Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2884–2890
Marseille, 20-25 June 2022

Overlooked Data in Typological Databases: What Grambank Teaches Us About Gaps in Grammars

Jakob Lesage[1], Hannah J. Haynie[2], Hedvig Skirgård[3], Tobias Weber[4], Alena Witzlack-Makarevich[5]

[1]Department of Asian and African Studies, Humboldt University Berlin; Berlin, Germany
[2]Department of Linguistics, University of Colorado Boulder; Boulder, USA
[3]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology; Leipzig, Germany
[4]Institute for Scandinavian Studies, Frisian and General Linguistics, Department of General Linguistics, Kiel University; Kiel, Germany
[5]Department of Linguistics, Faculty of Humanities, The Hebrew University of Jerusalem; Jerusalem, Israel
jakob.lesage@gmail.com, hannah.haynie@colorado.edu, hedvig_skirgard@eva.mpg.de, tweber@isfas.uni-kiel.de, witzlack@gmail.com

Abstract

Typological databases can contain a wealth of information beyond the collection of linguistic properties across languages. This paper shows how information often overlooked in typological databases can inform the research community about the state of description of the world's languages. We illustrate this using Grambank, a morphosyntactic typological database covering 2,467 language varieties and based on 3,951 grammatical descriptions. We classify and quantify the comments that accompany coded values in Grambank. We then aggregate these comments and the coded values to derive a level of description for 17 grammatical domains that Grambank covers (negation, adnominal modification, participant marking, tense, aspect, etc.). We show that the description level of grammatical domains varies across space and time. Information about gaps and uncertainties in the descriptive knowledge of grammatical domains within and across languages is essential for a correct analysis of data in typological databases and for the study of grammatical diversity more generally. When collected in a database, such information feeds into disciplines that focus on primary data collection, such as grammaticography and language documentation.

Keywords: Less-Resourced/Endangered Languages, Linked Data, Typological Databases

1. Introduction

There exist grammar sketches and grammars of approximately 4,000 of the world's languages (Hammarström et al., 2021), but not all languages are described in equal detail. The past years have seen some assessments of the breadth of our documentation of global linguistic diversity (e.g. Seifart et al., 2018). In this paper, we explore how typological databases, typically used to study linguistic diversity, can be used to examine differences in the coverage of grammars and grammar sketches. Using information often overlooked in typological databases, we identify what grammatical domains are covered grammatical descriptions and how the coverage of grammatical descriptions patterns in time and space. We use the global typological database Grambank which covers 2,467 language varieties based on 3,951 grammatical descriptions (grammars) (Skirgård et al., submitted; The Grambank Consortium, 2021).

By studying the coded values for typological features and the comments provided by the coders, we can estimate the description level of a given grammatical domain in the available resources for a language (see section 3 for details on the metric). After combining this score with information about the location and genealogy of a language and the publication date of the grammatical descriptions, we can explore areal, genealogical, and temporal patterns in grammar writing. The results can guide descriptive efforts and increase our understanding of biases in typological datasets.

2. Grambank: A Global Morphosyntactic Database

Grambank is a typological database that encodes morphosyntactic information for 2,467 languages and language varieties across the world. Grambank covers languages from 215 families as well as 101 isolates (see Figure 1). Grambank represents 42% (558) of all languages currently described in Africa, 55% (137) in Australia, 42% (546) in Eurasia, 48% (242) in North America, 68% (718) in Papunesia, and 58% (222) in South America.

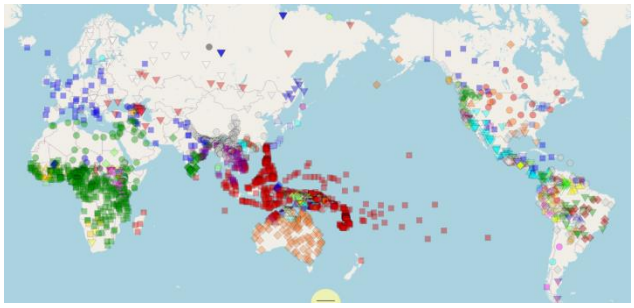

Figure 1: Geographical distribution of the languages coded in Grambank. Different colors and symbols represent different language families.

The morphosyntactic information is coded as answers to yes-no questions for 195 features which cover most core grammatical domains (negation, tense, aspect, word order, argument marking, case, gender, number, etc.).[1] The database is compiled by a team of over 70 coders who have filled in 441,663 values. Grambank aims to be a resource for investigating questions about deep language prehistory,

[1] Six of the features are in fact multi-valued, e.g. the feature encoding the order of numeral and noun has the possible values 1 "numeral-noun", 2 "noun-numeral" and 3 "both orders are possible".

the identification of linguistic areas, the stability of grammatical sub-systems, and interactions between different areas of grammar.

Examples of questions asked in Grambank are "Are there definite or specific articles?" and "Are verbs reduplicated?". Each of these questions can be answered with 1 "yes", 0 "no" or ? if the answer is unclear. Additionally, coders are encouraged to add comments if a case is not straightforward, requires reinterpretation of the data, or if the answer value 0 was chosen because a source does not provide an explicit answer. A 0 in the value field and "not mentioned" in the comment field, for instance, reflect a coder's judgment that in a particular case, a source does not mention a feature for a language (e.g. plural marking) because it does not exist in that language (e.g. because there are no plural markers). In other words, in this case the coder treats the absence of evidence as evidence of absence.

Coders fill in the answers to these questions by reading the most comprehensive publication on each language's grammar. They may supplement the information with additional publications if questions remain unresolved in the first source they consulted. Sources range from nineteenth century missionary grammars to reference grammars published in the 2020s. Occasionally, coders also consult articles on specific grammatical domains based on primary data.

3. Inferring the Level of Description of a Grammatical Feature

The comments provided by coders can reflect the complexity of some grammatical phenomena and the difficulty of striking a balance between fully describing a feature in a particular language and reducing it to a discrete typological variable. They also provide information on whether or not a feature is described for a language. By exploring which features are most often answered with a question mark and which features are most often provided with the comment "not mentioned", we can investigate how well different topics are described. To address this question, we conceptualize the level of description as a measure of whether there is positive evidence of presence or absence of a feature in a grammar: whether a feature is coded a certain way because a grammatical topic is treated explicitly, because there is not enough information in a source, or because a source does not include information on a certain topic.

To quantify the information provided in comments, we established five comment classes and assigned the comments to them. We then assessed the level of description by examining the coded values in combination with the comment classes (see Sections 4.1 and 4.3).

4. Dataset

The data used for the current study are of three types: the coded values for the 195 typological features, the optional comments clarifying the coding, and the bibliographic information for the source grammar. Below we outline the composition and processing of these three sources of data.

4.1 Classifying Comments

For each coded value the coder has the possibility to provide comments. The original dataset contains 117,041 free-form comments. Many of these comments are recurrent and generic, e.g. "not mentioned" or "not found", whereas others are unique and provide very specific additional information on the coding, e.g. "found in some positional and movement verbs" or "the causative has the same marking as the benefactor -*ɛ*". Before pre-processing, there were 42,276 comment types. Our goal was to assign each of these comments to one of five comment classes.

The comments were processed in R using the functionality of the stringR package (R Core Team, 2022; Wickham, 2021). We first applied a number of common text preprocessing techniques in the following order: removing punctuation, lowercasing, and stripping extra whitespace (trimming and squishing). This reduced the number of comment types to 41,464. We next implemented a number of further text-cleaning tasks using regular expressions: we removed comments which contained only references to the numbered examples in the cited grammar, as well as comments reflecting the history of coding of a specific value (e.g. comments with the string "autotranslated", which tag entries inherited from other typological projects). This procedure further reduced the number of comment types to 40,338.

We then proceeded with the classification of comments into five types using a rule-based approach. We first identified a number of relevant keywords and then incrementally built a system which combines this list of keywords with the information about the length of the comments to classify all comments into one of the five classes. For instance, all comments containing the character strings "not mentioned" or "no mention" were classified as the comment class "not mentioned". The five classes of comments are the following:

1. **"not mentioned"**: used when a specific grammatical phenomenon is not explicitly discussed in the grammar (e.g. "not mentioned, no reason to expect them", "no evidence found", "data very limited, unknown");
2. **"no category"**: used when a grammatical category does not exist in a language (e.g. "no noun classes", "nouns do not take any affixes", "there are only numerals for 1, 2, and 3");
3. **"specific"**: used for comments which provide further details or justification for the coding (e.g. "the majority of inalienable nouns occur with a nominal prefix", "only for the first person singular");
4. **"note on references or variety"** (e.g. "Table 1.2");
5. **"passim"**: presumably used by coders to highlight the fact that the feature does not have a dedicated discussion, but there are enough examples and side remarks to support a specific coding decision (e.g. "passim no evidence in data").

After the classification, the dataset contained 306,471 entries without comments ("NA"), 55,331 "not mentioned" comments, 39,917 "specific" comments, 7,852 "no category" comments, 1,734 comments with "note on references or variety", and 1,280 "passim" comments.

4.2 Bibliographic Data

The bibliographic information about the resources used for languages coded in Grambank is stored in the BibTeX format. It contains 3,951 references, which is the subset of references from Glottolog (Hammarström et al., 2021) that is used in Grambank. The sources include reference grammars, grammar sketches, dictionaries, and research articles. The BibTeX bibliography file was first parsed in R with the package bib2df (Ottolinger et al., 2019; R Core Team, 2022). For the purposes of the current study, we extracted the information about the publication date and cleaned it (e.g. removed any erroneous entries which do not represent the year, extracted the earliest date in case of reprints, etc.).

4.3 Data from Applying the Metric

The 441,663 coded values (i.e. the answers to the 195 questions) are distributed as in Table 1.[2]

Value	Meaning	Frequency
?	Unclear	74,540
0	No	226,809
1	Yes	104,671
2	Feature-specific	5,153
3	Feature-specific	1,412

Table 1: Frequency of coded values in Grambank

We aggregated the coded data and the classified comments to derive a level of description for a particular grammatical feature in a particular language. The aggregation was done in the following way:

- If the coded value is 0, 1, 2 or 3 (i.e. not ?) and if the coding is not accompanied by a comment, or it is accompanied by the classes "specific" or "note on the reference or variety", it is counted as "described". Every feature that is described is assigned the value 1.
- If the comment class is "not mentioned", we count the feature as "not described", irrespective of the coded value. For the purpose of our analysis, these were assigned the value 0.
- The rest of the combinations of coded values and comment classes do not affect the level of description (i.e. they were coded as NA).
- We further aggregated these values across languages. For each of the 195 features, we derived a 'description level' between 0 and 1 by taking the mean value of all 1s and 0s we assigned to that feature.

5. Results

5.1 General Results

After calculating the description level for each of the 195 features in our dataset (see Section 4.3), we grouped the features into 17 grammatical domains to obtain aggregated scores that are easier to interpret than individual feature scores. Grammatical domains consist of 3 to 19 features. For instance, the grammatical domain 'number' groups together 19 features covering productive nominal singular, dual, and plural marking, as well as associative plurals. The grammatical domain 'complex predicates' combines three features: a feature on light-verb constructions, a feature on verbal compounds, and a feature on serial verb constructions. Figure 2 provides a density plot of description level scores for the 17 domains in our dataset. It also indicates the median level of description for each grammatical domain. We ordered the domains from highest to lowest median scores.

Negation, adnominal modification, and core participant marking are the best described grammatical domains in our sample. On the other hand, comparative constructions, predicative possession, and complex predicates are the lowest-ranking grammatical domains according to our description metric. Not coincidentally, some of the grammatical domains that score highest are those that have seen a lot of typological research over the years. These typological efforts produced questionnaires that have also assisted language description. Some classic examples of this type of work are Kahrel & van den Berg (eds., 1994) on negation (see also Miestamo, 2016) and Dahl (1985) on tense, aspect, and mood. These grammatical domains are good examples where fruitful collaboration between typology and description has yielded both more insightful typologies and more comprehensive descriptions. Some grammatical domains that score lowest for description level are those where there are challenges in the interaction between typology and description. For instance, different types of complex predicates have proven difficult to define cross-linguistically, with various solutions proposed but no general consensus (e.g. Butt, 2003; Anderson, 2011; Haspelmath, 2016). Other lower scoring domains may be of interest to a typologist but may appear as less of a priority to someone describing a language, e.g. comparative constructions and predicative possession.

5.2 Variation Across Macro-Areas

Following the classification adopted by Glottolog (Hammarström et al., 2021) we assigned the languages in our sample to six macro-areas (see Hammarström & Donohue, 2014 for the discussion of the principles of this classification, see also Dryer, 1989). Figure 3 shows the variation in description per grammatical domain across macro-areas. Eurasia scores highest on almost all grammatical domains.

Considering that comparative constructions score low on level of description overall, Eurasia scores exceptionally high here. Australia tends to score on the lower end for most grammatical domains, with the exception of complex predicates, where it scores highest of all macro-areas. Papunesia generally scores similarly to Australia but has slightly better described comparative constructions and fewer descriptions of complex predicates. It scores lowest of all macro-areas on derivation and valency. Africa scores very low on complex predicates but has no clear outlier scores in other grammatical domains. Both North and South America score fairly high on almost all grammatical domains, except comparative constructions.

[2] In multi-state features, 1 does not mean "yes", but encodes a feature-specific meaning (see footnote 1).

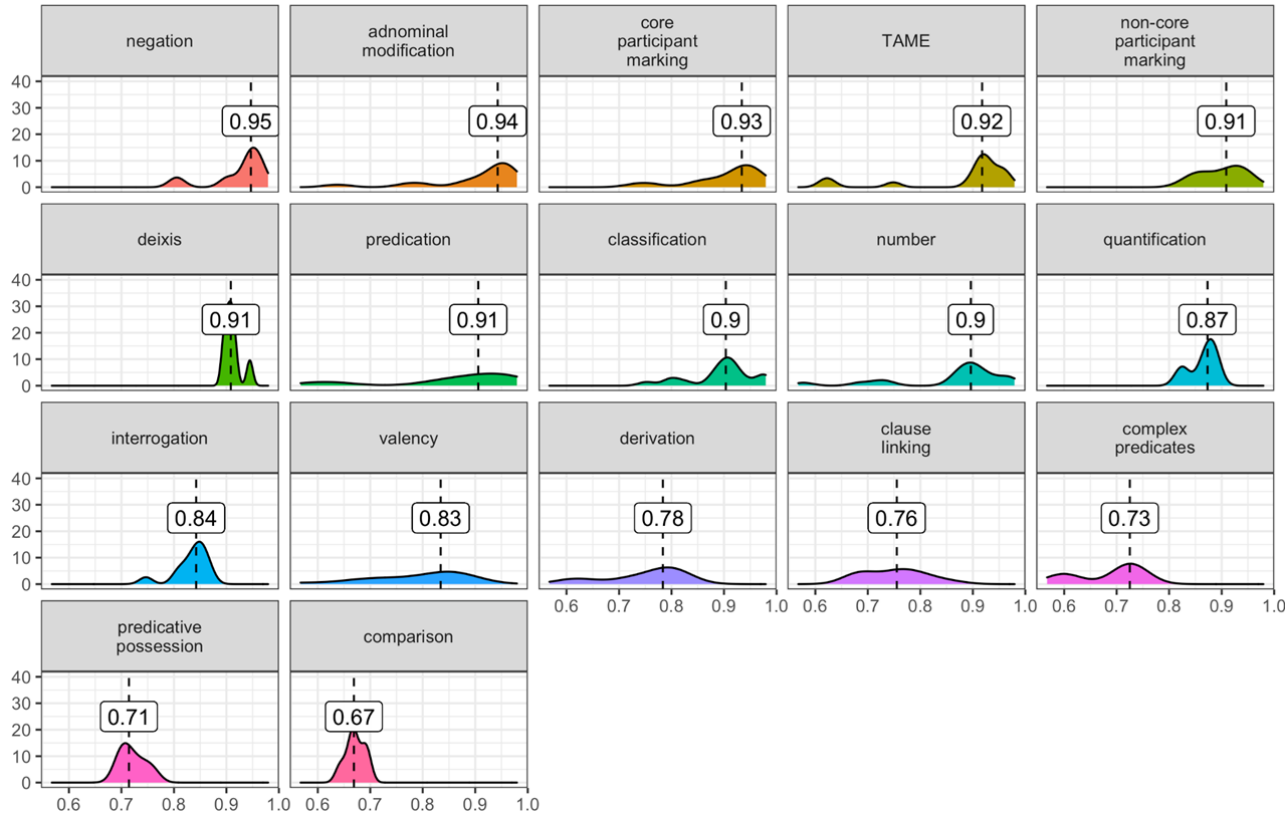

Figure 2: Density plot of description level scores per grammatical domain. The labeled scores are medians. The X axis displays description level scores and the Y axis displays density of the scores in the dataset.

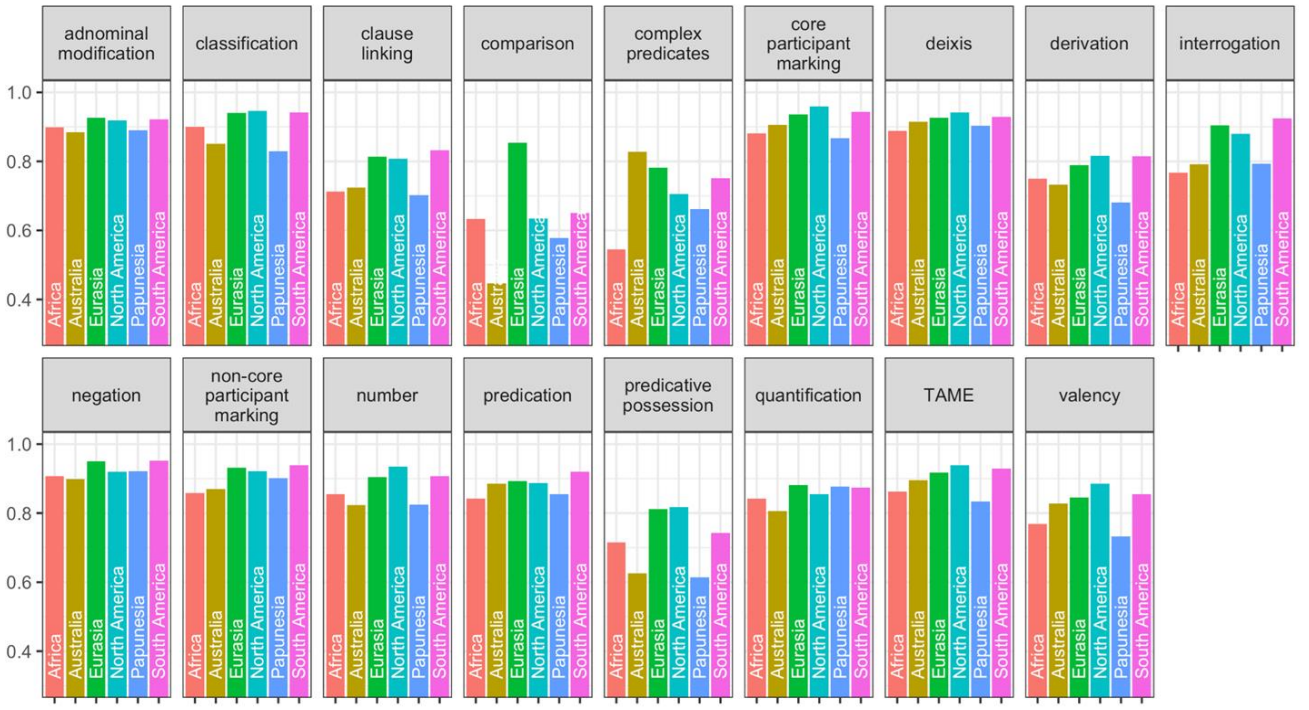

Figure 3: Bar plot of the mean description level scores per grammatical domain, per macro-area.

Differences across macro-areas may reflect a different focus of descriptive traditions. Comparative constructions are usually well described in Eurasian languages, possibly because some well-studied Eurasian languages express comparison with a dedicated marker (e.g. *than* in English and *mint* 'than, like' in Hungarian) (cf. Stassen, 1984, 2013). When describing a new language, linguists working on languages of Eurasia may actively look for such dedicated markers. In other areas, such as Australia, Papunesia, and North and South America, some languages do not have a dedicated marker to express comparatives, but instead use a so-called 'conjoined comparative' construction, expressing 'this house is bigger than that house' as *this house is big, that house is small* (e.g. Roberts, 1987: 135; cf. Stassen, 1984, 2013). Authors do not always consider such periphrastic constructions to be an integral part of the grammar of a language. When a grammar writing tradition develops, such considerations can make a grammatical domain seem less of a priority for future descriptions of languages.

5.3 Change Over Time

In this section we focus on two grammatical domains, viz. complex predicates and comparative constructions, to illustrate how Grambank can reveal changes in linguistic description over time.

Figure 4 shows the level of description of complex predicates per macro-area over time. In most macro-areas, there is a relatively steady increase in the description of complex predicates. They seem to remain relatively undescribed until they are picked up in North America in the 1960s and in Australia in the 1970s. After the 70s, there is an increase in their description around the world. In South America the description of complex predicates declines in the 2000s.

Figure 5 shows the level of description of comparative constructions per macro-area over time. The description of comparative constructions is quite uneven across the world and over time. The data from Africa and Papunesia suggest that comparative constructions received more attention in the early 20th century than after the 1950s. In Australia, research on comparative constructions has never been a priority. The description level has remained quite low in the 2000s in most parts of the world, with the exception of Eurasia, where comparatives are more commonly described and where the description level of comparatives has increased since the 1950s.

6. Discussion and Conclusion

The description level of the 195 grammatical features in Grambank varies across grammatical domains, across macro-areas, and over time. This variation can be interpreted in different ways, e.g. with reference to historically prominent researchers or trends, to different descriptive traditions or methodologies or to language structure – comparative constructions may simply be more common in Eurasia than in Australia, for instance, or serial verb constructions may be conspicuous in one language but difficult to detect in another. We leave these interpretations for future research. We divided the features into 17 grammatical domains, but it is possible that other ways of aggregating features yields results that are easier to interpret. It is our aim in this paper to illustrate that variation in description levels exists and can be revealed using typological databases.

We want to highlight that we detected this variation by focusing on often overlooked data in typological databases, viz. indications of uncertainty and comments left by the people who entered the data points. Major typological databases (including the World Atlas of Language Structures, Dryer & Haspelmath, 2013) do not collect or choose not to release the information on coding (un)certainty and comments on the individual data points. Quantifying uncertainty can be essential to a correct analysis and interpretation of the data in databases. For instance, we might not know for certain what a typical comparative construction looks like across languages if a substantial amount of data on comparative constructions is missing. But we might be able to quantify this uncertainty. Acknowledging and emphasizing uncertainty also feeds into the disciplines that focus on primary data collection: in this case, grammar writers can use Grambank as a source for specific unanswered questions they can focus on in their work.

Apart from these points, these types of data from Grambank can give us an idea of the topics which grammar writers have focused on in different traditions of grammar writing across space and through time. It can be used as a data source for historians of grammaticography to study grammar writing trends and the impact of pivotal researchers, institutions, and events on the discipline.

7. Author Contributions

All authors designed the study and wrote the final manuscript. JL and TW developed the measurement metric. AWM and JL classified the comments and processed bibliographic data. AWM, HJH, and HS aggregated the data. HJH and HS produced the visualizations. JL supervised the project.

Figure 4: Boxplots of the description scores of complex predicates over time (per decade) per macro-area. Boxplots show the median, first quartile and third quartile of description scores. Each language in the dataset is associated with one publication year.

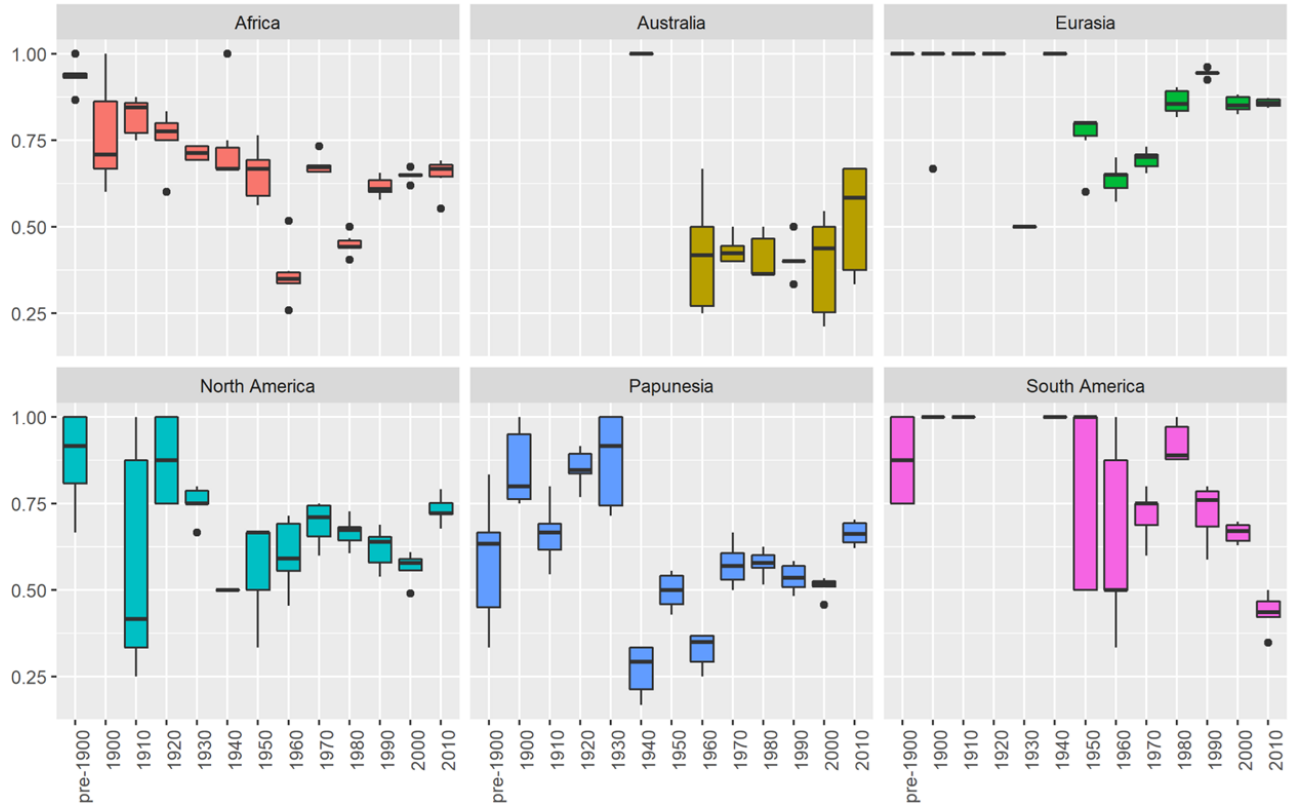

Figure 5: Boxplots of the description scores of comparative constructions over time (per decade) per macro-area. Boxplots show the median, first quartile and third quartile of description scores. Each language in the dataset is associated with one publication year.

8. Bibliographical References

Anderson, G. D. S. (2011). Auxiliary verb constructions (and other complex predicate types): A functional-constructional overview. *Language and Linguistics Compass* 5(11). 795–828.

Butt, M. (2003). The light verb jungle. *Harvard Working Papers in Linguistics* 9. 1–49.

Dahl, Ö. (1985). *Tense and aspect systems*. Oxford ; New York, NY: B. Blackwell.

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292.

Hammarström, H. & Donohue, M. (2014). Some principles on the use of macro-areas in typological comparison. In H. Hammarström & L. Michael (eds.), *Quantitative approaches to areal linguistic typology*, 167–187. Leiden: Brill.

Haspelmath, M. (2016). The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics* 17(3). 291–319.

Kahrel, P. & van den Berg, R. (eds.) (1994). *Typological studies in negation* (Typological Studies in Language 29). Amsterdam: John Benjamins.

Miestamo, M. (2016). Questionnaire for describing the negation system of a language. http://tulquest.huma-num.fr/sites/default/files/questionnaires/134/negation-questionnaire-for-SWL-volume-190207.pdf.

Ottolinger, P., Leeper, T., Salmon, M., Egeler, P & Esposito, E. X. (2019). *bib2df: Parse a BibTeX file to a data frame*. https://github.com/ropensci/bib2df.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, J. R. (1987). *Amele*. (Croom Helm Descriptive Grammar Series.) London: Croom Helm.

Seifart, F., Evans, N., Hammarström, H. & Levinson, S. C. (2018). Language documentation twenty-five years on. *Language* 94(4). e324–e345.

Stassen, L. (1984). The comparative compared. *Journal of Semantics* 3(1–2). 143–182.

Stassen, L. (2013). Comparative constructions. In Dryer, M. S. & Haspelmath, M. (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/121, Accessed on 2022-04-25.)

Wickham, H. (2021). *stringr: Simple consistent wrappers for common string operations*. http://stringr.tidyverse.org.

9. Language Resource References

Dryer, M. S. & Haspelmath, M. (eds.) (2013). *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. (2021). *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642 (Available online at http://glottolog.org, Accessed on 2022-01-06.)

Skirgård, H., Haynie, H. J., Hammarström, H., Blasi, D. E., Collins, J., Latarche, J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Maurits, L., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Robbeets, M., Abbas, K., Auer, D., Bakker, N., Barbos, G., Borges, R., Danielsen, S., Dorenbusch, L., Dorn, E., Elliott, J., Falcone, G., Fischer, J., Ghanggo Ate, Y., Gibson, H., Göbel, H., Goodall, J., Gruner, V., Harvey, A., Hayes, R., Heer, L., Herrera Miranda, R., Hübler, N., Huntington-Rainey, B., Ivani, J., Johns, M., Just, E., Kashima, E., Kipf, C., Klingenberg, J., König, N., Koti, K., Kowalik, R., Krasnoukhova, O., Lindvall, N., Lorenzen, M., Lutzenberger, H., Martins, T., Mata German, C., Meer, S., Montoya Samamé, J., Müller, M., Muradoglu, S., Neely, K., Nickel, J., Norvik, M., Oluoch, C. A., Peacock, J., Pearey , I., Peck, N., Petit, S., Pieper, S., Poblete, M., Prestipino, D., Raabe, L., Raja, A., Reimringer, J., Rey, S., Rizaew, J., Ruppert, E., Salmon, K., Sammet, J., Schembri, R., Schlabbach, L., Schmidt, F., Skilton, A., Smith, W. D., Sousa, H., Sverredal, K., Valle, D., Vera, J., Voß, J., Witte, T., Wu, H., Yam, S., Ye 葉婧婷, J., Yong, M., Yuditha, T., Zariquiey, R., Forkel, R., Evans, N., Levinson, S. C., Haspelmath, M., Greenhill, S. J., Atkinson, Q. D. and Gray, R. D. (submitted). Grambank data reveal global patterns in the structural diversity of the world's languages.

The Grambank Consortium (eds.). (2021). *Grambank*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://grambank.clld.org, Accessed on 2022-01-10.)