Bangime: Secret Language, Language Isolate, or Language Island?

Abbie Hantgan, Johann-Mattis List

Bangime: Secret Language, Language Isolate, or Language Island?

Abbie Hantgan[1]* and Johann-Mattis List[2]
[1] Dynamique du Langage, Lyon
[2] Max Planck Institute for the Science of Human History, Jena
* corresponding author: ahantgan@gmail.com

Draft, September 2018, to appear in *Journal of Language Contact*

Abstract

We report the results of a qualitative and quantitative lexical comparison between Bangime and neighboring languages. Our results indicate that the status of the language as an isolate remains viable, and that Bangime speakers have had different levels of language contact with other Malian populations at different time periods. Bangime speakers, the Bangande, claim Dogon ancestry, and the language has both recent borrowings from neighboring Dogon varieties and more rooted vocabulary from Dogon languages spoken to the east from whence the Bangande claim to have come. Evidence of multi-layered long-term contact is clear: lexical items have even permeated even core vocabulary. However, strikingly, the Bangande are seemingly unaware that their language is not intelligible with any Dogon variety. We hope that our findings will influence future studies on the reconstruction of the Dogon languages and other neighboring language varieties to shed light on the mysterious history of Bangime and its speakers.

Keywords: comparative method, historical linguistics, African language isolates, sound correspondences, computational linguistics

1 Introduction

Spoken by an estimated 1,500 people among seven villages situated in a cove at the western edge of the Bandiagara escarpment in central-eastern Mali, *Bangime* translates as 'secret language' in many of the Dogon languages spoken along the eastern-edges of the cliff range. Bangime speakers, correspondingly depicted as *Bangande*, identify themselves and their language as Dogon, and yet they are unaware that the terms used for their group and language are actually exonyms with the meaning of 'secret'. While the reason Bangande essentially do not have an endonym remains part of their mystery, their unfamiliarity with the lexical root [bang-], meaning 'secret, hidden, furtive' is due to the fact that the Dogon speakers in their immediate vicinity do not use the same lexical root, which alone suggests that Bangime-speakers once co-existed with Dogon-speaking populations who differ from those near them now.

This paper employs qualitative and quantitative methods for historical lexical comparison to propose scenarios to explain Bangime and its speakers' relationship to surrounding Dogon languages and their speakers along Mali's well-known Bandiagara escarpment. Scholars also contest foregone conclusions about Dogon ethnographic and linguistic affiliations. Thus, we compare portions of both the Dogon and Bangime lexica with those from neighboring Mande, Atlantic, and Songhay languages in order to rule out

any genetic affiliation between Bangime and languages beyond the Dogon escarpment. Placing Bangime in the context of other languages in Mali further supports its status as a linguistic isolate. Not only is it disparate from the Dogon varieties, it also shares little, if anything, with the surrounding languages - Fulfulde, Songhay, or those of the Mande group - although our findings support contact with all of them, and yet at different stages in history.

Our quantitative studies are based on state-of-the-art methods for automatic word comparison across various languages (List et al. 2017b). We propose a combination of conservative approaches that closely model the traditional comparative method (Ross and Durie 1996) while taking regular sound correspondences across all languages in the dataset into account, as in the LexStat method (List 2012a), and using simpler approaches that mainly pick up surface similarities between words, as in the Sound-Class-Based Alignment method (List 2012b). By combining these approaches and comparing their results directly, we can automatically identify potential layers of contact between Bangime and its neighboring varieties that reflect different contact periods. These layers can then be further analyzed qualitatively by comparing the automatic findings in detail.

In this study, we limit our analyses to lexical similarities but note that many grammatical features in Bangime, discussed in detail by Heath and Hantgan (2018), are also distinct from those found at least among Niger-Congo and Nilo-Saharan families. By investigating socio-linguistic, cultural, and historic implications, we hope to suggest avenues for future research.

2 Background

2.1 Previous Research on Bangime

Bangime is considered to be one of only four undisputed African isolates (Blench 2017, p. 167).[1] However, as Campbell (2016, 2017) advises, a language's confirmed classification as an isolate should not preclude a thorough investigation into its origins. Although Bangime was once classified as Dogon, those who have discussed the Dogon languages and peoples have long recognized the Bangime speech community as an outlier (Calame-Griaule 1956; Hochstetler et al. 2004; Plungian and Tembine 1994). While Bertho (1953, pp. 413–414) went so far as to state that Bangime was distinct, not only from Dogon, but also from Fulfulde and Bozo, Blench (2005) was the first researcher to suggest it was an isolate. According to the Dogon and Bangime Linguistics Project (Moran et al. 2016, http://dogonlanguages.org), within the variation attested among the now estimated 22 distinct Dogon languages, the lowest limit for mutual intelligibility based on lexical estimates is 32 percent (Prokhorov et al. 2012). In contrast, we now estimate that Bangime shares less than 20 percent of its core vocabulary with Dogon, and even these few lexical items, such as numerals, were likely borrowed long-ago from Dogon languages for the sake of identity-inclusion.

Bangime has been called by many other names in the literature. Among those discussed by Hantgan (2013, p. 5), one remained a mystery. Blench (2005, p. 1) describes the "intrusive -ri-" as it appears in */báŋeri mé/* (Calame-Griaule 1956, p. viii). Table 1 shows, in Dogon words for 'hide, conceal', the [-ri-] suffix and its allomorphs represent the Dogon causative, or transitive, morpheme.
The suffix [-mɛ] or [-jɛ] in Bangime denotes a language from the name of the speakers (Hantgan 2013, p. 112), and yet [-jɛ] also corresponds with the medio-passive suffix among the Dogon languages. Thus, the word *bangime* could be seen as a mix between the Dogon root [bang] and the Bangime suffix [-mɛ], or simply an alternate (and often attested) pronunciation of the medio-passive form of the verb [bang-i-jɛ], meaning 'it is hidden' among the Dogon languages with this root.

In either case, what is surprising is that, shown in Figure 1, the languages with words that most resemble the term listed by Calame-Griaule (first column, Table 1) are those that are currently spoken furthest away

[1] Campbell (2016) and Simons and Fennig (2018) list ten isolate or unclassified languages, only four of which are considered to be undisputed.

Language	IPA	Language	IPA
Ben Tey	bàŋgì-rí	Bunoge	jɔ́gɛ̀
Gourou	bàŋà-rã́	Tiranige Diga	dʒíná-ŋgó
Jamsay	bàŋà-rã́	Mombo	dábú-rè
Tebul Ure	bàŋgì-rí	Penange	kúj-rɛ̀
Togo Kan	bàŋú-rù	Donno So	dʒɔ̀ɔ̀-rɔ́
Tommo So	bàŋú-ndá	Nanga	dăw-rí
Yanda Dom	báá-ndɛ́	Najamba	síbí-r
Yorno So	bàŋá-rá	Toro Tegu	sútù
Toro So	bàŋì-rí	Perge Tegu	súgú-ró

Table 1: Dogon terms for concept ‘HIDE, CONCEAL SOMETHING’

from where Bangime is spoken. Among the languages spoken closest to where Bangime is spoken (third column), none uses anything like the term that most closely resembles the name of their language and people. Therefore, it is likely that the name of the language and speakers was given at a time when, and place where, the ancestors of the Bangande and the now eastern Dogon were in contact.

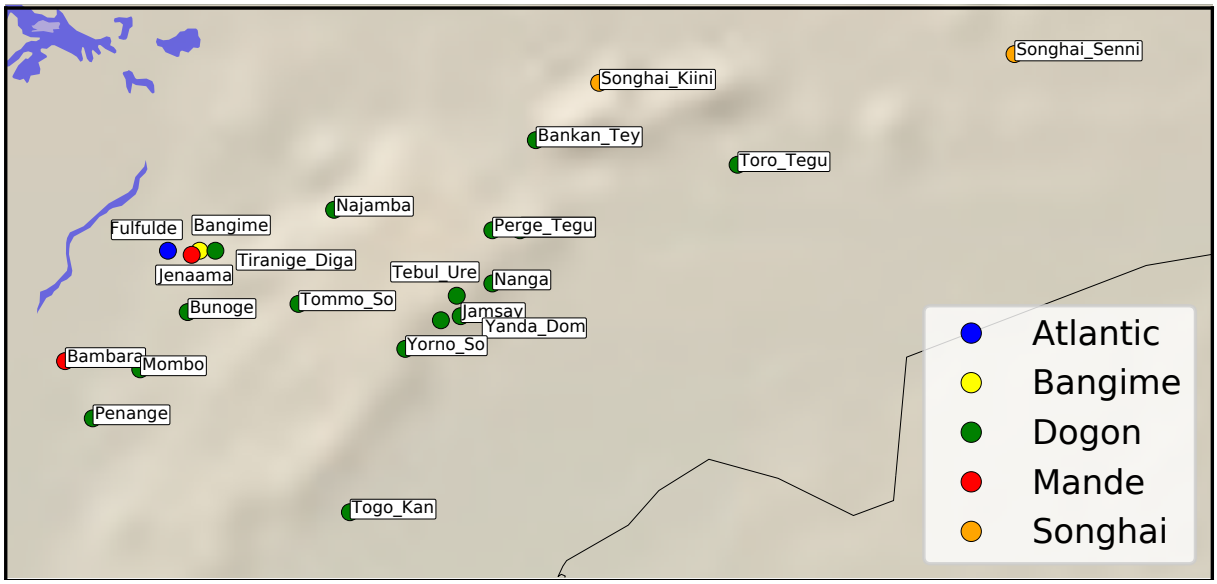

Figure 1: Bangime and its surrounding languages.

The map in Figure 1 also shows the area in Mali where the other languages that impact Bangime are spoken.[2] Geographically, Mali encompasses three regions - Sahara, Sahel, and sub-Sahel - each with its own heterogenous and independent ethnological-linguistic societies. Tuareg and Songhay live in the north; Dogon, Fulani, and Bozo Soninke in the center; and Bambara live primarily in the south and west, but some have migrated to other regions. The only genetic overlap among these groups is between Bozo Soninke and Bambara. Even with all this diversity, Bangime stands out as a linguistic isolate. Moreover, genetic studies demonstrate that its speakers represent a genetic deme (Babiker et al. 2018). Bangime is spoken among seven villages at the end of a valley that runs east, perpendicular to the Niger river, directly into the central-western edge of the Bandiagara escarpment. Two villages sit atop the mountain range, and the rest reside just below on the sandy soil. The Gueou valley stretches into the Sahelian plains

[2]Coordinates were given for the ethnolinguistic areas closest to the seven Bangande villages; the data used can be found in the supplemental materials.

and, approximately 250 kilometers north, emerges at Timbuktu in the Sahara. The Gueou valley's other inhabitants speak a Bozo variety called Jenaama. Most other Bozo languages are spoken along the Niger river.

The location of Bangime speakers necessitates constant contact with Fulfulde, Bozo-Jenaama, and Dogon-Tiranige speakers. In terms of wider contact areas, Bangande travel to the weekly markets in the Fulani city Konna (35 kilometers northwest) and the Soninke town Sambere (25 kilometers west), mainly to sell wild-grown and gathered fruits but also tobacco and other cultivated crops. Younger people travel to the regional capital Sevare-Mopti and the country's capital Bamako, where they learn to speak Bambara. Few Bangande travel north and thus rarely speak Songhay, though Songhay has made a historical impact on the Dogon languages, discussed in Section 2.2.

Fulani herders and their livestock roam throughout northern Africa, building temporary camps, many in the mountains to prevent their animals from destroying crops planted in the fields on top of the cliff and in the plains. Songhay groups live adjacent to the Dogon, directly south of the fabled city Timbuktu and further east along the single paved road that stretches south from the Malian capital, Bamako, to Gao. Most Bangande are fluent in Fulfulde and use it as a language of wider communication with groups other than Fulani since Bangime is unintelligible to the speakers of any neighboring languages, despite the Bangande people's insistence that it is Dogon.

Although the single term, 'Dogon' implies a linguistic and cultural homogeneity, substantiated by descriptions of a unified Dogon language (Bendor-Samuel et al. 1989) and genetically cohesive people (Tishkoff et al. 2009), even early depictions of the Mali Bandiagara Escarpment's inhabitants illustrate their diversity and heterogeneity (Desplagnes 1907). According to the Dogon and Bangime Linguistics Project (Moran et al. 2016, http://dogonlanguages.org), Dogon as a group constitutes at least 22 separate languages and upwards of 60 dialects. Dogon languages are primarily spoken among villages located at varying levels of a rocky cliff range known as the Bandiagara Escarpment, which is bordered on both sides by Sahelian plains. Some Dogon villages have descended to the plains within the past decade and a few Dogon languages are spoken across the border with Burkina Faso. Each secluded cliff village speaks its own variety, plus a reduced and often mixed Dogon variety for the purposes of wider communication such as Tommo So or Jamsay.

2.2 Historical Context

Those known as Dogon are thought to have moved to the Bandiagara escarpment from its eastern edges some 700–500 years ago. The 500-year estimate was proposed by Griaule (1938, p. 28), who claims to have uncovered nine Sigui masks at a site in the village Ibi in 1933. The Sigui is an eastern-Dogon ritual that is performed every 60 years, with one mask made for each festival. Thus, Griaule estimated a date of the first Sigui mask at 1430 A.D. Methodological concerns aside, note that not all Dogon villages have a tradition of mask dancing. Bedaux (1972, p. 41) asserts that the Dogon did not originally constitute a single culturally homogeneous group; they may have migrated to the cliffs from various directions.

Paleo-climatological investigations and archeological evidence amassed by Mayor and Huysecom (2016) and Mayor et al. (2005), Macdonald (1997) and Bedaux (1972) place the date of Dogon occupation of the Bandiagara Escarpment from the 16th to 20th centuries. One scenario presented by Mayor et al. (2005) depicts the Dogon fleeing the area's invaders sometime between 1230 and 1430 AD. We especially see linguistic evidence of the effect of the Songhay Empire on currently occupied Dogon areas. Most of the Najamba-speaking villages on the north-western portions of the Escarpment are still known by outsiders only by their Songhay names. Some include, Koira Beiri [kòjrà bé:rì] 'big town', Ibissa [í-bìs-â] 'I have passed', Borko [bòrkìn] 'noble', and Tondifere [tóndì fèrè] 'stone brick'. However, the contact has not been unidirectional as the Songhay village Kikara [kì-kárá], adjacent to where Dogon language Bankan Tey is spoken, translates to 'armpit' in Jamsay and many other Dogon varieties.

Dogon oral traditions describe a migration from Mande (Dieterlen 1955; Izard 1970; Marchal 1978), thought to be located in present-day southwest Mali or northeast Guinea to a village called Kani-Gogouna,

located adjacent to Ibi, the village where Griaule found the Sigui masks. Bangande also recount that they moved from Mande with the Dogon, settled in the village of Kani-Gogouna, and then relocated northwest, over and down the escarpment, to their current villages. However, a more likely scenario based on recent fieldwork (Hantgan 2013, pp. 407-410), proposes that instead of traveling south-east and then over the arduous cliff range, the move went in a straight line east, through a series of villages located along the Gueou valley. Since the Bangande also recount many stories of war with the Jenaama-Bozo, they were probably driven back down the valley to finally settle in the current location at the edge of the Escarpment.

According to a widely-known Malian legend, the Jenaama-Bozo, along with the other groups subsumed under the larger Mande branch, are said to have traveled with the Dogon from Mande but split away due to a disagreement between brothers. Indeed, the name Dogon, pronounced by its speakers as [dɔ̀gɔ̀-nɔ́] or [dɔ̀gɔ̌-n] derives from the term [dɔ́gɔ́-nín] which means 'little brother' in many Mande languages. One of the consequences of the split between the brothers was a law forbidding Bozo and Dogon to intermarry. As the Bozo ethnicity is defined by its trade as fishermen, the Jenaama who live in the Gueou valley are technically outsiders: they do not fish but live off subsistence millet farming like the Bangande and Dogon, yet neither Bangande nor Dogon intermarry with them.

High above Bara, one of the Bangime-speaking villages, is a group of buildings carved directly into the rock face. Found throughout the escarpment, these constructions within the caves are currently used by Bangande and Dogon as both burial grounds and granaries for storing millet, but the local populations believe they were housing built by a group known in most Dogon languages spoken in the eastern portions of the Escarpment as Tellem [télɛ̀-m], and translated to mean 'those before' by (Mayor et al. 2014).[3] First identified by archaeological excavations by Bedaux (1972), the Tellem are said to have occupied the cliffs from the 11th to 16th centuries AD. More recent work by Mayor and Huysecom (2016) depicts pre-Dogon populations as a cultural "mosaic" but with a distinct change in material cultural practices occurring around 400–600 years ago. They speculate that the populations who inhabited the escarpment prior to the Dogon either died or were subsumed into a new culture, that of the present-day Dogon, around 400–600 years ago.

Our results, given in Section 4, support evidence for the supposition that the Bangande were among the pre-Dogon groups; that they occupied the cliffs before the Dogon arrived. Further, we propose that there were separate waves of Dogon settlement, with indications of recent contact between Bangime-speakers and their immediate neighbors, and yet longer-term contact with Dogon groups that live to east of their current location. While we do not provide proof of certain dates, we suggest that the different Dogon waves occurred at different times in history, and perhaps from different directions. We explore these findings in the sections below.

3 Materials and Methods

All of the data were gathered independent of this study. The primary data were obtained through lexical elicitation for a Dogon-comparative word-list (Heath et al. 2015), with the addition of data collected for Bangime (Heath and Hantgan 2018), and Jenaama (Heath 2016), and Mande language data from the RefLex database (Segerer and Flavier 2016). Fulfulde lexical data are drawn from (Osborn et al. 1993) and Songhay Kiini and Sinni data from Heath (2005, 2015), respectively.

The Dogon language data were gathered by a team of eight researchers as part of the ongoing US National Science Foundation (NSF)-funded Dogon and Bangime Linguistics Project. Lexical, grammatical, and geographic information about the Dogon languages may be found at The Dogon and Bangime Linguistics website (dogonlanguages.org, Moran et al. 2016). Language identifiers, classification hierarchies, and geographic data for the remainder of the languages used in the sample are drawn from

[3] While the suffix [-m] can be assumed to be the animate noun class morpheme, used in Ben Tey and Bankan Tey, only the root [tɛ́lɛ́] is listed in the Dogon comparative wordlist, with the meaning of 'almost'. Thus, we cannot confirm the translation of 'those before' however, we do note the similarity to the Mande root [téli] meaning 'quickly'.

Glottolog (http://glottolog.org/, Hammarström et al. 2018).

3.1 Lexical Data Preparation

Given the diversity of the original sources, we had to normalize the data in various ways to render it comparable. Our first step was to convert each lexical spreadsheet to a file that could be read and interpreted by LingPy (List et al. 2017a), a Python library for automatic tasks in historical linguistics (http://lingpy.org), EDICTOR (List et al. 2017b), a web-based tool for creating, editing, and inspecting etymological datasets (http://edictor.digling.org), and similar packages that are part of recently proposed tool chains for computer-assisted language comparison (List 2016), developed in close collaboration with the Cross-Linguistic Data Formats (CLDF) initiative (https://cldf.clld.org, Forkel et al. forthcoming).

Thus, we converted the various lexical spreadsheets organized by rows into separate, five-columned, tab-separated files. In the first column of each file, we specified a unique ID to each row, followed by the language name, and then the individual lexical form in IPA adjacent to its English and French gloss. In cases where more than one form was listed for a particular lexical item, we treated them as individual entries in that language. These entries are not duplicates; rather, they represent important synonyms that may lead to deeper connections among the related languages.

Next, due to the multitude of glosses used to elicit the same meanings in different contexts, we used the semi-automatic tools provided by the Concepticon project (http://concepticon.clld.org, List et al. 2016, 2018b) to systematically identify equivalent concepts across all datasets by linking all elicitation glosses to Concepticon concept sets. The linking procedure is strictly semi-automatic: in the first pass, we used the Concepticon API to link all elicitation glosses automatically to one or more Concepticon concept sets (e.g., linking 'arm (of hand)' and 'arm (body)' to #1673 ARM). In a further step, we manually checked and corrected all automatic linkings.

The automatic cognate detection methods included with the LingPy software package require that the orthographies in the data be error-free and comparable; that is, every sound must be represented in a consistent manner. Confusion can arise when one dataset does not follow traditional IPA conventions. For example, the Africanist tradition of writing IPA [j] as *y*, and [dʒ] as [j] can be particularly confusing. If one dataset follows traditional IPA conventions and another does not, the discrepancies can lead to a depiction of the lack of a sound change between IPA [j] to [dʒ] among cognates as in [dija] to [didʒa] in Bangime, 'eat'. The algorithms for automatic cognate detection rely heavily on consistent transcriptions. If transcriptions are inconsistent, the methods may either cease to work at all, or the results may be largely problematic.

Therefore, in an additional step to render the data comparable, we used orthography profiles (Moran and Cysouw 2017), as implemented in the Segments package (Moran and Forkel 2017) in order to segment the transcriptions in the data into units representing single sounds in the languages under investigation and to convert the data to a plain IPA representation accepted by the automatic cognate detection methods provided by LingPy. Data in Table 2 show an example of how orthography profiles work: much of the Mande words listed in the RefLex database use an underscore tilde [◌̲] to represent nasalization while that of the Dogon comparative spreadsheet uses a superscript [ⁿ] following the segment, and then Bangime data were transcribed with a tilde above the segment [◌̃].
Diacritics and tonal markings on vowels and sonorants were converted to a number corresponding to the 5-tonal step pattern used for Chinese, followed by a tilde in the cases of nasalized segments. Once properly defined, orthography profiles were used to automatically tokenize unsegmented transcriptions in the varying orthographies and convert them to our desired target transcription system. The tokenized representation is usually indicated by inserting spaces between graphemes (consisting of one or more characters) to indicate the start of a new unit that could phonetically or phonologically be perceived as a sound on its own. Without these implementations, potential cognates would likely be missed due to mismatched orthographic conventions.

Language	Source Transcription	Tokenized Representation	Source Gloss	Concepticon Concept
Bangime	kĩ̃ĩ	k ĩː 53 ~	boat	BOAT
Mombo	kíːⁿ	k ĩː 5 ~	boat	BOAT
Tiranige Diga	kŭːⁿ	k ũː 15 ~	boat	BOAT
Jenaama	kūⁿ	k ũ 3 ~	boat	BOAT
Bambara	kúrṵ	k u 5 r u 5 ~	boat (skiff)	BOAT

Table 2: Illustration of data conversion in our workflow.

After we created compatible transcriptions and translations, we combined our various wordlists into one multilingual comparative file that essentially contains a (new) unique ID, a language name without spaces or special characters, both the original IPA transcription and the tokenized form produced by our orthography profile, original glosses, and concept identifiers from the Concepticon concept sets. This method resulted in a wordlist consisting of 315 concepts translated into 38 of the estimated 68 Malian languages (Simons and Fennig 2018). To make sure that languages were equally represented, showing high *mutual coverage* (see List et al. 2018a and Rama et al. (2018) for details on the concept of average mutual coverage in wordlist data) and as few missing translations per concept set as possible, we further extracted a selection of 22 languages and 300 concepts (see the Appendix for the entire list of 300 concepts). Table 3 illustrates the overlap of our 300-item concept list with other well-known concept lists. The overlap with the larger lists is above 50 percent and even higher with the smaller lists.

List	Source	Coverage
Blust-2008-210	Greenhill et al. 2008	126 (60%)
Swadesh-1952-200	Swadesh 1952	117 (59%)
Matisoff-1978-200	Matisoff 1978	117 (59%)
Gregersen-1976-217-1	Gregersen 1976	115 (53%)
Swadesh-1955-100	Swadesh 1955	73 (73%)
Tadmor-2009-100	Tadmor 2009, (Leipzig-Jakarta)	70 (70%)

Table 3: Overlap with popular basic vocabulary lists.

Our initial selection procedure aimed to achieve a good mix of cultural and cross-linguistically interesting concepts, while selecting language data that would elucidate the ancestry of Bangime with the overall goal of providing a balanced sample of high coverage and interesting concepts.

3.2 Lexical Data Comparison

Identifying Layers of Contact

Using automatic methods for cognate detection on languages that we know are unrelated such as Bangime and its neighbors, can produce many false positives that reflect neither recent contact nor ancient relations. To address this problem, we propose a new workflow for automatic word comparison inspired by general approaches to exploratory data analysis (Morrison 2014). Our main idea is not to restrict our analysis to one method only but to take advantage of the methods LingPy offers for automatic cognate detection, which are quite different in their underlying basic models and strategies.

The LexStat-Infomap approach (List et al. 2017b) has been shown to outperform earlier approaches, coming quite close to expert judgments on cognates in multilingual wordlists.[4] Its strategy closely resembles the classical comparative method, searching the data for *regular sound correspondences* before

[4]The method by Jäger et al. (2017) seems to outperform LexStat-Infomap on certain datasets, but according to Rama et al. (2018), the difference is minimal and may be in favor of LexStat.

assigning words to common cognate sets. Regular sound correspondences usually reflect deep genetic signals but can also result from intensive language contact. The family of LexStat approaches tends to single out sporadic borrowings to some degree (List 2012a), but borrowings can be easily confused with cognates where language contact is intensive (List 2014). We expect that applying the LexStat-Infomap approach will propose cognate sets that reflect (a) a true genetic signal among closely related languages, and (b) ancient layers of contact intense enough to surface (potentially weakly) as regular sound correspondences.

In contrast to the conservative and highly sophisticated LexStat-Infomap approach, LingPy offers simpler methods that are especially useful for quick data exploration, especially when the number of languages and concepts is large. The Sound-Class-Based Alignment (SCA) approach (List 2012b) derives pairwise word similarity scores from pairwise phonetic alignments between all word pairs in a given concept slot without taking regular sound correspondences into account. As a result, it may select true genetic cognates, or it may assign to the same cognate sets words that are similar only due to spurious borrowings or coincidental similarities. We can use this seeming disadvantage to our advantage when dealing with complex linguistic situations like the one we encounter with Bangime.

By comparing the findings for the conservative, yet rather accurate LexStat-Infomap approach with the SCA method, we can systematically search for discrepancies between our *genotypic* and *phenotypic* (Lass 1997) cognate detection approaches. As a rule of thumb, we can say that when both algorithms identify certain words as cognates, they generally are, notwithstanding certain erroneous judgments that arise for several reasons. We expect to find most of these cases in languages already known to be related. However, if the SCA method identifies certain words as cognate, and LexStat-Infomap does not, we may assume we are dealing with either chance similarities or rather recent instances of borrowing.[5] The more cases of borrowing we find in specific language pairs, the stronger the argument for recent borrowing.

Although our approach is relatively simple, we think it offers several improvements over previously proposed approaches to the automatic identification of borrowings. In contrast to Mennecier et al. (2016) and Ark et al. (2007), for example, who use a version of the edit distance (Levenshtein 1965) to search for borrowings between unrelated language varieties only, or phylogeny-based approaches (List 2015; List et al. 2014) that can only be applied to related languages, our approach can be applied to both related and unrelated languages. Since we generally assume that LexStat-Infomap is sufficient to detect very clear cases of cognates, we are also confident that cognates it does not accept but SCA does are probably true instances of borrowings.

Note that we do not expect the cognate sets proposed by LexStat-Infomap for the unrelated languages given in 4 to reflect true, deep cognates. Given the spuriousness of these findings, LexStat-Infomap is probably capturing ancient layers of contact, which is all the more interesting given the unknown history of Bangime. Hence, our approach offers an automatic *stratification* analysis (Lee and Sagart 2008): by applying methods with different degrees of conservatism, we can extract different layers of shared words. The ones LexStat-Infomap recognizes represent the oldest layer, and the ones identified only by SCA represent more recent layers.

Shared Vocabulary Statistics

We wanted to know the degree to which the languages of different genetic origins in our sample shared words, especially with respect to Bangime. We viewed the problem as similar to admixture analyses in population genetics (Pritchard et al. 2000), although analyses in biology automatically determine which genes are most likely to represent a certain ancestral population. We are in a much more comfortable position; thanks to classical approaches to linguistic reconstruction, we often have an independent account on the words that were used in a given proto-language, so we do not need sophisticated algorithms to

[5] As our dataset shows, in recent borrowings, donor and recipient word tend to resemble each other more than older borrowings do since borrowings are usually nativized over time and adjusted to the phonotactic system of the target language.

determine which words in our Bangime sample are shared with other language families. Instead, we can simply consider all inferred cognate sets in our data for a given language or group of languages and count how many are shared exclusively between the given language or group and the other languages and groups in the sample. Again, by contrasting LexStat-Infomap and SCA results when calculating the inferred cognate set statistics, we can better assess the degree to which the 'admixture' of a given language or group of languages differs.

Implementation

Our approach is implemented in the form of Python scripts, which are available in the supplementary material. They use the parameters indicated in List et al. (2017b) for the LexStat-Infomap and the SCA approach to cognate detection. The data output are given in the supplemental materials in tabular form for manual inspection (e.g., by using the EDICTOR tool) or as matrices that can be fed to phylogenetic software like SplitsTree (Huson 1998) to compute splits networks (e.g., with the NeighborNet algorithm, Bryant and Moulton 2004) and as plots that visualize specific aspects of the data.

4 Results

4.1 Word-list Statistics

Table 4 summarizes the coverage statistics for each of the 22 languages selected for our study. Subgroupings for the Mande and Songhay languages are based on Glottolog (Hammarström et al. 2018). The Dogon grouping sub-classification follows Moran and Prokić (2013) and Prokhorov et al. (2012), except that our findings, indicated by the asterisk in the table and discussed below, align Najamba with the western rather than the eastern group.

Language ID	Language name	Items	Coverage	Subgroup	Source
1	Bambara	212	0.71	Western Mande	Dumestre 2011
2	Bangime	300	1.00	Isolate	Hantgan and Heath 2016
3	Bankan Tey	297	0.99	Eastern Dogon	Heath et al. 2015
4	Ben Tey	276	0.92	Eastern Dogon	Heath et al. 2015
5	Bunoge	272	0.91	Western Dogon	Heath et al. 2015
6	Fulfulde	282	0.94	Northern Atlantic	Osborn et al. 1993
7	Jamsay	286	0.95	Eastern Dogon	Heath et al. 2015
8	Jenaama	237	0.79	North-western Mande	Heath 2016
9	Mombo	292	0.97	Western Dogon	Heath et al. 2015
10	Najamba	293	0.98	*Eastern Dogon	Heath et al. 2015
11	Nanga	299	1.00	Eastern Dogon	Heath et al. 2015
12	Penange	280	0.93	Western Dogon	Heath et al. 2015
13	Perge Tegu	296	0.99	Eastern Dogon	Heath et al. 2015
14	Songhay Senni	267	0.89	Eastern Songhay	Heath 2015
15	Songhay Kiini	249	0.83	Eastern Songhay	Heath 2005
16	Tebul Ure	286	0.95	Eastern Dogon	Heath et al. 2015
17	Tiranige Diga	290	0.97	Western Dogon	Heath et al. 2015
18	Togo Kan	288	0.96	Eastern Dogon	Heath et al. 2015
19	Tommo So	292	0.97	Eastern Dogon	Heath et al. 2015
20	Toro Tegu	297	0.99	Eastern Dogon	Heath et al. 2015
21	Yanda Dom	295	0.98	Eastern Dogon	Heath et al. 2015
22	Yorno So	294	0.98	Eastern Dogon	Heath et al. 2015

Table 4: Coverage and sources of our data

We note that, based on the genetic relationship to Jenaama, Soninke might have been a more appropriate choice for an additional Mande language spoken in the area of Bangime, but the available Soninke wordlist did not cover our selected concepts well enough to be included in the subsample. Further, Bambara has a somewhat surprising influence on the Bangime lexicon as found in our sample and explored below.

Using the selected languages and concepts with the coverages shown in Table 4, we performed both methods, SCA and LexStat-Infomap, on the condensed 300-item wordlist. Both models require thresholds that determine at which level of similarity or distance words are considered to be cognate. For our analysis, we employed the thresholds reported by List et al. (2017b), who determined on empirical data of six language families, manually coded for cognates by experts, which thresholds yield the best results on average. This yielded a threshold of 0.55 for the LexStat-Infomap approach and 0.45 for the SCA approach.

4.2 Shared Similarities

Bangime is undoubtedly a language isolate, but it is not an insular language; we expect to see clear effects of contact with its neighbors. The most likely impact is from the Dogon languages, yet questions remain about when and which varieties left their mark. The LexStat-Info method shows us the deeper levels of contact. We can then compare these findings to the surface levels SCA brings out.

LexStat

The dark blue patterning along the rows adjacent to Bangime in the heat map (Fig. 2) indicates that the LexStat method finds practically no notable cognates between Bangime and the other languages in the sample. We still see more similarities between Bangime and Dogon than between Dogon and Fulfulde or Songhay. That is, while Bangime is certainly not genetically related to any of the surrounding languages, it has deep lexical affiliations, particularly with the Dogon languages. We explore the implications of this finding in detail below.

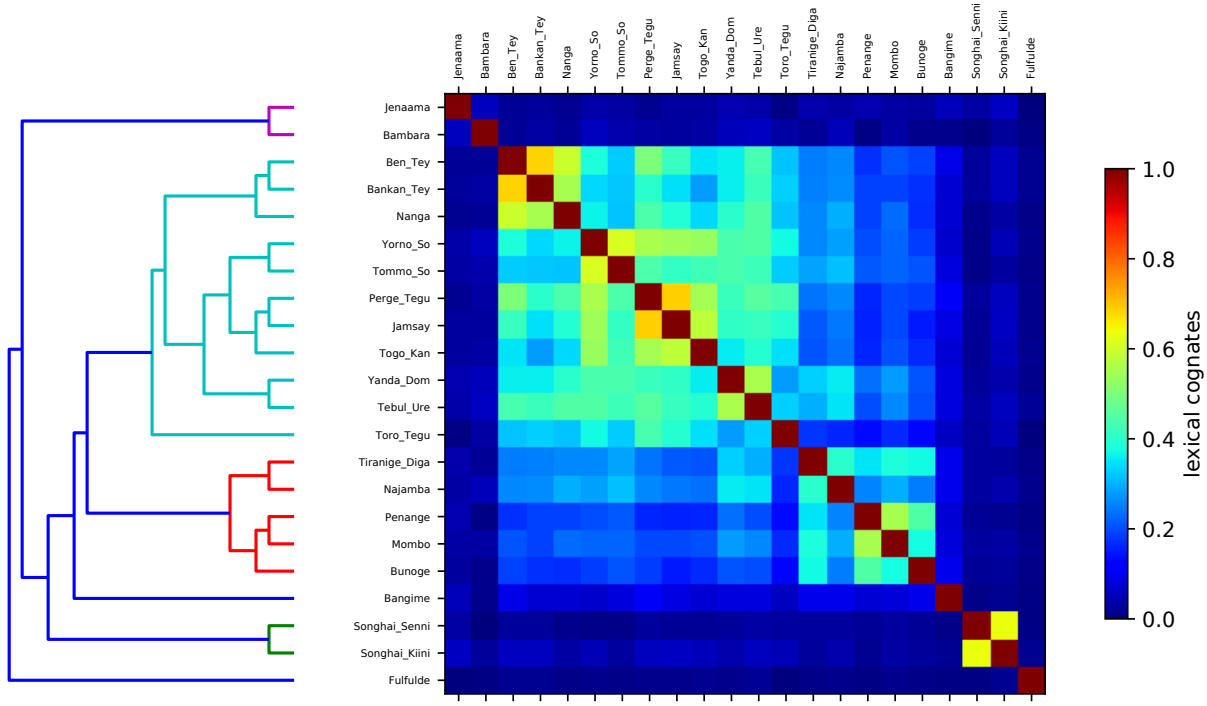

Figure 2: Heat Map generated from the cognate percentages inferred by the LexStat-Infomap approach. The cognates chosen by the LexStat-Infomap approach which appear here as lighter colors on the spectrum towards red in the middle, where each language is compared with itself, can be interpreted either as true cognates or deep levels of contact.

We observe that the phylogenetic tree produced by the model accurately captures the accepted rela-

tionships among the languages in the sample. Furthermore, we see below that this unrooted tree shows contributions to both Bangime and the Dogon languages from unrelated Songhay and Fulfulde donor words that are borrowings from Arabic as well as, although more recently, the Mande languages Bambara and Jenaama.

Therefore, we see a connection between Bangime and the Dogon languages through, not phylogeny, but contact. Most interesting, as noted in Section 2.1, we see a connection with Dogon languages spoken at quite a distance from where Bangime is spoken, making the internal reconstruction of the Dogon languages important to the relationship with Bangime. While we see the same east-west split first depicted by Prokhorov et al. (2012), here, we propose a departure from the grouping described by Moran and Prokić (2013). Even though we use the same methods, SCA and LexStat, they place Najamba within the eastern Dogon subgroup, while we place it in the western branch, where it is geographically located.

In fact, the phylo-genetic tree pictured in Figure 3 essentially reproduces the location of the languages shown in the map in Figure 1, verifying the clarity of the signal among the lexical data used in this study. The only difference between the grouping in the figure below from that above is the additional subgrouping of Nanga with its subbranches Bankan Tey and Ben Tey.

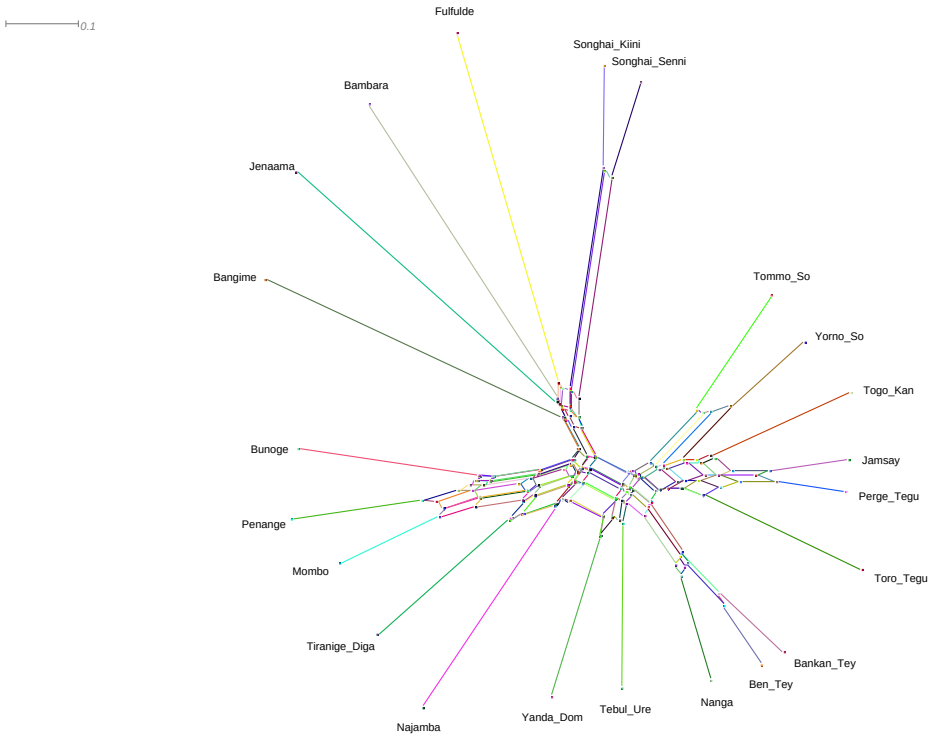

Figure 3: An equal-angle Neighbor-Net tree created in Splits Tree (Huson and Bryant 2006) from an alignment site matrix produced by the LexStat-Info method.

While Fulfulde may appear as somewhat of an outlier, we attribute its diversion to the lack of related languages in the sample; Maasina Fulfulde is the only Atlantic, and also non-tonal, language spoken in Mali. In many instances, both computational methods rejected similar words between Fulfulde and the other languages in the sample due to the lack of tones in Fulfulde, even though many borrowings are known to be found among the languages with Fulfulde as the source. Jenaama is not thought to be closely related to Bambara within the Mande language group, thus it is not entirely surprising that deep levels of shared cognates are not picked up by the LexStat-Info method. Data from languages more closely related to Jenaama, among the Bozo subgroup, can be viewed in the larger dataset from which the subset was

drawn but due to coverage gaps these languages were omitted for the purposes of the current study.

Figure 4 shows percentages of shared vocabulary as judged by both our methods. Examining the specific relationships between Bangime and the other languages in the sample shows that, despite the fact that Bangime is the smallest single language represented, it is the least homogenous. The graphs confirm our expectations that Bangime has been heavily influenced by both Dogon and Mande languages.

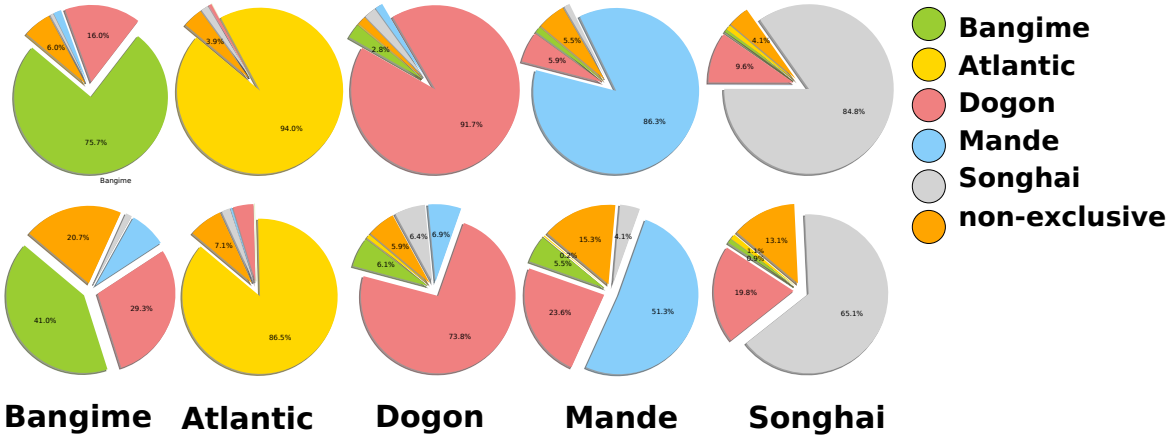

Figure 4: Comparing exclusively shared cognates across language families in our sample for LexStat-Infomap (top row) and SCA (bottom row). The "non-exclusive" group refers to cases where inferred cognate sets are attested in more than two families.

The first row of pie charts was produced based on the cognates found by the LexStat-Info method. As discussed in Section 3, the lexical similarities found by LexStat are not necessarily true cognates, though they are likely to reflect long-term contact between languages. We see that even the conservative estimate shows that only 75 percent of Bangime lexical items are unique and the generous estimate posits a mere 41 percent. Not surprisingly, the largest contribution of affinities in the lexicon are from the Dogon languages. Again, although, we find no evidence to support a genetic relationship between Bangime and the other languages used in the sample, shown in the following subsection, the differences between LexStat and SCA results are striking, splitting recent borrowings and ancestral contact into different categories of language contact.

SCA

The overall SCA-generated results represented by the heat map in Figure 5 are similar to those shown in the LexStat-generated heat map. Here, they have a lighter shade overall, portraying more recent stages of contact, that is, mostly through borrowing.

Compared with the LexStat-Infomap heat map above, the SCA generated heat map here is far lighter in color, illustrating the tendency of SCA to reflect much more recent stages of language contact. While the Dogon languages' affiliations remain largely the same in both figures, Bangime is more clearly aligned with its neighboring languages such as those in the western group and Jenaama, than was shown in the darker borders of the LexStat-Infomap diagram.

Another representation of the results generated by the SCA-method through a distance matrix tree further illustrates the place of Bangime somewhat in between the Western Dogon languages and the Mande languages. Indeed, the speakers' position geographically is precisely between these two populations. We do note once again the split between the Dogon languages along east-west lines is maintained in both diagrams; the results of both methods indicate a separation of the Dogon languages into two distinct subgroupings, furthering our postulation that the Dogon dispersal of the Bandiagara Escarpment could have happened from different directions, and/or distinct time periods.

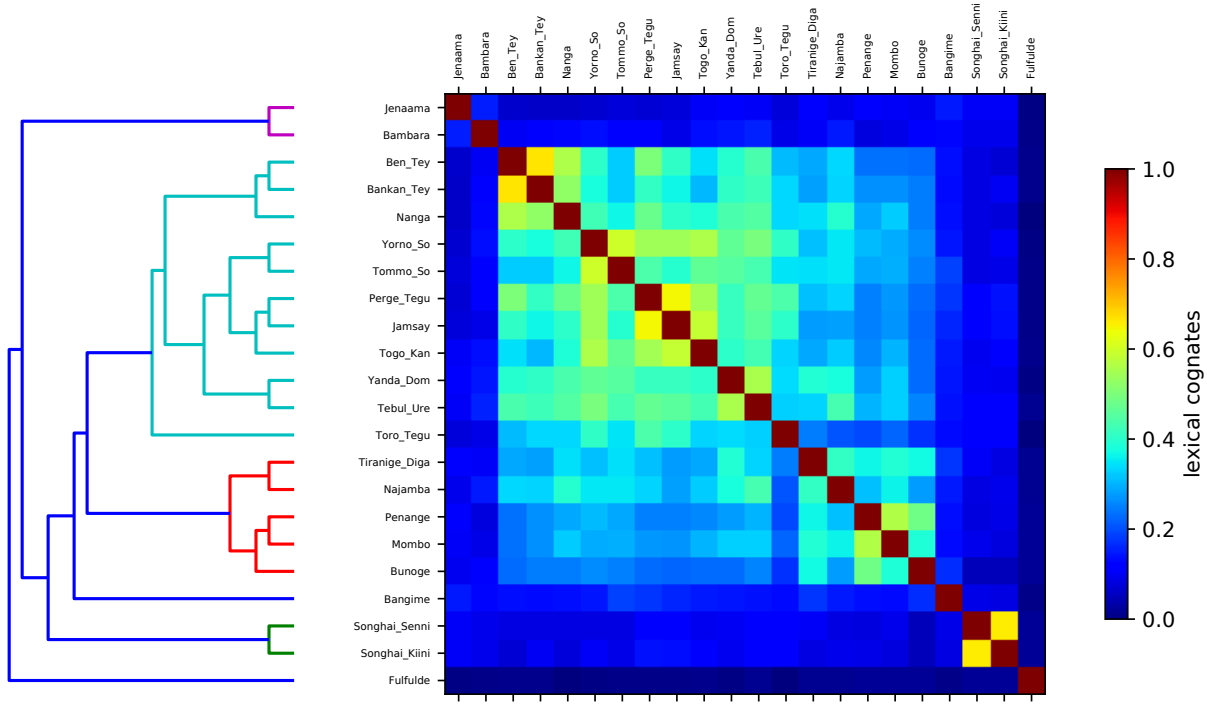

Figure 5: Heat Map generated from the cognate percentages inferred by the SCA approach. The "cognates" selected by the SCA approach which appear here as lighter colors on the spectrum towards red in the middle, where each language is compared with itself, can be interpreted either as borrowings rather than leading us towards genetic affinities.

Interestingly, we see that Jenaama and Bambara are still not considered to be as closely related as, for example, Songhay Kiini and Songhay Senni. This finding is reasonable in that the two Mande groups have neither shared proximate ancestry nor frequent language contact, whereas the two Songhay groups have both. Thus, both of the computational methodologies we used found very similar end-results in terms of language relatedness, with the locus of the largest difference between methods being found in the Bangime results. The congruence of the comparison between the two methodologies, shown in Figure 7 clearly leads us to identify Bambara and Tommo So as likely sources of both borrowings and cognates.

What is fascinating about the graph in Figure 7 is the strong indication towards an affiliation between Bangime and Bambara, on the one hand, and Tommo So on the other. Many young Bangande now speak Bambara as a result of traveling to the southern part of the country, and thus the language is having an inevitable effect on their speech, especially if they depart the village at a young age without having the opportunity to learn Bangime.

We ascribe the influence of Tommo So to three factors: (1) all Dogon and Bangande perform songs in Tommo So, regardless of whether the people speak the language; (2) legend describes the Dogon migration arriving first at the currently Tommo So-inhabited village Kani-Gogouna, so the language may express a type of ancestral lineage; and (3) Tommo So is centrally spoken, and many use it as a lingua-franca, especially in the region surrounding Bandiagara.

Considering the Bangande-projected Dogon identity, we are not entirely surprised to find that Dogon languages have lexically influenced even core Bangime vocabulary, specifically we will see in the next section, low numerals, body parts, and culturally significant items. However, what is somewhat surprising is the geographical distance at which we find evidence of this deep language contact. We explore specifics in the next section.

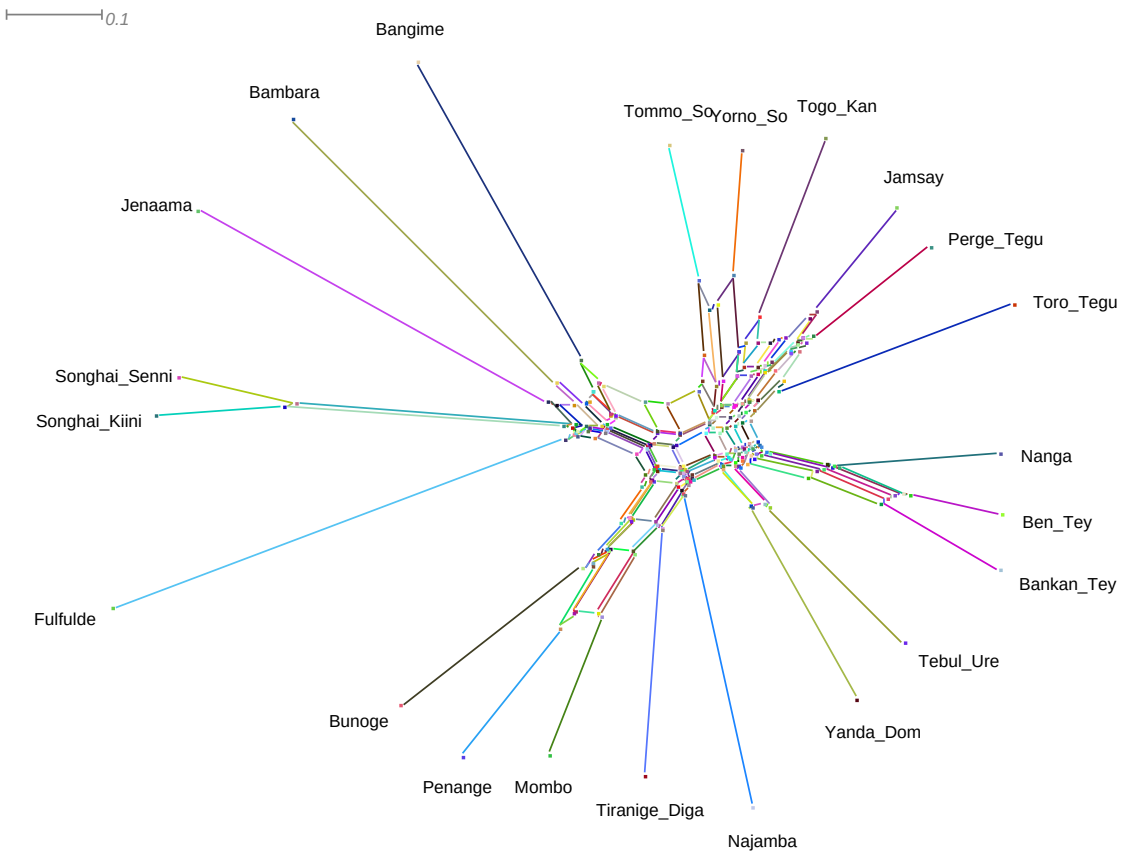

Figure 6: An equal-angle Neighbor-Net tree created in Splits Tree (Huson and Bryant 2006) from an alignment site matrix produced by the SCA method.

5 Discussion

5.1 Areal Comparisons

As explained in Section 3, both the SCA and LexStat methods may indicate borrowings, but the latter certainly reveals both true cognates and deeper levels of contact. Here, we focus on those instances when both methods select a word as 'cognate' with Bangime. In this way, we can concentrate on long-term contact between Bangande and surrounding speakers. As Campbell (2017) points out, one way to find the lost ancestors of a language isolate is through areal comparisons within the Sprachbund. Almost all instances of LexStat cognates are shared with the SCA method, though obviously the reverse does not hold true. Since Bangime is an isolate, it does not have cognates per se with surrounding languages, but it certainly shares vocabulary worth exploring in further detail.

First, we illustrate the distribution of numerals Bangime shares with the languages in our sample (Table 5) because we note that this area of the language's core vocabulary greatly resembles languages from different groups spoken in Mali. Check marks indicate where both the SCA and LexStat-Info method selected the concept and language has having a 'cognate' with Bangime. The grey shading highlights the

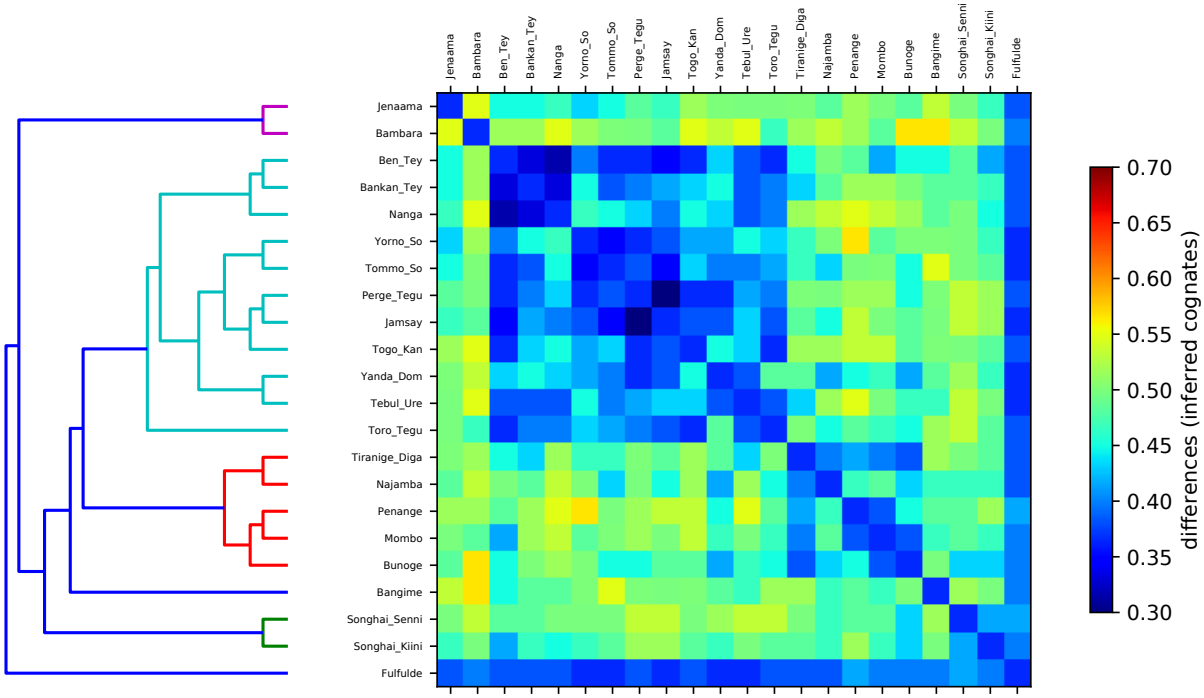

Figure 7: Comparing the discrepancies between cognate percentages inferred by LexStat-Infomap and SCA. If no difference is encountered, we set the value in the heatmap to 0.5. If the proportions differ, we first subtract the LexStat-Infomap value from the SCA value and then add this value to our baseline of 0.5. As a result, comparing one language with itself will always yield the baseline, and values larger than 0.5 indicate to which degree SCA infers more cognates for the respective language pair, while values less than 0.5 indicate that SCA infers less cognates. We suspect that those cases in which SCA infers a larger number of more cognates point to recent borrowings between the varieties.

most prevalent languages and concepts.

Both of our methods highlight the numerals THREE, FOUR, and SIX as areal; that is, they are found spanning several language families in the region. Bangime's most common source languages are found in the eastern Dogon subgroup; only one western neighboring language, Bunoge, shares four of the ten lower numerals with Bangime, as compared to five eastern Dogon languages. However, the numeral SEVEN is shared only with the distantly spoken Songhay languages, not any of the nearby Dogon varieties, as might be expected. The numeral EIGHT appears to come from the Mande group, though it could also be from Najamba or Yanda Dom. The only numerals exclusive to Bangime are TWO, NINE, and TEN. The numerals ONE through FIVE are the most essential to both the Bangande and Dogon; the week is five days with a different village's market rotating on one day of the five-day week. This traditional method of keeping track of the days is being lost in favor of the seven day week, but the date of many performances and events crucially rely on the five-day week.

Among other core vocabulary items, we do witness, although less often, similarities for certain body parts as shown in Table 6.

Blench (2005, p. 15) states that the Bangime words for 'ear' and 'mouth' are commonly viewed roots among Niger-Congo languages, yet neither of our analyses find a significant connection to any languages of Mali used in our sample. Among numerals and body parts, we see many core Bangime words are shared with Jamsay, including the word for 'chest', also found in Songhay. Even some of the more integral body parts, such as HEAD and HAND, were selected by the SCA method as being potential cognates/borrowings from both the Dogon languages and also Bambara. Table 7 shows examples of

DOCULECT	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	TEN
Bambara								✓		
Bankan Tey			✓	✓		✓				
Ben Tey			✓	✓		✓				
Bunoge	✓		✓	✓		✓				
Fulfulde			✓	✓						
Jamsay	✓		✓	✓		✓				
Jenaama								✓		
Mombo			✓			✓				
Najamba			✓			✓		✓		
Nanga			✓	✓		✓				
Penange			✓			✓				
Perge Tegu	✓		✓	✓		✓				
Songhay Senni							✓			
Songhay Kiini							✓			
Tebul Ure	✓		✓			✓				
Tiranige Diga			✓			✓				
Togo Kan	✓		✓			✓				
Tommo So	✓		✓	✓		✓				
Toro Tegu	✓		✓	✓		✓				
Yanda Dom			✓			✓		✓		
Yorno So	✓		✓	✓		✓				

Table 5: Shared Numerals

words for body parts that appear unique to Bangime, and yet are found in some Dogon languages, albeit with a different meaning; these items were hand-selected as neither of our automatic methods detected them.

There is a precedent for switched meanings in Bangime. Hantgan (2013) describes portions of the Bangime lexicon with purposely reversed meanings such as the naming of plants used in traditional medicines by reversed colors than they appear to be. Even though practically no one outside the Bangande community speaks Bangime, they claim that the use of reverse meanings further protects them from potential eavesdropping.

Other terms shared areally between Bangime across language families are shown in Table 8. Concepts COW and MEAT, along with HUNDRED, GOLD, FORGIVE, and MOSQUE are likely attributed to the wide distribution Fulfulde lexical items, some of which are originally from Arabic, while CASSAVA probably has a source among the southern-spoken Mande languages from where the crop originated. The fact that both our methods selected MAIZE as being shared between Bangime and Jamsay is somewhat unexpected given the word in Jamsay [pòrò: ɲú:] bears little resemble to Bangime [bìrɔ̀ ndɔ́ⁿ]. However, the Jamsay word differs from that of the other Dogon languages, thus this warrants further inspection in the future.

Our methods somewhat incorrectly attribute the source of some Arabic borrowings to Songhay rather than Fulfulde. We attribute this error to the fact that Songhay shares the tonal properties with the other Malian languages whereas Fulfulde does not. The only clear cases of a borrowing from Songhay into nearly all the surrounding languages are COUNTRY and CLOUD. The original source of these Songhay shared words is likely not from Arabic.

DOCULECT	BEARD	CHEST	JAW	JOINT	LUNG	VAGINA
Bankan Tey					✓	✓
Ben Tey					✓	✓
Bunoge					✓	
Jamsay		✓	✓	✓	✓	✓
Jenaama	✓					
Mombo			✓	✓	✓	
Najamba				✓	✓	✓
Nanga				✓		
Penange				✓		
Perge Tegu		✓		✓	✓	
Songhay Senni		✓				
Tebul Ure				✓		
Tiranige Diga				✓		
Togo Kan			✓		✓	
Tommo So				✓		
Toro Tegu				✓		
Yanda Dom				✓		
Yorno So				✓		

Table 6: Shared Body Parts

Bangime	Bondu So
SKIN	HEAD
kíndʒē	kíŋgè

Bangime	Tiranige Diga
HAIR	SKIN
kwìì	gwíí

Bangime	Toro-Tego
WING (shoulder)	FOOT OR LEG
kúwó	kúwó

Table 7: Reverse Borrowed Body Parts

Curiously, neither the Bangande nor many of the more geographically isolated Dogon speakers, has not adopted the term ANIMAL from Arabic [√dbb] via Fulfulde [dabba], as most of the plains-speaking Dogon villages and even Jenaama spoken in the valley have. This fact, coupled with the assertion of Blench (2005, p. 2) who states that Bangime crop names are distinctive, supports the hypothesis that the Bangande lived in their current location before the Dogon. Our results concur that the most essential crop to the area, MILLET, is indeed unique to Bangime, while many of the Dogon languages may share the term with southern Mande varieties. In the table above, only Jamsay was selected by both of our methods as sharing the term MAIZE with Bangime. CASSAVA, with a likely source from Bambara and other Mande varieties, is a

DOCULECT	CASSAVA	COW	GRASS	HORN	MAIZE	MARROW	MEAT
Fulfulde	✓	✓					
Bunoge	✓	✓	✓	✓			✓
Mombo	✓	✓		✓			✓
Najamba	✓	✓		✓			✓
Penange		✓	✓	✓			✓
Tiranige Diga	✓	✓		✓		✓	✓
Bankan Tey		✓					✓
Ben Tey	✓	✓					✓
Jamsay	✓	✓			✓		✓
Nanga						✓	✓
Perge Tegu	✓	✓		✓			✓
Tebul Ure	✓	✓					✓
Togo Kan	✓	✓		✓			✓
Tommo So	✓	✓		✓			✓
Toro Tegu							
Yanda Dom	✓	✓					✓
Yorno So		✓		✓			✓
Songhay Kiini	✓	✓				✓	
Songhay Senni	✓	✓				✓	
Bambara	✓						
Jenaama	✓	✓				✓	

Table 8: Shared Agriculture Vocabulary

recently introduced crop, selected by our methods as being a shared term.

5.2 Group-specific Comparisons

While words from Bambara are shown in Figure 7 to be particularly prevalent in Bangime, we actually attribute many of the words in our dataset, particularly those to do with the caste system, to Soninke. The Mande influence of the Mali Empire on Bangime and the Western-Dogon languages is evidenced through culturally significant vocabulary shown in Table 9.

Recently acquired vocabulary was often not picked up by either of our methods, and yet manual inspection reveals that there is clear evidence of a connection between the above listed words and the languages spoken to the west of the Bangande speakers. We attribute these relatively recent borrowings to the Mali Empire. The Mali Empire, with Mande rule of the western portions of Mali from the 1200's to 1600's A.D., is evidenced by relatively recent borrowings with specific cultural meanings found in Bangime and the languages spoken to the west of the Bangande-speaking region.

While most researchers assume that the Dogon villages came to be occupied as a result of escaping from prior empires, Nunn and Puga (2012) provide statistical support for the unique benefits of living in rocky terrain in Africa in order to avoid the repercussions of the slave trades from 1400 and 1900 A.D. As explained in Hantgan (2013), the Bangande separate themselves

DOCULECT	NOBLE	SLAVE	TONGS	FIGHT	FOAM	NEW	LOUSE	WORK
Jenaama	*✓	*✓	*✓	✓	✓		✓	
Bambara	*✓	*✓	*✓	✓		✓		
Bunoge	*✓	✓	*✓		✓	✓	✓	✓
Mombo			*✓		✓	✓	✓	✓
Penange	*✓		*✓		✓	✓	✓	
Tiranige Diga	*✓	✓			✓	✓	✓	✓

Table 9: Recent Shared Cultural Vocabulary

into two classes: slave and royal, with depictions of slave raids happening in the villages and surroundings told through oral histories. At least at the western side of the Bandiagara Escarpment where Bangime is spoken, the results of language contact appear to be recent. Specific concepts within the domain of castes, central to Malian cultural hierarchies, such as NOBLE and SLAVE are likely from Soninke, passed into Bangime and western Dogon languages, yet not the Dogon languages of the east.

Another caste concept, BLACKSMITH, among the Dogon languages likely derives from [zèmò] 'to forge' in Songhay Senni. Despite that many Tiranige Diga speaking blacksmiths live among the Bangande villages, the word for blacksmith in Bangime bears no resemblance to any form in Dogon. In fact, the term most closely resembles that of Bambara in our sample, but since neither of our automatic detection methods determines the source, it is likely a recent acquisition into the language. The concept CHIEFTAIN also comes from Bambara. The form for God, misattributed to Jenaama [ālà], more likely comes from the widely-used Mande [ŋàlà], the pre-Islamic form for God. Other recently acquired concepts in our sample include PIG, from Jenaama, DONKEY, likely from Bunoge or Tiranige Diga, and SHEEP from Bunoge or Penange.

We can propose that many recently introduced domesticated animal and crops names are equally recent borrowings; this coupled with the fact that only the SCA method and sometimes neither method discovers the target concepts. Further, recent borrowings into Bangime are overwhelmingly from the languages that surround the location of the Bangime speakers. However, both our methodologies concur that the concepts given in Table 9 belong to Bangime and the Dogon languages from the eastern, rather than the more geographically proximate, western, sub-grouping.

In addition to the accepted knowledge that verbs are less commonly borrowed than nouns (Myers-Scotton 2002), the nouns in the sample that are selected as being shared between Bangime and the Eastern Dogon languages have significance, and are thus more susceptible to being borrowed into Bangime vocabulary to further portray their projected Dogon identity. The Dogon were made famous by the assertion of Griaule and Dieterlen (1965) that the Dogon Sigui ritual celebrates the presence of a star that is unable to be viewed without a telescope. Although it is unlikely that Dogon possess the ability to view the particular star in question, stars are an integral part of daily Dogon life, providing navigational guidance in the barren plains and details by which they plan events such as rituals and plantings. As mentioned above in Section 4, throughout the Dogon and Bangande communities, songs are almost always performed in Tommo So, independent of whether or not the person singing speaks or understands the language. Even the word in our dataset for the concept ROOF, listed as [taŋa] and attributed to Jenaama by Blench (2005, p. 12), is more likely from the languages listed in Table 10, and yet Bangande building style with two-story houses is not shared with Dogon, nor any other village construction in Mali.

DOCULECT	SONG	STAR	ROOF	GO DOWN	HELP	PUSH	SHOW
Bankan Tey	✓		✓	✓	✓	✓	✓
Ben Tey	✓	✓	✓		✓	✓	✓
Jamsay	✓	✓	✓	✓	✓		✓
Nanga	✓		✓	✓	✓	✓	✓
Perge Tegu	✓	✓	✓	✓	✓	✓	✓
Tebul Ure	✓			✓	✓	✓	✓
Togo Kan	✓	✓		✓	✓	✓	✓
Tommo So	✓		✓	✓	✓	✓	✓
Toro Tegu	✓		✓	✓	✓	✓	✓
Yanda Dom	✓		✓	✓	✓	✓	✓
Yorno So				✓	✓	✓	✓

Table 10: Ancient Shared Cultural Vocabulary

Albaugh (2018) makes correlations that languages spoken in Africa's secluded and mountainous regions are more likely to be preserved than those spoken in more easily accessible areas. Bangime, and the Dogon languages, are excellent examples of this trend. The mountains have undoubtedly contributed to the cohesiveness of the Dogon languages as a whole, and yet Bangime has not succumbed to its speakers pressures to preserve a Dogon identity.

6 Conclusion

Our study confirms that Bangime is a language isolate, yet the effects of both recent and distant contact, crucial to solving the mysteries surrounding and positing a reasonable history of these 'secret' people, are difficult to observe using automatic cognate detection methods alone. The more conservative method rejects borrowings, while even the lenient method sometimes misses them, requiring careful manual inspection of the wordlists informed by additional knowledge of Bagande culture and their own understanding of their ethnological and linguistic identity. Alinei (2004) speaks of "alloglottic linguistic colonies", and notes that language enclaves survive because their speakers want to retain an identity associated with their homeland. In the case of Bangime, the opposite holds true - Bangande claim that their homeland lies with the Dogon in Mande, and yet one of their most important sources of perceived identity, their language, speaks a different story.

While Bangande consider themselves Dogon, and Dogon consider themselves to be the young brothers of the Mande, there is no lexical evidence to suggest that any of these groups are related to each other. More likely, the Bangande escaped whatever caused the annihilation of their ancestors and came to settle in the secluded Geou valley prior to Dogon settlement. As the influence of the western Dogon languages seems relatively more recent than that of the eastern groups, the Dogon peoples seem to have come to the Bandiagara escarpment from different directions and settled at different times. No dating or archeological research has been conducted in the area where the Bangande reside, but if they did live in the cliffs before the Dogon, then perhaps they are 'those who came before' -- the Tellem.

But, if the Bangande have been, not only in contact, but practically immersed in the surrounding Dogon cultures for at least half a millennium then it is that much more surprising

that they have managed to keep their language intact. Samar and Bhatia (2017, p. 62), "...we are not sure if there is such a case where two compatible languages have been in contact for more than, say, 500 years, without any of them dying". If so, then Bangime represents an interesting and vanishingly rare counterexample. Not only is it not dead, it is thriving; around 1,500 speakers use it on a daily basis in their homes and with their children.

Given the neighboring Dogon groups' disdain for the Bangande, we can propose that a reason why Bangande have not mixed genetically with their Dogon neighbors (cf Babiker et al. 2018) is because Dogon village men do not accept to marry Bangande wives; though the opposite holds true. In traditional West African society, arranged marriages forbid the joining of certain groups to one another, such as between the Bozo and the Dogon. However, given the geographic proximity, language contact is inevitable. Bangime is a 'secret language' to its neighbors; linguists classify it as an isolate; and since it is likely currently spoken in a remote location, at a great remove from its original, ancestral speakers, it is a language island. In identifying the words that have washed up on its shores over time, we may someday be able to track its passage, explain its practices, and gain from its knowledge and experiences. Future interdisciplinary research should be pursued, aided by computational resources that can decipher the linguistic, genetic, and anthropological clues.

Supplemental Material

The supplementary material contains the Python code along with the data that are needed to replicate the analyses discussed in this study along with usage instructions. The supplementary material has been submitted to the Zenodo, where it can be accessed at https://zenodo.org/record/1407141. The data is curated at GitHub, where it can be accessed at https://github.com/lingpy/language-islands-paper.

Acknowledgements

We gratefully acknowledge the support and generosity of Professor Russel Gray, Hiba Babiker, and the Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History, without which the present study could not have been completed. Abbie Hantgan's work was additionally supported by Jeffrey Heath and the Dogon and Bangime Linguistics Project under NSF Grant award #BCS-1263150. Johann-Mattis List was funded as part of the European Research Council Starting Grant 'Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics'. We immensely appreciate the assistance and support provided by MPI colleagues Heidi Colleran, Anne-Maria Fehn, Tom Güldemann, Simon Greenhill, Martin Haspelmath, Ezequiel Koile, Ana Kondic, Adam Powell, Martine Robbeets, Christoph Rzymski, Robert Spengler, Natalie Uomini, Annemarie Verkerk, as well as the organizers and participants of the Workshop on Linguistic Islands in Africa. Responsibility for any errors in the resulting work remains our own.

References

Albaugh, E. (2018). "Language movement and civil war in West Africa". In: *Tracing Language Movement in Africa*. Ed. by E. A.K. M. de Luna. Oxford University Press, pp. 187–212.

Alinei, M. (2004). "Conservation and Change in Language". In: *Origin of European Languages*.

Ark, R. van der, P. Mennecier, J. Nerbonne, and F. Manni (2007). “Preliminary identification of language groups and loan words in Central Asia”. In: *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. (Borovets), pp. 13–20.

Babiker, H., F. Reed, and J. Heath (2018). *The genetic identity behind the masks: Bangande and Dogon of Western Africa*. Unpublished manuscript.

Bedaux, R. M. A. (1972). “Tellem, reconnaissance archéologique d’une culture de l’Ouest africain au Moyen-Age: recherches architectoniques”. In: *Journal de la Société des Africanistes* 42.2, pp. 103–185.

Bendor-Samuel, J., E. Olsen, and A. White (1989). “Dogon”. In: *The Niger-Congo languages*. Ed. by J. Bendor-Samuel. Lanham MD/New York/London: University Press of America, pp. 169–177.

Bertho, J. (1953). “La Place des Dialectes dogon (Dogõ) de la Falaise de Bandiagara parmi les autres Groups Linguistiques de la Zone Soudanaise”. In: *Bulletin de l’Institut Francais d’Afrique Noire* 15.1, pp. 405–441.

Blench, R. (2005). *Bangime, a language of unknown affiliation in northern Mali*. online.

– (2017). “African Language Isolates”. In: *Language Isolates*. Ed. by L. Campbell. London and New York: Routledge, pp. 162–192.

Bryant, D. and V. Moulton (2004). “Neighbor-Net. An agglomerative method for the construction of phylogenetic networks”. In: *Molecular Biology and Evolution* 21.2, pp. 255–265.

Calame-Griaule, G. (1956). “Les dialectes Dogon”. In: *Africa* 26.1, pp. 62–72.

Campbell, L. (2016). “Language Isolates and Their History, or, What’s Weird, Anyway”. In: *Berkeley Linguistics Society* 36.01, pp. 16–31.

– (2017). “Introduction”. In: *Language Isolates*. Ed. by L. Campbell. London and New York: Routledge, pp. 01–18.

Desplagnes, L. (1907). *Le Plateau Central Nigerien*. Paris: La Rose.

Dieterlen, G. (1955). “Mythes et organisation sociale au Soudan français”. In: *Journal de la Société des Africanistes* 25 (1,2), pp. 39–76.

Dumestre, G. (2011). *Dictionnaire bambara–français*. Paris: Karthala.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (forthcoming). “Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics”. In: *Scientific Data*.

Greenhill, S. J., R. Blust, and R. D. Gray (2008). “The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics”. In: *Evolutionary Bioinformatics* 4, pp. 271–283.

Gregersen, E. A. (1976). “The glottochronological performance of African languages”. In: *Cahiers de l’Institut de Linguistique de Louvain* 3.5-6, pp. 107–146.

Griaule, M. (1938). *Masques Dogons*. Paris: Institut d’Ethnologie.

Griaule, M. and G. Dieterlen (1965). *Le Renard Pâle*. Paris: Institut d’Ethnologie.

Hammarström, H., S. Bank, R. Forkel, and M. Haspelmath (2018). *Glottolog 3.2*. Max Planck Institute for the Science of Human History. Jena.

Hantgan, A. (2013). “Aspects of Bangime Phonology, Morphology, and Morpho-syntax”. Doctoral dissertation. Indiana University.

Hantgan, A. and J. Heath (2016). “Bangime lexicon”. unpublished wordlist.

Heath, J. (2005). *Tondi Songway Kiini (Songhay, Mali) : reference grammar and TSK-English-French dictionary*. Stanford, California: CSLI Publications.

– (2015). *Dictionary Humburi Senni (Songhay of Hombori, Mali) - English - French*.

– (2016). “Jenaama lexicon”. unpublished wordlist.

Heath, J. and A. Hantgan (2018). *A Grammar of Bangime*. Berlin, Boston: De Gruyter Mouton.

Heath, J., L. McPherson, K. Prokhorov, and S. Moran (2015). “Dogon Comparative Wordlist”.

Hochstetler, J., J. Lee Durieux, and E. Durieux-Boon (2004). “Sociolinguistic Survey of the Dogon Language Area”. In: *SIL International*.

Huson, D. H. and D. Bryant (2006). “Application of Phylogenetic Networks in Evolutionary Studies”. In: *Molecular Biology and Evolution* 23.2, pp. 254–267.
Huson, D. H. (1998). “SplitsTree: analyzing and visualizing evolutionary data”. In: *Bioinformatics* 14.1, pp. 68–73.
Izard, M. (1970). *Introduction à l’histoire des royaumes mossi*. Paris and Ouagadougou: Eds du CNRS-CVRS.
Jäger, G., J.-M. List, and P. Sofroniev (2017). “Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*. EACL 2017. Valencia: Association for Computational Linguistics, pp. 1204–1215.
Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.
Lee, Y.-J. and L. Sagart (2008). “No limits to borrowing: The case of Bai and Chinese”. In: *Diachronica* 25.3, pp. 357–385.
Levenshtein, V. I. (1965). “Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov”. In: *Doklady Akademij Nauk SSSR* 163.4, pp. 845–848.
List, J.-M. (2012a). “LexStat. Automatic detection of cognates in multilingual wordlists”. In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*. Stroudsburg, pp. 117–125.
– (2012b). “SCA: Phonetic alignment based on sound classes”. In: *New directions in logic, language, and computation*. Ed. by M. Slavkovik and D. Lassiter. Berlin and Heidelberg, pp. 32–51.
– (2014). “Investigating the impact of sample size on cognate detection”. In: *Journal of Language Relationship* 11, pp. 91–101.
– (2015). “Network perspectives on Chinese dialect history”. In: *Bulletin of Chinese Linguistics* 8, pp. 42–67.
– (12/2016). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Tech. rep. Jena: Max Planck Institute for the Science of Human History.
List, J.-M., S. Nelson-Sathi, H. Geisler, and W. Martin (2014). “Networks of lexical borrowing and lateral gene transfer in language and genome evolution”. In: *Bioessays* 36.2, pp. 141–150.
List, J.-M., M. Cysouw, and R. Forkel (2016). “Concepticon. A resource for the linking of concept lists”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC 2016. (Portorož). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. European Language Resources Association (ELRA), pp. 2393–2400.
List, J.-M., S. Greenhill, and R. Forkel (2017a). *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena: Max Planck Institute for the Science of Human History.
List, J.-M., S. J. Greenhill, and R. D. Gray (2017b). “The potential of automatic word comparison for historical linguistics”. In: *PLOS ONE* 12.1, pp. 1–18.
List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018a). “CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats”. In: *Linguistic Typology* 22.2, pp. 277–306.
List, J.-M., M. Cysouw, S. Greenhill, and R. Forkel (2018b). *Concepticon. A Resource for the linking of concept list*. Version 1.1. URL: http://concepticon.clld.org/.
Macdonald, K. C. (1997). “More forgotten tells of Mali: an archaeologist’s journey from here to Timbuktu”. In: *Archaeology International* 1, pp. 40–42.
Marchal, J.-Y. (1978). “Vestiges d’occupation ancienne au Yatenga (Haute-Volta): une reconnaissance du pays Kibga”. In: *Cahiers ORSTOM. Série Sciences Humaines* 15.4, pp. 449–484.
Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The ‘organic’ approach to linguistic comparison*. Institute for the Study of Human Issues.

Mayor, A. and E. Huysecom (2016). ""Toloy", "Tellem", "Dogon": une réevaluation de l'histoire du peuplement en Pays dogon (Mali)". In: *Regards scientifiques sur l'Afrique depuis les indépendances*. Ed. by M. Lafay and E. C. F. Le Guennec-Coppens. Paris: Karthala, pp. 333–350.

Mayor, A., E. Huysecom, A. Gallay, M. Rasse, and A. Ballouche (2005). "Population dynamics and Paleoclimate over the past 3000 years in the Dogon Country, Mali". In: *Journal of Anthropological Archeology* 24, pp. 25–61.

Mayor, A., E. Huysecom, S. Ozainne, and S. Magnavita (2014). "Early social complexity in the Dogon Country (Mali) as evidenced by a new chronology of funerary practices". In: *Journal of Anthropological Archeology* 34, pp. 17–41.

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). "A Central Asian language survey". In: *Language Dynamics and Change* 6.1, 57–98.

Moran, S. and M. Cysouw (2017). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Zürich: Zenodo.

Moran, S. and R. Forkel (2017). *cldf/segments: segments 1.1.1*.

Moran, S. and J. Prokić (2013). "Investigating the Relatedness of the Endangered Dogon Languages". In: *Literary and Linguistic Computing* 28.4.

Moran, S., R. Forkel, and J. Heath, eds. (2016). *Dogon and Bangime Linguistics*. Jena: Max Planck Institute for the Science of Human History.

Morrison, D. A. (2014). "Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis". In: *WIREs Data Mining and Knowledge Discovery*.

Myers-Scotton, C. (2002). *Language contact: Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.

Nunn, N. and D. Puga (2012). "Ruggedness: The Blessing of Bad Geography in Africa". In: *The Review of Economics and Statistics* 94.1, pp. 20–36.

Osborn, D. Z., J. I. Donahoe, and D. J. Dwyer (1993). *A Fulfulde (Maasina)-English-French lexicon: a root-based compilation drawn from extant sources followed by English-Fulfulde and French-Fulfulde listings = Lexique Fulfulde (Maasina)-Anglais-Francais*. East Lansing: Michigan State University Press.

Plungian, V. A. and I. Tembine (1994). "Vers une description sociolinguistique du pays Dogon: attitudes linguistiques et problèmes de standardisation". In: *Stratègies communicatives au Mali: langues régionales, bambara, francaise*. Ed. by G. Dumestre. Paris: Didier Erudition, pp. 163–195.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). "Inference of population structure using multilocus genotype data". In: *Genetics* 155, 945–959.

Prokhorov, K., J. Heath, and S. Moran (2012). "Dogon classification". In: *Proto-Niger-Congo: Comparison and Reconstruction International Congress*. Paris.

Rama, T., J.-M. List, J. Wahle, and G. Jäger (2018). "Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?" In: *Proceedings of the North American Chapter of the Association of Computational Linguistics*. NAACL 18. (New Orleans).

Ross, M. and M. Durie (1996). "Introduction". In: *The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. New York: Oxford University Press, pp. 3–38.

Samar, R. G. and T. K. Bhatia (2017). "Predictability of language death: Structural compatibility and language contact". In: *Language Sciences* 62, pp. 52–65.

Segerer, G. and S. Flavier (2016). *RefLex: Reference Lexicon of Africa*. Version 1.1. Paris, Lyon.

Simons, G. F. and C. D. Fennig, eds. (2018). *Ethnologue: Languages of the World, Twenty-first edition*. URL: http://www.ethnologue.com.

Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts". In: *Proceedings of the American Philosophical Society* 96.4, pp. 452–463.

– (1955). "Towards greater accuracy in lexicostatistic dating". In: *International Journal of American Linguistics* 21.2, pp. 121–137.

Tadmor, U. (2009). “Loanwords in the world’s languages”. In: *Loanwords in the world’s languages. A comparative handbook*. Ed. by M. Haspelmath and U. Tadmor. Berlin and New York: de Gruyter, pp. 55–75.
Tishkoff, S. A. et al. (2009). “The Genetic Structure and History of Africans and African Americans”. In: *Science (New York, N.Y.)* 5930.324, pp. 1035–1044.