Unified Knowledge-Graph for Brazilian Indigenous Languages: An Educational Applications Perspective

Gustavo Polleti [1], **Fabio Cozman** [1], **Fabricio Gerardi** [2]

[1]Universidade de São Paulo, Brazil

[2]Universität Tübingen, Germany

gustavo.polleti@gmail.com

Abstract. *In this paper we present an unified knowledge-graph for Brazilian indigenous languages (BIL) from the perspective of potential applications, with a particular focus to the educational domain. We present BILGraph, a prototype we built for Bororo and Tupian languages, such as Guajajara, Munduruku and Akuntsu. Then we describe the knowledge extraction and entity linking process to build the graph from a dependency treebank and a lexical database for Tupian and Bororo languages. We discuss the limitations of BILGraph, highlighting ethical and practical implementation concerns.*

Resumo. *Este artigo apresenta um grafo de conhecimento unificado para as línguas indígenas brasileiras (BIL) a partir da perspectiva de aplicações potenciais, com foco particular no domínio educacional. Apresentamos o BILGraph, um protótipo construído para o Bororo e línguas tupis, como Guajajara, Munduruku e Akuntsu. Em seguida, descrevemos o processo de extração de conhecimento e ligação de entidades para construir o grafo a partir de um banco de árvores de dependências e de um banco de dados lexical para línguas Tupi e Bororo. Discutimos as limitações do BILGraph, destacando questões éticas e práticas de implementação.*

1. Introduction

The development of applications for Brazilian Indigenous languages (BIL) is severely limited by the lack of resources and tools. As is often the case with endangered languages, available resources are both scarce and dispersed [Pinhanez et al. 2023]. For some languages, such as Guajajara, Asurini, and Bororo, dictionaries are now available [Harrison and Harrison 2013, Cabral and Rodrigues 2003, Ferraz Gerardi]. For other languages, treebanks are available through the Universal Dependencies Project (UD) [Nivre et al. 2020a], though they vary in length and quality. Some languages, however, have only a handful of miscellaneous resources [Monserrat 2000]. This lack of standardization and proper linked data poses a significant barrier to developing tools and methods that could support language revitalization initiatives and accelerate the production of pedagogical material.

Recent efforts to unify Brazilian Indigenous language resources, such as TuLeD [Gerardi et al. 2022a] and the TuDeT treebanks on UD — a lexical database and a dependency treebank for several Tupian languages (still in their initial phase), respectively — have been pivotal in the development of language-learning applications targeted at Indigenous communities [Polleti 2024]. Additionally, the recent publication of

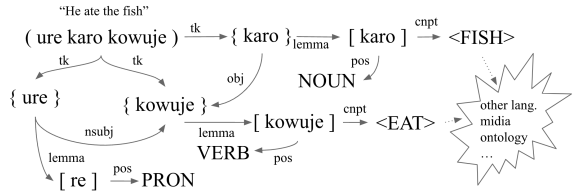

Figure 1. BILGraph toy example displaying a sampled subgraph associated with the sentence Bororo "ure karo kowuje", i.e. "He ate the fish".

the Bororo Corpus [Ferraz Gerardi et al. 2024], which is connected to the UD Treebank [Ferraz Gerardi 2024], has enabled the use of various computational tools to develop educational materials and other online resources; notably a language-learning app for the Bororo language.[1] UD-treebanks [Nivre et al. 2020b] are an important resource since the standardized type of annotation for all languages facilitate the development of new applications. On the other hand, heterogeneous and complex network structures, such as knowledge graphs, are known for their flexibility in incorporating linguistic characteristics [Cong and Liu 2014, Miller 1994] and can be effectively utilized to power sophisticated applications, including recommendation systems, information retrieval, and educational assistants.

In this work, we introduce a preliminary version of an unified knowledge graph for Brazilian indigenous languages, which we will refer as "BILGraph", and we describe its knowledge extraction pipeline. We developed a prototype for Tupian languages available in Tuled and Tudet [Gerardi et al. 2022b], and the Bororo language [Ferraz Gerardi et al. 2024]. We discuss the knowledge graph prototype with a focus on the potential applications. We managed to develop a natural language processing pipeline to build BILGraph that can handle semi-structured data from several sources, such as annotated phrases from treebanks and dictionaries. We discuss the pipeline challenges and limitations. The main contribution of this work is to present a prototype version of BILGraph as a case of study on building an unified knowledge graph for BIL. We hope the knowledge graph and the methods presented in this work can support the development of sophisticated applications.

The paper is organized as follows. Section 2 describes BILGraph's design and its development, including their data sources and knowledge extraction pipeline. Section 3 discusses the challenges and limitations of our prototype, analyses our processes and resources from both a practical implementation and potential applications perspective, and offers concluding remarks.

2. BilGraph: Linguistic Knowledge Graph

We have developed a knowledge extraction pipeline to structure and link language resources for Brazilian Indigenous Languages (BIL) available in Universal Dependencies (UD) treebanks and lexical databases, such as TuLeD and the Bororo dictionary. The result of this effort is "BILGraph", a knowledge graph for BIL that contains four principal types of nodes: (1) sentence, (2) token, (3) lemma, and (4) concept. Consider the example depicted in Figure 1. The sentence node represents the Bororo treebank phrase *ure karo*

[1] https://bilingo-4388e.web.app/

kowyje ‘He ate the fish’. This sentence node connects to its token nodes, which represent the individual words composing the sentence and their syntactic dependencies. In this example, “kowyje” is the root, with the object “karo” and the nominal subject “ure” linking to it. Each token node is connected to a single sentence node. Each token is further linked to a lemma node, which represents the word’s base form and its relationships to linguistic classes, including any applicable synonyms. Up to this point, the entities and relationships described are those typically found in dependency treebanks. However, the lexical database or dictionary adds another layer by linking lemma nodes to concept nodes. Concept nodes represent high-level abstractions that convey meaning across different languages and domains. In our example, the lemmas “karo” and “kowyje” are linked to the concepts “fish” and “eat,” respectively. The goal is to establish the concept nodes as a semantic layer that enables interoperability between the treebank sentences and other knowledge bases, such as ontologies, multimedia resources (e.g., phonetic or image databases), and other languages. Using BILGraph, one could easily search for sentences in other languages with similar structures or themes by fetching all sentence nodes connected to a given concept node. For example, a search engine could retrieve the Guajajara sentence *u?u ipiratetea?u* ‘It eats many fishes’ because it is connected to the concept node “FISH” as the similar sentence in Bororo *ure karo kowyje*. Note that the graph structure is flexible enough to encode N-N relationships between lemmas and concepts.

The relationships between sentences, tokens and lemmas can be extracted directly from UD treebanks, as the treebank sentences are annotated with attributes that allow a straightforward graph representation. To build BILGraph, the real challenge lies in linking lemma to concept nodes. In our preliminary version, we applied a simple entity linking process as follows. For each lemma, we generated a neighborhood set of similar words by changing and trimming characters based on rules. For example, in the Bororo language, we have different spellings where some words exchange “u” for “y”, and words like “boe” are often applied, so some of our neighborhood generation rules involved in adding or removing prefixes and changing exchangeable letters. The size of the neighborhood was defined considering a similarity threshold based on the Leveshnstein distance. Next, we select from all the vocabulary in our database the words that display high similarity, considering again a threshold based on leveshnstein distance, with at least one instance in our neighborhood. Finally, we test if dictionary entry or description for each candidate has at least one word in the sentence. So, for example, consider we are trying to link the lemma “karo”, from the sentence “He ate the fish”, to its appropriate concepts. Additionally, consider the dictionary description for a word candidate “kabo” is “a type of river fish”. In this case, we will establish the link due to the lexical similarity between “karo” and “kabo”, and due to the word “fish” that is present in both the dictionary entry and the sentence. Note that relying on lexical similarity may lead to innacuracies. For example, the Bororo words “apido” (palm heart) and “apodo” (toucan) have high lexical similarity while there meanings are not related at all. If a dictionary entry contains both words, such as “palm hearth, edible for many animals like toucans”, this would lead to incorrect links being added to the graph. BILGraph’s knowledge extraction pipeline code, with the used Leveshentein distance thresholds for each language, and the knowledge graph itself is available in Github.[2] We adopted the RDF format, where each edge in the graph is represented as a triple.

[2]https://github.com/gpadpoll/bilgraph

3. Discussion and Concluding Remarks

The preliminary version of BILGraph introduced in this work represents a significant step forward in advancing resources for Brazilian Indigenous languages. We envision that BILGraph could power typical applications such as information retrieval from texts written in these languages, with a particular emphasis on its educational potential. The process of creating educational resources often involves organizing texts based on their linguistic characteristics, themes, and complexity levels. For instance, one might search for specific sentences to teach someone how to ask for food. BILGraph simplifies this task by allowing queries for sentences linked to specific concept nodes. To find sentences that include food-related vocabulary, one can attach a generic ontology to BILGraph's concept nodes and search for sentences associated with food-related concepts. Moreover, BILGraph makes it easy to query sentences based on linguistic features, such as those using possessive pronouns, verb forms, plurals, adverbs, and more. We believe that BILGraph's ability to query and organize sentences can enhance the use of treebanks and other available BIL resources in the development of educational materials. By organizing resources in a standardized and unified format, we can develop applications that scale across multiple languages. For example, a query that searches for food-related concepts in sentences for one language can be reused for other languages included in BILGraph. We are already leveraging BILGraph to develop a curriculum for a Bororo language course, which will be released as a language-learning app. We aim to extend this approach to other languages as they are incorporated into the knowledge graph.

At this point, our BILGraph prototype falls short in several aspects and remains a work in progress, from the difficulties of working with limited sources of data to innacuracies and ethical concerns. BILGraph was built from TuLeD, TuDet and the Bororo treebank and dictionary. All these data sources were developed by compiling several sources from the literature, without a proper structured data gathering process. As a result, it suffers from incompleness, notably when we consider coverage of dependency trees with translation to portuguese. We only have have Portuguese translations for "Bororo", "Guajajara", "Munduruku" and "Akuntsu" out of the 9 languages available. The lack of Portuguese translations limits the application of these resources, as for educational purposes for example. Furthermore, it reasonable to expect that some innacuracies may have been introduced as part of the entity linking and knowledge extraction process. We haven't evaluated the correctness in a comprehensive manner yet, except for limited manual inspection by the researchers. Finally, it is worth mentioniong ethical concerns. BILGraph has been developed without the involvement of indigenous community [Pinhanez et al. 2023], except for the case of Bororo, so it is hard to enforce ethical guidelines [Lewis et al. 2020], as for example proposed by the Los Pinos Declaration,[3] before BILGraph can be properly inspected and validated by actual indigenous speakers.

We recognize a limitation in distinguishing similar forms that map to different lemmas. While various solutions exist, the most effective approach tend to be probabilistic, improving in accuracy with larger datasets. We also focus on further research in developing a pipeline which only uses the target language, without relying on the use af a dictionary. Overall, we hope BILGraph represents a positive step towards an unified source for BIL resources so that more tools and applications can be developed for them.

[3]https://unesdoc.unesco.org/ark:/48223/pf0000374030

Acknowledgements

The second author was partially supported by CNPq grant 305753/2022-3. We also thank support by CAPES -Finance Code 001. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and from the IBM Corporation.

References

Cabral, A. S. and Rodrigues, A. (2003). Dicionário da língua asurini do tocantins. *Belém-Pará: UFPA/IFNOPAP/UnB: IL/LALI*.

Cong, J. and Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598–618.

Ferraz Gerardi, F. *Bororo Dictionary*. Forthcoming. Available upon request.

Ferraz Gerardi, F. (2024). Universaldependencies/ud$_{b}ororo - bdt$.

Ferraz Gerardi, F. M., Sollberger, D., and Toribio Serrano, L. (2024). Corpus bororo (corbo) (v0.1.1).

Gerardi, F. F., Reichert, S., Aragon, C., Wientzek, T., List, J.-M., and Forkel, R. (2022a). *TuLeD. Tupían Lexical Database*. Zenodo.

Gerardi, F. F., Reichert, S., Aragon, C., Wientzek, T., List, J.-M., and Forkel, R. (2022b). TuLeD. Tupían Lexical Database (v0.12).

Harrison, C. and Harrison, C. (2013). *Dicionário Guajajara-Português*. SIL.

Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P.-L., Kealiikanakaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque, M., Mahelona, K., Moses, C., Nahuewai, I. I., Noe, K., Olson, D., Parker Jones, Ō., Running Wolf, C., Running Wolf, M., Silva, M., Fragnito, S., and Whaanga, H. (2020). Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis. English Language Version of "Ka?ina Hana ?Ōiwi a me ka Waihona ?Ike Hakuhia Pepa Kūlana" available at: https://spectrum.library.concordia.ca/id/eprint/990094/.

Miller, G. A. (1994). WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Monserrat, R. F. (2000). *Vocabulário Amondawa-Português, Vocabulário e frases em Arara e Português, Vocabulário Gavião-Português, Vocabulário e frases em Karipuna e Português, Vocabulário e frases em Makurap e Português, Vocabulário e frases em Suruí e Português, Pequeno dicionário em Tupari e Português*. Universidade do Caixas do Sul.

Nivre, J., Abrams, M., Agić, Z., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, A., Asahara, M., Ateyah, L., Attia, M., et al. (2020a). Universal dependencies v2: An evergrowing multilingual treebank collection. `https://universaldependencies.org/`. Accessed: 2024-08-27.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020b). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Pinhanez, C. S., Cavalin, P., Vasconcelos, M., and Nogima, J. (2023). Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In Elkind, E., editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Polleti, G. (2024). Building a language-learning game for Brazilian indigenous languages: A case study. Technical report, arXiv:2403.14515.