Journal of Open Humanities Data

KAHD: Katukinan-Arawan-Harakmbut Database (Pre-release)

RESEARCH PAPER

FABRÍCIO FERRAZ GERARDI
CAROLINA COELHO ARAGON
STANISLAV REICHERT

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Katukinan, Arawan, and Harakmbut are small language families spoken in south-western Amazonia. These families have received some attention, but there are no consistently transcribed and machine-readable datasets available for them. We address this lacuna by introducing the first publicly available linguistic dataset of Arawan languages as the first part of the Katukinan-Arawan-Harakmbut Database, created with the goal of providing and regularly updating a list of lexical items in a consistent transcription and with cognacy annotation. The database is being developed to be used in quantitative and genealogical investigations.

CORRESPONDING AUTHOR:
Fabrício Ferraz Gerardi
Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, DM
fabricio.gerardi@uni-tuebingen.de

KEYWORDS:
Arawan languages; Amazonian languages; lexical database; historical linguistics; computational linguistics; language documentation

TO CITE THIS ARTICLE:
Gerardi, F. F., Aragon, C. C., & Reichert, S. (2022). KAHD: Katukinan-Arawan-Harakmbut Database (Pre-release). *Journal of Open Humanities Data*, 8: 18, pp. 1–11. DOI: https://doi.org/10.5334/johd.80

1 OVERVIEW

One of the most prominent trends in the linguistics of the 21st century is the unparalleled growth of machine-readable resources. While some legacy databases are steadily being converted into standardized formats like CLDF (Forkel & List, 2020), more and more new datasets get published every year. These datasets, while often including well-described languages, increasingly add languages that have so far enjoyed little to no scholarly attention, or have not been aggregated and made publicly accessible for various reasons (Dellert et al., 2020; Dellert, Daneyko, & Münch, 2019; Kassian, 2020). An example of the latter is TuLeD (Tupían Lexical Database) (Gerardi, Reichert, Aragon, List, & Wientzek, 2021; Gerardi, Reichert, & Aragon, 2021) which grew out of a wide variety of sources on Tupían languages (living and extinct) and was subsequently used in a phylogenetic classification of the Tupí-Guaraní branch (Gerardi & Reichert, 2021).

However, there is a clear need for further resources that would ideally capture even more of the linguistic and cultural diversity of South America. Our overarching goal is not only to continue providing sources to spread the knowledge on Amazonian languages and thus broaden our understanding of linguistic typology, but also to do so in a way that would enable us to empirically test some of the hypotheses put forth in the research literature. One such hypothesis suggests that two of these language families (Katukinan and Harakmbut) are genetically related (Adelaar, 2000, 2007). We also conjecture a macrofamily which adds the Arawan family to Katukinan[1] and Harakmbut, following a proposal by dos Anjos (2011); Jolkesky (2016). For these reasons we are working on the Katukinan-Arawan-Harakmbut Database (KAHD) by aggregating the published sources and making sure the data is consistently transcribed, aligned, and enriched with information on cognacy. Far from being a purely lexical database, KAHD is planned to encompass phonetic-phonological and morphological information as well.

At present, the size of the database can neither support nor refute the genetic relationship between these languages. Our goal in this paper is thus to introduce the database as an instrument which could, among other things, be employed in attempts to answer this question of genetic relatedness. The quantitative methods presented in Section 2 are intended to demonstrate the current status of the database.

1.1 THE ARAWAN LANGUAGE FAMILY

The Arawan family is roughly known since 1891, when Brinton recognized similarities between Arawá and Paumari, and consists of six languages:[2] Paumari, Madi (and its dialects Jarawara, Jamamadí, and Banawá) (see Dixon 2004), Sorowaha, Deni, Kulina, and the extinct Arawá (Ehrenreich, 1897). The number of speakers varies, as well as their social vulnerability, and consequently the status of their language: vigorous for Sorowaha with less than 200 speakers, but threatened for Kulina, with 2500 speakers.

Most of the Arawan speakers were contacted during the end of the nineteenth century and some of them, as is the case of the Sorowaha, escaped from the intensive Indigenous territorial invasion process which took place in the middle Purus River, and (they) still live as a recently contacted group (Aparicio, 2015; Huber, 2012). Others remain isolated like the groups who live in the Hi-Merimã Indigenous Area in the middle Purus (Shiratori, Cangussu, & Furquim, 2021). The Table 1 presents information on ethnic population, speakers and status of the language according to (Eberhard, Simons, & Fennig, 2021) which stem from a source dating back to 2012, and the Figure 1 shows the location of the languages. These numbers do not necessarily reflect the current situation, but they offer a general picture of the state of the Arawan language family. The ethnic population for Kulina, for example, differs significantly from that given by Dienst 2014 (5500 in Brazil and 600 in Peru), while a source from 2015 cites a comparable figure for the Sorowaha (Aparicio, 2015). In the lack of official or more precise and recent sources, Ethnologue (Eberhard et al., 2021) seems to be the most reliable source to quote.

1 There is a Panoan language called Katukina (Glottocode pano1254, ISO knt), which bears no relation to the Katukinan family in spite of the homonym. For the ethnonym Katukina, see Carvalho (2019).

2 Perhaps seven, if the language of the discovered isolated group Hi-Merimã is a still unknown Arawan language, or it is a dialect of an already known Arawan language.

DOCULECT	VARIETY	ISO	GLOTTOCODE	ETHNIC POPULATION	SPEAKERS	STATUS
Arawá		aru	arua1263	0	0	Extinct
Dení		dny	deni1241	880	740	Developing
Madí	Banawá	Jaa	bana1307	(780)	100	Educational
	Jamamadí		jama1261	780	450	Educational
	Jarawara		jara1276	(780)	230	Educational
Kulina		cul	culi1244	3500	3000	Threatened
Paumari		pad	paum1247	890	290	Moribund
Sorowaha		swx	suru1263	140	140	Vigorous

Table 1 The Arawan languages in KAHD. Information on ethnic population, speakers and status taken from Eberhard et al. (2021).

The Arawan communities are located in Brazil, in the south-western Amazonia, except for the Kulina speakers, who live near the Peruvian border (Ucayali). The Purus basin and the Juruá river are the historical seats of the Arawan groups. Their presence on the margins of the Purus and Juruá rivers, especially in the middle course of the Purus (which extends from the surroundings of the Acre River to the surroundings of the city of Tapauá, between the Acre River and the Tauamirim stream), was marked by the continuous exploration of rubber and the presence of proselytizing missionaries (Aparício, 2019). Only after the 1990s, their territories have started to be delimited and recognized by the Brazilian authorities (Aparício, 2011), although not soon enough to avoid the devastating effects of genocide and epidemics which happened since the rubber extraction had been introduced in the Purus (Kroemer, 1985).

The Arawá, a group whose name is now used for the Arawan language family, are a case in point. Their presence on the Juruá River was first signaled by Castelnau (1851, 87). The tribe was reported to have been exterminated by an epidemic of measles, introduced by the first migration of people from the north-eastern state of Ceará on the east coast of Brazil which was caused by the drought of 1877. The few survivors sought refuge with the Kulina, speakers of a language from the same family, who are said to have massacred them (Rivet & Tastevin, 1938, 72). Little is known about Arawá language and it is possible that the remnants of the group were incorporated into the Kulina, whose language they may have influenced.

1.2 THE HARAKMBUT-KATUKINAN LANGUAGE FAMILY

Harakmbut is spoken along the Madre de Dios River and its upper tributaries in Peru. There are several dialects which fall into two large clusters (Helberg Chávez, 1984, xv,50) (Helberg Chávez & Solís Fonseca, 1990, 227–228). Toyoeri and Huachipaeri form one cluster, while the other is formed by Sapiteri, Arasaeri and Amarakaeri, which is the best known and has the largest number of speakers (see also van Linden (2022)). It was initially classified as belonging to the Arawak family (Matteson, 1972; McQuown, 1955), but more recently, based on lexical evidence, Adelaar (2000, 2007) has proposed that it is genetically related to the Brazilian Katukina family. Wise (1999) seems to consider it an isolate.

The Katukinan family is known thanks to the work of Tastevin (1920) (see also Rivet 1920; Rivet and Tastevin 1921, 1923) and Natterer (1817–1835). Rodrigues takes it for granted (Rodrigues 1986, 79–81). It was Adelaar (2000) who first proposed the link between Harakmbut and the Katukinan languages, which has not been challenged and seems to be widely accepted by now. Of the two dialects of Kanamari, Katukina is probably the only surviving of the family, since Katawixi was already said to have disappeared in 1926 (dos Anjos, 2011, 16–17). Table 2 offers a brief overview over the current situation of these two language families.

2 METHOD

In the era of rapidly growing number of linguistic resources, arriving at comparable results in cross-linguistic research entails working with comparable datasets and standardized sets of tools and specifications. Despite the proliferation of datasets, they often fail to conform to the data FAIRness (Findable, Accessible, Interoperable, Reproducible) principles outlined in Wilkinson et al. (2016) and may require laborous and costly preprocessing before any analysis can take place. In order to address this need, we decided to follow the standards of the CLDF

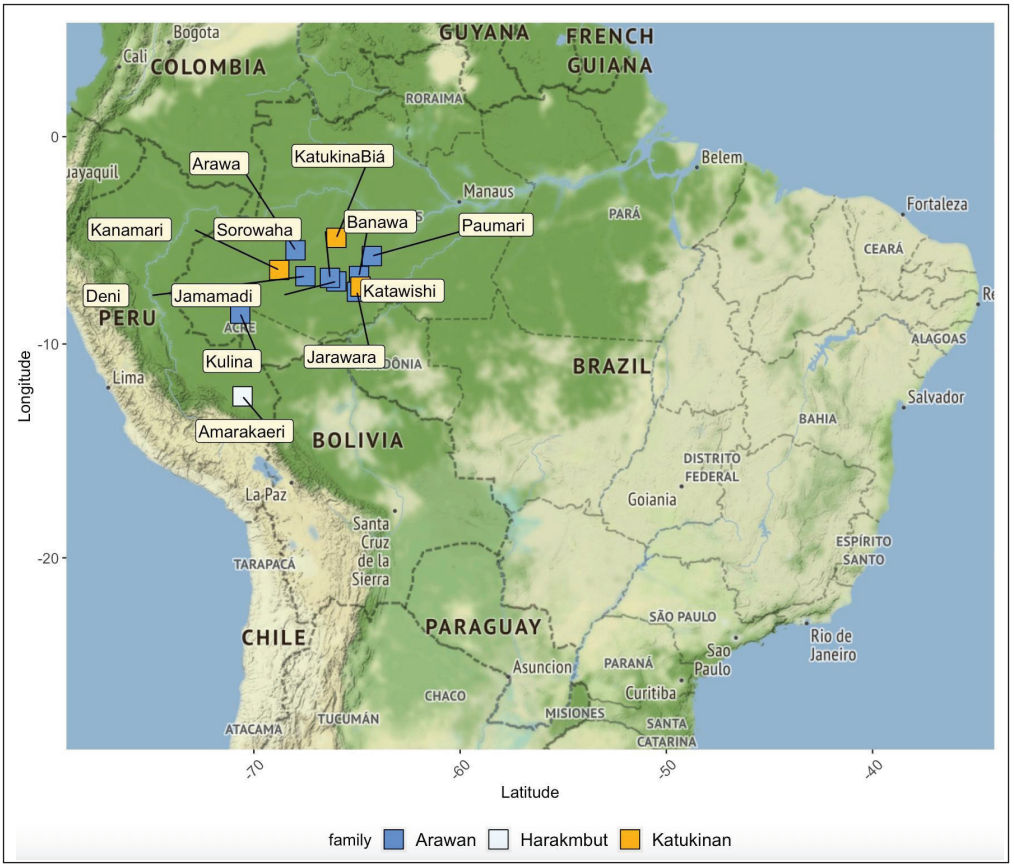

Figure 1 Location of the Arawan languages according to Hammarström et al. (2021).

DOCULECT	VARIETY	ISO	GLOTTOCODE	ETHNIC POPULATION	SPEAKERS	STATUS
Harakmbut		aru	hara1260	2090	1910	Threatened
Katukina						
	Kanamari	knm	cuti1242	?	1700	Vigorous
	Katukina Biá	knm	katu1276	?	550	Vigorous

Table 2 The Harakmbut and Katukinan languages in KAHD. Information on ethnic population, speakers and status taken from dos Anjos (2011); Eberhard et al. (2021).

(Cross-Linguistic Data Formats) initiative that enjoys growing popularity in the (computational) linguistic community (Forkel et al., 2018). CLDF offers ways to ensure the integrity of the data, its connection to the major reference catalogs like Glottolog (Hammarström et al., 2021) and Concepticon (List et al., 2022), as well as scripts written specifically for (historical) linguists to get the most out of their data. CLDF works with simple text formats that can be read and modified in any environment and allows for automatic validation of datasets against the specifications. Additionally, projects based on CLDF specification, like CLICS (Database of Cross-Linguistic Colexifications) (List et al., 2018) or the CLTS (Cross-Linguistic Transcription Systems) initiative, which endorses the use of unified phonetically transcribed forms (Anderson et al., 2018; List, Anderson, Tresoldi, & Forkel, 2021), constantly add new ways to explore available data and increase its cross-linguisitic interoperability. This framework alongside its tools as well as the agreed upon workflow was used in preparation of the TuLeD dataset (Gerardi, Reichert, & Aragon, 2021).

Similarly, in the case of the Arawan dataset, the data harvested from numerous sources is being curated and expanded using the Javascript graphical application EDICTOR (List 2017, 2021) from where it can be easily exported in *csv* format and used for further processing with various modules within LingPy, a state-of-the-art computational suite of computational tools for historical linguistics (List & Forkel, 2021, July 29).

The pre-release version of the dataset, which this paper describes, consists of 8 doculects (Good & Cysouw, 2013) and 556 concepts across 2503 forms.[3] The lexical coverage for each language in the dataset is given in Table 3.

3 The list of all the concepts and sources can be viewed in our GitHub repository at: https://github.com/LanguageStructure/ KAHD_pre_release.

DOCULECT	LEXICAL COVERAGE
Amarakaeri	51
Arawa	36
Banawa	309
Deni	400
Jamamadi	294
Jarawara	419
Kanamari	37
Katawishi	56
KatukinaBiá	18
Kulina	405
Paumari	268
Proto-Arawan	386
Sorowaha	423

Table 3 Lexical coverage for each language in the database.

The choice of concepts respects the established lists like Swadesh (1955, 2017) as well the Leipzig-Jakarta list (Tadmor, Haspelmath, & Taylor, 2010), but also adds multiple concepts whose inclusion is motivated by their cultural prominence in the (daily) life of the native speakers.[4] These concepts cover a variety of semantic domains: food and drink, kinship, the physical world, agriculture and vegetation, basic actions and technology, emotions and values, as well as fauna and flora, among others. The specifics for each concept, including semantic domains, except for some fauna and flora items can be accessed on Concepticon (List et al., 2022), since the names of concepts in our database are based on this source.

Cognacy was at first obtained through the five methods for automatic cognate detection implemented in LingPy and discussed in List, Greenhill, and Gray (2017) using the default parameters with the number of permutations set to 10,000 for each method, thus closely following the workflow of the original paper. The B-Cubed scores used for evaluation of each analysis are given in Table 4.

METHOD	PRECISION	RECALL	F-SCORE
SCA	0.952	0.963	0.944
LexStat	**0.972**	**0.931**	**0.951**
InfoMap	0.960	0.942	0.951
EditDistance	0.973	0.884	0.926
Turchin	0.985	0.810	0.889

Table 4 Comparison of tests using B-Cubed scores.

Initially, we relied on the LexStat method because of how it performed (see Table 4) in cognate assignment and subsequently manually improved the results using expert judgment. This did not lead to any significant improvement, because the family appears to be quite shallow, as indicated by the low number of cognate diversity of *cogids*: 0.169 and for *cogid*: 0.186 (see List et al. (2017) for cognacy diversity in other families). Even though we have assigned *cogids* for partial cognacy and added morpheme glosses, partial cognacy will only be thoroughly addressed in the next release. This means that morphological segmentation will be made available as well.

LingPy also implements an alignment algorithm which was used for this pre-release version of the dataset.[5] It should be noted that the resulting alignments have not been manually checked and no changes have been added to the output of LingPy. An example of the alignment for the concept "shoot with blow-gun" is given in Figure 2.

4 The CLLD web-application which will make the concepts available upon the release of version 1.0, will also contain links to Concepticon List et al. (2022) for each concept in the database and provide their respective semantic field.

5 We refer to the ksl.html file in our GitHub repository for the list of alignments of each cognate class.

Concept: *(shoot with) blow-gun* (ID: 4)			
CogID	**Language**	**Entry**	**Aligned Entry**
104	Banawa	karabowa	k a r a b o w a
104	Deni	karibaha	k a r i b a h a
104	Jamamadi	karabowa	k a r a b o w a
104	Jarawara	karabowa	k a r a b o w a
104	Kulina	karibehe	k a r i b e h e
104	Sorowaha	karokoba	k a r o k o b a

Figure 2 Example of alignment of from the KAHD Database.

We have further computed maximal mutual coverage[6] for all doculects in the dataset. The result is 6 doculects with an average mutual coverage of 219.

We have conducted a simple attempt of classification in order to compare the results with the proposed classification of Arawan by Dienst (2008), shown in Figure 3. We are not proposing a classification, but testing the validity of the automated cognacy against an already existing classification.

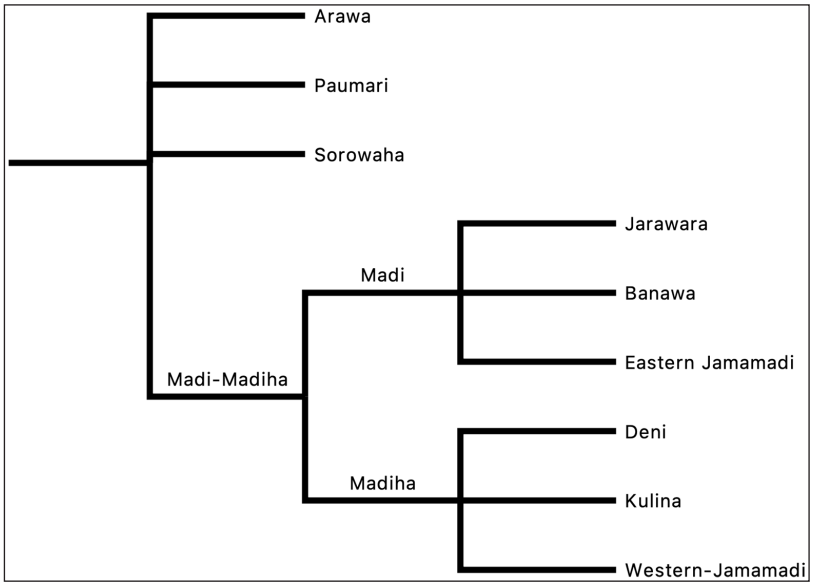

Figure 3 Classification of Arawan from Dienst (2008).

We obtained a similar classification using our cognates as input to the UPGMA algorithm (Sokal & Michener, 1958). The result of this classification, an unrooted tree, is given in Figure 4.

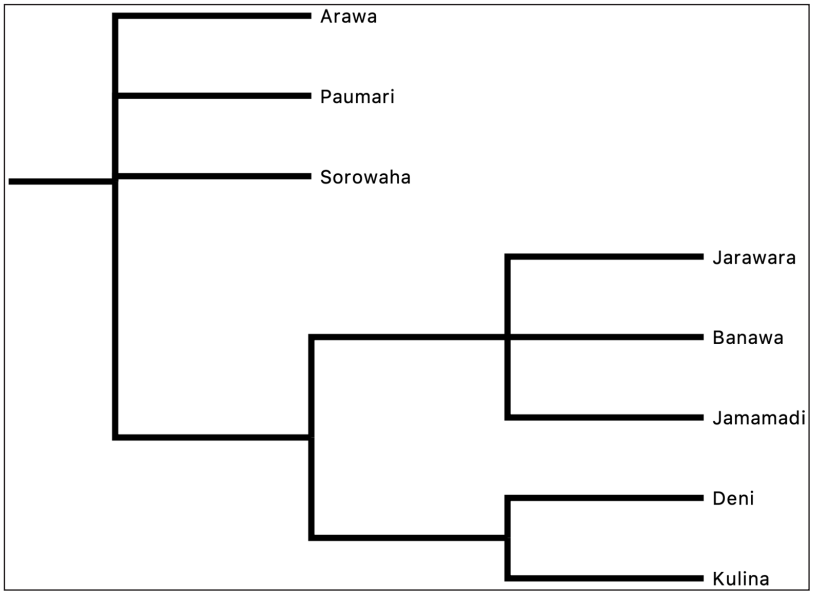

Figure 4 UPGMA classification of Arawan from KAHD data.

6 The number of doculects for which the coverage could be found as well as a list of all pairings in which this coverage is possible.

3 RESULTS AND DISCUSSION

Despite the various ways of hosting scientific datasets on the web, the process of data validation and curation may require considerable time and cost investment alongside technical skills and acumen. An additional consideration is the need to increase interoperability between datasets for typological and phylogenetic analyses, among others. In the case of South American language families, having freely available data in standardized transcription and enriched with information on linguistic features like cognacy would bring together the many valuable contributions from ethnographers and linguists alike. We believe that the next crucial step can be made much easier by using the toolset built around the CLDF datasets. The effort involved in checking the data's integrity is minimal and the steadily growing number of datasets published in adherence to these standards attests to its robustness and utility for (primarily) linguistic purposes. In making our database open-access, we rely on the *cldfbench* framework that greatly reduces the cost of the FAIR data curation by providing ways to read, write, and validate standardized CLDF datasets (Forkel & List, 2020).

This pre-release version is not yet hosted in the CLLD web-application despite being publicly available. The official release is planned to include a suitable graphical user interface, but the dataset can be accessed in its entirety via a permanent link in the EDICTOR which offers various search and analysis tools (List, 2017) as well as an option to download the full dataset.[7]

With the publication of the pre-release, we now begin to focus on the primary official release (version 1.0) which will contain enough data in all three families with cognacy assignment to preliminary test an interesting hypothesis regarding the relation between these families (Adelaar, 2000, 2007, Jolkesky, 2011; 2016). The inclusion of morphological items will provide valuable insights for comparison and allow for better typological description of languages, for which few resources are available.

We submitted our dataset to Zenodo[8] for archiving.

4 IMPLICATIONS/APPLICATIONS

The last decades have witnessed a growing amount of phylogenetic classifications of language families thanks to the use of lexical databases with cognacy assignment (Heggarty, 2021; Kolipakam et al., 2018; Sagart et al., 2019; Walworth, 2017; Zhang, Yan, Pan, & Jin, 2019). Such databases, beside elucidating the internal classification of language families, play a role in the understanding of displacement and linguistic contact, for example, through borrowing. Words of a language are valuable for understanding the culture where it is spoken (Harrison, 2008), even more so when the whole family is considered. In addition, culturally relevant lexical items offer us insights into possible genetic relations between individual languages, and it is even possible to putatively reconstruct items that were part of a proto-culture (Corrêa-da Silva, 2013; Rodrigues, 2010).

Apart from its value for (computational) historical linguistics mentioned in the previous section, the KAHD database also serves as language documentation and preservation effort for Amazonian language families since, as shown in Section 1.1, the number of speakers for some of the languages is diminishing at a fast rate (see e.g. D'Ávila 2019). Lehmann (2001, 5) affirms that the primary purpose of language documentation is to "represent the language for those who do not have direct access to the language itself." KAHD strives to achieve this goal by collecting primary data and making it publicly available after careful pre-processing, e.g. by performing cognacy judgment. Aside from language documentation (Romaine, 2015), the preparation of the Arawan dataset reveals the vast amount of work which is still to be done. The relative scarcity of published linguistic research on this language family underscores the necessity for a project like KAHD that would become the central hub for collaboration and

7 The following link leads to the complete dataset as seen in EDICTOR: https://lingulist.de/edictor/?file=arawa&remote_dbase=arawa&publish=true&preview=500&css=menu:hide|textfields:hide&columns=DOCULECT|CONCEPT|FORM|TOKENS|COGID|COGIDS|MORPHEMES|ALIGNMENT|NOTES&basics=DOCULECT|CONCEPT|FORM|TOKENS|COGID|COGIDS|MORPHEMES|ALIGNMENT|NOTES.

8 https://zenodo.org.

research into the lexical richness of these three underdescribed language families. Access to further sources to include in the dataset is essential in substantiating any theories on these language families.

An important future direction of the project is its use as a source for creating learning materials for the Indigenous communities, helping them raise their language vitality and providing an authentic context for the language acquisition. Dictionaries, for instance, are one type of pedagogical materials whose compilation could be made easier and more cost efficient by relying on a database like KAHD. An obvious advantage of an online database is the quick and effortless addition of new concepts and words. Thus, KAHD is being prepared with an eye toward wedding technology with ongoing language revitalization efforts. Moreover, as with KAHD's precursor TuLeD, we intend to actively involve community members in shaping KAHD into a useful and free tool for a variety of purposes starting with the preparation of educational resources locally. We welcome any kind of contributions to the project.

SUPPLEMENTARY FILES

All data relevant to the creation of this pre-release version of the Arawan dataset can be accessed and downloaded from our GitHub repository (https://github.com/LanguageStructure/KAHD_pre_release). All output files produced by running LingPy scripts are uploaded into the folder LingPy.

ACKNOWLEDGEMENTS

We would like to thank the following people: Dr. Johann-Mattis List (Max Planck Institute) for the Python scripts and for his assistance with curating the data; Tatiana Merzhevich (Universität Tübingen) for helping with the data-collection and adapting the scripts for the purposes of the present dataset.

FUNDING INFORMATION

The research presented in this paper is supported by the by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834050).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

FFG: conceptualization, data curation, formal analysis, methodology, project administration, writing (original draft).

CCA: data curation, investigation, methodology, validation, writing (original draft).

SR: data curation, investigation, methodology, validation, writing (original draft, review, and editing).

AUTHOR AFFILIATIONS

Fabrício Ferraz Gerardi orcid.org/0000-0002-1438-7336
Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, DE

Carolina Coelho Aragon orcid.org/0000-0001-9459-9939
Departamento de Língua Portuguesa e Linguística, Universidade Federal da Paraíba, João Pessoa, BR

Stanislav Reichert orcid.org/0000-0002-8330-1954
Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, DE

REFERENCES

Adelaar, W. F. (2000). Propuesta de un nuevo vínculo genético entre dos grupos lingüísticos indígenas de la Amazonía occidental: Harakmbut y Katukina. In *Actas del i congreso de lenguas indígenas de sudamérica, 2*, 219–236.

Adelaar, W. F. (2007). Ensayo de clasificación del Katawixí dentro del conjunto Harakmbut-Katukina. In *Lenguas indüígenas de américa del sur: Estudios descriptivo-tipológicos y sus contribuciones para la lingüística teórica* (pp. 159–169). Universidad Católica Andrés Bello Caracas.

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., & **List, J.-M.** (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4(1), 21–53. DOI: https://doi.org/10.2478/yplm-2018-0002

Aparício, M. (2011). Panorama contemporâneo do Purus indígena. *Álbum Purus*. Manaus: EDUA, 113–130.

Aparicio, M. (2015). *Presas del veneno: Cosmopolítica y transformaciones Suruwaha (Amazonía occidental)*. Editorial Abya-Yala.

Aparício, M. (2019). A relação banawá. *Socialidade e transformação nos Arawá do Purus*.

Carvalho, F. (2019). On the etymology of the ethnonym Katukina. *Revista Brasileira de Línguas Indígenas, 2*(1), 05–16. DOI: https://doi.org/10.18468/rbli.2019v2n1.p05-16

Castelnau, F. d. (1851). *Expédition dans les parties centrales de l'Amérique du Sud, de Rio de Janeiro à Lima, et de Lima au Para, exécutée par ordre du gouvernement français pendant les années 1843 a 1847* (Vol. 5). Paris: Bertrand. DOI: https://doi.org/10.5962/bhl.title.61493

Corrêa-da Silva, B. C. (2013). O mundo a partir do léxico: reconstruindo a realidade social Mawé-Awetí-Tupí-Guaraní. *Revista Brasileira de Linguística Antropológica, 5*(2), 385–400. DOI: https://doi.org/10.26512/rbla.v5i2.16271

Dellert, J., Daneyko, T., & **Münch, A.** (2019). NorthEuraLex 0.9 [dataset]. *Lang Resources and Evaluation*. DOI: https://doi.org/10.1007/s10579-019-09480-6

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., … **others.** (2020). NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation, 54*(1), 273–301. DOI: https://doi.org/10.1007/s10579-019-09480-6

Dienst, S. (2008). The internal classification of the Arawan languages. *LIAMES: Línguas Indígenas Americanas, 8*(1), 61–67. DOI: https://doi.org/10.20396/liames.v0i8.1471

Dienst, S. (2014). *A grammar of Kulina*. De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110341911

Dixon, R. M. (2004). Proto-Arawá phonology. *Anthropological Linguistics, 46*(1), 1–83.

dos Anjos, Z. (2011). *Fonologia e gramática Katukina-Kanamari* (Unpublished doctoral dissertation). Vrije Universiteit Amsterdam.

D'Ávila, A. (2019). Estratégias de resistência e a língua Paumari: uma breve reflexão glotopolítica. *Revista Versalete, 7*(13), 74–92.

Eberhard, D. M., Simons, G. F., & **Fennig, C. D.** (2021). *Ethnologue: Languages of the world. Twentyfourth edition* (Vol. 16). Dallas, TX: SIL international. Retrieved from http://www.ethnologue.com

Ehrenreich, P. (1897). Materialien zur sprachenkunde Brasiliens. Vokabulare von Purus-Stämmen. *Zeitschrift für Ethnologie, 29*, 59–71.

Forkel, R., & **List, J.-M.** (2020). CLDFBench: Give your cross-linguistic data a lift. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6995–7002). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.864

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., … **Gray, R. D.** (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data, 5*(1), 1–10. DOI: https://doi.org/10.1038/sdata.2018.205

Gerardi, F. F., & **Reichert, S.** (2021). The Tupí-Guaraní language family: A phylogenetic classification. *Diachronica, 38*(2), 151–188. DOI: 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, 55*(4), 997–1015. DOI: https://doi.org/10.1007/s10579-020-09521-5

Gerardi, F. F., Reichert, S., Aragon, C., List, J.-M., & **Wientzek, T.** (2021). TuLeD: Tupían lexical database, version 0.11. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.4629306

Good, J., & **Cysouw, M.** (2013). Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language documentation & conservation, 7*, 331–359.

Hammarström, H., Forkel, R., Haspelmath, M., & **Bank, S.** (2021). *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: https://doi.org/10.5281/zenodo.5772642

Harrison, K. D. (2008). *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

Heggarty, P. (2021). Cognacy databases and phylogenetic research on Indo-European. *Annual Review of Linguistics, 7*, 371–394. DOI: https://doi.org/10.1146/annurev-linguistics-011619-030507

Helberg Chávez, H. A. (1984). *Skizze einer Grammatik des Amarakaeri* (Unpublished doctoral dissertation). Eberhard-Karls-Universität.

Helberg Chávez, H. A., & **Solís Fonseca, G.** (1990). Analisis funcional del verbo amarakaeri. In *Temas de lingüística amerindia (primer congreso nacional de investigaciones lingüístico-filolíogicas)* (pp. 227–249).

Huber, A. (2012). *Pessoas falantes, espíritos cantores, almas-trovões. história, sociedade, xamanismo e rituais de auto-envenenamento entre os Suruwaha da Amazônia ocidental* (Unpublished doctoral dissertation). Universität Bern.

Jolkesky, M. (2011). *Arawá-Katukína-Harakmbet: correspondências fonológicas, morfológicas e lexicais.* Unpublished. (Encontro Internacional: Arqueologia e Linguística Histórica das Línguas Indígenas Sul-Americanas Brasília).

Jolkesky, M. (2016). *Estudo arqueo-ecolinguístico das terras tropicais Sul-Americanas* (Unpublished doctoral dissertation). Universidade de Brasília.

Kassian, A. (Ed.). (2020). *Moscow lexical database [online]*. Accessed: 20.05.2022. http://moslex.org/

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., & **Verkerk, A.** (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society open science, 5*(3), 171504. DOI: https://doi.org/10.1098/rsos.171504

Kroemer, G. (1985). *Cuxiuara: o Purus dos indígenas.* São Paulo: Edições Loyola.

Lehmann, C. (2001). Language documentation. a program. In W. Bisang (Ed.), *Aspects of typology and universals* (pp. 83–99). Berlin: Akademie Verlag.

List, J.-M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the software demonstrations of the 15th conference of the european chapter of the association for computational linguistics* (pp. 9–12). DOI: https://doi.org/10.18653/v1/E17-3003

List, J.-M. (2021). Using EDICTOR 2.0 to annotate language-internal cognates in a German wordlist. *Computer-Assisted Language Comparison in Practice, 4*(4), 1–7.

List, J.-M., Anderson, C., Tresoldi, T., & **Forkel, R.** (2021). Cross-Linguistic Transcription Systems (version v2.1.0). *Data set*. DOI: http://doi.org/10.5281/zenodo.4705149

List, J.-M., & **Forkel, R.** (2021, July 29). LingPy. a Python library for historical linguistics, version 2.6.9. *Zenodo*. https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy

List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., & **Forkel, R.** (2018). CLICS2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, *22*(2), 277–306. DOI: https://doi.org/10.1515/lingty-2018-0010

List, J.-M., Greenhill, S. J., & **Gray, R. D.** (2017). The potential of automatic word comparison for historical linguistics. *PloS one, 12*(1), e0170046. DOI: https://doi.org/10.1371/journal.pone.0170046

List, J. M., Tjuka, A., Rzymski, C., Greenhill, S., Schweikhard, N., & **Forkel, R.** (Eds.) (2022). *Concepticon 2.6.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: https://doi.org/10.5281/zenodo.6560398

Matteson, E. (1972). Proto Arawakan. In A. Wheeler, F. L. Jackson, N. R. Waltz, & D. R. Christian (Eds.), *Comparative studies in amerindian languages* (pp. 160–242). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110815009.160

McQuown, N. A. (1955). The indigenous languages of Latin America. *American Anthropologist, 57*(3), 501–570. Retrieved from http://www.jstor.org/stable/665445. DOI: https://doi.org/10.1525/aa.1955.57.3.02a00080

Natterer, J. (1817–1835). *Unpublished manuscripts*. Unpublished word lists (Sprachproben), literary estate of Johann Jakob Tschudi, Basel, University Library, Manuscript T.2.b.19.

Rivet, P. (1920). Les Katukina.étude linguistique. *Journal de la Société des Américanistes, 12*, 83–89. DOI: https://doi.org/10.3406/jsa.1920.2884

Rivet, P., & **Tastevin, C.** (1921). Les tribus indiennes des bassins du Purús, du Juruá et des régions limitrophes. *La Géographie, 35*(5), 449–482.

Rivet, P., & **Tastevin, C.** (1923). Les langues du Purús, du Juruá et des régions limitrophes. i° le groupe arawak pré-andin (fin). *Anthropos*(H. 1./3), 104–113.

Rivet, P., & **Tastevin, C.** (1938). Les langues arawak du Purus et du Juruá (groupe arauá). *Journal de la Société des Américanistes, 30*(1), 71–114. DOI: https://doi.org/10.3406/jsa.1938.1966

Rodrigues, A. D. (1986). *Línguas brasileira: para o conhecimento das línguas indígenas.* São Paulo: Ed. Loyola.

Rodrigues, A. D. (2010). Linguistic reconstruction of elements of prehistoric Tupi culture. In *Linguistics and archaeology in the americas* (pp. 1–10). Brill. DOI: https://doi.org/10.1163/9789047427087_002

Romaine, S. (2015). The global extinction of languages and its consequences for cultural diversity. In H. F. Marten, M. Rießler, J. Saarikivi, & R. Toivanen (Eds.), *Cultural and linguistic minorities in the russian federation and the european union* (pp. 31–46). Springer. DOI: https://doi.org/10.1007/978-3-319-10455-3_2

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., & **List, J.-M.** (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences, 116*(21), 10317–10322. DOI: https://doi.org/10.1073/pnas.1817972116

Shiratori, K., Cangussu, D., & **Furquim, L.** (2021). Notas botánicas sobre aislamiento y contacto. Plantas y vestigios hi-merimã (río Purús/Amazonía brasileña). *Anthropologica, 39*(47), 339–376. DOI: https://doi.org/10.18800/anthropologica.202102.014

Sokal, R. R., & **Michener, C. D.** (1958). A statistical method for evaluating systematic relationships. *University of Kansas, 38*, 1409–1438.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics, 21*(2), 121–137. DOI: https://doi.org/10.1086/464321

Swadesh, M. (2017). *The origin and diversification of language*. Chicago: Routledge. DOI: https://doi.org/10.4324/9781315133621

Tadmor, U., Haspelmath, M., & **Taylor, B.** (2010). Borrowability and the notion of basic vocabulary. *Diachronica, 27*(2), 226–246. DOI: https://doi.org/10.1075/dia.27.2.04tad

Tastevin, C. (1920). *Notebooks*. Unpublished. (Archives générales de la Congrégation du Saint Esprit, Chevilly-Larue).

van Linden, A. (2022). Harakmbut. In P. Epps & L. Michael (Eds.), *Amazonian languages: An international handbook*. Berlin: Mouton de Gruyter.

Walworth, M. (2017). Classifying old Rapa: Linguistic evidence for contact networks in Southeast Polynesia. *Issues in Austronesian Historical Linguistics*(Special publication 1), 102–122.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … **Mons, B.** (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data, 3*(1), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

Wise, M. R. (1999). Small language families and isolates in Peru. In R. M. W. Dixon & A. Aikhenvald (Eds.), *The amazonian languages* (pp. 307–340). Cambridge University Press.

Zhang, M., Yan, S., Pan, W., & **Jin, L.** (2019). Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature, 569*(7754), 112–115. DOI: https://doi.org/10.1038/s41586-019-1153-z

TO CITE THIS ARTICLE:
Gerardi, F. F., Aragon, C. C., & Reichert, S. (2022). KAHD: Katukinan-Arawan-Harakmbut Database (Pre-release). *Journal of Open Humanities Data*, 8: 18, pp. 1–11. DOI: https://doi.org/10.5334/johd.80

Published: 03 August 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.