scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Grammars Across Time Analyzed (GATA): a dataset of 52 languages

Frederic Blum[1], Carlos Barrientos[1,2], Adriano Ingunza[3], Damián E. Blasi[1,4,5,6,7,8] ✉ & Roberto Zariquiey[3,8] ✉

Grammars Across Time Analyzed (GATA) is a resource capturing two snapshots of the grammatical structure of a diverse range of languages separated in time, aimed at furthering research on historical linguistics, language evolution, and cultural change. GATA comprises grammatical information on 52 diverse languages across all continents, featuring morphological, syntactic, and phonological information based on published grammars of the same language at two different time points. Here we introduce the coding scheme and design features of GATA, and we describe some salient patterns related to language change and the coverage of grammatical descriptions over time.

Background & Summary

There are approximately 6500 mutually unintelligible languages in the world[1]. Their varied social, ecological, and cultural setups have allowed us to explore fundamental questions about language and its relation to other domains of the study of humans. The world's linguistic diversity constitutes a unique resource for understanding the cognitive basis of the human capacity to learn and use languages (e.g.[2–4]), for untangling human history at a global and regional scale (e.g.[5,6]), and for making inferences about language diversification and change (e.g.[7]). Contemporary approaches to the study of linguistic diversity rely extensively on databases with information about hundreds and even thousands of languages[7–10]. However, most of these databases display information about individual languages either at specific points in their history, or -more problematically- by combining reference sources from different points in time. This limits the study of dynamic processes of language change, as indirect inferences about the past history of languages need to be supplemented (for instance through phylogenetic histories[11]). Grammars Across Time Analyzed (GATA) is a novel resource that aims at full-filling the need for diachronic information about languages based on published descriptions of the world's languages. GATA includes information for 52 diverse languages through the independent coding of two (or more) grammatical descriptions of the same language in different points of their histories.

The study and research on language change is of foremost importance across human sciences. Naturally, language change is the main source of information in historical linguistics, as it informs us about the biases, tempo, and dynamics of the linguistic system. GATA allows the exploration of fundamental questions in the field, including e.g. the speed of grammatical change[12] and the presumed co-evolutionary processes that dominate language change[9]. Yet, more broadly, language change can be put in relation to non-linguistic questions about the human mind, culture, and history. For instance, languages are transmitted along traditions, social structures, and genes, so tracing changes in one domain can inform about processes that have taken place elsewhere[6]. At the same time, language structures and their associated patterns of change might reflect specific societal and cultural pressures directly. For instance, languages that are adopted by a large and diverse community of speakers with different linguistic backgrounds have been claimed to change in the direction of simplified morphology[13] (c.f.[14]). In a similar line, languages that are not transmitted to the newer generations have been claimed to undergo intense language change, typically resulting in a significant simplification and reduction of their grammatical inventories[15–18]. GATA offers a unique resource for testing these hypotheses and other related claims with relevant and adequately coded information on language change.

[1]Department for Linguistic and Cultural Evolution, Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [2]Institut für Linguistik, Universität Leipzig, Leipzig, Germany. [3]Pontificia Universidad Católica del Perú, Lima, Peru. [4]Department of Human Evolutionary Biology, Harvard University, Cambridge, USA. [5]Harvard Data Science Initiative, Harvard University, Cambridge, USA. [6]Center for Brain and Cognition, Pompeu Fabra University, Barcelona, Spain. [7]Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain. [8]These authors contributed equally: Damián E. Blasi, Roberto Zariquiey. ✉e-mail: dblasi@fas.harvard.edu; rzariquiey@pucp.edu.pe

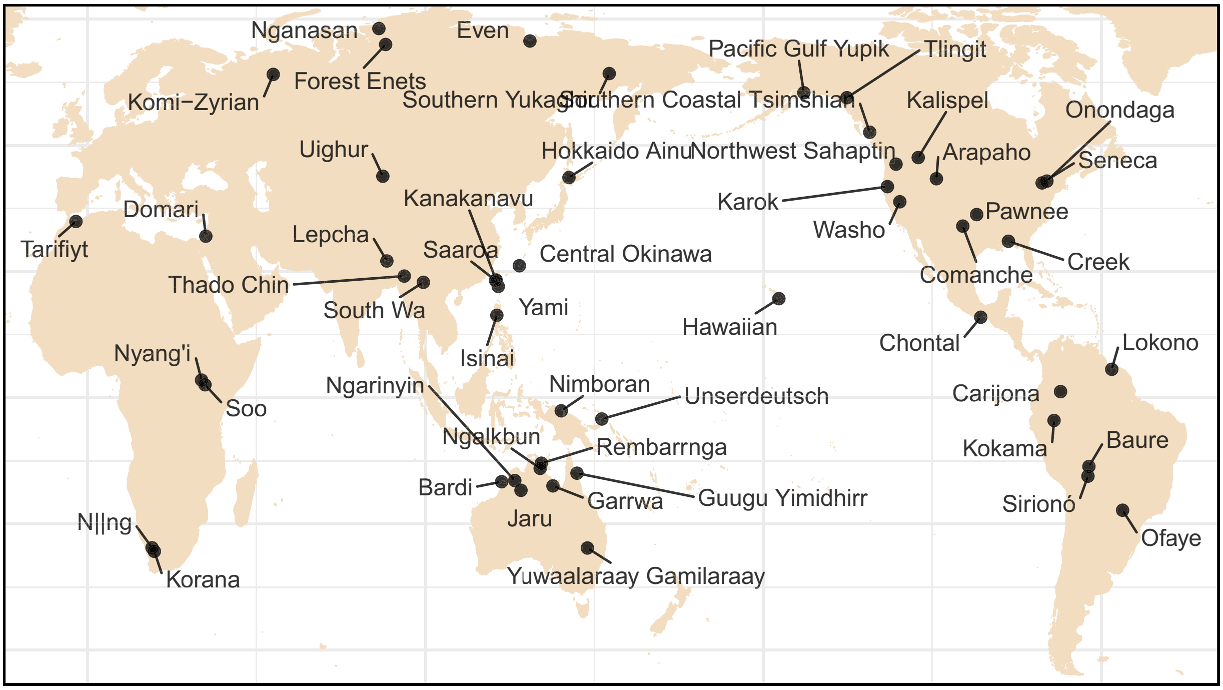

Fig. 1 Approximate location of the languages included in the first release of GATA, based on Glottolog[1].

Methods

Sample creation. The main design principle of GATA is providing a diverse set of language histories based on published scholarship. For this purpose, we tap on a thorough collection of digitized literature covering over 37,000 digitized books and articles on descriptive linguistics[19]. The collection comprises: (1) out-of-copyright texts digitized by libraries, scientific societies, and Google books; (2) texts posted online with a license that allows them to be used for research; and (3) texts under publisher copyright where quotations of short extracts are legal. A listing of the collection can be accessed via the open-access bibliography Glottolog[20]. All the documents in this collection have been digitized into machine-readable text through ABBYY Finereader 14, an optical character recognition (OCR) software, using the metalanguage as the recognition language. This collection comprises some 12,000 grammatical descriptions[21]. Based on this collection, we assembled a sample of grammars which were selected following these criteria:

1. There are two accessible grammars of the same language (at least) 25 years apart from each other. This guarantees that there is minimally a generation between the two snapshots of the same language.
2. The languages were chosen evenly with respect to geographic and genealogical distribution. This is for the purpose of providing a balanced perspective of language dynamics across the widely varying circumstances of different regions of the world and their language families.

Following these guidelines, GATA includes 52 languages coded for two reference times. Their geographic distribution can be observed in Fig. 1, and the time interval between the two grammars coded for each language is presented in Fig. 3.

Features. We selected 31 grammatical features divided into six typological categories: grammatical relations, nominal categories, phonology, pronominal systems, verbal categories, and word order. Features are classified into three types: binary (b), numeric (n) and multi-state (m). They cover various grammatical domains ranging from phonology (e.g., number of vowels, consonants and tones) to morphology (e.g., number of cases, alignment types,tense-aspect-mood markers) and syntax (e.g., word order, interrogatives constructions, alignment types). More specifically, GATA includes 4 phonological features (n = 4), 18 morphological features (n = 7 and b = 11), and 9 syntactic features (m = 3 and b = 6). Table 1 lists all the grammatical features included in GATA.

The criteria for selecting GATA's 31 grammatical features are twofold. Firstly, we included salient grammatical features whose presence/absence would be easy to determine from the description and/or the examples in each state of language (particularly in the older one, which may be associated with a relatively old grammatical description, which has not benefited from contemporary advances in descriptive linguistics). Thus, we avoid grammatical categories that have not been typically discussed until recently (such as applicative, mirative, frustrative, and engagement). Other categories, associated, for instance, with person, case, tense, vowels and consonants would be expected to show up even in more traditional grammars.

A second criterion for feature selection relates to the stability of the features[9]. We selected features that have been singled out, as being particularly labile for change (e.g., various types of classifier sets and the relative

Domain	Feature	Stability	Shortname	Type
Pronominal system	1 P pronoun distinctions	Relatively stable[33]	Pron1P	n
Pronominal system	2 P pronoun distinctions	Relatively stable[33]	Pron2P	n
Pronominal system	3 P pronoun distinctions	Relatively stable[33]	Pron3P	n
Nominal categories	Number of cases	Relatively stable[34]	Cases	n
Nominal categories	Spatial demonstratives	NA	Dem	n
Nominal categories	Numeral classifiers	Relatively unstable[34]	ClassNum	b
Nominal categories	Genitive classifiers	Relatively unstable[35]	ClassGen	b
Nominal categories	Grammatical gender	Relatively stable[33]	ClassNoun	b
Nominal categories	Instrumental vs comitative	Relatively stable[34]	InstrCom	b
Nominal categories	Alienable and inalienable possession	NA	Poss	b
Nominal categories	Locative vs directional	Possibly stable[34]	LocDir	b
Nominal categories	Temporal locative vs spatial locative	Possibly stable[34]	TempLoc	b
Nominal categories	Ditransitive argument marking	Possibly stable[34]	DitransMarking	b
Nominal categories	Agreement between noun and adjectives	NA	NP_Agr	b
Grammatical relations	Core arguments via case	Possibly unstable[34]	CA_case	b
Grammatical relations	Core arguments via head marking	Possibly unstable[34]	CA_head	b
Grammatical relations	Core arguments via word order	Possibly unstable[34]	CA_wordorder	b
Grammatical relations	Pronominal alignment	Possibly unstable[34]	PronAlign	b
Grammatical relations	Nominal alignment	Possibly unstable[34]	NomAlign	m
Verbal categories	Tense and aspect markers	NA	TA_marks	n
Verbal categories	Evidential markers	NA	Evid	n
Verbal categories	Interrogatives	NA	Questions	m
Verbal categories	Causative	Relatively unstable[33]	MorphCaus	b
Verbal categories	Agreement between verb and argument	Possibly unstable[33]	VP_Agr	b
Phonology	Number of consonants	NA	Consonants	n
Phonology	Number of tones	Stable[33,34]	Tones	n
Phonology	Number of oral vowels	NA	Vowels	n
Phonology	Number of nasal vowels	Stable[35,36]	NasVowels	n
Word order	Adj-N order	Relatively unstable[37]	AdjN	m
Word order	Gen-N order	Relatively stable[37]	GenN	m
Word order	Basic word order	Possibly stable (some)[33]	BasicWO	m

Table 1. List of GATA features organized by domain. Shortnames and Type and references to illustrative claims on relative stability are also included.

position of the adjective in relation to the noun), as well as others that have been claimed to exhibit extraordinary stability through time (e.g., gender markers, distinctions in first person pronouns, and case systems).

Coding. The coding was carried on by a careful evaluation of the grammatical descriptions selected for each language, based on the collection described in the section on Sample creation. Each grammatical feature was coded for each of the two reference grammars resulting from our search, which we refer to as 'states'. For each feature and each state, we included the following domains:

- Value. Introduces the data point for each feature, based on the typology presented in the Features subsection: binary (b), numeric (n) and multi-state (m).
- Reliability. Provides an assessment of the reliability of the original source in relation to the feature. Only for those cases in which the evidence was conclusive, (2) was coded, while non-conclusive evidence was coded with (1). (1) was mostly used for cases that presented a lack of coverage in the grammatical description, or for instances of explicit uncertainty expressed by the original author.
- Reference. The course of each data point is coded following the format author (year: page). A complete list of references is provided in the first release of GATA.
- Comments. Open-ended entry dedicated to relevant information not captured by the other fields.

Each language description underwent a detailed scrutiny. Three independent coders were assigned each other the grammars, and two senior researchers revised and curated their initial coding. The process involved a number of decisions in relation to the quality of grammatical descriptions as well as assumptions about unspoken conventions on grammar-making. We illustrate the nature of this process by highlight a handful of cases.

Illustrative cases. **Lokono** is an Arawakan language which received extensive documentation in the 19th century as well as a more recent description[22,23]. The author of the grammar corresponding to 'state 1' did not describe the Tense-aspect-mood (TAM) markers systematically or exhaustively, although evidence for TAM markers is present in glossed examples included in the grammar. Given this, we conclude that the author might have in fact missed altogether some TAM markers which are described in the 'state 2' grammar. This led us to code (1) 'non-conclusive' in the certainty column for 'state 1' in relation to TAM markers. The same author presents the notion of 'letters' instead of phonemes, a recurring issue in colonial documents, so no reliable inference regarding the phonological inventory of the language can be drawn either, which resulted in a further 'non-conclusive' judgment in relation to phonology.

Central Okinawan is a Japonic language with two grammatical descriptions[24,25]. The author of the earlier publication provides a list of personal pronouns, in which two sets based on politeness are proposed for the second and the third person. The author extensively discusses the pragmatic differences between various of these pronominal forms in a footnote. The total set of personal pronouns in Central Okinawan according to the author was 18 (including up to 12 third person pronouns with different honorific meanings). A value of (2) was assigned in the certainty column as the evidence seemed conclusive. In the more recent grammatical description of Central Okinawa, however, the author only documents two pronominal forms for the third person. The set of honorific pronouns in the third person paradigm seems to have disappeared in between both documents. For the two other person paradigms, in turn, the pronominal sets did not change significantly. The number of first person pronouns increased by one, while the number for second person pronouns remained the same. Despite the differences attested between the two grammatical descriptions, the latter one offers a detailed discussion of the pronominal set listing and illustrating all the attested free forms. We then coded a value of (2) in the certainty column for'state 2' too.

Kukama-Kukamiria is a Tupi-Guaraní language spoken in Peru. For this language, there are two grammatical descriptions available: a textbook with abundant grammatical information[26] and a contemporary reference grammar[27]. The first source does not incorporate any discussion on evidentiality, perhaps because this term was not widespread enough by the publication time of the source, and the markers that the more recent source describes as evidentials appear very superficially analyzed as modal markers. This led to a coding of evidentiality that assigns (1) 'non-conclusive' in the certainty column for the first source, and (2) 'conclusive' for the second one. A very similar situation is found regarding TAM markers. The older source describes only four tense suffixes, while the newer source lists eight clitics encoding both tense and aspect. It turned out that Kukama-Kukamiria aspect markers were also listed in the first source, but as independent words and not as bound morphemes. This may relate to the sometimes elusive morphosyntactic nature of clitics, which may manifest as dependent markers that exhibit phonological and prosodic properties of independent words. Clitics are somewhere in-between more clear-cut morphological categories like affixes and words. Thus, both sources have the same number of TAM markers, and the differences between the two states may be the result of a grammaticalization process, according to which independent words became grammatical clitics. Note that these difference may also be linked just to two distinct analysis on the morphological nature of the elements under discussion. It must be taken into consideration that the first source is one of the first studies of the language and was not oriented towards a linguistic audience, but rather to Spanish speakers who would like to learn the language. In turn, the second source is a contemporary functional oriented referential grammar. In the second source, one would expect a more detailed discussion of morphological elements in Kukama-Kukamiria. In any case, the data discussed here do not reveal a process of morphological reduction in association with the second stage, and both stages reveal basically the same number of forms.

The cases listed in this section show the importance of a careful qualitative analysis of the data, particularly in those instances where discrepancies between the two states of a language are identified.

Data Records

The dataset is stored on Zenodo (https://doi.org/10.5281/zenodo.8250217)[28] and curated on Github (https://github.com/cldf-datasets/gata). The current release of the repository is Version 1.0.0 and was peer-reviewed in 2023. All data is available under a CC-BY 4.0 license. In total, the dataset contains 3224 observations across 52 languages. We present the two sources per language in Table 2. In order to make GATA maximally compatible with other cross-linguistic databases, we adopt the Cross-Linguistic Data Formats (CLDF)[29,30]. This framework supports sharing, re-use and comparison of data in a cross-linguistic framework. One of the central underlying principles of CLDF is the use of csvw-files. Instead of storing all the data in a single file, it is stored in separate but linked tables. For example, GATA is directly linked to Glottolog[20], so that all languages are uniquely identified. This allows us to align our code and data with the FAIR[31] principles: Findable, Accesible, Interoperable, and Reusable.

The main folder of the dataset that is intended for re-use is 'cldf/', which consists of several linked csvw-files. As required by the CLDF model, our dataset has four central entities, each in its own file: Languages ('languages.csv'), Parameters ('parameters.csv'), Values ('values.csv'), and Sources ('sources.bib'). A descriptive JSON file ('StructureDataset-metadata.json') links the tables together while defining the relation between them. The 'requirements.txt' file indicates the necessary Python packages for reproducing the conversion into CLDF.

The 'languages.csv' file contains the columns 'ID', 'Name', 'Macroarea', and coordinates ('Latitude', 'Longitude') of each language. It also includes the 'Glottocode' as well as information on the endangerment status of the language ('AES'). The 'parameters.csv' file stores the information about the 31 parameters of GATA. Each parameter is given an 'ID' and 'Name', as well as a 'Shortname'. The 'Description' of the parameter is given in English and Spanish ('Description_esp'). The 31 different parameters are sorted into six different linguistic categories ('Category', 'Category_esp'). The columns 'Shortname' and 'Variable_type' contain information on how the parameter is coded in the data table. The 'Comments'-column includes the instructions that were given to the coders for filling out the questionnaires

The final component of the data is stored in 'cldf/values.csv' and contains one observation per row. Each entry has its own 'ID'. The columns 'Language_ID' and 'Parameter_ID' link the observation to the respective language

Language	Grammar 1	Grammar 2	Time lapse
Arapaho	Sallzmann 1963[38]	Cowell and Moss 2008[39]	45
Bardi	Metcalfe 1972[40]	Bowern 2012[41]	40
Chontal	Waterhouse 1967[42]	O'Connor 2004[43]	37
Kukama-Kukamiria	Faust 1972[26]	Vallejos 2016[27]	44
Kalispel	Vogt 1940[44]	Speck 1977[45]	37
Karok	Bright 1957[46]	Halpern 1997[47]	40
Lepcha	Mainwaring 1876[48]	Plaisier 2003[49]	127
Ngarinyin	Coate and Oates 1970[50]	Spronck 2015[51]	45
Ofaye	Gudschinsky 1971[52]	De Oliveira 2006[53]	35
Onondaga	Zeisberger 1888[54]	Chafe 1970[55]	88
Seneca	Holmer 1954[56]	Chafe 2015[57]	51
Siriono	Schermair 1949[58]	Dahl 2014[59]	55
Uighur	Nadzhip 1971[60]	Yakup 2005[61]	34
Washo	Kroeber 1907[62]	Jacobsen 1964[63]	57
Yami	Asai 1936[64]	Rau and Dong 2006[65]	70
Baure	Adam and Leclerc 1880[66]	Danielsen 2007[67]	127
Central Okinawan	Loveless1963[24]	Miyara 2015[25]	52
Comanche	Pimentel 1865[68]	Charney 1993[69]	128
Creek	Nathan 1977[70]	Innes, Alezander and Tilkens 2004[71]	27
Garrrwa	Furby and Furby 1977[72]	Mushin 2012[73]	35
Guugu	Roth 1901[74]	Haviland 1979[75]	78
Hidatsa	Mathews 1965[76]	Park 2012[77]	47
Jaru	Tsunoda 1981[78]	Senge 2015[79]	34
Kanakanavu	Tsuchida 1975[80]	Wild 2018[81]	43
Nimboran	Anceaux 1965[82]	May 1997[83]	32
Sahaptin	Jacobs 1931[84]	Jansen 2010[85]	79
Pawnee	Parks 1976[86]	Cruickshanks 2011[87]	35
Rembarrnga	McKay 1975[88]	McKay 2000[89]	25
Saaroa	Tsuchida 1975[80]	Pan 2012[90]	37
Wa	Drage 1907[91]	Seng 2012[92]	105
Tarifiyt	Sarrionandia 1905[93]	McClelland 1996[94]	91
Thado	Krishian 1980[95]	Haokip 2014[96]	34
Yuwaalaraay	Williams 1980[97]	Giacon 2014[98]	34
Carijona	Koch-Grünberg 1908[99]	Moreno 2000[100]	92
Domari	Macalister 1914[101]	Matras 2012[102]	98
Even	Benzing 1955[103]	Kim 2011[104]	56
Forestenets	Castrén 1854[105]	Siegl 2013[106]	159
Hawaiian	Andrews 1854[107]	Elbert and Pukui 1979[108]	125
Hokkaido	Refsing 1986[109]	Bugaeva 2012[110]	26
Isinai	Scheerer 1918[111]	Perlawn 2015[112]	97
Komi	Von der Gabelentz 1841[113]	Hausenberg 1998[114]	157
Korana	Meinhof 1930[115]	Maingard 1962[116]	32
Lokono	Crevaux, Sagot and Adam 1882[22]	Pet 2011[64]	129
Ngalkbun	Capell 1962[117]	Cutfield 2011[118]	49
Nganasan	Castren 1966[105]	Wagner-Nagy 2019[119]	165
Niing	Maingard 1937[116]	Collins and Namaseb 2011[120]	74
Nyangi	Heine 1975[121]	Beer, McKinney and Kosma[122]	34
Southern Coastal Tsimishian	Boas 1911[123]	Dunn 1979[124]	68
Yukaghir	Jochelson 1905[125]	Maslova 1999[126]	94
Tiingit	Boas 1917[127]	Naish 1979[128]	62
Unserdeutsch	Volker 1983[129]	Leindensfelser and Maitz 2017[130]	34

Table 2. List of the grammars sample per language and the time apart from each other.

and parameter. The value of the observation is given in the column 'Value'. Further, all observations include the bibtex-key of the 'Source' (linked to 'sources.bib'), the 'Certainty' of the observation, specific page-references ('Reference') for examples, and the 'Year' of the publication of the data. In some cases, the coders have added a

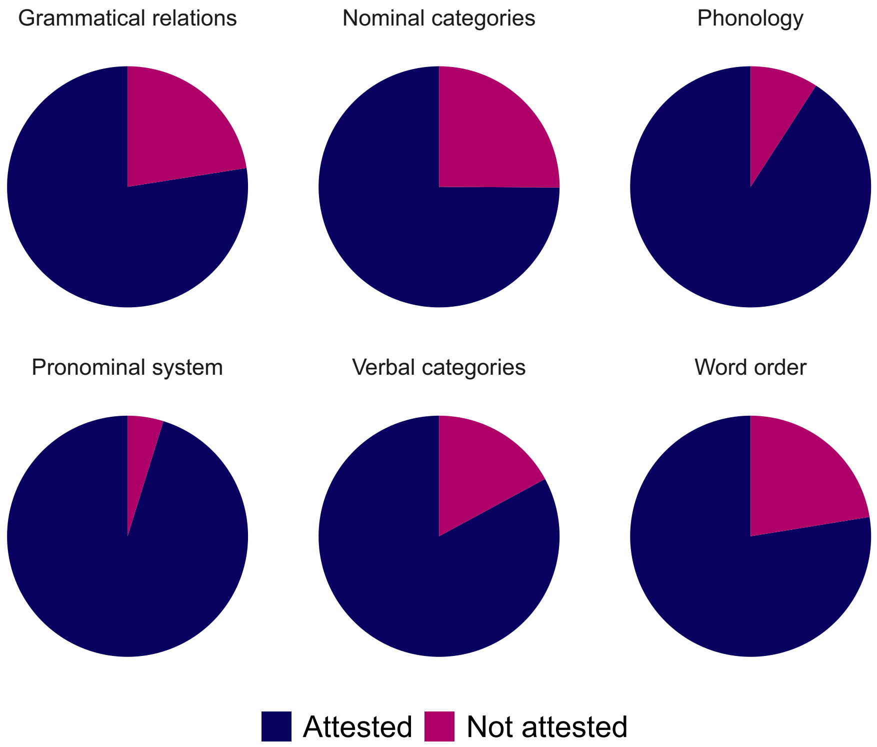

Fig. 2 Feature coverage for each typological domain in GATA.

'Comment' to the observation, which provides further information on the judgments and assumptions that were made during the analysis.

We rigorously followed the workflow and examples provided by the CLDF documentation. This included the usage of the CLDFBench package[32] to customize and create the dataset in CLDF automatically (see the Code availability section). The individual files in the repository are part of the CLDF workflow and describe the different contents, such as the cldfbench-script ('cldfbench_gata.py'), the contributors table ('Contributors.md'), the license, or the metadata. Dataset-specific files that were used in order to convert the data into CLDF are stored in the folders 'etc/' and 'raw/'. Those two folders are only used for creating the CLDF dataset and should not be used as data source. The 'raw/'-directory contains the combined raw data ('gata_raw.csv'), a bib-file with all sources, the original questionnaires, and all scripts that have been used during pre-processing. The 'etc/'-directory stores metadata that has been used during the conversion to CLDF. This includes a list of all parameters ('parameters.csv') as well as the list of languages ('languages.csv'), incorporating also information on their vitality status. In addition to the standard CLDF folders, we also created a folder 'plots/' which contains descriptive plots of the data and the code to create them.

Technical Validation

We assess GATA in the light of general desiderata that apply to all cross-linguistic and cross-cultural databases: balanced sampling and feature coverage. In addition, we discuss the temporal distribution of the time gaps between grammars.

The **balanced sampling** design principle entails that, to the extent possible, the resource should provide an accurate perspective of the diversity and the variation present across the world's languages. The current version of GATA[28] comprises information on 52 languages, representing in a balanced manner, by design, all major linguistic regions in the world, as shown by the map presented in Fig. 1.

Reference grammars might not specify the features coded in GATA for a number of reasons, which would lead to uneven **feature coverage**. A visualization of the feature coverage for each typological domain is featured in Fig. 2, where we can see that no typological domain in GATA has a feature coverage under 75% of the total number of languages, being phonology and pronominal systems the better covered features with a coverage of over 90%.

The **temporal distribution** between the two states of a language coded in GATA varies significantly across languages, as observed in Fig. 3. One language in the sample, Rembarrnga, includes two states separated by 25 years, while there are three languages, Komi-Zyrian, Forest Enets, and Nganasan, for which the time interval between states is approximately 160 years (see Fig. 3).

Finally, we can assess the aggregated amount of language change emerging from each of these domains across time. Figure 4 showcases, within each grammatical domain, the fraction of all features who have changed over a given period of time. There is substantial variation in the total amount of change witnessed across domains, partially due to the differing temporal stability of linguistic features[9]. It should also be noted that, larger time

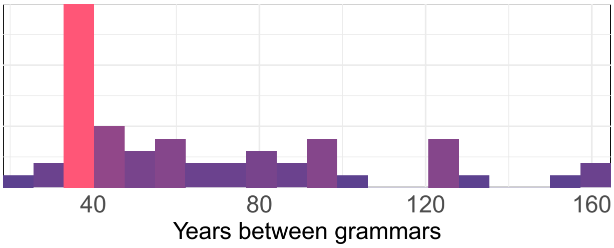

Fig. 3 Time intervals between states (published grammars) in the languages coded in GATA.

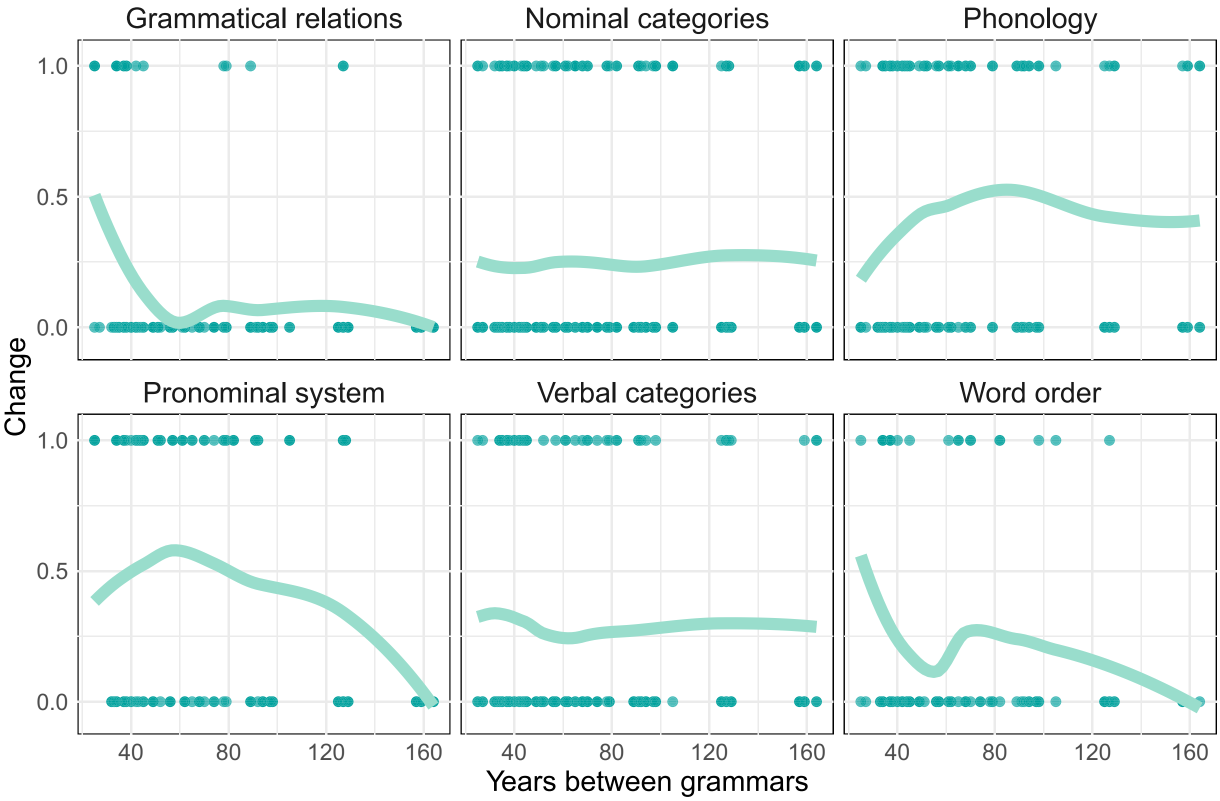

Fig. 4 An example analysis of change across the different domains. The amount of change within each domain is plotted against the difference in years between both states of documentation.

intervals are not necessarily associated with larger amounts of language change - within the relatively narrow time span between the grammars analyzed in this paper. This can be understood as (partially) reflecting how language change is widely modulated by varying social and cultural factors, but it could point as well to a more subtle effect associated with the perceived utility of grammars. In a nutshell: given that a grammar of a language exist, the need for a second grammar would increase if its considered that sufficient language change has taken place (other factors, such as the theoretical frame and the coverage of the former grammar, being equal).

Usage Notes

Following CLDF standards, the GATA dataset[28] is published as linked CSVW-files. It can easily be accessed either with CSV-reading applications or with designated tools developed from the CLDF-community. For example, designated programming packages for retrieving CLDF data have been developed for Python (https://github.com/cldf/pycldf) and R (https://github.com/SimonGreenhill/rcldf). The tabular CLDF format permits easy comparison with other CLDF-datasets. For example, it is possible to use the commandline interface of the pycldf-Python package to access the data or to create a SQlite database out of this dataset[29]. In general, the tabular-format makes it easy to reuse the data in various ways (associated with a large list of potential research questions), and with many different programs.

Code availability

The dataset is stored on Zenodo (https://doi.org/10.5281/zenodo.8250217)[28] and curated on Github (https://github.com/cldf-datasets/gata). The current release of the repository is Version 1.0.0 and was peer-reviewed in 2023. All data is available under a CC-BY 4.0 license. All scripts that have been used during the pre-processing of the data are made available within the repository. Specifically, Python-scripts were used after the manual annotation of the data for the standardization of all annotations, as well as for the aggregation of the individual spreadsheets. The script that was used for the conversion into CLDF is also part of this repository.

Received: 14 March 2023; Accepted: 18 October 2023;
Published online: 28 November 2023

References

1. Hammarström, H. *et al*. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation* **12**, 723–729 (2018).
2. Blasi, D. E., Michaelis, S. M. & Haspelmath, M. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour* **1**, 723–729, https://doi.org/10.1038/s41562-017-0192-4 (2017).
3. Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F. & Christiansen, M. H. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* **113**, 10818–10823, https://doi.org/10.1073/pnas.1605782113 (2016).
4. Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences* https://doi.org/10.1016/j.tics.2022.09.015 (2022).
5. Koile, E., Greenhill, S. J., Blasi, D. E., Bouckaert, R. & Gray, R. D. Phylogeographic analysis of the bantu language expansion supports a rainforest route. *Proceedings of the National Academy of Sciences* **119**, e2112853119, https://doi.org/10.1073/pnas.2112853119 (2022).
6. Matsumae, H. *et al*. Exploring correlations in genetic and cultural variation across language families in northeast asia. *Science Advances* **7**, eabd9223, https://doi.org/10.1126/sciadv.abd9223 (2021).
7. Skirgård, H. *et al*. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* **9**, https://doi.org/10.1126/sciadv.adg6175 (2023).
8. Dryer, M. & Haspelmath, M. The World Atlas of Language Structures Online, *Zenodo*, https://doi.org/10.5281/zenodo.3607047 (2013).
9. Dediu, D. & Cysouw, M. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLOS ONE* **8**, 1–20, https://doi.org/10.1371/journal.pone.0055009 (2013).
10. List, J.-M. *et al*. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* **9**, 316, https://doi.org/10.1038/s41597-022-01432-0 (2022).
11. Greenhill, S. J., Heggarty, P. & Gray, R. D. Bayesian phylolinguistics. In Richard D. Janda, B. S. V., Brian D. Joseph (ed.) *The Handbook of Historical Linguistics*, chap. 11, 226–253, https://doi.org/10.1002/9781118732168.ch11(2020).
12. Greenhill, S. J. *et al*. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* **114**, E8822–E8829, https://doi.org/10.1073/pnas.1700388114 (2017).
13. Lupyan, G. & Dale, R. Language structure is partly determined by social structure. *PloS One* **5**, e8559, https://doi.org/10.1371/journal.pone.0008559 (2010).
14. Shcherbakova, O. *et al*. Societies of strangers do not speak less complex languages. *Science Advances***9**. https://doi.org/10.1126/sciadv.adf7704 *(2023).*
15. Dorian, N. Maintenance and loss of same-meaning structures in language death. *Word* **31**, 39–45, https://doi.org/10.1080/00437956.1980.11435680 (1980).
16. Sasse, H.-J. Theory of language death. In Brenzinger, M. (ed.) *Language Death: Factual and Theoretical Explorations with Special Reference to East Africa*, 7–30, https://doi.org/10.1515/9783110870602.7 (Cambridge University Press, Berlin, 1992).
17. Palosaari, N. & Campbell, L. Structural aspects of language endangerment. In Austin, P. & Sallabank, J. (eds.) *The Cambridge Handbook of* Endangered Languages, 100–119, https://doi.org/10.1017/CBO9780511975981.006 (Cambridge University Press, Cambridge, 2011).
18. Aikhenvald, A. Y. Language change in language obsolescence. In Janda, R. D., Joseph, B. D. & Vance, B. S. (eds.) *The Handbook of Historical Linguistics*, 447–467, https://doi.org/10.1002/9781118732168.ch21 (Wiley Blackwell, Croyden, UK, 2021).
19. Hammarström, H. Gramfinder: Human and machine reading of grammatical descriptions of the languages of the world. In Dragut, E. C., Li, Y., Popa, L. & Vucetic, S. (eds.) *3rd Workshop on Data Science with Human in the Loop, DaSH@KDD, Virtual Conference, August 15, 2021* (2021).
20. Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. glottolog/glottolog-cldf: Glottolog database 4.8 as CLDF, *Zenodo*, https://doi.org/10.5281/zenodo.8131091 (2023).
21. Virk, S., Hammarström, H., Borin, L., Forsberg, M. & Wichmann, S. From linguistic descriptions to language profiles. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, 23–27 (European Language Resources Association, Marseille, 2020).
22. Crevaux, J., Sagot, P. & Adam, L. Grammatik der arawakischen sprache. In *Roucouyenne, Arrouague, Piapoco et d'autres langues de la région des Guyanes*, vol. VIII of *Bibliothèque Linguistique Américaine*, 166–240 (Librairie-Éditeur J. Maisonneuve, Paris, 1882).
23. Pet, W. J. A. *A Grammar Sketch and Lexicon of Arawak (Lokono Dian)*, vol. 30 of *SIL e-Books* (SIL International, Dallas, Texas, 2011).
24. Loveless, R. *The Okinawan Language*. Ph.D. thesis, University of Michigan (1963).
25. Miyara, S. Shuri Okinawan Grammar. In Heinrich, P., Miyara, S. & Shimoji, M. (eds.) *Handbook of the Ryukyuan Languages: History, Structure, and Use*, vol. 11 of *Handbooks of Japanese Language and Linguistics*, 379–404, https://doi.org/10.1515/9781614511151 (DeGruyter Mouton, Berlin, 2015).
26. Faust, N. *Gramática Cocama: Lecciones para el aprendizaje del idioma Cocama*, vol. 6 of *Serie Lingüística Peruana* (Instituto Lingüístico de Verano, 1972).
27. Vallejos, R. *A Grammar of Kukama-Kukamiria*. Brill's Studies in the Indigenous Languages of the Americas (Brill, Leiden, Boston, 2016).
28. Blum, F., Barrientos, C., Ingunza, A., Blasi, D. E. & Zariquiey, R. CLDF dataset for the GATA (Grammars Across Time Analyzed) dataset. *Zenodo* https://doi.org/10.5281/zenodo.8250217. Version 1.0.0 (2023).
29. Forkel, R. *et al*. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* **5**, https://doi.org/10.1038/sdata.2018.205 (2018).
30. Forkel, R. *et al*. Managing historical linguistic data for computational phylogenetics and computer-assisted language comparison. In Berez-Kroeker, A., McDonnell, B., Koller, E., Collister, L. & Thomason, S. (eds.) *The Open Handbook of Linguistic Data Management*, https://doi.org/10.7551/mitpress/12200.003.0033 (MIT Press, 2022).
31. Wilkinson, M. *et al*. The fair guiding principles for scientific data management and stewardship. *Scientific Data* **3**, https://doi.org/10.1038/sdata.2016.18 (2016).

32. Forkel, R. & List, J.-M. CLDFBench: Give your cross-linguistic data a lift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6995–7002 (European Language Resources Association, Marseille, France, 2020).
33. Nichols, J. Diachronically stable structural features. In Andersen, H. (ed.) *Historical Linguistics 1993*, vol. 124 of *Current Issues in Linguistic Theory*, 337–355 (John Benjamins, Amsterdam, 1995).
34. Nichols, J. Diversity and stability in language. In Brain D. Joseph, R. D. J. (ed.) *The Handbook of Historical Linguistics*, 283–310, https://doi.org/10.1002/9781405166201.ch5 (Wiley Online Library, 2017).
35. Croft, W. *Typology and Universals*, 1 edn (Cambridge University Press, 1996).
36. Croft, W. *Typology and Universals*, 2 edn (Cambridge University Press, 2003).
37. Hawkins, J. A. *Word order universals*, vol. 3 of *Quantitative Analyses of Linguistic Structure* (San Diego: Academic Press, 1983).
38. Salzmann, Z. *A Sketch of Arapaho Grammar*. Ph.D. thesis, Indiana University, Ann Arbor (1963).
39. Cowell, A. & Moss, A. *The Arapaho language* (University Press of Colorado, Boulder, 2008).
40. Metcalfe, C. D. *Bardi Verb Morphology*. Ph.D. thesis, https://doi.org/10.25911/5d73925c95ab5 Australian National University, Canberra (1975).
41. Bowern, C. *A grammar of Bardi*, vol. 57 of *Mouton Grammar Library* (De Gruyter Mouton, Berlin, Boston, 2012).
42. Waterhouse, V. Huamelultec Chontal. In McQuown, N. A. (ed.) *Linguistics*, vol. 5 of *Handbook of Middle American Languages*, 349–368 (University of Texas Press, Austin, 1967).
43. O'Connor, L. *Motion, transfer, and transformation: The grammar of change in Lowland Chontal*. Ph.D. thesis, https://doi.org/10.1075/slcs.95 University of California at Santa Barbara (2004).
44. Vogt, H. *The Kalispel Language: An Outline of the Grammar with Texts, Translations and Dictionary* (Det norske videnskaps-akademi i Oslo, Oslo, 1940).
45. Speck, B. J. *An edition of Father Post's Kalispel grammar*. Master's thesis, University of Montana (1977).
46. Bright, W. *The Karok Language*, vol. 13 of *University of California Publications in Linguistics* (University of California Press, Berkeley and Los Angeles, 1957).
47. Halpern, A. M. *Kar?úk: Native Accounts of the Quechan Mourning Ceremony*, vol. 128 of *University of California Publications in Linguistics* (University of California Press, 1997).
48. Mainwaring, C. G. B. *A Grammar of the Róng (Lepcha) Language* (Baptist Mission Press, Calcutta, 1876).
49. Plaisier, H. Lepcha. In Thurgood, G. & LaPolla, R. J. (eds.) *The Sino-Tibetan languages*, 705–716, https://doi.org/10.4324/9780203221051.ch43 (Routledge, London, New York, 2003).
50. Coate, H. H. J. & Oates, L. *A Grammar of Ngarinjin, Western Australia*, vol. 25 of *Australian Aboriginal Studies* (Australian Institute of Aboriginal Studies, Canberra, 1970).
51. Spronck, M. *Reported speech in Ungarinyin: grammar and social cognition in a language of the Kimberley region, Western Australia*. Ph.D. thesis, https://doi.org/10.25911/5d6c394b1bc12 Australian National University (2015).
52. Gudschinsky, S. C. *Estudos sôbre línguas e culturas indígenas: Trabalhos linguísticos realizados no Brasil (edição especial)* (Instituto Lingüístico de Verão, Brasília, 1971).
53. das Dores de Oliveira, M. *Ofayé, a língua do povo do mel: Fonologia e Gramática*. Ph.D. thesis, Maceió: Universidade Federal de Alagoas (2006).
54. Zeisberger, D. *Essay of an Onondaga grammar*, vol. 11, 12, 12, 12 (Pennsylvania Magazine of History and Biology, Philadelphia, 1888).
55. Chafe, W. L. *A Semantically Based Sketch of Onondaga*, vol. 25 of *Indiana University Publications in Anthropology and Linguistics* (Waverly Press, Baltimore, Maryland, 1970).
56. Holmer, N. M. *The Seneca Language (A Study in Iroquoian)*, vol. 3 of *Upsala Canadian Studies* (Lundequistska Bokhandeln, Upsala, 1954).
57. Chafe, W. *A Grammar of the Seneca Language*, vol. 150 of *University of California Publications in Linguistics* (University of California Press, Berkeley and Los Angeles, 2015).
58. Schermair, A. *Gramática de la lengua sirionó* (Talleres Gráficas de A. Gamarra, La Paz, 1949).
59. Muysken, P., Crevels, M. (eds.). *Oriente, vol. 3 of Lenguas de Bolivia* (pp. 99–133. Plural Editores, La Paz, 2014). Östen Dahl. Sirionó.
60. Nadzhip, E. *Modern Uigur* (Nauka Publishing House, Moscow, 1971).
61. Yakup, A. *The Turfan dialect of Uyghur*, vol. 63 of *Turcologica* (Harrassowitz, Wiesbaden, 2005).
62. Kroeber, A. L. *The Washo language of East Central California and Nevada* (University of California Press, Berkeley, 1906-1907).
63. Jacobsen, W. H. Jr. *A Grammar of the Washo Language*. Ph.D. thesis, University of California at Berkeley (1964).
64. Asai, E. *A study of the Yami language: an Indonesian language spoken on Botel Tobago Island*. Ph.D. thesis, Rijksuniversiteit te Leiden, Leiden (1936).
65. Rau, V. D. & Dong, M.-N. *Yami texts with Reference Grammar and Dictionary*, vol. 10 of *Language and linguistics monograph series* (Academica Sinica, Taipei, 2006).
66. Adam, L. & Leclerc, C. *Gramática de la Lengua de los Indios Baures de la provincia de los Moxos*, vol. VII of *Bibliothèque Linguistique Américaine* (Librairie-Éditeur J. Maisonneuve, Paris, 1880 [1749]).
67. Danielsen, S. *Baure: An Arawak Language of Bolivia*, vol. 6 (CNWS Publications, Leiden, 2007).
68. Pimentel, D. F. *Cuadro descriptivo y comparativo de las lenguas indígenas de México* (Andrade y Escalante, 1865).
69. Charney, J. O. *A Grammar of Comanche*. (University of Nebraska Press, Lincoln, 1993). Studies in the Anthropology of North American Indians.
70. Nathan, M. *Grammatical Description of the Florida Seminole Dialect of Creek*. Ph.D. thesis, Tulane University, Ann Arbor (1977).
71. Innes, P., Alexander, L. & Tilkens, B. *Beginning Creek* (University of Oklahoma Press, Norman, 2004).
72. Furby, E. S. & Furby, C. E. *A preliminary analysis of Garawa phrases and clauses*, vol. 42 of *Pacific Linguistics: Series B* (Research School of Pacific and Asian Studies, Australian National University, Canberra, 1977).
73. Mushin, I. *A grammar of (Western) Garrwa*, vol. 637 of *Pacific Linguistics* (De Gruyter Mouton, Berlin, Boston, 2012).
74. Roth, W. E. *The structure of the Koko-Yimidir language, North Queensland*, vol. 2 of *North Queensland Ethnography Bulletin* (Government Printer, Brisbane, 1901).
75. Haviland, J. Guugu yimidhirr. In Dixon, R. & Blake, B. (eds.) *Handbook of Australian Languages*, vol. 1, 27–182, https://doi.org/10.1075/z.hal1.06hav (John Benjamins, Amsterdam, Philadelphia, 1979).
76. Matthews, G. H. *Hidatsa syntax*, vol. 3 of *Papers on formal linguistics* (Mouton & Co, The Hague, 1965).
77. Park, I. *A grammar of Hidatsa*. Ph.D. thesis, Indiana University (2012).
78. Tsunoda, T. *The Djaru Language of Kimberley, Western Australia*, vol. 78 of *Pacific Linguistics: Series B* (Pacific Linguistics Press, Canberra, 1981).
79. Senge, C. *A Grammar of Wanyjirra, a language of Northern Australia*. Ph.D. thesis, Australian National University (2015).
80. Tsuchida, S. *Reconstruction of Proto-Tsouic Phonology*. Ph.D. thesis, New Haven: Yale University (1975).
81. Wild, I. *Voice and transitivity in Kanakanavu*. Ph.D. thesis, Universität Erfurt (2018).
82. Anceaux, J. C. *The Nimboran language: Phonology and morphology*, vol. 44 of *Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde, 44* (Nijhoff, 's-Gravenhage, 1965).
83. May, K. *A Study of the Nimboran Language: Phonology, morphology, and phrase structure*. Master's thesis, LaTrobe University (1997).

84. Jacobs, M. A sketch of northern sahaptin grammar. *University of Washington Publications in Anthropology* **4**, 85–292 (1931).
85. Jansen, J. W. *A grammar of Yakima Ichishkíin/Sahaptin*. Ph.D. thesis, Eugene: University of Oregon (2010).
86. Parks, D. R. *A grammar of Pawnee*. Garland studies in American Indian linguistics (Garland, New York, 1976).
87. Cruickshanks, A. *Complementation Strategies in South Band Pawnee*. Master's thesis, University of Sydney (2011).
88. McKay, G. R. *Rembarnga, A language of central Arnhem Land*. Ph.D. thesis, Australian National University (1975).
89. McKay, G. Ndjébbana. In Dixon, R. M. W. & Blake, B. J. (eds.) *Handbook of Australian Languages 5*, 154–354 (Oxford University Press, Oxford, 2000).
90. jung Pan, C. *A Grammar of Lha'alua, an Austronesian Language of Taiwan*. Ph.D. thesis, James Cook University (2012).
91. Drage, G. *A Few Notes on Wa* (Superintendent, Government Print, 1907).
92. Seng Mai, M. *A descriptive grammar of Wa*. Master's thesis, Chiang Mai: Payap University (2012).
93. Sarrionandia, P. *Gramática de la lengua rifeña*, 2 edn (Imprenta Hispano-Arabiga, Tánger, 1905).
94. McClelland, C. W. *Interrelations of prosody, clause structure and discourse pragmatics in Tarifit Berber*. Thesis, University of Texas at Arlington (1996).
95. Krishan, S. *Thadou: Grammatical Sketch* (Anthropological Survey of India, Government of India, Calcutta, 1980).
96. Haokip, M. *Grammar of Thadou-Kuki: A Descriptive Study*. Ph.D. thesis, New Delhi: Jawaharlal Nehru University (2014).
97. Williams, C. J. *A Grammar of Yuwaalaraay* (Australian National University, Canberra, 1980).
98. Giacon, J. *A grammar of Yuwaalaraay and Gamilaraay: a description of two New South Wales languages based on 160 years of records*. Ph.D. thesis, Australian National University (2014).
99. Koch-Grünberg, T. Die Hianákoto-Umáua, *Anthropos*, 83-124, 297-335, 952-982, **3** (1908).
100. Moreno, C. A. R. Avance sobre morfología carijona. In González de Pérez, M. S. & Rodríguez de Montes, M. L. (eds.) *Lenguas indígenas de Colombia: una visión descriptiva*, 171–180 (Instituto Caro y Cuervo, Santafé de Bogotá, 2000).
101. Macalister, R. S. *The Language of the Nawar or Zutt, the Nomad Smiths of Palestine, vol. 3 of Gypsy Lore Society Monograph*. (Bernard Quaritch, London, 1914).
102. Matras, Y. *A Grammar of Domari*, vol. 59 of *Mouton Grammar Library* (De Gruyter Mouton, Berlin, Boston, 2012).
103. Benzing, J. *Lamutische Grammatik mit Bibliographie, Sprachproben und Glossar*, vol. VI of *Akademie der Wissenschaften und der Literatur: Veröffentlichungen der Orientalischen Kommission* (Franz Steiner, Wiesbaden, 1955).
104. Kim, J. *A grammar of Ewen*, vol. 6 of *Altaic languages series* (Seoul National University Press, Seoul, 2011).
105. Matthias Alexander, C. Grammatik der samojedischen Sprachen, 7 (Kaiserliche Akad. der Wissenschaften, 1854).
106. Siegl, F. *Materials on Forest Enets, an indigenous language of Northern Siberia*, vol. 267 of *Mémoires de la Société Finno-Ougrienne* (Suomalais-Ugrilainen Seura, Helsinki, 2013).
107. Andrews, L. *Grammar of the Hawaiian language* (Mission Press, Honolulu, 1854).
108. Elbert, S. H. & Pukui, M. K. *Hawaiian Grammar* (University of Hawaii Press, Honolulu, 1979).
109. Refsing, K. *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect*. (Aarhus University Press, Aarhus, 1986).
110. Bugaeva, A. Southern hokkaido ainu. In Tranter, N. (ed.) *The Languages of Japan and Korea*, 461–509 (Routledge, New York, 2012).
111. Scheerer, O. *The Particles of Relation of the Isinai Language* (KITLV, The Hague, 1918).
112. Perlawan, S. E. Grammatical Sketch of Isinay Dupax. Ms. (2015).
113. von der Gabelentz, H. C. *Grundzüge der syrjänischen Grammatik* (Pierer, Altenburg, 1841).
114. Hausenberg, A.-R. Komi. In Abondolo, D. (ed.) *The Uralic Languages*, 305–326 (Routledge, London, 1998).
115. Meinhof, C. *Der Koranadialekt des Hottentottischen*, vol. 12 of *Beihefte zur Zeitschrift für Eingeborenen-Sprachen* (Verlag von Dietrich Reimer (Ernst Vohsen), Berlin, Hamburg, 1930).
116. Maingard, L. F. The khomani dialect of bushman: Morphology and other characteristics. In Jones, J. D. R. (ed.) *Bushmen of the Southern Kalahari*, 237–275 (University of the Witwatersrand Press, Johannesburg, 1937).
117. Capell, A. Linguistic research needed in australia. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* **5**, 23–28 (1962).
118. Cutfield, S. *Demonstratives in Dalabon: A language of southwestern Arnhem Land*. Ph.D. thesis, Monash University (2011).
119. Wagner-Nagy, B. B. *A grammar of Nganasan* (Brill, Leiden, 2019).
120. Collins, C. & Namaseb, L. *A grammatical sketch of N/uuki with stories*, vol. 25 of *Quellen zur Khoisan-Forschung/Research in Khoisan Studies* (Rüdiger Köppe, Cologne, 2011).
121. Heine, B. Tepes und nyang'i, zwei ostafrikanische restsprachen. *Afrika und Übersee* **58**, 263–300 (1975).
122. Beer, S., McKinney, A. & Kosma, L. The So Language: A grammar sketch. Ms. (2009).
123. Boas, F. Tsimshian. In Boas, F. (ed.) *Handbook of American Indian languages: Volume 1*, vol. 40 of *Bulletin of American Ethnology*, 283–422 (Government Printing Office, Washington, 1911).
124. Dunn, J. A. *A Reference Grammar for the Coast Tsimshian Language*, vol. 55 of *National Museum of Man, Mercury Series: Canadian Ethnology Service Paper* (National Museums of Canada, Ottawa, 1979).
125. Jochelson, W. Essay on the grammar of the yukaghir language. *Annals of the New York Academy of Sciences* **16**, 97–152 (1905).
126. Maslova, E. *A grammar of Kolyma Yukaghir* (Universität Bielefeld, Bielefeld, 1999).
127. Boas, F. *Grammatical notes on the language of the Tlingit Indians*, vol. 8(1) of *Anthropological publications/University Museum, University of Pennsylvania* (The University Museum, Philadelphia, 1917).
128. Naish, C. M. *A syntactic study of Tlingit*, vol. 6 of *Language Data Amerindian Series*, microfiche ed. edn (Summer Institute of Linguistics, Dallas, 1979).
129. Volker, C. A. *An Introduction to Rabaul Creole German (Unserdeutsch)*. Master's thesis, University of Queensland (1983).
130. Lindenfelser, S. & Maitz, P. The creoleness of unserdeutsch (rabaul creole german): A typological perspective. In Maitz, P. & Volker, C. A. (eds.) *Language Contact in the German Colonies: Papua New Guinea and beyond*, LLM Special Issue, 91–142 (LSPNG, Port Moresby, 2017).

Acknowledgements

We thank Pontificia Universidad Católica del Perú for funding the research that gave rise to GATA and this paper. We thank Johann-Mattis List who assisted with the conversion of the dataset into CLDF.

Author contributions

Senior authors R.Z. and D.E.B. conceived the GATA database and the experiment(s) conducted for this paper. F.B., A.I. and C.B. coded the languages included in the fist release of GATA. F.B. aggegrated and preprocessed the data. D.E.B. conducted the experiments and produced the figures, R.Z. and D.E.B. analysed the results. F.B. and C.B. exported the GATA data into CLDF format. R.Z. and D.E.B. wrote the Background & Summary, and Technical Validation. All authors wrote the Methods section. F.B. and C.B. wrote the Data Records and Usage Notes. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.E.B. or R.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.