From Isolates to Families: Using Neural Networks for Automated Language Affiliation

Frederic Blum
DLCE / MCL Chair
MPI-EVA / University of Passau
Leipzig / Passau, Germany
frederic_blum@eva.mpg.de

Steffen Herbold
Chair of AI Engineering
University of Passau
Passau, Germany
steffen.herbold@uni-passau.de

Johann-Mattis List
MCL Chair
University of Passau
Passau, Germany
mattis.list@uni-passau.de

Abstract

In historical linguistics, the affiliation of languages to a common language family is traditionally carried out using a complex workflow that relies on manually comparing individual languages. Large-scale standardized collections of multilingual wordlists and grammatical language structures might help to improve this and open new avenues for developing automated language affiliation workflows. Here, we present neural network models that use lexical and grammatical data from a world-wide sample of more than 1,200 languages with known affiliations to classify individual languages into families. In line with the traditional assumption of most linguists, our results show that models trained on lexical data alone outperform models solely based on grammatical data, whereas combining both types of data yields even better performance. In additional experiments, we show how our models can identify long-ranging relations between entire subgroups, how they can be employed to investigate potential relatives of linguistic isolates, and how they can help us to obtain first hints on the affiliation of so far unaffiliated languages. We conclude that models for automated language affiliation trained on lexical and grammatical data provide comparative linguists with a valuable tool for evaluating hypotheses about deep and unknown language relations.

1 Introduction

One of the central tasks in historical linguistics is the grouping of languages into families. While this is often done to propose new language families, this task also includes the affiliation of individual languages into existing families. The affiliation of languages to a common language family or a subgroup within that family is usually carried out manually and relies on comparing individual languages in depth.

The traditional *comparative method* in historical linguistics uses lexical and grammatical features to assign individual languages to one of the more than 200 language families proposed so far (Osthoff and Brugmann, 1878; Anttila, 1972; Durie and Ross, 1996). The main part of this comparison starts with the initial assumption that two languages are related. However, such family relationships do not come as given and must first be established (Hoenigswald, 1978; Nichols, 1996; Campbell and Poser, 2008; Campbell, 2017). While the rise of computational methods in historical linguistics has largely replaced traditional techniques for subgrouping with computational approaches in phylogenetic reconstruction (Greenhill et al., 2020; Wu et al., 2020; Blum and List, 2023), the classical workflow for language *affiliation* of the comparative method is still considered the state-of-the-art in the field. But given that the affiliation of languages to families can be viewed as a computational *classification task*, it is possible to model the task in a setting that benefits from a large number of digital cross-linguistic datasets that have recently been published (List et al., 2022; Seifart et al., 2022; Wichmann et al., 2022; Skirgård et al., 2023; Blum et al., 2024b).

We show how *automated language affiliation* can be implemented computationally, how this approach can recover known long-distance genealogical relationships, and how it can shed new light on the affiliation of linguistic isolates and small language families. We train models on known language families and affiliate languages of unknown classification with the existing ones while aiming for world-wide coverage. When reporting the results, we pay special attention to the success of classifying small language families. The supervised training approach enables us to build upon the vast existing knowledge of historical linguistics. Having a large baseline of classes corresponding to true language families makes it possible to go beyond individual comparisons and to compare against all families simultaneously.

2 Background

Starting with the detection of the Indo-European (Bopp, 1816) and Uralic (Gyarmathi, 1799) families towards the beginning of the 19th century, linguists have been able to group the more than 7000 languages still spoken today into several hundred language families. However, while major parts of the traditional workflow of the comparative method have been intensively discussed and partially formalized, the first step of this workflow, the *affiliation* of languages to a family by *proving* their genetic relationship ("proof of relationship", see Durie and Ross, 1996) still lacks formalization. Although many quantitative and qualitative approaches have been proposed throughout the 20th and 21st centuries, only the comparative method has gained general acceptance. Some methods, such as the application of superficial "mass comparison" techniques to large wordlists (Greenberg, 1957), have been heavily criticized because of their lack of standardized data and clear criteria, and ultimately fell out of vogue (Campbell, 1988). Other methods, such as the proposal by Dolgopolsky (1964), who suggested looking for matching consonant classes to identify potential cognates, were ignored by most scholars for a long time before they were re-adopted in different contexts.

While scholars working in the traditional paradigm of the comparative method usually agree that lexical and grammatical evidence combined is best to prove language relationship (Campbell and Poser, 2008), the former is given preference in those cases where grammatical evidence is hard to obtain (Dybo and Starostin, 2008). Quantitative and statistical methods typically restrict themselves to either lexical *or* grammatical evidence.

Quantitative methods that take lexical data as their primary source can be divided into two basic types, depending on the evidence they try to obtain. Some approaches concentrate on the regularity of sound correspondences, trying to show that pairs of related languages exhibit significantly more matches in sound correspondences than unrelated ones (Ringe, 1992; Kessler, 2001; Blevins and Sproat, 2021). Other methods do not use specific sounds as observed in the languages in question and convert them to broader classes (*sound classes* or *consonant classes*, as originally proposed by Dolgopolsky 1964, see List 2014) to identify cognate words. These methods argue that words that share direct matches in a certain number of sound classes are likely to be etymologically related and that languages for which a certain number of matches can be observed are likely to be genetically related (Baxter and Manaster Ramer, 2000; Turchin et al., 2010; Kassian et al., 2023). Among the latter approaches, the *Automated Similarity Judgment Program* (ASJP, https://asjp.clld.org, Wichmann et al. 2022) deserves special mention, given that it can be seen as a first attempt to automatically classify as many of the languages of the world as possible with the help of phylogenetic methods. Using a specific sound class alphabet by which speech sounds are reduced to 40 classes, ASJP computationally compares word forms in language pairs, using traditional methods for sequence alignment (Wagner and Fischer, 1974), to infer distances between language pairs. These are later used to reconstruct a phylogenetic tree with the help of the *Neighbor-joining* algorithm (Saitou and Nei, 1987).

Quantitative methods that exclusively use grammatical data as their primary evidence have less frequently been proposed than their lexical counterparts. Despite this, Dunn et al. (2005) suggest that analyzing grammatical data could lead further back in time than the traditional comparative method. Today, these claims have lost supporters due to other studies indicating that grammatical features alone are less well suited for language classification because they diffuse easily in cases of language contact (Gray et al., 2010; Greenhill et al., 2010). The high potential for such diffusion is due to the limited amount of variation that grammatical features exhibit (Wichmann, 2017). However, these dynamics remain understudied, and we lack further case studies to analyze the behavior of grammatical data in large-scale classification settings.

We can now refine those early automated classification methods thanks to the release of new databases (Skirgård et al., 2023; Blum et al., 2025). The key difference between automated language affiliation and previous computational approaches to language classification is adopting a supervised learning approach. Our method directly benefits from previously established classifications based on the comparative method. It allows us to affiliate previously unclassified languages to existing language families, therefore strictly following an incremental approach to language classification. This approach also allows us to test hypotheses of deep language families, as we will show in our case studies, or to re-consider the affiliation of language isolates to other language families.

We test the model predictions in three case studies (Indo-European, Sino-Tibetan, and Uto-Aztecan) to evaluate the model classification on established language families sharing a long common history. Further, we test the affiliation of four language isolates: Basque, Bangime, Kusunda, and Mapudungun. We also show how this method can contribute to affiliating historical data of unknown classification to existing language families. At the same time, we preserve a conservative approach by including linguistic isolates in the training to restrain the model from unsubstantiated speculation in the form of false positives.

3 Materials and Methods

3.1 Cross-Linguistic Data

We use Lexibank (Version 2.1, Blum et al., 2025) and Grambank (Version 1.0.3, Skirgård et al., 2023), the two currently largest collections of standardized lexical and grammatical data, to train our model. Both databases are created and published using the *Cross-Linguistic Data Formats* (CLDF, Forkel et al. 2018, https://cldf.clld.org), in which common linguistic constructs, such as *language*, *concept*, and *sound*, are linked to reference catalogs, such as Glottolog (https://glottolog.org, Hammarström et al., 2024) for languages, Concepticon (https://concepticon.clld.org, List et al., 2025b) for concepts, and CLTS (https://clts.clld.org, List et al., 2024) for speech sounds. This ensures the standardization and comparability of data both within and across datasets. To test the quality of the data provided by Lexibank, we train an additional lexical model using the ASJP data (Version 20, Wichmann et al., 2022), which is also available in CLDF.

3.2 Data Vectorization

3.2.1 Lexibank and ASJP

For Lexibank, we use the concepts that are part of the Swadesh-100 list (Swadesh, 1955), and filter for languages which have a coverage of at least 50 such concepts across the dataset. However, the method can be used with any concept list that has been standardized in Concepticon (List et al., 2025b). For ASJP, we use their original 40-item wordlist (Holman et al., 2008).

We convert the lexical forms into vectors that can be used as input for the neural network. This processing step is illustrated in Table 1. Each segment in Lexibank and ASJP is standardized phonetically. To improve the comparability, we convert the sounds to their corresponding *Dolgopolsky class*, based on the sound classes proposed by Dolgopolsky (1964). Those ten classes are based on the likelihood of sound change between individual sounds. Such sound change happens frequently between sounds within a class of similar articulation, but not between sounds of different classes (for details, see List, 2014). Following the original proposal, only the first two consonant classes are considered. All additional consonants as well as vowels – except for word-initial vowels, which are considered a special consonant class H – are removed from the representation. We then convert the classes to indices of a one-hot vector of length 10, the number of Dolgopolsky sound classes, with the index of the attested class set to 1. All indices remain zero if a word is not attested in the language. All individual word vectors of a language are concatenated and given as input to the model.

3.2.2 Grambank

Grambank is a database of 195 typological features from more than 2,000 languages (Skirgård et al., 2023). Since most features are coded as binary, converting them into vectors is more straightforward than with Lexibank. Each value is mapped to an index in a one-hot vector of length two, and the index related to the attested value in a language is set to 1. Some word order features have three possible values, where '3' represents the meaning 'both 1 and 2 are attested'. Therefore, both indices in the vector are set to 1.

3.3 Baseline

We designed a fast and simple test to have a baseline comparison for our neural network models, based on the idea that the number of matching consonant classes can give hints on cognate words (Dolgopolsky, 1964; Turchin et al., 2010). In this baseline, we compare one language with unknown affiliation L_1 against all the other language varieties in the training data and assign it to the language family of the language L_{max} that shares the largest number of words matching in the two first consonant classes with L_1.

3.4 Neural Network Model Training

We carefully selected our model based on the properties of the data and compare the baseline with a neural network analysis. The multi-dimensional structure of the output data with 31 different language families leads to problems with simpler

Segments	Sound Cl.	Cons. Cl.	Vector
d e	TV	T –	[1000000000]
			[0000000000]
n a l a	NVRV	NR	[0100000000]
			[0010000000]
n e r a	NVRV	NR	[0100000000]
			[0010000000]
k o k o n	KVKVN	KK	[0001000000]
			[0000100000]

Parameter	Type	Value	Vector
Fixed order S/A/P	Bin	0	[10]
Fixed order S/A/P	Bin	1	[01]
Fixed order S/A/P	Bin	-	[00]
Order of NUM and N	1-3	1	[10]
	1-3	2	[01]
	1-3	3	[11]
	1-3	-	[00]

Table 1: Vectorization of words and grammatical features into vectors. The left table presents the conversion of segments in lexical forms into Dolgopolsky classes. Each class is assigned an index value in the vector, and the corresponding index set to 1. The same procedure applies to the binary Grambank features in the right table. In some cases, the value '3' represents the meaning 'both orders are attested', hence both vector indices are set to '1'.

model architectures like decision trees: to have reasonable tree structures that lead to a meaningful separation of 31 classes, the trees need to become very deep. While the issues are not purely additive, that both is the case at the same time further aggravate the problems with decision trees. Thus, while such an algorithm would have been desirable from an interpretability point of view, the resulting trees would have had to be very large, directly counteracting the interpretability. At the same time, large trees come with a high risk of overfitting. Many more powerful algorithms are based on such decision trees, e.g., random forests (Breiman, 2001). Moreover, the interpretability of such models is not very high: we could identify which features are important, but a detailed analysis of the relationship between features is not easily possible.

Other common modeling choices would be Naïve Bayes models or Multinomial Logistic Regression models. Arguably, these would have been suitable baselines. However, there is no reason to believe that such models, that are only capable of capturing linear decisions, would be sufficient to identify something as complex as language families. Distance-based approaches like k-Nearest Neighbor would degenerate to simply finding the most similar language in the sense of 'most binary markers are equal', which is also too simplistic. Instead, we opted for the simple matching of consonant classes as presented in the previous section as our baseline.

This leaves the neural networks, more specifically multi-layer perceptrons as 'standard', fairly shallow, modelling technique for fixed-length input that does not have a specific structure between features (like, e.g., images) and that does not require huge amounts of data. Multi-layer perceptrons nat-

urally build relationships between all features and the weighted sums that are computed at the layers are well-suited to deal with binary input data (as we, e.g., know from BoW-tokenization as well). Having many output neurons is also not a problem for this modelling approach, as the layers naturally untangle the data already into linearly separable dimensions, which can then be exploited at the last layer by, in fact, learning nothing more than logistic regression models for each class.

The neural network models are trained on the input vector of lexical, grammatical, or combined data as well as the labels of the language families. We train the models on an 80% sample of the data stratified by family labels and evaluate the performance of each model on the remaining 20% to find the best-performing one based on the balanced accuracy across language families (Brodersen et al., 2010). This split is necessary to avoid over-fitting the model to the training data (Dietterich, 1995). Even though a split into an additional development- or tuning-set would improve the model training even further (van der Goot, 2021), the data scarcity makes this a difficult enterprise. To test the robustness of our results despite the small sample size, we run each model 100 times with random seeds, so that different train/test sets are used (Gorman and Bedrick, 2019; Vabalas et al., 2019; Çöltekin, 2020). As an additional measure against over-fitting, we implement an early-stopping strategy during training (Ying, 2019).

3.5 Evaluation

We conduct four different experiments. The first experiment (§ 4.1) consists of a model comparison using a common subset of languages attested in ASJP, Lexibank, and Grambank. This comparison

17918

tests our baseline and the neural network models using an identical selection of 1227 languages. In the second experiment (§ 4.2) we evaluate how well our models can affiliate entire subgroups with the correct language family. We select three language families – Indo-European, Sino-Tibetan, and Uto-Aztecan – in which larger branches have split off considerably early during their evolution. We then train our models without the languages corresponding to these branches and test how automated language affiliation succeeds in assigning these languages to the correct family.

In the third experiment (§ 4.3), we investigate how our models affiliate language isolates, that is, languages which could so far not been convincingly assigned to *any* established language family. We use Bangime, Basque, Kusunda, and Mapudungun as exemplary case studies. In the fourth and final experiment (§ 4.4), we demonstrate with the example of Cararí (Natterer, 1817) how historical languages that have not been affiliated with any language family so far can be investigated with our automated language affiliation models.

3.6 Implementation

We implemented our models with PyTorch (Version 2.5.1, Ansel et al. 2024) in a feed-forward neural network with two hidden layers with a ReLU activation function. The hidden layers have a size of four times the number of language families. We process the datasets with SQLite (https://www.sqlite.org/) after conversion from CLDF via PyCLDF (Version 1.40.4, Forkel et al. 2025). We converted the segments to sound classes with LingPy (Version 2.6.13, List and Forkel 2023).

We use a weighted CrossEntropy loss function to better account for the many small language families present in our data (Zhang and Sabuncu, 2018). We used an Adam optimizer with a learning rate of 1e-3. The batch size we used is 2048. The hyperparameters were chosen on individual model comparisons between common values. Each training run consists of 5,000 epochs and is canceled if no improvement is made for 500 epochs. The models were trained on a V100 GPU node on a high-performance cluster, taking approximately 90 minutes. They can also be trained on ordinary computers due to the small size of the underlying data. We confirmed this in several test runs on ordinary laptops without GPU nodes with a working memory of 16GB.

4 Results

4.1 Initial Model Comparison

We compared our two baseline models (ASJP and Lexibank) to four neural network models for automated language classification (ASJP, Grambank, Lexibank, and Grambank/Lexibank combined). In this test, we selected all 1227 languages common to ASJP, Lexibank, and Grambank, which belong to language families with at least five members in the data. The classification target consisted of 31 different language families, including one for *isolates* (languages not assigned to any family).

The results in Figure 1 show the performance of all six models, based on the balanced accuracies from each of the 100 runs. The lexical models strongly outperform the model relying exclusively on grammatical data. The baseline models perform similarly to the Lexibank neural network model, while the neural network model using ASJP data falls off. The combined Lexibank/Grambank neural network model outperforms all models by about four points in accuracy.

Our results show that the simple baseline and computationally fast models perform on par or even slightly better than the models with more complex structures. From these results we can conclude that those baselines do a surprisingly good job in comparison with neural network models based on lexical data alone. The combined neural network model is the only exception, outperforming all other models in the comparison. This suggests that language affiliation benefits from a holistic approach combining lexicon and grammar data, confirming traditional assumptions from historical linguists. To further explore the potential of neural network approaches to automated language affiliation, we concentrate on the models based on Grambank, Lexibank, and their combined data in the following experiments.

4.2 Finding Deep Genealogical Relations

4.2.1 Indo-European

We conducted three case studies testing the affiliation of entire subgroups. The summary results across the three language families are presented in Table 2. Our first test is based on the Indo-European language family, spoken mainly in Europe, Northern India, and the Iranian plateau. The time depth for the initial split of the first branches, Anatolian and Tocharian, from the rest of the language family, is contested, with individual propos-
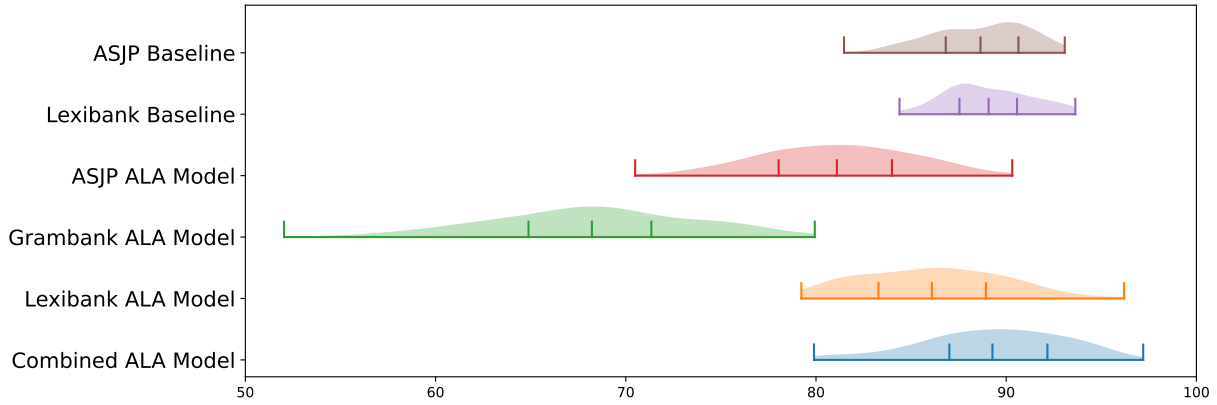
Figure 1: Classification results for all models, based on 100 runs with random seeds for the train/test split. Vertical lines indicate minimum, maximum, 25th, and 75th percentile, as well as the mean. The comparison is based on the balanced accuracy across language families to account for the difficulty of classifying small language families.

Family	Lexibank	Combined	Grambank
Indo-European	**100%**	/	/
Sino-Tibetan	90%	**95%**	38%
Uto-Aztecan	15%	**23%**	14%

Table 2: Results of the experiments for the deep language families across the Neural network models.

als ranging from 6,000 years ago (Anthony and Ringe, 2015) up to 8,000 years before present (Heggarty et al., 2023). We separate the languages from those two branches from the training data. The Lexibank model correctly classifies the languages as Indo-European (100%). Additional tests could not be carried out, since the languages in question are not coded for Grambank.

4.2.2 Sino-Tibetan

Estimates for the age of the Sino-Tibetan language family range between 5,900 years (Zhang et al., 2019) and 7,200 years (Sagart et al., 2019). In our test, we separate the languages of the Sinitic branch (the Chinese languages), commonly believed to be one of the earliest branches to split off from the ancestral proto-language, from the training data, and train our models without them.

Similar to the test on Indo-European languages, the Lexibank model successfully classifies the Sinitic branch to be part of the Sino-Tibetan language family, with an overall accuracy of 90%. The combined model surpasses the Lexibank model in accuracy, reaching 95.5%. The Grambank model classifies only one variety (wutu1241) primarily as Sino-Tibetan, whereas the other varieties from Sinitic tend to be classified either as Hmong-Mien (mand1415, wuch1236) or Austroasiatic (hakk1236). At least in the case of Hmong-

Mien, the classification of the grammatical model seems to point to an important role of areality when affiliating a language to a family based on grammatical data.

4.2.3 Uto-Aztecan

Uto-Aztecan is one of the largest language families spoken in North America, located primarily on the west coast of the Pacific (Campbell, 1997). It consists of two main branches, the northern and the southern languages. This split is estimated to have occurred between 3,258 and 5,025 years ago (Greenhill et al., 2023). Both the lexical and the grammatical model fail in this task, correctly classifying the northern languages only in 15% and 14% of the cases, respecively. Instead, the Lexibank model has a high percentage of Unclassified (63%), whereas the Grambank model classifies the languages as Arawakan, Nuclear-Macro-Je, Cariban, or Pama-Nyungan. The combined model achieves 23%, surpassing both other models. Overall, the results for Uto-Aztecan are worse than for the other two language families we have tested.

4.3 Affiliation of Isolates

We test our models on four language isolates that have found recent attention in the literature. Bangime, a language spoken in central-eastern Mali, is considered an isolate that has been in contact with many surrounding languages for such a long time that the speakers consider it related to the neighboring Dogon languages (Hantgan and List, 2022). The data on Bangime in Lexibank is based on Hantgan and List (2022). Basque has been recently hypothesized to be part of a Proto-Euskarian-Indo-European language family using
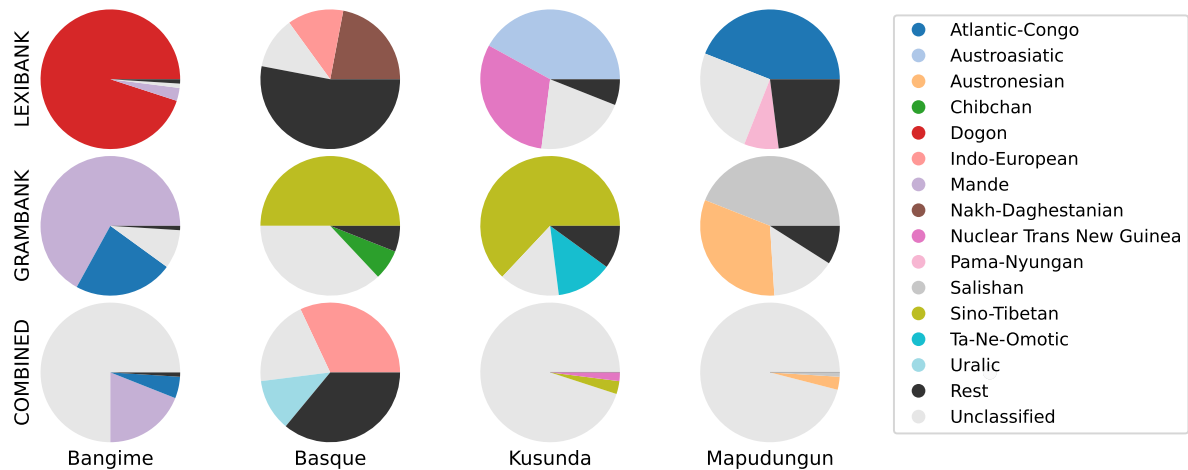
Figure 2: Results for the experiment on isolate affiliation. Results are limited to the first three families to which an isolate is affiliated, showing the proportion of the remaining families under the label *Rest* in the charts.

both traditional and computational methodology (Forni, 2013; Blevins, 2018; Blevins and Sproat, 2021). The data on Basque in Lexibank is taken from (Dellert et al., 2020). Kusunda is spoken in Nepal, but has previously been hypothesized to be related to Papuan languages (Whitehouse et al., 2004), the Dene family, or Yenisseien (Gerber, 2017; van Driem, 2014). However, no classification has found broad acceptance, and the language remains classified as isolate. A new wordlist recorded in 2020 by Aaley and Bodt (2020) has been included in Lexibank. Finally, Mapudungun is a language spoken in south-eastern South America. The available evidence suggests that its genealogical relations with other languages are restricted to close relatives that have since become dormant, with no indication of deeper connections. The data for Mapudungun is based on Tadmor (2009) with phonetic mappings provided by Miller et al. (2020).

The results for all isolates are presented in Figure 2. Bangime is classified consistently as Dogon in the Lexibank model (95%) and as Mande (67%) or Atlantic-Congo (23%) in the Grambank model. Consequently, the combined model primarily has Bangime unclassified (75%).

Basque on the other hand remains unclassified only in few iterations in both the Lexibank (12%) and the Grambank model (37%). The former suggests an affiliation with Indo-European (13%) or Nakh-Dagestanian (22%), whereas the latter tends to propose an affiliation with Sino-Tibetan (50%). The combined model proposes an affiliation with Indo-European (32%) or Uralic (12%) but also includes the unclassified affiliation (20%).

In the Lexibank model, Kusunda is affiliated either with Nuclear Trans-New Guinea (31%), Austroasiatic (41%), or left unclassified (21%). The Grambank model mostly affiliates Kusunda with the Sino-Tibetan family (63%). The combined model mostly proposes no affiliation of Kusunda with other language families (95%).

Mapudungun is split between Atlantic-Congo (44%) or Unclassified (25%) for the Lexibank model. The Grambank model mostly suggests a Salishan affiliation (44%) or no affiliation (15%). The combined model, again, finds no clear affiliation pattern (96%).

Given that the status of all four languages concerning their affiliation with other language families has been disputed without a result for a long time, it would go too far to speculate on any particular finding presented in the charts here. What we can see, however, is a tendency for the combined model to affiliate the four isolates with the large group of unclassified (i.e., isolate) languages in our sample. The Lexibank and Grambank models differ quite remarkably in this regard, often giving preference to particular language families that often reflect areal closeness or previously proposed theories on the origin of certain varieties.

The Lexibank model classifies Bangime as a Dogon language, reflecting the well-known fact that the lexicon shares many words with this family (Hantgan et al., 2022), whereas the grammatical model suggests an affiliation with Mande languages. If we consider the results from the previous tests indicative and the affiliation of the grammatical model as being prone to contact phenom-

ena, heavy grammatical restructuring for Bangime based on Mande languages seems likely. This mirrors previous arguments that typological features reflect geographical distributions, rather than genealogical relations (Greenhill et al., 2010; Gray et al., 2010; Donohue et al., 2011), and contributes directly to our understanding of the linguistic layers of Bangime, which so far focused on the lexicon (Hantgan and List, 2022). Given that genetic studies show that the speakers are unique in the region concerning their genetic history (Babiker et al., 2020), it is reasonable to classify Bangime as an isolate, as does the combined model.

For Kusunda, we find a lexical link to Nuclear Trans-New Guinea languages, while the grammar fits well into the Sino-Tibetan neighborhood where the language is spoken. This finding aligns with previous research that suggests that Kusunda is genealogically a Trans-New Guinea language that has recently migrated to its current location (Whitehouse et al., 2004; van Driem, 2014). This is where Kusunda would have come into contact with Sino-Tibetan languages, adopting several grammatical features from this language family.

The closeness between Basque and Indo-European suggested by the combined model and to some degree the lexical model has been recently suggested using both traditional and computational methods (Forni, 2013; Blevins, 2018; Blevins and Sproat, 2021). On the other hand, the connection of Basque with Sino-Tibetan suggested by the Grambank model and the connection with Nakh-Daghestanian suggested by the Lexibank model would fit with the far-ranging proposal of a Sino-Caucasian macro-family, in which scholars at times include Basque (Starostin, 2017). Both proposals, however, are far from accepted in current literature. The results show how such conflicting evidence can also be reflected in quantitative models.

For Mapudungun, few prominent classification hypotheses exist in the literature (Campbell, 2012). All our models tend to leave it unaffiliated, with some suggesting a Atlantic-Congo (lexical), Austronesian (grammatical), or Salishan (grammatical) affiliation. We are not aware of previous mentions of those affiliations. One way to explore them is to analyze the individual shared data points to evaluate the possibility of chance similarities. This points also to the major drawbacks of the neural network approach, as it does not allow us to directly determine the concrete words or grammatical features that contribute to a particular decision.

4.4 Classification of Unaffiliated Languages

The lexical model can also easily affiliate newly identified or historical languages documented in ancient sources to language families. To illustrate this, we affiliated data from Cararí, a language documented during the early 19th century by Johann Natterer at the confluence of the Mucuim River and the Purús in the Brazilian Amazon (Natterer, 1817). According to Adelaar and Brijnen (2014), some of the languages documented by Natterer could not yet be classified – Cararí being one of them. However, they suggest that an Arawak affiliation might be identifiable with more comparative work done.

In the first step, we digitized the data and converted the wordlist into the Lexibank format. Figure 3 presents the original and the standardized versions of the documented word forms. Through this formatting, we can easily input this data into the lexical model by adding the database with the same workflow as the rest of the data. The results are clear: the Lexibank model strongly suggests an Arawak affiliation of the language (91%).

5 Discussion and Conclusion

We have presented a new approach to *automated language affiliation* in this study. The lexical and combined models show very good classification results despite their simple architecture. The baseline models also performed extremely well, only being surpassed by the combined lexical+grammatical neural network model.

Given the strong performance, we can use the models for downstream tasks such as testing long-distance relationships between subgroups of language families and the affiliation of unclassified language varieties. The first experiment shows that our models correctly identify such long-distance relationships at a time depth of 5,000 years and more. When testing the affiliation of linguistic isolates, our models reflect the actual discussions in the linguistic literature. This shows that our method can extract information from sparse data in a way that can be compared with language-specific studies.

From our findings, we can make three conclusions, (a) automated language affiliation achieves promising results even for language relations way back in time, (b) grammar alone is not sufficient for a successful affiliation, and (c) combined models seem to work very well, reflecting that languages are best affiliated by using lexicon plus a bit of grammar (Campbell and Poser, 2008). However,
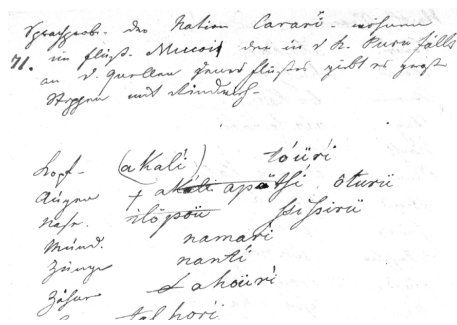
Figure 3: The left shows some of the original Cararí data published by Natterer (1817). The right shows our standardization of the same entries using the EDICTOR tool (List et al., 2025a), with the original transcriptions given in the column 'Form'.

the combined data is only available for a small subset of languages. In cases of data scarcity, the lexical models are almost on par with the combined model and strongly outperform comparable models based on grammatical data (Holman et al., 2008).

A particular use case of our method is the affiliation of historical data with contemporary language families. In many cases, the material is so scarce that cannot be affiliated with any language family based on a traditional analysis alone. Our models provide a quantitative perspective, and the case of Cararí shows that it might even be possible to provide strong arguments for a specific affiliation.

We do not see automated language affiliation replacing the traditional comparative method. While our model of language affiliation can be used to evaluate hypotheses about long-range genealogical relationships between languages, it cannot provide conclusive proof in favor or against such relationships. The strength of our approach is a principled comparison of the data across a large range of languages that can find hints at a shared descent between languages. This can be a starting point for a linguistic evaluation of such hypotheses, which would be verified, for example, through the means of cognate reflex prediction (Blum et al., 2024a) or other traditional workflows (Durie and Ross, 1996). The task of evaluating those relations will have to remain with the comparative method, which could now target specific proposals to shed further light on the history of human languages.

Data Availability

Code and data underlying this study are curated on GitHub (v1.0, https://github.com/lexibank/ala/releases/tag/v1.0) and archived with Zenodo (10.5281/zenodo.15524321, Version 1.0).

Limitations

Data for many small language families is scarce, and all models showed that classification is much more difficult for small language families. Considering the availability of training data in those settings, this is expected. While projects like ASJP (Wichmann et al., 2022) opt for smaller concept lists from more languages due to this reason, we are convinced that there are several advantages of using the explicit orthography conversion from Lexibank, even though this means a trade-off in terms of languages available.

This also influences the comparability of our models. Only a subset of around 1,200 languages is available in all three major datasets. We only use this set of common languages to compare all models using the same data. Ideally, we would have a larger subset with a better representation of the world-wide linguistic diversity.

There are many options for further baseline comparisons, of which we could only present one. The great success of the baseline shows that it would be worthwhile to explore other approaches that fit the data, and that it might not be necessary to design complex model architectures for a successfull automated language affiliation.

Competing Interests

The authors declare no competing interests.

Ethical Considerations

We see no potential risks involved, except for the potential of creating unwarranted hypotheses about long-distance genealogical relationships between languages. We call for all users of our models to sensibly analyze the linguistic data behind the proposed classifications.

Acknowledgements

References

Uday Raj Aaley and Timotheus A. Bodt. 2020. New data on Kusunda: A list of 250 concepts. *Computer-Assisted Language Comparison in Practice*, 3(4):1–4.

Willem F. H. Adelaar and Hélène B. Brijnen. 2014. Natterer's linguistic heritage. *Archiv Weltmuseum Wien*, 63–64:162–183.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, pages 929–947. ACM.

David W. Anthony and Don Ringe. 2015. The Indo-European homeland from linguistic and archaeological perspectives. *Annual Review of Linguistics*, 1(1):199—219.

Raimo Anttila. 1972. *An Introduction to Historical and Comparative Linguistics*. The Macmillan Company, New York.

Hiba Babiker, Jeffrey Heath, Floyd Reed, Stephan Schiffels, and Russell D. Gray. 2020. Striking genetic diversity among populations of West Africa uncovers the mystery of a language isolate. *SSRN Current Biology*.

William H. Baxter and Alexis Manaster Ramer. 2000. Beyond lumping and splitting: Probabilistic issues in historical linguistics. In April McMahon Colin Renfrew and Larry Trask, editors, *Time Depth in Historical Linguistics*, pages 167–188. McDonald Institute for Archaeological Research, Cambridge.

Juliette Blevins. 2018. *Advances in Proto-Basque Reconstruction with Evidence for the Proto-Indo-European-Euskarian Hypothesis*. Routledge, New York.

Juliette Blevins and Richard Sproat. 2021. Statistical evidence for the Proto-Indo-European-Euskarian hypothesis: A word-list approach integrating phonotactics. *Diachronica*, 38(4):506–564.

Frederic Blum, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List. 2025. Lexibank 2: precomputed features for large-scale lexical data [version 1; peer review: awaiting peer review]. *Open Research Europe*, 5:1–19.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Johann-Mattis List. 2024a. Cognate reflex prediction as hypothesis test for a genealogical relation between the Panoan and Takanan language families. *Scientific Reports*, 14(30636):1–12.

Frederic Blum, Carlos Barrientos, Roberto Zariquiey, and Johann-Mattis List. 2024b. A comparative wordlist for investigating distant relations among languages in Lowland South America. *Scientific Data*, 11(92):1–9.

Frederic Blum and Johann-Mattis List. 2023. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.

Franz Bopp. 1816. *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der*

griechischen, lateinischen, persischen und germanischen Sprache. Windischmann, Karl Joseph Hieronymus, Andreäische Buchhandlung, Frankfurt am Main.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*, pages 3121–3124, Istanbul, Turkey. IEEE.

Lyle Campbell. 1988. Review Article: Language in the Americas. By Joseph H. Greenberg. Stanford, California: Stanford University Press, 1987, Pp. x.,438. *Language*, 64(3):591–615.

Lyle Campbell. 1997. *American Indian Languages: The Historical Linguistics of Native America*. Oxford University Press, New York.

Lyle Campbell. 2012. Classification of the indigenous languages of South America. In Lyle Campbell and Verónica Grondona, editors, *The Indigenous Languages of South America*, pages 59–166. De Gruyter, Berlin, Boston.

Lyle Campbell. 2017. How to show languages are related: Methods for distant genetic relationship. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 262–282. Blackwell Publishing, Malden, Oxford, Melbourne, Berlin.

Lyle Campbell and William John Poser. 2008. *Language classification: History and method*. Cambridge University Press, Cambridge.

Çağrı Çöltekin. 2020. Verification, reproduction and replication of NLP experiments: a case study on parsing Universal Dependencies. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 46–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel 1 · Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. Northeuralex: a wide-coverage lexical database of Northern Eurasia. *Language Resources & Evaluation*, 54:273–301.

Tom Dietterich. 1995. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327.

Aron B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Voprosy Jazykoznanija*, 2:53–63.

Mark Donohue, Simon Musgrave, Bronwen Whiting, and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. *Language*, 87(2):369–383.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Mark Durie and Malcolm Ross. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press, New York, Oxford.

Anna Dybo and George S. Starostin. 2008. In defense of the comparative method, or the end of the Vovin controversy. In I. S. Smirnov, editor, *Aspekty komparativistiki*, volume 3, pages 119–258. RGGU, Moscow.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Robert Forkel, Christoph Rzymski, and Sebastian Bank. 2025. *PyCLDF [Software, Version 1.40.4])*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gianfranco Forni. 2013. Evidence for Basque as an Indo-European language. *Journal of Indo-European Studies*, 41(1/2):39.

Pascal Gerber. 2017. The Dene-Kusunda hypothesis: a critical account. *Man in India*, 97(1):111–204.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Russell D. Gray, David Bryant, and Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3923–3933.

Joseph H. Greenberg. 1957. *Essays in linguistics*. The University of Chicago Press, Chicago.

Simon J. Greenhill, Quentin D. Atkinson, Andrew Meade, and Russell D. Gray. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693):2443–2450.

Simon J. Greenhill, Hannah J. Haynie, Robert M. Ross, Angela M. Chira, Johann-Mattis List, Lyle Campbell, Carlos A. Botero, and Russell D. Gray. 2023. A recent northern origin for the Uto-Aztecan family. *Language*, 99(1):81–107.

Simon J. Greenhill, Paul Heggarty, and Russell D. Gray. 2020. Bayesian phylolinguistics. In Barbara S. Vance Richard D. Janda, Brian D. Joseph, editor, *The Handbook of Historical Linguistics, II*, chapter 11, pages 226–253. John Wiley & Sons, Hoboken.

Sámuel Gyarmathi. 1799. *Affinitas lingua Hungaricae cum linguis Fennicae originis grammatice demonstrata. Nec non vocabularia dialectorum Tataricarum et Slavicarum cum Hungarica comparata.* Dieterich, Göttingen.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog database [Dataset, Version 5.1]*. Max-Planck Institute for Evolutionary Anthropology, Leipzig.

Abbie Hantgan, Hiba Babiker, and Johann-Mattis List. 2022. First steps towards the detection of contact layers in Bangime: A multi-disciplinary, computer-assisted approach [version 2; peer review: 2 approved]. *Open Research Europe*, 2(10):1–27.

Abbie Hantgan and Johann-Mattis List. 2022. Bangime: secret language, language isolate, or language island? a computer-assisted case study. *Papers in Historical Phonology*, 7:1–43.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(eabg0818):1–12.

Henry M. Hoenigswald. 1978. The annus mirabilis 1876 and posterity. *Transactions of the Philological Society*, 76(1):17–35.

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3–4).

Alexei S. Kassian, George Starostin, Mikhail Zhivlov, and Sergey A. Spirin. 2023. Calibrated weighted permutation test detects ancient language connections in the circumpolar area (Chukotian-Nivkh and Yukaghir-Samoyedic). *Journal of Historical Linguistics*.

Brett Kessler. 2001. *The significance of word lists*. CSLI Publications, Stanford.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List, Cormac Anderson, Christph Tresoldi, Tiagoand Rzymski, and Robert Forkel. 2024. *CLTS. Cross-Linguistic Transcription Systems [Dataset, Version 2.3.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Frederic Blum, and Kellen Parker van Dam. 2025a. *EDICTOR 3. A Web-Based Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]*. MCL Chair at the University of Passau, Passau.

Johann-Mattis List and Robert Forkel. 2023. *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.13]*. MCL Chair at the University of Passau, Passau.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):1–16.

Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos, Christoph Rzymski, Simon Greenhill, and Robert Forkel. 2025b. *CLLD Concepticon [Dataset, Version 3.4.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists. *PLOS One*, 15(12):e0242709.

Johann Natterer. 1817. *Wortlisten von Indianersprachen in Brasilien*. E-Manuscripta, Zurich.

Johanna Nichols. 1996. The comparative method as heuristic. In Mark Durie and Malcolm Ross, editors, *The comparative method reviewed*, pages 39–71. Oxford University Press, New York.

Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, volume 1. Hirzel, Leipzig.

Donald A. Ringe. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82(1):1.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Frank Seifart, Ludger Paschen, and Matthew Stave. 2022. *Language Documentation Reference Corpus (DoReCo) [Dataset, Version 1.2]*. Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin & Lyon.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

George Starostin. 2017. Sino-Caucasian and Sino-Yeniseian. In Rint Sybesma, editor, *Encyclopedia of Chinese language and linguistics*. Brill, Leiden, Boston.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Uri Tadmor. 2009. Loanwords in the world's languages. findings and results. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the world's languages. A comparative handbook*, pages 55–75. de Gruyter, Berlin, New York.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, 3:117–126.

Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11):e0224365.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Georg van Driem. 2014. A prehistoric thoroughfare between the Ganges and the Himalayas. In Tiatoshi Jamir and Manjil Hazarika, editors, *50 Years after Daojali-Hading: Emerging Perspectives in the Archaeology of Northeast India*, pages 60–98. Research India Press, New Delhi.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Paul Whitehouse, Timothy Usher, Merritt Ruhlen, and William S.-Y. Wang. 2004. Kusunda: An Indo-Pacific language in Nepal. *Proceedings of the National Academy of Sciences*, 101(15):5692–5695.

Søren Wichmann. 2017. Genealogical classification in historical linguistics. *Oxford Research Encyclopedia of Linguistics*.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. *The ASJP database (v20)*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mei-Shin Wu, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data*, 6(1):2.

Xue Ying. 2019. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022.

Menghan Zhang, Shi Yan, Wuyun Pan, and Li Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the late Neolithic. *Nature*, 150(569):112–115.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8792–8802, Red Hook, NY, USA. Curran Associates Inc.