ERC Consolidator Grant 2021

The Mysterious Bang: A Language and Population Isolate Unlocks the Secrets of Interior West Africa's Lost Ethnolinguistic Diversity

BANG

- **Principal investigator (PI):** Dr. Abbie Hantgan
- **Host institution:** Langage, Langues et Cultures d'Afrique (LLACAN, UMR 8135)
- **Proposal duration:** 60 months

Species are often the most diverse at their origins, and all modern humans can trace their origins back to Africa. Our ability to communicate through language defines us; every human being speaks (or signs) at least one language. We may never know when our current forms of language were first spoken, or what speakers sounded like, but we can be assured that primordial languages were spoken in Africa. Yet, a paradox presents itself with respect to the apparent lack of linguistic heterogeneity in Africa: the languages comprise only four of the world's more than 150 families. Isolates, languages with no known living relatives, represent the missing links between today's homogeneity and the past's diversity.

Bangime may represent the only confirmed language isolate spoken today in interior West Africa. Its speakers, the **Bang**ande, have resisted genetic admixture with neighboring populations for upwards of 9.000 years. The root of their eponym [**BANG**], means 'hidden' or 'secret' in surrounding Dogon languages whose linguistic and genetic ancestors are almost as mysterious.

ERC-**BANG** will search for evidence of contact or inheritance between Bangime and a hitherto unexplored set of West African languages and populations using computationally-supported methods. Our research team will conduct an in-depth study of culturally significant and historically relevant loan words, so that a statistician can incorporate these findings into stochastic models of language diffusion. A geneticist will process the entire Bangande genome to compare with West African populations from a wider range than thus far considered. Together, we will perform a character-based Bayesian phylogenetic analysis of the Dogon languages to estimate the time depth of the group in order to test settlement hypotheses and proposed migration patterns. Completion of this project will uncover traces of vanished ethnolinguistic varieties in West Africa, and our methods can be replicated to solve similar questions.

c. **State-of-the-art and objectives**

c.1. Research Questions

ERC-**BANG** aims to further confirm that isolate **Bang**ime represents remnants of former linguistic diversity in West Africa (cf. Dimmendaal (2008)), while pursuing preliminary findings that the **Bang**ande are a population isolate (Babiker et al. 2020; Hantgan, List, and Babiker 2020). Our computer-assisted comparison of borrowing patterns compared with admixture events will uncover the lost migratory paths of the Bangande people. Although many studies have compared language contact with genetic drift, even addressing an African isolate (Tishkoff, Gonder, et al. 2007), to my knowledge, ERC-**BANG** is the first large-scale project leveraging cutting-edge, cross-disciplinary methods and the results of extensive fieldwork to reconstruct the history of a language isolate and its speakers, with a focus on the lateral and vertical signals in the data.

Thus, the proposed project is designed to answer the following questions:

- How do Bangime/Bangande fit within the wider networks of West African languages/populations?
- From where, and when, did the Bangande migrate to, and settle in, their current home among Dogon peoples atop the Bandiagara Escarpment?
- How and why do language/genetic isolates like Bangime/Bangande persist in the face of so many challenges and contradictions?

In order to answer these questions, the proposed ERC-**BANG** team will amass *big data* to compare the Bangime language and Bangande people with geographically proximate and distant languages and populations to trace historical West African diversity. Specifically, the project is designed to achieve the following objectives:

c.2. Research Objectives

The overall goal of ERC-**BANG** is to unravel the **Bang**ime/**Bang**ande mystery using an innovative computationally supported methodology and drawing on cross-disciplinary expertise.

- To broaden the search beyond immediately neighboring groups to further confirm Bangime and the Bangande as a linguistic and genetic isolate or to discover a temporally and geographically distant ancestor, using the latest computer-assisted language technologies and character-based Bayesian phylogenetic analysis;
- To document the Bangande's path to their remote geographic location and the conditions that drove them there using a range of cross-disciplinary sources and methods, including an in-depth study of culturally significant and historically relevant loan words to produce stochastic models of language diffusion;
- To determine the genealogical and genetic relationships of the surrounding languages and peoples of contested affiliation by analyzing and comparing previously collected genomic data;
- To share our findings with the Bangande in an ethically sensitive and culturally appropriate manner;
- To offer a methodology other researchers/communities can use to answer similar intransigent questions.

c.3. Impact

As opposed to the Big Bang that purportedly created our universe, the **BANG**-mystery represents a relatively small origin story. However, the potential for discovery is great. Dimmendaal (2008) states that language isolates and languages with contested affiliations, with specific reference to Songhay in the Nilo-Saharan language phylum, Dogon and Mande in Niger-Congo, represent a now-lost linguistic diversity in West Africa. The unknown origin of the Bangime language and Bangande people is an impasse to our understanding of West African ancient history, and the development of language families. The above listed research objectives will serve a wider impact not only for linguists and geneticists interested in West Africa, but also by providing a new means by which researchers from any discipline may attempt to understand the origins of a language and population isolate. ERC-**BANG** will contribute to the scientific study of linguistic and population isolates by complementing evidence from different perspectives with innovative models of lexical distribution to detect events and periods of population and linguistic contact and therefore sketch a migration map of the Bangande people and their language. The next sections outline previous linguistics-genetics studies concentrated in Africa and preliminary findings regarding the Bangime language and Bangande people.

c.4. Hypotheses

Two competing hypotheses regarding the Bangime mystery have been proposed. Blench (2015) advanced the first: "…[the Bangande] must represent one of the layers of population on the Plateau prior to the expansion of the Dogon. There is some evidence for this in the presence of lexemes that resemble Bangime in the Dogon languages immediately adjacent to it, suggesting that there were formerly other languages related to Bangime which were assimilated by the Dogon" (p. 74). Hantgan, List, and Babiker (2020) and Hantgan and List (2018a) attribute apparent similarities between Bangime and adjacent Dogon languages to recent loans from neighboring Mande languages based on semantic distributions and the absence of such forms in more geographically distantly spoken Dogon languages.

The second and, for the moment at least, more plausible, hypothesis aligns with oral histories: like the Dogon, the Bangande were driven to seek refuge from slave raids, religious persecution, and imperial domination, but they arrived at their current location relatively recently and were subsequently prevented from settling farther along the Escarpment by Dogon groups to the east.

Neither hypothesis accounts for Bangime words at deeper levels of the lexicon, such as those for body parts and lower numerals, which strongly resemble those found in Dogon languages that are not spoken in the immediate area and with whose speakers the Bangande have infrequent, if any, contact today.

The most perplexing part of the **BANG** mystery is their deliberate retention of their endogamous marriage patterns and otherwise unique language despite their imitation of Dogon culture and their claim of Dogon integration. Babiker et al. (2020) estimate a 9,000 year bottleneck, which is striking given the proposed 1,000 year timeline of Bangande-Dogon co-habitation of the Bandiagara Escarpment.

Yet, another, delicate, piece of the puzzle has not been considered. After several years of intense fieldwork, the Bangande revealed another secret to me: they comprise two distinct societies, or castes: free and slave. Complex caste networks are prevalent throughout interior West Africa, particularly in Mali (Tamari 1991). The Bangande's revelation to me is supported by Babiker et al. (2020) who show that Bangande lineage is evidenced either by homozygous alleles or some contribution of neighboring groups. To a large extent, the admixture is mitochondrial. The society is patrilineal, and exogamy is practiced only within the "slave" caste. Naturally, Bangande with Dogon mothers are more apt to be conversant in a Dogon language than those who were raised by two monolingual Bangime-speaking parents.

Thus, this study will test a third hypothesis: Over the generations, "mixed" Bangande - those with Dogon ancestry - have contributed words to the Bangime lexicon that are now completely integrated into the speech of all Bangande. A thorough examination of borrowing patterns in comparison with whole-genome sequencing to examine migration and settlement patterns will shed light on this theory.

c.5. Previous Studies Combining Linguistic and Genetics

Partially due to an outdated over-conglomeration of language groupings (Greenberg 1948), the linguistic diversity in Africa pales in comparison to its genetic diversity, which is greater than anywhere else in the world (Retshabile et al. 2018). In West Africa there are only three linguistic phyla (Eberhard et al. 2021); Niger-Congo is the dominant language phylum, which includes the well-known Bantu sub-grouping.

Another contributing factor to this apparent paradox is that the majority of studies comparing languages with their speakers' genetic profiles have focused on Bantu-speaking groups, who are known to be both linguistically, as well genetically, homogeneous (Schlebusch and Jakobsson 2018). West African regions north of the Bantu expansion's presumed origin in Cameroon have been largely ignored. Lipson et al. (2020) illustrate how so-called "ghost populations" who contributed to modern humans' DNA occupied the upper regions of West Africa before the Bantu Expansion took place. Choudhury et al. (2020) confirm that genetically, Niger-Congo speakers outside the Bantu spread-zone pattern differently than those within it. They further show that in Mali, where the Bangande live today, individuals are even more divergent than those living in lower regions of West Africa. In the past three decades, our understanding of the relationships between individual languages and their speakers has been greatly enhanced by collaborations between linguists and geneticists. Cavalli-Sforza's group (Cavalli-Sforza et al. 1988) was among the first to show the resemblances between linguistic and genetic genealogical trees, but their conclusions are debated (MacEachern 2000). A more recent study of similar magnitude, Fan et al. (2019) shows corresponding linguistic and genetic relationships in Africa. A natural conclusion is that speaking one another's language is favorable for finding a marriage partner. Yet, in the case of the Bangande, spouses from other communities are forced to learn the complex Bangime language that is unlike any other spoken in the area.

Comparing languages with speakers' DNA involves many compatible mechanisms. Across the world, the main method of classifying languages into groups is based on *regular sound correspondences* across shared, *basic vocabulary*. Genomic sequences are compared in a similar fashion. Thus, Creanza et al. (2015) conducted a world-wide study comparing phoneme distribution (retention and spread) to genetic variation. They found that, unlike gene flow, phonemes in Africa tend to stay in Africa; the out-of-Africa expansion did not affect languages' phonemic inventories as it did populations' genomes. The authors further state that phoneme inventories are affected only by recent language contact and do not reflect ancient states. Researchers comparing gene flow and language dispersal - again within the Bantu expansion (Filippo C. et al. 2012; Veen et al. 2009) - found that even when a language is replaced, speakers retain some aspects, particularly the phonemic inventory. Idiatov and van de Velde (2021) found that the relatively rare cross-linguistic presence of labio-velar phonemes, which are robust in sections of Western Africa, represent, not an inherited trait, but one spread through close, long-standing language contact. It is noted in §a.6 that Bangime has an unusual phoneme inventory for the area.

One drawback of DNA sampling projects is that researchers have a preconceived notion of population groups based on established linguistic groupings. That is, when Tishkoff, F. A. Reed, et al. (2009) first sampled and depicted their results for "Dogon" populations, they were using a label assigned by linguists and anthropologists. Even for relatively small communities like the Dogon, linguistic diversity is sufficient to delineate 21 separate, mutually unintelligible languages. Dogon speakers must be multilingual to participate in life outside of their language community. Studies of genetic admixture often overlook speakers' linguistic practices beyond their ethno-linguistic label. In West Africa, occupation is important to understanding ethnicity.

Both Schlebusch and Jakobsson (2018) and Skoglund et al. (2017) highlight hunter-gatherers' importance to our understanding of prehistory on the African continent, but no pre-farming population studies include data from the area in which Bangime and the Dogon languages are spoken. As Babiker et al. (2020) note, Bangime is the only known language isolate spoken by a non-hunter-gatherer population in West Africa. The Kalash, a population isolate who speak an Indo-European language and live in Hindu Kush mountain ranges of present-day Pakistan have recently been found to be diverged from ancient Siberian hunter-gatherers (Ayub et al. 2015). Hunter-gatherers' languages are also said to have unique features (Bickel and Nichols 2020). Typological comparisons between Bangime and other isolates across the globe are highly anticipated.

An advantage of supplementing linguistics data with genetics data is to determine and confirm time spans. Phylogenetic analyses enable us to move beyond the famous declaration by A. McMahon and R. McMahon (2005), "linguists don't do dates", but the 10,000-year before the present (BP) boundary of linguistic reconstruction stated by Ringe (1995) still stands. Some scholars, such as Greenhill et al. (2010), have pushed the date back using typological traits that are thought to be more ingrained in, and interconnected to, a language system and therefore to change less rapidly than other aspects of the grammar. Further, Pagel et al. (2007) found that frequency of word usage increased 'half-life', at least for the family of languages they studied. Even with these advanced methods, especially given the recent surge in ancient DNA research, only geneticists can provide linguists with time depths far beyond that of language data.

c.6. Preliminary Collaborative Findings

Because the Bangande and the Dogon are the only occupants of the forbidding Bandiagara cliff range, early linguistic researchers considered Bangime an outlier Dogon language (Bertho 1953; Calame-Griaule 1956). The language received little special attention until Hochstetler et al. (2004) conducted a Dogon language mutual intelligibility survey and found cognates between Bangime and the Dogon languages to be less than 10%. Blench (2007) followed up and declared that, based on his knowledge of common Niger-Congo lexical cognates - the language phylum that dominates West Africa -, the language is an isolate.

Bangime remains classified as one of five African isolates (Eberhard et al. 2021). Dr. Johann-Mattis List and I used computationally-assisted methods, outlined below, to show that Bangime displays no affiliation to any of 22 Malian languages based on a comparison of 300 concepts (Hantgan and List 2018a), but the effect of language contact, especially with the Dogon, is extensive; preliminary estimates find upwards of 60 percent of even the Bangime core lexicon is borrowed (Hantgan and List 2018a), and some loan words come from languages with which the speakers have infrequent contact today (Hantgan, List, and Babiker 2020).

The map in Figure 1 illustrates the position of Bangime and surrounding ethnolinguistic groups along and adjacent to the Bandiagara Escarpment in Central-eastern Mali. Note that each point is but one village or settlement; groups such as Fulfulde-speaking Fulani are found throughout West Africa.

Today, the Bangande occupy seven small villages at the end of a valley on the Bandiagara Escarpment in central-eastern Mali, around 300 kilometers south of the legendary town of Timbuktu and just east of the Niger River. They are surrounded to the north and east by Dogon speakers in whose languages the root [**BANG**] translates as 'secret' or 'hidden'. Ironically, although the Bangande consider themselves and their language to be Dogon, only some speak a Dogon language. Furthermore, the speakers of the Dogon languages from which the lexeme [**BANG**] is drawn have little contact with Bangime speakers today; they live more than 40 kilometers away in the northeastern portions of the cliffs.

To the west of the Bangande villages, in the valley between the cliffs, live a group of farmers who speak a language called Jenaama. Jenaama belongs to the wide-spread Mande language family. Both Mande and Dogon constitute their own branches within Niger-Congo, and neither of their positions within the phylum is well defined (Güldemann 2018). Ethno-occupationally, all other Bozo practice fishing. The Bandiagara Escarpment blocks the spread of the Sahara Desert to the north, so the area is quite lush compared to the surrounding Sahel, making it an ideal area for farming. In fact, the valley floods in the rainy season, causing the Bangande area to be almost inaccessible for at least three months of the year. Since the Bangande identify culturally with the Dogon, they adhere to a long-standing feud that bars intermarriage between Bozo and Dogon peoples. Babiker et al. (2020), whose research is presented below, confirms that the Bangande and Bozo share no admixture.

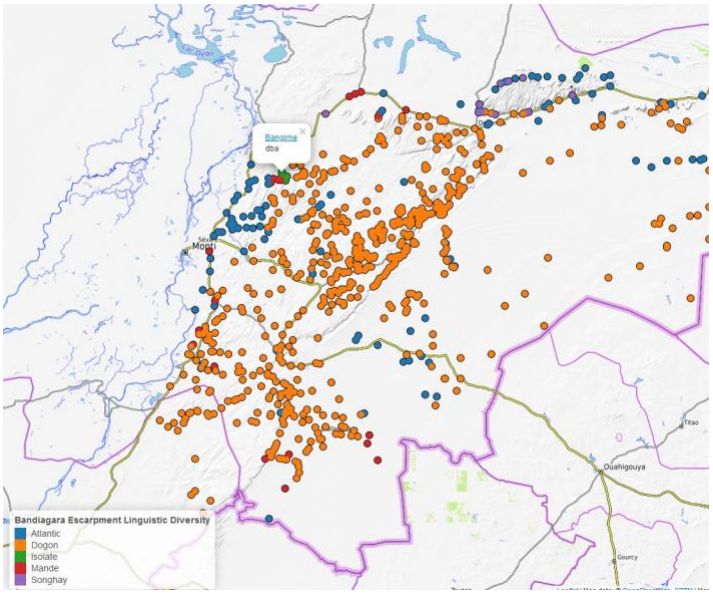

Figure 1: Bandiagara Escarpment Area (for further geographic detail, see the Interactive Bandiagara Escarpment Map)

The only other ethno-linguistic group with which the Bangande converse on a regular basis is the Maasina Fulfulde-speaking herders who graze their animals throughout the valley and cliff range. In terms of more distant language contact, the Bangande are forced to travel the 25-kilometer-long valley every week to visit the closest market town. At the market, they encounter various ethno-linguistic groups, including two additional Mande languages: the widely spoken Bambara of southern Mali and Soninke, which is relatively closely related to the Bozo languages, including Jenaama. To the north of the cliffs are found speakers of Songhay, and Berber speakers of Tamasheq, an Afro-Asiatic language, are found in the far north, but Bangande have infrequent contact with either of these groups. The Songhay languages are currently affiliated with a different phylum, Nilo-Saharan, but these are also of disputed classification (Dimmendaal 2014; Bendor-Samuel 2000). Songhay languages have been heavily influenced by surrounding speakers through language contact (Souag 2012), and their effects on Dogon languages are becoming apparent (Hantgan, List, and Babiker 2020).

Postdoctoral researcher Hiba Babiker from the Max Planck Institute for the Science of Human History (MPI-SHH) collected and analyzed 270 DNA samples from 12 villages in 2012, including 3 Bangande villages. Her preliminary findings (Babiker et al. 2020) show that the Bangande also represent a population isolate; they are only distantly genetically related to the populations thus far sampled in Africa and distinct from the 9 populations Babiker analyzed (Babiker, Hantgan, List, Heath, and R. Gray 2018). Similar to the Ari Blacksmiths of Ethiopia (Dorp et al. 2015), Babiker et al. (2020) attributes the genetic isolation of the Bangande to a bottleneck caused by their geographic location. However, the reason why they do not intermarry with neighboring Dogon populations with whom the Bangande claim linguistic and ethnic affinity is one of the mysteries ERC-**BANG** will investigate.

The Dogon peoples have been the topic of countless studies for nearly a century, but none have resolved many of their mysteries either. The first were ethnographic; the most infamous is the now largely debunked research of anthropologist Marcel Griaule (Griaule 1938). Archaeologists have been excavating the ancient caves of the so-called Tellem and Toloy in the Bandiagara Escarpment since the 1970s (Bedaux 1972). Recent paleoclimatic evidence combined with excavations focused on the eastern parts of the cliff range suggest a "mosaic" of pre-Dogon populations (Mayor and Huysecom 2016). Mayor et al. (2005) estimate that the Dogon arrived at the Bandiagara cliffs within the last thousand years, but their earlier history remains unknown.

Historians suggest that the Dogon settled the cliff range in flight from the Mali and Songhay empires from the 13th to the 17th centuries, and/or to escape religious persecution and enslavement (Brooks 1993). Nunn and Puga (2012) specifically discuss the advantages of the rough and remote Bandiagara Escarpment as a refuge. Based on our preliminary explorations of borrowings from neighboring Mande and Songhay languages into Bangime and Dogon Hantgan, List, and Babiker (2020) propose that the Bangande followed a similar trajectory.

The Dogon are a larger group than the Bangande; their languages number at least 21, and, contrary to depictions of their cultural homogeneity (Bouju 1995), their actual practices vary, depending on the area of the cliff range that they inhabit. Like the Bangande, their DNA offers little-to-no genetic admixture whatsoever; Babiker et al. (2020) show that the connection to South African populations put forth by Tishkoff, F. A. Reed, et al. (2009) is misleading. Similarly, the Dogon languages were once grouped with Gur languages spoken to the east of the cliffs, but Sands (2019) establishes this mis-classification was probably based on lexical borrowings through language contact rather than true cognates, words shared through descendence from a common ancestral language.

Mali is included in an area that has been known for the past decade as the "Macro-Sudan belt" (Güldemann 2008; Clements and Rialland 2008), based on a set of areally robust grammatical and phonological features. One of the Macro-Sudanic belt features is [±ATR] (advanced tongue root) vowel harmony. Many languages of the Akan group in Ghana display a ten-vowel system with [±ATR] contrasts at the high and low vowel heights but this is unusual for central-Mali. Still, Hantgan and Davis (2012) argue for a ten-vowel phonemic inventory based on [±ATR] vowel harmony in the Dogon language Bondu-so, even though the surface inventory is phonetically restricted to seven vowels. This patterning can be explained through typological and phonotactic universal constraints on the difficulty of simultaneously pronouncing [-ATR] high and [+ATR] low vowels. Today, no Dogon language has the full [±ATR] contrast for all vowel heights, but some, such as Bunoge, appear to pattern similarly to Bondu-so in exhibiting traces of such a system.

These findings mentioned in §a.5 concerning phoneme and gene distributions are of special interest because even the first superficial comparisons revealed that the sound structure of Bangime differs significantly from that of the Dogon languages. Bangime speakers contrastively use the phonemes /ɕ/, the voiceless alveolo-palatal fricative, and /ɥ/, the voiceless alveolo-palatal fricative, neither of which are found phonemically among Dogon languages and otherwise only occur as phonemes in the Mande languages Jenaama and Soninke, and languages of Ghana such as Akan. Similarly, the tripartite tonal system of Bangime is unlike any of the surrounding languages except some varieties of Bozo; otherwise the closest language is again Akan in Ghana. Although Soninke-Bozo speakers are found in geographic proximity Bangime, as noted above, there is no genetic admixture between the Bangande and Jenaama speakers (whose language is closely related to Soninke) sampled from the immediate vicinity (Babiker et al. 2020). These evidences of language transfer at the level of the phonemic inventory without any genome admixture will contribute to the continuing discussion of the impact of gene versus phoneme distribution among populations and languages.

d. **Methodology**

d.1. Computationally Supported Approaches to Language Comparison

Comparative and historical linguists examine and classify the world's ways of speaking into languages and dialects based on the relative ease with which people can communicate with each other. An expert can determine that divergent languages derive from one spoken source, the Proto-language. Language classification, in turn, helps us to understand the genealogical relationships between languages and, by consequence, speakers.

Our approach to language classification and comparison follows the standards advanced by the project Computer-Assisted Language Comparison *Computer Assisted Language Comparison* (CALC `http:calc.digling.org` ERC grant ID #715618). The workflow requires researchers to follow strict guidelines to prepare their data to be interpreted computationally (Wu et al. 2018).

This approach also advocates for the use of FAIR data principles in which data are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). To this end, all of our data are available online at digiling.org and we share the code for each project through organizations on Github. To date, we have relied on wordlists for cross-language comparison and computer-assisted cognate detection. Through the use of stable identification numbers associated with each lexical entry, we ensure that each datum is traceable back to its original source. The data that have been used were not necessarily gathered for the purpose of comparison. Thus, these linguistic data must be rendered comparable. To accomplish this goal, we adhere to the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al. 2018).

The traditional comparative method requires intense dedication, commitment, and extensive expertise. First, researchers who spend a great deal of time working with speakers of a particular language group (usually as defined geographically) compile data from different parts of the language; the lexicon is often the most robust area of study. From there, we seek to discover if *sound correspondences* - pronunciation-meaning patterns - exist. These patterns are most detectable in multilingual wordlists. Table 1 shows an example of sound correspondences between Dogon language subgroups.

GROUP	LANGUAGE	SOURCE TRANSCRIPTION	TOKENIZED REPRESENTATION	CONCEPTICON CONCEPT
W-Dogon	Bondu So	gòʒúú	g ¹o **ʒ** ⁵u:	BODY
E-Dogon	Ben Tey	gèsú	g ¹e **s** ⁵u	BODY
W-Dogon	Bunoge	kíʒè	k ⁵i **ʒ** ¹ɛ	ANSWER
E-Dogon	Bankan Tey	kísɛ́	k ⁵i **s** ⁵ɛ	ANSWER

Table 1: Regular sound correspondences [**ʒ**] ~ [**s**] between western and eastern Dogon

Like DNA base-pairs and sequences, phonemes, the smallest analyzable units of language, can be aligned and systematically compared across languages to detect whether the same changes occur with significant frequency. Each of the pairs of words for the concepts BODY and ANSWER in Table 1 can be considered cognate - words shared through descendance from a common ancestral language - not only because each one closely resembles the other in terms of its individual sounds (see the column representing the tokenized representation), but also because pronunciation patterns hold across concepts. These patterns occur between language pairs and across language subgroups, but not over language families. That is, the sound [ʒ] regularly appears among languages spoken in the western Dogon sphere, while [s] appears among those spoken in the east. Otherwise, the two pairs of words are nearly identical, rendering this regular sound correspondence: [ʒ] ~ [s] quite obvious. Other correspondences are less clear, with multiple stages of change, and they are also tedious to detect by hand. Furthermore, validating these correspondences requires adequate amounts of data. Computational approaches provide obvious range, scale, specificity, subtlety, and speed advantages.

However, simply having access to lexical wordlists is insufficient to perform comparative analyses. Primarily, transcription conventions differ widely among researchers. Phonetic transcriptions from the original source must be retained, yet converted to the standard, International Phonetic Alphabet (IPA). To implement conversion, we first create an orthography profile that lists each phonetic realization in the language as it was originally transcribed together with the corresponding, appropriate IPA symbol. Fortunately, Moran and Cysouw (2017) provide Python scripts that automatically convert the existing orthographies into the *tokenized representations*, which are essential for *phonetic alignment*. Table 2 illustrates a converted example.

GROUP	LANGUAGE	SOURCE TRANSCRIPTION	TOKENIZED REPRESENTATION	SOURCE TRANSLATION	CONCEPTICON CONCEPT
Isolate	Bangime	nàà	n ¹ a:	cow	COW
E Dogon	Ben Tey	nă:m	n ¹⁵ a: + m	cow, bull (any bovine)	COW
NW Mande	Jenaama	nà	n ¹ a	cow	COW
W Mande	Bambara	mìsí	m ¹ i s ⁵ i	vache, boviné domestique	COW

Table 2: Comparative lexemes for the concept COW across languages

For computational methods to be successful, it is crucial that we accurately segment every word to permit phonetic alignment. Diacritics marking tone, nasalization, and any other super-segmental or prosodic features are represented as separate segments to be compared apart from the segmental portions of the lexeme under consideration. This step is crucial for an algorithm to accurately assess whether sound correspondences are regular. Morpheme boundaries are represented with a [+]. Often, transcriptions are inconsistent with regard to morpheme boundaries. We must also ensure that semantic glosses are the same for a given concept. It is only through the tokenized representation and linking to universal Concepticon (List, Rzymski, et al. 2020) concepts that the data become readable to the computer-based methods. Thus far, our automatic cognate and borrowing detection methods have included only the lexical root, rather than the root plus additional morphemes; those delineated in Table 2 were previously ignored. Yet, affixes are a crucial aspect of comparative historical reconstruction. List et al. (2016) have shown that comparing words at both the level of the root as well as affixes provides deeper time depth and exposes unforeseen relationships among languages.

As an example that pertains to our study, a criterion for inclusion in the Niger-Congo language phylum is the presence of noun class affixes. Some Atlantic languages mark nouns with upwards of 15 class affixes and corresponding agreement on adjectives and numerals. Mande languages do not have noun classes, which is one of the reasons their affiliation is dubious. Although not discussed in the published literature, causing many experts to exclude Dogon from Niger-Congo (cf. Hepburn-Gray (2020)), some Dogon languages display a full noun-class system, while others have traces of, or residual, noun-class suffixes. Preliminary examples of computer-assisted comparative partial cognate recognition in progress are shown in Table 3.

GROUP	LANGUAGE	SOURCE TRANSCRIPTION	TOKENIZED REPRESENTATION	COGIDS	BORID
W Mande	Bambara	mìsí	m ¹ i s ⁵ i	840	0
NW Mande	Jenaama	nà	n ¹ a	92	1354
Isolate	Bangime	nàà	n ¹ a:	91	1354
NE Dogon	Bankan Tey	nàá-m	n ¹⁵ a: + m	841 842	1354
NE Dogon	Ben Tey	nàám	n ¹⁵ a: + m	841 842	1354
NW Dogon	Dogul Dom	nàá-g	n ¹⁵ a: + g	841 844	1354
Atlantic	Fulfulde	nag-ge	n a g + g e	845 846	1354

Table 3: Noun Class Affix Alignment and Comparison: Terms for COW

To determine if a word is cognate across languages, each phoneme is compared within and across sets to establish that sound correspondences are regular. The algorithm that determined the first COGID in Table 3 found that the phonemes and tones in the root for COW in the Dogon languages are not only in the same position across different concept sets, but that the sounds correspondences are regular, and thus, the forms are cognate. The second COGID the algorithm produces is based on the material following the [+]; that is, after the morpheme boundary, between the root and the noun class suffix in examples in Table 3. The Dogon languages have different suffixes for the classification of the noun COW, and thus represent two levels of potential cognancy: one at the root, and the second at the suffix. This is represented by the second number in the COGID column.

An additional advantage to using computer-assisted methods is their ability to discern the difference between cognates and borrowings. Historically, genealogy and influence by language contact are notoriously difficult to differentiate computationally. Yet, we, as well as List (2019) profit from the discrepancies between two algorithms: LexStat-InfoMap (List 2012a) and Sound-Class-Based Alignment (SCA) (List 2012b); only the former looks for regular sound correspondences. Cognates are confirmed when the results from the two models are the same; a lexical borrowing can be deduced from disagreement. The BORID column in Table 3 represents borrowings across unrelated languages. Note that the form in Bambara was not determined to be a borrowing.

However, the comparative process does not end at this point. The researcher must also provide input as to his/her awareness and knowledge of speakers' contemporary and historical situations. From a historical perspective, the root for COW as *nak is among the most stable reconstructed for Proto-Niger-Congo and is widely attested from central African languages to the far west (Pozdniakov 2013). From sources beyond language, we believe that cows have been raised by Fula herders in West Africa since at least 5000 BC (Blench 1993). The experienced researcher must decide whether the term has been profusely borrowed through the widespread influence of Fula speakers or represents an ancient state of the languages or a combination of both. In order to make these decisions, we must consult data from many angles.

These methods, together with the ability to automatically search for more distant borrowings among the over one million records included in the pan-African lexical database *RefLex* (Segerer and Flavier 2020), have also led to unanticipated contact. We have begun to detect contact with Dogon speakers through loan words in languages currently spoken in Nigeria (Hantgan in prep.). Language contact between Bangime and neighboring Bambara speakers is not surprising given their proximity, but commonalities between Yoruba spoken in Nigeria and the Dogon varieties are unexpected. While we know trade routes existed between Saharan and coastal peoples generally (Brooks 1993; Kea 2004), and ancient beads were exchanged between residents of Gao in Mali and the Igbo in Nigeria (Insoll and Shaw 1997), the specific impact of these migrant merchants on Dogon and Bangande speakers is as yet unknown. An example of this work in progress is represented in Table 4.

GROUP	LANGUAGE	SOURCE TRANSCRIPTION	TOKENIZED REPRESENTATION	COGIDS	BORID
Isolate	Bangime	gɛ́ŋgɛ̄	g ⁵ɛ ⁿg ³ɛ	206	57
W Mande	Bambara	nɛ̀gɛ̀	n ¹ɛ g ¹ɛ + f ⁵ĩ	2134 643	57
NE Dogon	Bankan Tey	ŋíɾɛ̌	ŋ ⁵i ~ ɾ̃ ⁵e + ~j	2135 2146	0
NE Dogon	Ben Tey	íírè-j̃	⁵i: ~ɾ̃ ⁵ɛ + ~j	2135 2146	0
SE Dogon	Jamsay	ííɾɛ́	⁵i: ~ ɾ̃ ⁵ ɛ	2135 2146	0
SE Dogon SE Dogon	Perge Tegu Tommo So	íínɛ́-m	⁵i: n ⁵ɛ + m	2136 2140	0
		ínɛ́	⁵i n ⁵ɛ	2136	0
Benue-Congo	Aku	i̱i̱re̱	–	–	–
Benue-Congo	Egba	i̱ri̱	–	–	–

Table 4: Cognate and Borrowing Comparison: Terms for IRON

Although Aku and Egba have not yet been incorporated into the cognate detection process, Hantgan and List (2018a) found the form shared between Bangime and Bambara to be a borrowing and the Dogon forms to be cognate, but split between groups. Bangime and Bambara forms are structurally and semantically similar, but sound correspondences are not reflected consistently across concepts. Speakers of minority African languages frequently borrow words from widely spoken languages (Batibo 2005). Widening the computer-assisted search beyond languages spoken in Central-eastern Mali will lead to evidence of long-distant contact.

Despite all of these advantages and fascinating findings, many researchers are wary of tools that are too far out of the expert's reach. A. McMahon and R. McMahon (2005) state, "If we are seeking an insight into trends, then computational methods can be of considerable help", but "...mathematical models and methods are not a substitute for careful and reasoned linguistic investigations..." (pg. 27, 36). The critical aspect of our workflow is that our methods are computationally *assisted* rather than strictly automatic or computer-based. All of our outputs are uploaded for hand-comparison in the web-based Edictor (List 2021) tool. Once we are satisfied with our comparative cognate judgments, we export the binary groupings that are automatically produced from Edictor to a Nexus file format for ease of visualization in programs such as SplitsTree (Huson and Bryant 2006). The traditional method of representing relationships using trees has been challenged since its instantiation. Jacques and List (2019) observe that not all language relationships can be accurately illustrated using tree models. Again, as A. McMahon and R. McMahon (2005, p. 17) state, "contact-induced changes cannot be shown in family trees". However, to uncover the past of languages with no known living relatives, such as Bangime, we must rely on language contact rather than shared inheritance (Campbell 2017). Alternatives to tree-like diagrams are under development (Kalyan and François 2019). Figure 2 illustrates an approach to visualizing the position of Bangime relative to other languages in our preliminary sample.

Even though the languages represented are spoken in a geographically proximate area, their diversity is apparent from the length of the lines that connect them; the longer the line, the further back in time the languages split. The diagram also shows that while Bangime has more affinities with the Dogon languages than with either the Mande grouping, Songhay, the length of the line connecting it to other languages in the sample shows it is the most distantly related. Dates are not shown, but Bangime is estimated to have diverged from the Dogon languages at least 4,000 years BP. The line between Fulfulde and the other languages is long because Fulfulde is the only language from the Atlantic group sampled thus far. The Dogon languages are quite closely connected, probably due to their speakers' position atop and within the Bandiagara Escarpment in comparative isolation from other speakers of the area.

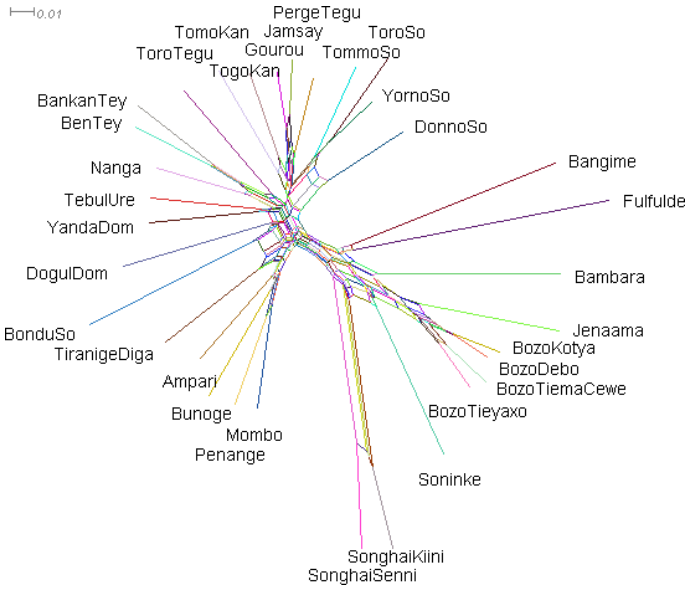

Figure 2: Equal-angle Neighbor-Net Diagram of Bangime and Neighboring Languages

Our final step has been to go beyond traditional methods in historical linguistics to utilize the most current evolutionary phylogenetic models to propose time depths and genealogical relations using a combination of linguistic data and evidence from other disciplines. Bayesian (Huelsenbeck and Ronquist 2001) phylogenetic models have been successfully tested and are now widely used across numerous types of linguistic data, including African Bantu languages (Janaki 2002). A phylogenetic character-based analysis of the Dogon languages incorporating both partial and full cognates is underway, but has yet to be performed as previous studies did not have access to partial cognate methodology. The preliminary tree in Figure 3 was created using Bayesian modeling software BEAST (R. et al. 2019). I inputted priors from archaeological (Mayor et al. 2005) and historical (Brooks 1986) dates. Dogon languages' internal relationships also correspond with oral histories gathered in the villages where the languages are spoken.

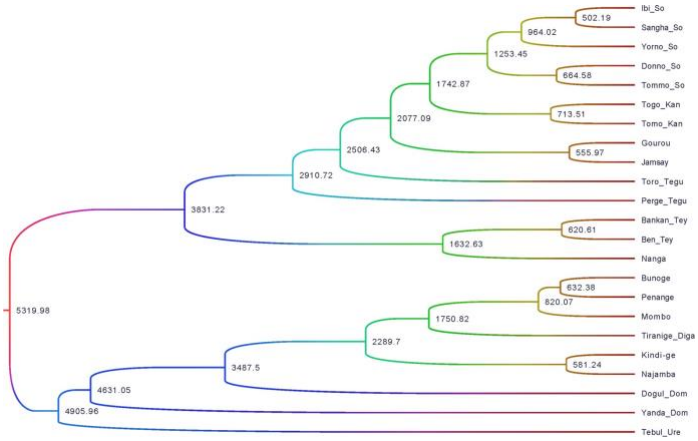

Figure 3: Phylogenetic tree of the Dogon language group

d.2. Proposed Methods

In the following subsections, I illustrate how the proposed methodology can be broken down into five steps, or *work packages*. ERC-**BANG** is primarily a data-driven linguistic project. Since I will be leading the proposed ERC-**BANG** researchers outlined in §b.3, this workflow reflects my skill set as a theoretical field linguist. Three additional members of the team, a geneticist, and two research engineers will support the project. Where applicable and especially innovative, I have added some possible avenues for their input.

Work package 1: Collection

The first *work package* is the *collection* of lexical wordlists. We will enlarge the comparative scope to include languages and speakers for which preliminary findings connote language contact with Bangime and/or the Dogon languages. Like the Nigerian example shown above, I have gathered preliminary data that point in intriguing directions but these additional languages and populations have yet to be integrated into our computational comparative models. Thus, the team can start by forming a dataset of 150 languages that have already been determined to have had an influence on Bangime and/or Dogon across a set of 500 core and culturally-significant concepts. In addition to 37,000 potential cognate sets, a database of loan words linked to archaeological evidence of dated material will be implemented. However, we will not compile data that will be too large for expert verification at each level of its inclusion.

Linguistic fieldwork has been conducted in West Africa for nearly a century; usable data are already abundantly available. The following databases are immediately available for linguistic data collection:

1. *Current Data Sources: Wordlists*
    - Bangime and Dogon Linguistics Project (Moran et al. 2016): Project website including upwards of 7,000 lexical entries for each of 21 Dogon languages, plus Bangime and Jenaama, as well as grammars, audio-visual materials, GPS coordinates, and bibliographic materials;
    - RefLex (Segerer and Flavier 2020): Pan-African lexical database housed at LLACAN featuring over one million records with built-in features for searching for borrowings, create atuomatic alignments, and assign cognates;
2. *Additional Data Sources: Corpora*
    - *Corpora Mandeica* for Mande languages at CNRS-LLACAN;
    - *CorpAfroAs* of 12 AfroAsiatic languages at CNRS-LLACAN;
    - *SpeechReporting* of 10 West African languages at CNRS-LLACAN;
    - *Pangloss* collection at LACITO;

- *Annis* Wolof corpus;
- *Kratylos* Hausa;
- *Crossroads* of 11 Senegalese languages;
- *ELAR*: World-wide collection of spoken language recordings with accompanying transcriptions and meta-data.

3. *Additional linguistic components*: nominal class markers, phoneme inventories, pronouns, and word order
    - *Grambank* (Skirgård and al. 2019): World-wide database of grammatical features (as a contributor, I have personal access)
    - *Phoible* (Moran and McCloy 2019): World-wide database of phoneme inventories

My connections with SOAS and the ELAR archive enable me enhanced access to 85 deposits from Africa spanning 82 different languages and a total of 21 countries, amounting to 2,374 transcribed files. I also have access to the *Crossroads* corpus which contains natural conversational data for languages of southwestern Senegal. I recently worked with researchers of West African languages to compile a carefully glossed corpus of discourse reporting for the on-going ERC-funded *Discourse reporting in African storytelling* project. Thus we have access to annotated data containing a range of features, across a variety of West African languages. Additional words which occur with high frequency in the spoken corpora will be selected for comparison.

These databases are specifically selected for their robustness and ease of data extraction. In close collaboration with LLACAN researcher Guillaume Segerer, the head developer of RefLex (Segerer and Flavier 2020), we will search for additional loan words into Bangime and potential cognates with the Dogon languages. Lexical items for each of the 150 targeted languages in the sample will be collected from core vocabulary in the *Swadesh* list (Swadesh 1952), but also Africa-specific word lists such as Gregersen (1976).

Beyond the lexicon, areal or typologically robust features that overlap language family boundaries will be necessary to account for in the comparative process because, as highlighted in the literature overview, shared evidence of attributes across languages does not necessarily mean that the languages are genealogically related. Rather, grammatical features such as fluctuating word order - attested in Bangime - can be a sign of a change in process probably due to language contact.

Languages found to have a significant influence on the Bangande or the Dogon will be targeted for genomic collection. Once the samples are fully genotyped, the Bangande genome can then be compared to individuals included in the Human Heredity and Health in Africa (H3Africa) consortium and the GenBank database. Before comparison, though, data *preparation* is necessary.

Work package 2: Preparation

The second work package will consist of the *preparation* of the lexemes collected in the first work package. Hitherto, much of this step has been implemented by hand, but the latest features of *LingPy* (List and Forkel 2021), a Python library that is designed to assist researchers with automating some of the tasks commonly done by hand in historical and comparative linguistics, has very recently been expanded so as to be able to handle the following sub-stages:

1. *Compilation* will adhere to the above outlined standards wherein each lexeme gloss is linked to universal concepts, each language is provided with a unique identifier. Original author and source metadata will be associated with permanent links to the raw data. The databases to which the compilation work package will be associated are listed as follows:
    - *Glottolog* (Hammarström et al. 2018): World-wide catalog of language affiliations, GPS coordinates, and bibliographical resources;
    - *Concepticon* (List, Rzymski, et al. 2020): Concept-labels and lists for basic vocabulary across cultures and geographical areas and visualization capabilities for semantic maps.
2. *Conversion* of phonetic transcriptions to IPA format will be undertaken using orthography profiles in the manner described above. Once the data are rendered readable, word forms will be tokenized and aligned in order to be ready for the next stage, *comparison*.

Work package 3: Comparison

The third work package is data *comparison*. For this step, there are two sub-steps: first the data are *automatically* treated by the two algorithms described above, then refined by language experts. Careful *correction* of computer-assisted correlations will be a central part of the *comparative* work-flow.

1. *Automatic* methods include comparison of phonetic alignments, cognate detection at the levels of the root and affixes, and borrowing detection.

2. *Correction* is performed by language experts. Algorithmic outputs will be uploaded for hand-comparison in the web-based Edictor (List 2017). We will add to the Bangime comparative cognate and loan word database (Hantgan and List 2018b) in order to collaboratively compare and edit our work in progress with other West African language experts. For this we will rely on two additional databases:
    - Edictor (List 2017): Website with built-in features for ease of cognate and alignment correction
    - CLICS (Rzymski et al. 2019): World-wide database of co-lexifications with visualization software

To complement this stage, a review of the relevant archaeological, ethnographic, and historical literature will be conducted to search for dates from external sources for the introduction of lexical items into West Africa. Additional interviews will also be collected from Bangande community members via telecommunication. All data will be stored in one comparative database designed by the project's proposed research technician.

Work package 4: Analysis

The fourth work package is the *analysis* of the data that have been *collected*, *prepared*, and *compared*. As with most studies of language traits, the majority of the phylogenetic and loan word analyses will be based on lexical distributions. However, we will also broaden these computer-assisted comparison with those from traditional methods incorporating phoneme inventories, pronouns, and morphological traits such as noun class affixes.

An essential aspect of the study will be to consider the genealogical grouping of the Dogon languages at a deeper level. For this, we will rely on our cognate judgments for distance matrices and producing Bayesian phylogenetic trees. An advantage to Bayesian phylogenetic methods is the ability to determine degrees of relatedness as well as time depths based on priors gathered from external sources. Specific to our purposes of inferring language and genetic contact through linguistic borrowings, combined with genome-wide genetic results and Y-chromosome haplogroups' distribution, we will use *approximate Bayesian computations* (ABC) such as implemented with *PopLingSim*, a simulation tool has been developed by colleagues at CNRS (Thouzeau et al. 2017). This method is perfectly suited for our needs of complementing - as opposed to strictly comparing - linguistics with genetics findings towards the goal of delving into populations' deep histories.

We will also utilize innovative corpus-based methodologies such as that proposed by Gamallo et al. (2020). They successfully implemented clustering methods on corpus data to measure the distance between isolated European languages and the Indo-European language family. Their results further substantiate those proposed for language isolate Basque by more traditionally based historical comparative methods.

Determining which factors contribute to a language's change and growth is a challenge, but not an insurmountable one. We have a plethora of information concerning Bangime and Dogon oral histories, as well as archaeological and historical corroboration of dates of introduction of culturally specific items or tools that we expect will have unique terminology. We will further source archaeological evidence in order to estimate dates of introduction for domestic animals and crops, thus providing us with an integrated picture of not only from which language, but also when and potentially where, a borrowing was introduced.

An approach that has we have yet to incorporate into our methods is phylogenetic relationships based on cognates as well as borrowings. As with horizontal gene transfer, we can use phylogenetic trees to determine lateral, as opposed to vertical, lexical trait data. Nearly all phylogenetic linguistic studies purposely eliminate known borrowings from the comparative lexical dataset. The problem with this is that while recent borrowings are easily detectable in the data, those which are phonologically integrated into the target language, that is, those which were acquired at a deeper time depth, are nearly impossible to recognize. Thus, we will follow the procedure put forth by Kelly and Nicholls (2017), the first study to include loan words into a likelihood-based approach of phylogenetic tree reconstruction. In this manner, the Stochastic Dollo model can determine disruptions in the phylogenetic signal that correspond to borrowing events. These events can then be compared with those in which gene transfer has taken place.

Somewhat surprisingly, means by which researchers can infer population histories from accompanying linguistic with genetic data are not yet numerous. Those concerning language and population isolates are that much less available. While most geneticists are versed in linguistic methodologies, the reverse generalization does not hold. Thus, the team will work closely with the project research engineers to create innovative ways to introduce not only lexical comparisons to gene distributions, but also our additional phonemic and morpho-syntactic data, into the analysis of the languages' lexemes to provide a full-spectrum image of the languages relationships to each other and to their speakers.

Work package 5: Dissemination

At each stage of the project, we will share results and *disseminate* our findings. Field-based research will include gestures to, and involvement from, the communities whose languages are studied. We will hold yearly workshops where we invite experts to examine our work for feedback and changes. We will consult with an ethics advisory board to plan for a culturally sensitivity manner to transmit our findings. We will seek approval for all of our Mali-based methods and data collection from the *Centre National de la Recherche Scientifique et Technologique* (CNRST) in the capitol city, Bamako. Bangande representatives recently confided in me that remain enthusiastic about the potential to learn more about their undiscovered origins.

d.3. Proposed Timeline

The following timeline outlines in detail the roles of each of the participants in the ERC-**BANG** team. The team will consist of five members in addition to the PI.

Year 1 of the project will be devoted to *collecting* available data. A postdoctoral researcher who has recently acquired his or her PhD with an interest in historical linguistics will be the first recruitment to begin expanding the current comparative lexical database.

Year 2 will consist of *comparing* the lexical data thus far *compiled* and *prepared*. We will also begin to compare the evidence of language contact between Bangime and Dogon speakers with those of West African populations to see if there are correlates with shared DNA. We will hold an introductory international conference at the beginning of Year 2 to familiarize the team with our data collection principles and methodologies. A second post-doctoral researcher with knowledge of, and experience using, computer-assisted comparative linguistics will be recruited for two and a half years to apply and further develop our computational approaches to automatic loan word detection and cognate recognition.

Year 3 will commence with data *analysis*. The project will recruit a research engineer - a statistician - who will support the researchers' needs by developing and employing tried and tested, as well as emerging, tools and methods as described above. As a team, we will triangulate our findings with cross-disciplinary sources to prepare a hypothetical migration map for the Bangande and relevant populations based on dates and areas of language contact between Bangime and speakers of other West African languages. These results will lead to modeling the position of Bangime relative to the Dogon and other West African peoples and languages. A PhD student will be recruited to commence reconstruction of the Dogon language group. The research thus far undertaken by the first two post-doctoral researchers will be integral to the phylogenetic study of the Dogon language group. An assistant research engineer will build a website and together we will organize a conference to share our methods and findings.

Year 4 will be targeted for sampling from populations among those who speak a language that has been found to have impacted Bangime or the Dogon languages, through contact or cognates, but for whom we lack genetic data. Additional data will be gathered from speakers in cases where evidence of language contact is likely through lexical evidence but for which grammatical descriptions are unavailable. These new data will be *prepared* and *compared*. We will hold a workshop to compare our findings with experts from the languages in our sample. Based on our current data and the findings of our comparative studies, we will make decisions as to whether to acquire thorough amounts of data from a small sample of additional languages and populations or to gather select data from a larger set of languages and peoples.

Year 5 will conclude the project by a workshop to finalize our findings in a monograph or book describing the history of the Bangime language and Bangande people in the context of West Africa. We will communicate the outcome of the project to the Bangande people, if possible in person, or if not, via a community representative and recorded video messages.

References

Ayub, Q., M. Mezzavilla, L. Pagani, M. Haber, A. Mohyuddin, S. Khaliq, S. Q. Mehdi, and C. Tyler-Smith (2015). "The Kalash genetic isolate: Ancient divergence, drift, and selection". In: *The American Journal of Human Genetics* 96, pp. 1–9. DOT: 10.1016/j.ajhg.2015.03.012.

Babiker, H., A. Hantgan, J.-M. List, J. Heath, and R. Gray (2018). *Insights into the population history of the "Hidden Ones": From oral history to genome-wide analysis*. Paper presented at the Society for Molecular Biology and Evolution conference, Yokohama, Japan. Yokohama, Japan.

Babiker, H., J. Heath, F. Reed, S. Schiffels, and R. D. Gray (2020). *Striking genetic diversity among populations of West Africa uncovers the mystery of a language isolate*. DOT: http://dx.doi.org/10.2139/ssrn.3631471.

Batibo, H. M. (2005). *Language decline and death in Africa: Causes, consequences and challenges*. Clevedon, UK: Multilingual Matters.

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. (2000). "Nilo-Saharan". In: *African Languages, An Introduction*. Ed. by B. Heine and D. Nurse. Cambridge: Cambridge University Press, pp. 43–73.

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.

Bickel, B. and J. Nichols (2020). "Linguistic typology and hunter-gatherer languages". In: *The Language of Hunter-Gatherers*. Ed. by T. Güldemann, P. McConvell, and R. A. Rhodes. Cambridge University Press, pp. 67–75. DOT: 10.1017/9781139026208.004.

Blench, R. (1993). "Ethnographic and linguistic evidence for the prehistory of African ruminant livestock, horses and ponies". In: *The Archaeology of Africa: Food, Metals and Towns*. Ed. by T. Shaw, P. Sinclair, B. Andah, and A. Okpoko. London: Routledge, pp. 71–103.

– (2007). "Bangime, a language of unknown affiliation in northern Mali". In: *Mother Tongue* XII, pp. 147–178.

– (2015). "Was there a now-vanished branch of Nilo-Saharan on the Dogon Plateau? Evidence from substrate vocabulary in Bangime and Dogon". In: *Mother Tongue, Journal of the Association for the Study of Language in Prehistory* 20. In Memory of Harold Crane Fleming (1926-2015), pp. 73–89.

Bouju, J. (1995). "Qu'est-ce que "l'ethnie" dogon?" In: *Cahiers des Sciences Humaines* 31.2.

Brooks, G. E. (1986). "A provisional historical schema based on climate periods for Western Africa based on seven climate periods (ca. 9000 B.C. to the 19th Century)". In: *Cahiers d'études Africaines, Milieux, histoire, historiographie* 26.101-102, pp. 43–62.

– (1993). *Landlords and strangers: ecology, society, and trade in Western Africa, 1000–1630. (African States and Societies in History.)* Boulder, Colorado: Westview.

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

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

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain (1988). "Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data". In: *Proceedings of the National Academy of Sciences* 85.16, pp. 6002–6006. TSSN: 0027-8424. DOT: 10.1073/pnas.85.16.6002. eprint: https://www.pnas.org/content/85/16/6002.full.pdf. URL: https://www.pnas.org/content/85/16/6002.

Choudhury, A. et al. (2020). "High-depth African genomes inform human migration and health". In: *Nature* 586.7831, pp. 741–748. TSSN: 1476-4687. DOT: 10.1038/s41586-020-2859-7.

Clements, G. and A. Rialland (2008). "Africa as a phonological area". In: *A linguistic geography of Africa*. Ed. by B. Heine and D. Nurse. Cambridge, UK: Cambridge University Press, pp. 36–85.

Creanza, N., M. Ruhlen, T. J. Pemberton, N. A. Rosenberg, M. W. Feldman, and S. Ramachandran (2015). "A comparison of worldwide phonemic and genetic variation in human populations". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1265–1272. TSSN: 0027-8424. DOT: 10.1073/pnas.1424033112. eprint: https://www.pnas.org/content/112/5/1265.full.pdf. URL: https://www.pnas.org/content/112/5/1265.

Dimmendaal, G. J. (2008). "Language ecology and linguistic diversity on the African continent". In: *Language and Linguistics Compass* 2.5, pp. 840–858. DOT: 0.1111/j.1749-818X.2008.00085.x.

– (2014). "Nilo-Saharan". In: *The Oxford handbook of derivational morphology*. Ed. by R. Lieber and P. Štekauer. Oxford: Oxford University Press, pp. 591–608.

Dorp, L. van et al. (2015). “Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference”. In: *PLOS Genetics*. DOT: :10.1371/journal.pgen.1005397.
Eberhard, D. M., G. F. Simons, and C. D. Fennig, eds. (2021). *Ethnologue: Languages of the World*. Twenty-fourth. Dallas, Texas: SIL International. URL: http://www.ethnologue.com.
Fan, S. et al. (2019). “African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations”. In: *Genome Biology* 20.82, pp. 2–14. DOT: 10.1186/s13059-019-1679-2.
Filippo C., de, B. K., S. M., and P. B. (2012). “Bringing together linguistic and genetic evidence to test the Bantu expansion”. In: *Proc Biol Sci*. 279.1741, pp. 3256–3263. DOT: 10.1098/rspb.2012.0318.
Forkel, R. et al. (2018). “Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics”. In: *Scientific Data* 5.1, p. 180205. TSSN: 2052-4463. DOT: 10.1038/sdata.2018.205. URL: https://doi.org/10.1038/sdata.2018.205.
Gamallo, P., J. R. Pichel, and I. Alegria (2020). “Measuring language distance of isolated european languages”. In: *Information* 11.4. TSSN: 2078-2489. DOT: 10.3390/info11040181. URL: https://www.mdpi.com/2078-2489/11/4/181.
Greenberg, J. H. (1948). “The classification of African languages”. In: *American Anthropologist* 50.1, pp. 24–30. DOT: https://doi.org/10.1525/aa.1948.50.1.02a00050.
Greenhill, S. J., Q. D. Atkinson, A. Meade, and R. D. Gray (2010). “The shape and tempo of language evolution”. In: *Proc. R. Soc. B* 277.1693, pp. 2443–50. DOT: 10.1098/rspb.2010.0051.
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.
Güldemann, T. (2008). “The ”Macro-Sudan belt”: Towards identifying a linguistic area in northern Sub-Saharan Africa”. In: *A linguistic geography of Africa*. Ed. by B. Heine and D. Nurse. Cambridge, UK: Cambridge University Press, pp. 151–185.
– (2018). “Historical linguistics and genealogical language classification in Africa”. In: *The Languages and Linguistics of Africa*. Ed. by T. Güldemann. Berlin, Boston: De Gruyter, pp. 58–444.
Hammarström, H., S. Bank, R. Forkel, and M. Haspelmath (2018). *Glottolog 3.2*. http://glottolog.org/ Max Planck Institute for the Science of Human History. Jena. URL: http://glottolog.org/%20accessed%202018-04-13.
Hantgan, A. (in prep.). *Bangime and Dogon language contact: New perspectives*. Unpublished manuscript.
Hantgan, A. and S. Davis (2012). “Bondu vowel harmony: A descriptive analysis with theoretical implications”. In: *Studies in African Linguistics* 41.2, pp. 2–26.
Hantgan, A. and J.-M. List (2018a). “Bangime: Secret language, language isolate, or language island?” In: *to appear in a Language Islands in Africa thematic issue of the Journal of Language Contact*. URL: https://hal.archives-ouvertes.fr/hal-01867003/file/bangime-secret-language.pdf.
– (2018b). *Dogon and Friends*. Edictor Database: http://tsv.lingpy.org/?remote_dbase=dogon.sqlite3&file=dogon&css=menu:show|textfields:hide|database:hide|&preview=10&basics=DOCULECT|CONCEPT|IPA|COGID|ALIGNMENT&pinyin=CHINESE&sampa=IPA|TOKENS&highlight=TOKENS|ALIGNMENT.
Hantgan, A., J.-M. List, and H. Babiker (2020). *First steps towards the detection of contact layers in Bangime: A multi-disciplinary, computer-assisted approach*. under review. DOT: http://dx.doi.org/10.17613/8bbr-6k95.
Hepburn-Gray, R. (2020). “Niger-Congo Noun Classes: Reconstruction, istorical implications, and morphosyntactic theory”. PhD. State University of New York at Buffalo.
Hochstetler, J., J. Lee Durieux, and E. Durieux-Boon (2004). *Sociolinguistic Survey of the Dogon Language Area*. SIL International.
Huelsenbeck, J. P. and F. Ronquist (Aug. 2001). “MRBAYES: Bayesian inference of phylogenetic trees”. In: *Bioinformatics* 17.8, pp. 754–755. TSSN: 1367-4803. DOT: 10.1093/bioinformatics/17.8.754. URL: https://doi.org/10.1093/bioinformatics/17.8.754.
Huson, D. H. and D. Bryant (2006). “Application of phylogenetic networks in evolutionary studies”. In: *Molecular Biology and Evolution* 23.2, pp. 254–267.
Idiatov, D. and M. V. de Velde (2021). “The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa”. In: *Language* 97, pp. 72–107. DOT: 10.1353/lan.2021.0002.

Insoll, T. and T. Shaw (1997). “Gao and Igbo-Ukwu: Beads, interregional trade, and beyond”. In: *African Archaeological Review* 14.1, p. 9. TSSN: 1572-9842. DOT: 10.1007/BF02968364.
Jacques, G. and J.-M. List (2019). “Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them)”. In: *Journal of Historical Linguistics* 9.1, pp. 128–166. DOT: 10.1075/jhl.17008.mat.
Janaki, H. C. (2002). “Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis”. In: *Proc. R. Soc. Lond. B* 269.1493, pp. 793–799. DOT: 10.1098/rspb.2002.1955.
Kalyan, S. and A. François (2019). “Freeing the comparative method from the tree model: A framework for Historical Glottometry”. In: *Let’s talk about trees: Tackling problems in representing phylogenic relationships among languages*. Ed. by R. Kikusawa and L. Reid. Osaka: National Museum of Ethnology, To appear.
Kea, R. (2004). “Expansions and Contractions: World-Historical Change And The Western Sudan World-System (1200/1000 B.C. - 1200/1250 A.D.)” In: *Journal of world-systems research* X.3, pp. 723–816.
Kelly, L. J. and G. K. Nicholls (2017). *Lateral transfer in Stochastic Dollo models*. arXiv: 1601 . 07931 [stat.AP].
Lipson, M. et al. (2020). “Ancient West African foragers in the context of African population history”. In: *Nature* 577.7792, pp. 665–670. TSSN: 1476-4687. DOT: 10.1038/s41586-020-1929-1.
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.
– (2017). “A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, pp. 9–12. URL: http://edictor.digling.org.
– (2019). “Automated methods for the investigation of language contact, with a focus on lexical borrowing”. In: *Language and Linguistics Compass* 13.e12355, pp. 1–16. DOT: 10.1111/lnc3.12355.
– (2021). *A web-based interactive tool for creating and editing etymological datasets*. Version 2.0.0. Leipzig. URL: https://digling.org/edictor/.
List, J.-M. and R. Forkel (2021). *LingPy. A Python library for historical linguistics*. With contributions by Greenhill, Simon, Tresoldi, Tiago, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel. Leizpig. URL: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.
List, J.-M., P. Lopez, and E. Bapteste (2016). “Using sequence similarity networks to identify partial cognates in multilingual wordlists”. In: *Association of Computational Linguistics*. Berlin: Association of Computational Linguistics, pp. 599–605. URL: http://anthology.aclweb.org/P16-2097.
List, J.-M., C. Rzymski, et al. (2020). *Concepticon: http://concepticon.clld.org/*. Jena: Max Planck Institute for the Science of Human History. URL: http://concepticon.clld.org/.
MacEachern, S. (2000). “Genes, tribes, and African history”. In: *Current Anthropology* 41.3, pp. 77–98. DOT: 0011-3204/2000/4103-0003.
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, F. L. Guennec-Coppens, and E. Coulibaly. 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.
McMahon, A. and R. McMahon (2005). *Language Classification by Numbers*. Oxford, UK: Oxford University Press.
Moran, S. and M. Cysouw (2017). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. DOI: 10.5281/zenodo.290662. Zürich: Zenodo.
Moran, S., R. Forkel, and J. Heath, eds. (2016). *Dogon and Bangime Linguistics http://dogonlanguages.org/*. Jena: Max Planck Institute for the Science of Human History. URL: http://dogonlanguages.org/.
Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0: https://phoible.org/*. Jena: Max Planck Institute for the Science of Human History. URL: https://phoible.org/.

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.
Pagel, M., Q. D. Atkinson, and A. Meade (2007). “Frequency of word-use predicts rates of lexical evolution throughout Indo-European history”. In: *Nature* 449, pp. 717–720. DOT: 10.1038/nature06176.
Pozdniakov, K. (2013). “НИГЕРО-КОНГОЛЕЗСКИЕ ЯЗЫКИ”. In: *Russian Great Encyclopaedia* 18.
R., B., V. T.G., B.-S. J., D. S., F. M., G. A., and et al (2019). “BEAST 2: 2.5: An advanced software platform for Bayesian evolutionary analysis”. In: *PLoS Computational Biology* 15.4.
Retshabile, G. et al. (2018). “Whole-Exome sequencing reveals uncaptured variation and distinct ancestry in the southern African population of Botswana”. In: *The American Journal of Human Genetics* 102.5. Publisher: Elsevier, pp. 731–743. TSSN: 0002-9297. DOT: 10.1016/j.ajhg.2018.03.010. URL: https://doi.org/ 10.1016/j.ajhg.2018.03.010 (visited on 04/18/2021).
Ringe, D. A. (1995). “‘Nostratic’ and the factor of chance”. In: *Diachronica* 12.1, pp. 55–74. DOT: 10.1075dia.12.1.04rin.
Rzymski, C. et al. (2019). “The Database of Cross-Linguistic Colexifications, reproducible analysis of crosslinguistic polysemies”. In: *Humanities Commons*. DOI: 10.17613/5awv-6w15, 5awv–6w15.
Sands, B. (2019). “Tracing language contact in Africa’s past”. In: *To appear in Cambridge Handbook of Language Contact*. Ed. by S. Mufwene. Cambridge: Cambridge University Press.
Schlebusch, C. M. and M. Jakobsson (2018). “Tales of human migration, admixture, and selection in Africa”. In: *Annual Review of Genomics and Human Genetics* 19.1. PMID: 29727585, pp. 405–428. DOT: 10.1146/ annurev-genom-083117-021759.
Segerer, G. and S. Flavier (2020). *RefLex: Reference Lexicon of Africa*. Version 2.0. Paris, Lyon. URL: http://reflex.cnrs.fr/.
Skirgård, H. and et al. (2019). *Grambank*. Max Planck Institute for the Science of Human History. in review. Jena.
Skoglund, P. et al. (2017). “Reconstructing prehistoric African population structure”. In: *Cell* 171.1, 59–71.e21.
Souag, L. (2012). “The subclassification of Songhay and its historical implications”. In: *Journal of African Languages and Linguistics* 33.2. DOT: 10.1515/jall-2012-0008.
Swadesh, M. (1952). “Lexico-statistic dating of prehistoric ethnic contacts”. In: *Proceedings of the American Philosophical Society* 96.4, pp. 452–463.
Tamari, T. (1991). “The development of caste systems in West Africa”. In: *The Journal of African History* 32.2, pp. 221–250.
Thouzeau, V., P. Mennecier, P. Verdu, and F. Austerlitz (2017). “Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations”. In: *Proc Biol Sci*. 284.1861, pp. 1–9. DOT: 10.1098/ rspb.2017.0706.
Tishkoff, S. A., M. K. Gonder, et al. (2007). “History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation”. In: *Molecular Biology and Evolution* 24.10.
Tishkoff, S. A., F. A. Reed, et al. (2009). “The genetic structure and history of Africans and African Americans”. In: *Science* 324.5930, pp. 1035–1044. TSSN: 0036-8075. DOT: 10.1126/science.1172257.
Veen, L. V. der, L. Quintana-Murci, and D. Comas (2009). “Linguistic, cultural and genetic perspectives on human diversity in west-central Africa”. In: *Becoming Eloquent: Advances in the emergence of language, human cognition, and modern cultures*. Ed. by F. d’Errico and J.-M. Hombert. Amsterdam, The Netherlands: John Benjamins Publishing Company, pp. 93–122.
Wilkinson, M. D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1, p. 160018. TSSN: 2052-4463. DOT: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.
Wu, M.-S., Y. Lai, and J.-M. List (2018). *The ancestry of Sino-Tibetan populations and languages*. Proceedings of the 51st International Conference on Sino-Tibetan Languages and Linguistics. DOI: 10.5281/zenodo.1306623.