.autores || ' (' || b.ano || ')' AS source
FROM   palavra p
JOIN   pronuncia   pr ON pr.fk_id_palavra    = p.id_palavra
JOIN   significado s  ON s.fk_id_palavra     = p.id_palavra
LEFT JOIN bibliografia b ON b.id_fonte       = s.fk_id_bibliografia
WHERE  p.termo_kanoe = 'p  r  ';
```

Listing 1. Example SQL query joining four tables to retrieve enriched lexical entries.

```
WITH ipa_per_source AS (
  SELECT p.id_palavra,
         p.termo_kanoe,
         pr.ipa,
         b.autores || ' (' || b.ano || ')' AS source,
         b.tipo AS source_type
  FROM   palavra      p
  JOIN   pronuncia    pr ON pr.fk_id_palavra      = p.id_palavra
  JOIN   significado  s  ON s.fk_id_palavra       = p.id_palavra
  JOIN   bibliografia b  ON b.id_fonte            = s.
     fk_id_bibliografia
  WHERE  pr.ipa IS NOT NULL
)
SELECT termo_kanoe,
       COUNT(DISTINCT ipa)  AS n_variants,
       STRING_AGG(DISTINCT ipa || ' [' || source || ']', ' | '
                  ORDER BY ipa || ' [' || source || ']') AS divergences
FROM   ipa_per_source
GROUP  BY id_palavra, termo_kanoe
HAVING COUNT(DISTINCT ipa) > 1
ORDER  BY n_variants DESC, termo_kanoe;
```

Listing 2. Resolving cross-source disagreement: lexemes with divergent IPA transcriptions, showing each variant alongside its bibliographic source so that curators can adjudicate.

the transparency policy described in Section 3. This workflow turns the relational schema into an active instrument for philological work, rather than a passive archive.

4.2. Quantitative Characterization

Table 1 summarizes the size of each table and the coverage of key annotation fields. The dataset contains 1,684 lexemes, 1,937 pronunciation records (of which 90.8% carry an IPA transcription), 2,134 sense-level annotations, 612 example sentences, and 18 documentary sources. In total, the five tables hold 6,385 records connected by referential integrity constraints.

Table 1. Quantitative summary of the dataset

Table	Rows	Notable coverage
`palavra`	1,684	296 entries with explicit POS tag (17.6%)
`pronuncia`	1,937	1,759 with IPA (90.8%); 251 lexemes with variants
`significado`	2,134	709 with cultural note (33.2%); 252 polysemous lexemes
`frase`	612	165 with contextual note (27.0%); all source-linked
`bibliografia`	18	1992–2025; 6 primers, 6 reports, 4 articles, 2 theses

The distribution of example sentences by source type (Table 2) shows that the pedagogical primers contribute the majority of contextualized sentences (522, or 85.3%), reflecting the didactic orientation of the PRODOCLIN materials, while peer-reviewed articles contribute 78 sentences (12.7%) and ethnographic reports contribute 11 sentences (1.8%).

Table 2. Distribution of sentences by type of documentary source

Source type	Sentences	Share
Primer (cartilha)	522	85.3%
Article (artigo)	78	12.7%
Report (relatório)	11	1.8%
Thesis (tese)	1	0.2%
Total	612	100.0%

Among the 296 lexemes whose part-of-speech is explicitly tagged, nouns predominate (153 entries, 51.7%), followed by verbs (100 entries, 33.8%), pronouns (22 entries, 7.4%), adverbs (8 entries, 2.7%), and adjectives (5 entries, 1.7%), with a residual tail of conjunctions, numerals, and interjections. The remaining 1,388 lexemes (82.4%) await systematic POS tagging, which is listed as an open task. The skew toward nouns mirrors the composition of the pedagogical primers, which privilege concrete vocabulary (plants, animals, tools, kinship terms) over functional categories.

4.3. Missing Data and Transparency

The project adopts an explicit policy of recording missing data as SQL `NULL` rather than imputing values. This design choice reflects two considerations. First, in extremely low-resource documentation projects, imputed values can be indistinguishable from attested ones once data are downloaded and reused, which would undermine philological traceability.

Second, the distribution of missing values is itself informative: gaps in `ipa` (9.2%), `nota_cultural` (66.8%), and `dominio_semantico` (100% of the current snapshot) signal where future documentation effort would have the greatest marginal impact. A companion coverage view (`view_dicionario_completo`) is provided to facilitate inspection of these gaps by end users.

5. Availability, Applications, and Limitations

This section outlines the dissemination and potential impact of the dataset. It begins by detailing the release platform and licensing conditions, then explores use cases in computational linguistics alongside the dataset's current limitations. The section concludes with directions for future expansion.

5.1. Public Availability

The dataset is publicly released on Hugging Face under the identifier [Alves Gomes, G., De Sousa Aragão, I. 2025] under a persisted DOI. Distribution includes (i) a SQL dump of the PostgreSQL schema, (ii) CSV exports of each table, and (iii) a materialized view (`view_dicionario_completo`) joining all five tables into a single flat file suitable for quick inspection. The license is Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0), which permits academic and educational reuse under attribution, forbids commercial exploitation, and requires derivative works to adopt the same terms. These constraints align with both the non-commercial orientation of the primary sources and with CARE Principles concerning collective benefit and authority to control [Carroll et al. 2020].

5.2. Research Applications and Limitations

The resource supports several classes of research activity. In *linguistics*, the corpus enables quantitative analysis of Kanoê phonology (for example, the distribution of nasal vowels and palatal consonants documented in [Bacelar 1992]), morphosyntactic description based on attested forms [Bacelar and Pereira 2009], and typological comparison with other South American isolates.

In *natural language processing*, the dataset offers a ground truth for experiments in (i) language identification of Brazilian indigenous languages [Cavalin et al. 2023], (ii) bilingual lexicon induction and cross-lingual alignment in extremely low-resource settings, (iii) phonetic modeling of under-resourced inventories [Moran and McCloy 2019], and (iv) evaluation of multilingual language models on languages absent from their pretraining data [Mager et al. 2021, Pinhanez et al. 2024]. In *education*, the structured form makes it feasible to generate flashcards, spaced-repetition decks, and simple digital primers that could complement the existing paper materials. In *cultural documentation*, the provenance-aware schema supports philological work that requires the ability to trace every attestation back to its source.

Four limitations of the present release should be noted. First, the dataset is derived entirely from written documentary sources; no original fieldwork or audio recordings were collected by the project team, and spoken-language phenomena not already represented in the sources are absent. Second, part-of-speech coverage is low (17.6%), reflecting inconsistent POS annotation in the primary sources; systematic POS tagging is an open task. Third, the `dominio_semantico` field is currently empty across all rows, as automated semantic-domain assignment was deferred to avoid introducing uncurated labels. Fourth, the 612 example sentences are biased toward didactic contexts (primers) and are not representative of spontaneous Kanoê discourse. None of these limitations affects the integrity of the data actually included, and all are recorded transparently in the dataset documentation.

5.3. Future Work

Planned extensions include: (i) development of a semi-automated human-in-the-loop ingestion pipeline combining OCR, IPA-aware normalization rules, and interactive validation, to reduce the manual bottleneck identified during curation and to make the engineering pattern transferable to other endangered languages documented in legacy PDFs; (ii) systematic POS tagging and morphological segmentation populated into the `tags_nosql` and `morfologia` JSON fields; (iii) enrichment of the `dominio_semantico` field with a controlled vocabulary derived from existing thesauri for Amazonian languages; (iv) publication of a versioned data-release protocol aligned with Hugging Face dataset revisions so that experiments conducted on specific snapshots remain reproducible; (v) acquisition and integration of spontaneous speech data, subject to community consent, to counterbalance the didactic bias of the current sentence sample (85.3% drawn from pedagogical primers) and broaden the dataset's coverage of natural discourse phenomena.

6. Conclusion

This paper introduced *I Speak Kanoê*, a structured relational dataset engineered to consolidate 18 documentary sources on the Kanoê language into a normalized, publicly acces-

sible, and provenance-aware resource. The dataset contains 1,684 lexical entries, 1,937 pronunciation records with IPA coverage of 90.8%, 2,134 sense annotations, and 612 example sentences, all linked by referential integrity to the bibliographic sources from which they were drawn. The resource fills a documented gap in the landscape of digital materials for Brazilian indigenous languages and offers a replicable engineering pattern, combining manual extraction, orthographic normalization, and relational modeling, that can be applied to other endangered languages whose documentation is similarly dispersed across legacy PDFs.

The contribution is specifically one of data engineering and curation, rather than of primary linguistic research. Its value lies in transforming a fragmented body of existing documentation into a queryable resource that respects provenance, exposes its gaps transparently, and operates under a license consistent with the non-commercial orientation of the primary materials and with the CARE Principles for indigenous data governance. By lowering the barrier to computational work on Kanoê, the dataset aims to support the linguistic, pedagogical, and cultural initiatives of the Kanoê community and of the research communities that work alongside it.

Use of Artificial Intelligence

Generative AI tools were used in the preparation of this manuscript for the following purposes: (i) language revision and translation assistance between Portuguese and English; (ii) generation of the TikZ code used to render the entity-relationship diagram in Figure **??**; and (iii) formatting of LaTeX tables from tabular data. All linguistic data, analytical decisions, curatorial choices, references, and substantive claims were reviewed and validated by the authors. No primary data were generated by AI tools, and no citations were produced without verification against their original sources.

References

Alves Gomes, G., De Sousa Aragão, I. (2025). I speak kanoê dataset. Available at https://huggingface.co/datasets/carpenterbb/i-speak-kanoe.

Bacelar, L. N. (1992). *Fonologia preliminar da língua Kanoê*. Dissertação de mestrado, Universidade de Brasília.

Bacelar, L. N. (2004). *Gramática da língua Kanoê: descrição gramatical de uma língua isolada e ameaçada de extinção, falada no sul do Estado de Rondônia, Brasil*. Tese de doutorado, Radboud Universiteit Nijmegen.

Bacelar, L. N. (2010). Kapeãw Kanoé vararoé! "vamos falar Kanoé!" cartilha escolar da língua Kanoé, vol. I. PRODOCLIN/Museu do Índio.

Bacelar, L. N. (2012). DOCKANOÊ: Etnografia e documentação da língua Kanoê. Relatório PRODOCLIN/Museu do Índio.

Bacelar, L. N. (2013). Kapeãw Kanoé vararoé! "vamos falar Kanoé!" cartilha escolar da língua Kanoé, vol. II. PRODOCLIN/Museu do Índio.

Bacelar, L. N. (2014a). Cartilha médica Kanoé: Cartilha de apoio linguístico para assistência médico-odontológica. PRODOCLIN/Museu do Índio.

Bacelar, L. N. (2014b). Vamos ler e escrever em português. cartilha de alfabetização para a etnia Kanoé (vols. I–III). PRODOCLIN/Museu do Índio.

Bacelar, L. N. (2015). Vocabulário ilustrado Kanoê. PRODOCLIN/Museu do Índio.

Bacelar, L. N. and Canoê, J. A. (2011). Xamanismo e etnomedicina Kanoé. Relatório PRODOCLIN/Museu do Índio.

Bacelar, L. N., Canoê, J. A., and Silva, G. F. d. (2011). Indumentária, arte plumária e adereços Kanoé. Relatório PRODOCLIN/Museu do Índio.

Bacelar, L. N., Canoê, J. A., and Silva, G. F. d. (2012). Artesanato Kanoé. Relatório PRODOCLIN/Museu do Índio.

Bacelar, L. N. and Pereira, C. d. S. (2009). Aspectos morfossintáticos da língua Kanoê. *Revista de Estudos da Linguagem*.

Bacelar, L. N. and Silva Junior, A. R. d. (2003). Tipologia da negação em Kanoê. *Signoética*, 15(2):259–272.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Cavalin, P., Domingues, P., Nogima, J., and Pinhanez, C. (2023). Understanding native language identification for Brazilian indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 12–18, Toronto, Canada. Association for Computational Linguistics.

FUNAI (2022). Brasil registra 274 línguas indígenas diferentes faladas por 305 etnias. Available at `https://www.gov.br/funai/pt-br/assuntos/noticias/2022-02/`.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2545–2568.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6282–6293.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 55–69, Santa Fe, New Mexico, USA.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutiérrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop*

on Natural Language Processing for Indigenous Languages of the Americas, pages 202–217. Association for Computational Linguistics.

Marcon, I. N., Galucio, A. V., and Brito, S. (2023). Dicionário multimídia Kanoê-Português, versão 2.0. Museu Paraense Emílio Goeldi. Available at `http://dicionarios.museu-goeldi.br/dic/kanoe/`.

Moran, S. and McCloy, D. (2019). PHOIBLE 2.0. Available at `https://phoible.org/`.

Pinhanez, C., Cavalin, P., Storto, L., Finbow, T., Cobbinah, A., Nogima, J., Vasconcelos, M., Domingues, P., de Souza Mizukami, P., Grell, N., Gongora, M., and Gonçalves, I. (2024). Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: Technologies and experiences. *arXiv preprint arXiv:2407.12620*.

UNESCO (2024). Languages, cultures, knowledge: UNESCO's action for indigenous peoples. Available at `https://unesdoc.unesco.org/ark:/48223/pf0000392089`.

Unternbäumen, E. H., Aquino, L. d. S., and Melo, L. B. d. (2025). Metáforas na língua Kanoé. *LIAMES: Línguas Indígenas Americanas*, 25:e025001.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.