BRILL

LANGUAGE DYNAMICS AND CHANGE 12 (2022) 155–223

brill.com/ldc

Sociogeographic correlates of typological variation in northwestern Bantu gender systems

Annemarie Verkerk | ORCID: 0000-0002-3351-8362
Department of Language Science and Technology, Saarland University, Saarbrücken, Germany;
Max Planck Institute for the Science of Human History, Jena, Germany
annemarie.verkerk@uni-saarland.de

Francesca Di Garbo | ORCID: 0000-0002-2499-8800
Department of Languages, University of Helsinki, Helsinki, Finland
francesca.digarbo@helsinki.fi

Abstract

This paper investigates the sociolinguistic factors that impact the typology and evolution of grammatical gender systems in northwestern Bantu, the most diverse area of the Bantu-speaking world. We base our analyses on a typological classification of 179 northwestern Bantu languages, focusing on various instances of semantic agreement and their role in the erosion of gender marking. In addition, we conduct in-depth analyses of the sociolinguistics and population history of the 17 languages of the sample with the most eroded gender systems. The sociohistorical factors identified to explain these highly eroded systems are then translated into a set of explanatory variables, which we use to conduct extensive quantitative analyses on the 179 language sample. These variables are population size, longitude, latitude, relationship with the Central African rainforest, and border with Ubangi/Central Sudanic languages. All these measures are relevant, with population size and bordering with Ubangi/Central Sudanic being the most robust factors in accounting for the distribution of gender restructuring. We conclude that fine-tuned variable design tailored to language and area-specific ecologies is crucial to the advancement of quantitative sociolinguistic typology.

Keywords

grammatical gender – northwestern Bantu – animacy distinctions – complexity – language contact – phylogenetics – phylogenetic comparative methods – generalized linear mixed effect models

1 Introduction

Grammatical gender is a relatively common feature in the languages of the world. It is found on all five continents and is characterized by substantial crosslinguistic variation at both the semantic and the morphosyntactic levels. In languages with grammatical gender, nouns are distributed into classes, which can be semantically, morphologically, and/or phonologically motivated, but also opaque (Corbett 2013; Dahl 2000, 2019; Wälchli & Di Garbo 2019). Gender assignment, that is, the mechanisms whereby nouns are allocated to gender classes, can be sex-based (as in German, which distinguishes between masculine, feminine, and neuter nouns) or non-sex-based (as in Swahili, which distinguishes between gender 1/2, which mostly contains human nouns; gender 3/4, which contains plants and body parts; and so on). Structural evidence for patterns of gender assignment comes from inflectional markers that are displayed beyond nouns, that is, typically, on adnominal modifiers, predicative constructions, and different types of pronouns. The inflections on these words point to the gender of the noun they are associated with. These patterns of displaced marking are referred to as instances of gender *agreement*; the word classes that carry agreement are labeled *agreement targets*, while the nouns that trigger agreement are labeled *controllers of agreement* (Corbett 1991).

The evolution of gender systems is said to be particularly sensitive to aspects of the social history of speech communities. Three remarks of this kind that are often repeated are:

1. Gender systems are highly grammaticalized and presuppose nontrivial historical developments (Corbett 1991; Dahl 2004).
2. Gender systems are very stable at the language-family level (Nichols 1992, 2003).
3. Gender systems are hard to master in non-native language acquisition (McWhorter 2007) and tend to break down under the pressure of intense language contact (Dahl 2019; Trudgill 1999).

In this paper we take (1) and (2) as our points of departure and put (3) to the test. We do this by conducting a sociolinguistic typological study of grammatical gender in the Bantu language family, focusing on 179 languages spoken in

the northwestern part of the Bantu area (our definition of northwestern Bantu is given in Section 2.3). Bantu languages are well known for their remarkable and generally highly stable gender systems (also known as noun class systems; Maho 1999; Katamba 2003), which are often portrayed as a conservative block of pervasive patterns of gender agreement and gender marking on nouns. However, significantly restructured and eroded systems of gender marking are also attested, most prominently in the northwestern zones of the Bantu-speaking world, where Bantu languages are and have been historically surrounded by non-Bantu groupings, such as Ubangi and Central Sudanic.

Bantu languages are distributed over a huge geographical area and are spoken by peoples whose sociohistorical profiles vary greatly, both in terms of number of speakers and present and past contact scenarios. Models that take population contact dynamics into account in order to explain the distribution of restructured and eroded systems of gender marking in the Bantu-speaking world have occasionally been proposed in the literature (Maho 1999; Güldemann 2018), but never really tested over a large language sample. Taken together, these factors make the Bantu language family an ideal test case to study how patterns of gender marking evolve in response to sociolinguistic and environmental factors. This is what we aim to do in this paper.

We investigate whether the distribution of restructured and eroded systems of gender marking can be explained through sociohistorical and environmental factors pertaining to the social history of the speech communities under study. We do this via a research design where state-of-the-art statistical modeling is combined with in-depth qualitative analyses of historical and ethnographic resources, which are used as input for statistical variable design. In developing this approach, we place special emphasis on the measures and methods that can be used to investigate sociolinguistic typological questions. This allows us to detect an effect of sociohistorical and environmental factors on the distribution of types of gender systems, which is in alignment with our hypotheses. However, while this is the first large-scale comparative study to demonstrate a link between the shape and fabric of gender systems and sociogeographic factors related to language contact, the nature of this study is exploratory rather than confirmatory. The conclusions we draw from our results are deeply entrenched in the unique ecologies of the languages we study, and additional factors, other than those considered here, may play a role in the distribution of the patterns that we observe.

The paper is organized as follows. Section 2 discusses relevant background literature on the typology and evolution of gender systems, and provides an overview of the Bantu language family and the typological profile of systems of gender marking in this family. Section 3 digs into variable design. Section 4

models the sociohistorical correlates of restructuring in northwestern Bantu gender systems using generalized linear mixed models. Section 5 discusses the results and provides some prospects for future research, while concluding remarks are offered in Section 6. Three appendices accompany the paper: Appendix A reports on the data collection procedure, Appendix B is a list of the sampled languages, and Appendix C is a table with further details on the sociolinguistic typology of seventeen languages with radically restructured gender systems. Supplementary Information 1, 2, and 3 contain an additional figure, further details and analyses related to the quantitative analyses reported on in Section 4, and all code and data.

2 Background

2.1 *The evolution of gender systems and its sociohistorical and environmental correlates*

The fact that, in spite of their overwhelming stability, gender systems are likely to change, and sometimes even break down under the influence of language contact is well known in the literature. Trudgill (1999) and McWhorter (2001, 2007) argue that while gender systems are highly stable in situations of native-like language transmission, they can undergo erosion and eventually fade away when adult learners start playing a significant role in the language ecology of a speech community. In a recent crosslinguistic survey, Di Garbo (2020) investigates the evolution of morphosyntactic complexity in the domain of gender marking and its sociohistorical correlates. The study finds that the reduction, loss, and emergence of patterns of gender marking tend to be distributed in areas at the crossroads between gendered and genderless languages and/or families, and that asymmetries in population structure and/or prestige dynamics between populations play an important role in the direction of change. These results support the claim by Nichols (1992, 2003) that both presence and absence of gender are reinforced by geography: languages with gender are expected to neighbor each other, and languages where complete gender loss occurs also neighbor each other or are surrounded by languages without gender.

Even though solidly established, these generalizations have not received statistical validation through the analysis of large crosslinguistic data sets. On the contrary, attempts at investigating the relationship between gender systems and sociohistorical or environmental factors from a quantitative point of view have so far produced only negative evidence. Grammatical gender is not included in the list of morphological features that Lupyan & Dale (2010)

demonstrate to be sensitive to the effect of social structures. Dahl (2019) reports that this omission is not due to lack of testing, but to the fact that, when tested for the effect of population data, the gender features included in the World Atlas of Language Structures (WALS) yield inconsistent results. He further replicates the results of Lupyan & Dale (2010) by correlating each of the gender features in WALS with the logarithm of the number of speakers for each language in the sample, which is also based on WALS. Similar findings are reported by Sinnemäki & Di Garbo (2018), who test whether population dynamics, which they model through the combination of the log number of native speakers and the proportion of second language speakers, have any effect on the distribution of number of genders based on a sample of about 300 languages. The study finds no significant effect of population size and population contact on the inventory size of gender distinctions. Finally, when testing the generalization that creole languages should exhibit radical simplification in several domains of morphosyntax as compared to non-creoles, Blasi et al. (2017) find no indication of such an adaptive response in the domain of gender marking on adnominal modifiers nor on personal pronouns. In the languages of their sample, presence/absence of gender marking in these morphosyntactic domains is a sheer reflection of ancestry, that is, of the type of system attested in the lexifier and/or the substratum language.

In this paper, we argue that the discrepancy between qualitative crosslinguistic research and large-scale quantitative analyses that investigated the impact of language contact on the evolution of gender systems is at least in part due to inadequate variable design in studies of the latter type. All studies cited above except for Blasi et al. (2017) use the gender features in WALS ("Number of gender distinctions," "Systems of gender assignment," "Sex-based vs. Non-sex-based gender") as measures of complexity. The three features have proven to be highly relevant to classify the diversity of grammatical gender systems. However, we suspect that they are not a particularly good fit to answer questions concerning the sociolinguistic typology of gender systems, because they are not directly concerned with gender inflections, that is, with what has been argued to be most sensitive to the pressure of social and environmental factors (see also Cysouw [2005] on the prevalence of simplistic categorical measures in typology).

In order to study the sociolinguistic typology of gender systems, and to seek statistical validation of the generalizations mentioned at the beginning of this section, we thus suggest that a shift in focus, from the WALS features to the morphosyntactic encoding of grammatical gender, is needed. While this is what Blasi et al. (2017) partly do in their study by looking at patterns of gender marking on adnominal modifiers and personal pronouns, we contend that a more

fine-grained research design is needed. This should be based on comprehensive accounts of *domains* (adnominal modifiers, predicative expressions, pronouns), *targets* (e.g., attributive and predicative adjectives, verbs, personal pronouns), and *patterns of gender agreement* (syntactic, semantic), and on their relation with *gender assignment*. Under syntactic agreement, gender marking is consistent with the gender a noun is lexically assigned to. Under semantic agreement, gender marking is based on the referential semantics of a noun, which may ultimately clash with its lexical gender assignment. For instance, in German, the noun *Mädchen* 'young woman' is lexically neuter, but semantically denotes a female human being. Agreement patterns with this noun alternate between neuter (syntactic) and feminine (semantic) in such a way that agreement on adnominal modifiers is often in the neuter, while agreement on personal pronouns is often in the feminine (Corbett 1991). It is a known fact that in those cases in which semantic agreement becomes generalized to all agreement targets and to a large class of nouns, this may also lead to a substantial restructuring of gender assignment rules, and of the overall makeup of a gender system (Corbett 1991: Chap. 8).

Here, we consider the expansion of semantic agreement as an instance of increased transparency (in line with Vihman et al. [2018]), and attempt to identify the linguistic and nonlinguistic scenarios under which these semantically transparent patterns of gender marking rise and spread. This approach tallies with a recent suggestion by Kempe & Brooks (2018), who argue that in order to investigate the impact of population dynamics (and second language learning in particular) on the evolution of morphological complexity, studying the emergence of transparency and compositionality is more revealing than, say, counting the sheer number of morphological distinctions that are made within a given morphosyntactic domain.

In addition, we argue that the issue of variable design also applies to the sociohistorical and environmental variables that are fed to the modeling of linguistic adaptation. To date, demographic variables related to population size and population structure have been at the fore of research in this field (see Lupyan & Dale 2010; Bentz & Winter 2013; Sinnemäki & Di Garbo 2018; among others). While we do not a priori object to the validity of these variables in the task of detecting nonlinguistic correlates of linguistic distributions, we advocate that exploratory studies of this kind should as much as possible aim to enhance the ecological validity of the sociolinguistic factors that are featured in the models. This requires an understanding of the language ecology and history of human populations that goes beyond census counts, as well as translating documented characteristics of the nonlinguistic environments under study into well-suited statistical variables. While such a fine-tuned approach to non-

linguistic variable design may be difficult to attain in worldwide typological investigations, it is certainly more feasible in genealogically and areally focused studies, like the one we present here.

2.2 *Bantu gender systems: General characteristics*

(Narrow) Bantu is a group of over 500 languages belonging to the Bantoid division of the Benue-Congo subfamily, a subgroup of the Atlantic-Congo family (Hammarström et al. 2018; Nurse & Philippson 2003). Bantu languages are spoken in sub-Saharan Africa; their northern border can be approximated by a line drawn from Nigeria in the west across the Central African Republic and the Democratic Republic of the Congo (DRC) to southern Somalia in the east. Except for the Khoe-Kwadi, Tuu, and Kx'a families and other languages previously described as "Khoisan" in the south, most of the language communities between this northern border and the Cape of Good Hope are Bantu (Nurse & Philippson 2003: 1), spoken alongside Cushitic and Nilotic languages. Traditionally, Bantu languages are classified into fifteen zones, which are labeled A through S and are referred to as Guthrie zones after Malcom Guthrie, who introduced them (Guthrie 1967–1971; Maho 2003, 2009).[1] This study focuses on northwestern Bantu languages (henceforth NWB), which we define as languages within Guthrie zones A, B, C, D, and H, and approximately covering southern Cameroon, southern Central African Republic, Equatorial Guinea, Gabon, the Republic of Congo, northern DRC, and northwestern Angola (see Grollemund et al. [2018] and references therein for an overview of multiple definitions of NWB). NWB is unanimously held to be the typologically most diverse area of the Bantu-speaking world, as well as one in which linguistic phenomena that may substantially diverge from those attested in the rest of the family are also encountered. It is in the northern borderlands of the northwestern Bantu area, for instance, that a variety of highly eroded systems of gender marking can be found (Maho 1999; Di Garbo & Verkerk 2022).

1 While the original classification proposed by Guthrie (1967–1971) consisted of fifteen zones, the current system includes sixteen zones, with a later addition (zone J) by Tervuren scholars. Each zone is divided into up to nine groups roughly consisting of ten languages each, and named after codes ending in zeroes (for instance, groups A10, A20, A30, etc. within zone A). Individual languages within each group are labeled after non-round digits, such as A11 within A10 and A21 within A20. Guthrie's classification is a referential system based on similarities between geographically close languages, and is in principle independent of genealogy. However, varying degrees of genealogical relatedness can be identified within the different zones. For instance, zones A and B are much less genealogically coherent than zone S (Philippson & Grollemund 2019).

A fully grammaticalized gender system has been reconstructed for Proto-Bantu (Meeussen 1967; Maho 1999; Van de Velde 2019), and is in turn conceived of as inherited from Proto-Niger-Congo. Typically, Bantu gender systems consist of two sets of markers: (1) the overt gender markers, which encode gender distinctions on nouns, and (2) the agreement markers, which encode gender distinctions on adnominal modifiers, pronouns, verbs, and so on. Both nominal class forms and agreement markers also encode singular and plural number distinctions. In most descriptions of Bantu gender systems, one gender consists of combinations (or *pairings*) of singular and plural classes.[2] Traditionally, reference is made to cognate noun classes via a numbering system where odd numbers stand for singular classes and even numbers for plural classes. For instance, class 1 is the singular human class reconstructed as *mò- in Proto-Bantu and class 2 is the plural human class reconstructed as *bà- in Proto-Bantu. Pairings of singular and plural classes are labeled using Roman numerals and referred to as genders (e.g., pairing 1/2 is known as gender I). In this paper, we prefer using the notation 'gender 1/2' because we want to make reference to the internal structuring of each pairing. The total number of classes and genders varies quite widely across the languages of the family. While up to twenty-four distinct classes are reconstructed for Proto-Bantu (Katamba 2003; Van de Velde 2019), none of the contemporary languages retains all twenty-four classes, and some languages have lost them all. Gender marking in Bantu is almost always prefixal (Katamba 2003: 111).[3] Examples of genders 1/2 and 7 from Akoose, spoken in Cameroon, are given in (1).[4]

(1) Gender marking in Akoose (Hedinger 2008: 16–17)

a. Agreement with a Class 1 noun
aw-í *mw-aád* *a-'só* *aw-é* *a-nsóg*
CL1-his CL1-wife CL1-first CL1-which CL1-PST.fat
'his first wife, who was fat'

b. Agreement with a Class 2 noun
áb-é *b-ăn* *bé-kal-e* *me-tóm*
CL2-those CL2-children CL2-tell-IPFV CL6-lies
'those children tell lies'

2 In addition to genders consisting of pairings of singular and plural classes, some genders can be invariant for number.

3 One exception are the Bua languages including Lika; see Boone & Olson (1995).

4 Examples are glossed following the Leipzig Glossing Rules. Abbreviations used: AM associative marker; CL class; IPFV imperfective; LOC locative; PFV perfective; PL plural; PRO pronoun; PST past; SG singular.

c. Agreement (across clauses) with a Class 7 noun
A-tédé e-hid é nyam, á-kɛɛné ch-á́
CL1-take.PFV CL7-bone CL7.AM CL7.animal CL1-take.PFV CL7-PRO
áhîn tê
LOC.CL5.bush in
'He took the animal bone and took it to the forest.'

Bantu gender systems are non-sex-based and characterized by a combination of semantic and formal, as well as opaque, gender assignment (Katamba 2003). A typical Bantu gender system is built upon the following semantic notions (Denny & Creider 1976; Katamba 2003; Contini-Morava 2000):

– Animacy: nouns that denote humans are typically assigned to gender 1/2, animal nouns are associated with gender 9/10, and plant and tree names with gender 3/4.
– Size: several classes (typically 7, 8, 11, 12, 13, 20, and 21) are associated with the encoding of diminutive and augmentative meanings.
– Infinitive marking is associated with class 15.
– Locative meanings are associated with classes 16, 17, and 18.

Not all Bantu languages closely adhere to these generalizations, but, for the purposes of this paper, we assume that the typical Bantu gender system is non-sex-based and only partly conditioned by the overt expression of animacy distinctions.

2.3 *Bantu gender systems: Animacy-based semantic agreement, with a focus on NWB*

One of the major sources of variation in Bantu gender systems is the spreading of semantic agreement (Maho 1999; Van de Velde 2019; Wald 1975). The type of semantic agreement that is most widely attested across Bantu is known in the Bantu literature as "animate concord," which we call here *animacy-based semantic agreement* or simply *animacy-based agreement*.[5] As mentioned above, in Bantu languages, human nouns are typically assigned to gender 1/2. Under animacy-based agreement, the agreement patterns associated with gender 1/2 are also applied to animate nouns from other genders, for instance the many animate nouns from gender 9/10. This pattern of semantic agreement is most often optional and restricted to some agreement targets only. It typically

5 Animacy-based agreement is not the only type of semantic agreement attested in Bantu languages. Other types of semantic agreement include diminutive, augmentative, and locative agreement. For an overview, see Van de Velde (2019).

occurs on verbs, in the form of subject agreement, as well as on a variety of pronouns. Conversely, adnominal modifiers are more likely to maintain syntactic agreement, that is, to agree with the lexical gender of a noun. Semantic and syntactic agreement may coexist as relatively stable variants in one and the same language for long periods of time (see example (3) below for an illustration of alternation between syntactic and animacy-based agreement).[6]

In those languages of the family that exhibit highly eroded gender systems, basic animacy contrasts (such as animate vs. inanimate) are often the only surviving type of gender distinction (see Table 1 below for an illustration from the Bantu language Bila). This raises the question of whether the erosion of grammatical gender and the expansion of animacy-based agreement may be in some way diachronically connected to each other. While Maho (1999) considers this as a plausible scenario, the relationship between semantic agreement and the erosion of gender marking has only recently been more closely investigated by Di Garbo & Verkerk (2022) in a sample of 179 NWB languages. The findings of this study constitute the point of departure for the questions we address in the present paper, and will be summarized in the remaining of this section. Di Garbo & Verkerk (2022) looked at the distribution of types of gender agreement (syntactic vs. animacy-based) on fifteen different agreement targets: attributive adjectives, copula-like constructions, demonstrative modifiers, demonstrative pronouns, genitives/connectives, independent third person pronouns, numerals, quantifiers, possessive pronouns, predicative adjectives, question words, reflexive pronouns, relative constructions, verbs, and other targets. This list of agreement targets captures the general morphosyntactic characteristics of (northwestern) Bantu gender systems and is also reflective of the general typological literature on (gender) agreement systems, and, in particular, on the relation between syntactic and semantic agreement. Data collection was coordinated through a questionnaire that is given in Appendix A. In this paper, we use the data collected by Di Garbo & Verkerk (2022) to test hypotheses about sociogeographic correlates of restructuring in the gender agreement systems of NWB languages.

In order to best capture the diversity of the gender systems attested in the languages of the sample and to avoid committing a priori to a specific model of diachronic change, Di Garbo & Verkerk (2022) identify different profiles of

6 Similar processes are attested in other branches of the Atlantic-Congo family (see Faraclas 1986; Marchese 1988; Good 2012; Güldemann & Fiedler 2019, amongst others). Semantic agreement has been shown to play an important role in the evolution of gender systems around the world (Igartua & Santazilia 2018).

TABLE 1 Animate/inanimate agreement in Bila (Kutsch Lojenga 2003: 462)

	Adjectives	**Numerals 2, 3, 5**	**Demonstratives**
Inanimate	*á-*	*é-/ɛ́-*	*Ń-*
Animate SG	*ma-*		*mɔ́-/mú-*
Animate PL	*ɓá-*	*ɓɔ́-/ɓú-*	*ɓɔ́-/ɓú-*

gender systems depending on whether gender agreement is syntactic (that is, based on the lexical gender of the noun), animacy-based, or both, or, alternatively, whether gender marking is missing altogether. Importantly, in the terminology introduced in that paper, the label *animacy-based agreement* is used for any agreement pattern that singles out the degree of animacy of the noun referent, independently of how this matches specific cutoff points along the Animacy Hierarchy (Smith-Stark 1974). Thus, we apply this label both to agreement patterns that feature an animate vs. inanimate contrast, as is the case of subject agreement in Lika (Augustin 2010: 18–19), and to agreement patterns that are based on a human vs. nonhuman contrast, as is the case for the third person pronouns in Nzadi (Crane et al. 2011: 75).

An example of a language with solely syntactic agreement is Bakole (Glottocode: bako1250), spoken in Cameroon. Syntactic agreement on the verb is shown in (2) for the animate nouns for 'child' and 'sheep'.

(2) Syntactic gender agreement in Bakole (Asobo 1989: 89)
  a. Subject-verb agreement with an animate Class 1 noun
  mw-ánà à mádà
  CL1-child CL1 ate
  'The child ate.'

  b. Subject-verb agreement with an animate Class 3 noun
  mù-ròŋgi mú mádà
  CL3-sheep CL3 ate
  'The sheep ate.'

Bila (bila1255), spoken in the DRC, is a language which displays solely animacy-based agreement. The only targets of gender agreement are adjectives, some numerals, and demonstratives, whose inflections are illustrated in Table 1. Inanimate gender markers are always invariant for number, while the markers of the animate genders distinguish between singular and plural.

In languages characterized by the co-existence of syntactic and animacy-based agreement, verbs are the word class where animacy-based agreement most commonly surfaces, whereas other agreement targets may agree syntactically. Such a combination of syntactic and animacy-based agreement is found in Bomboma (bomb1262), spoken in the DRC. Bomboma has syntactic agreement as described for Bakole on a range of different agreement targets. In addition, it has animacy-based agreement on the genitive markers that are used to mark possession both adnominally and pronominally as well as for verbal subject agreement. Example (3) shows that singular nouns denoting animals may take Class 1 subject agreement prefixes, independently of their lexical gender, while plural nouns for animals may not. The same applies to human nouns from Gender 7/10 (Toronzoni 2004: 70).

(3) Syntactic and animacy-based agreement in Bomboma (Toronzoni 2004: 65)

a. Animacy-based subject-verb agreement with a singular animate Class 9/10 noun
N-va *a-wei*
CL9-dog CL1-died
'The dog is dead.'

b. Syntactic subject-verb agreement with a plural animate Class 9/10 noun
N-va *i-yato* *i-wei*
CL10-dog CL10-three CL10-died
'Three dogs are dead.'

Di Garbo & Verkerk (2022) discuss three languages where animacy-based agreement also extends to the domain of inanimate nouns, leading to what they call *generalized animacy-based semantic agreement*. These languages are Lika (lika1243), Mpiemo (mpie1238), and the Bibaka variety of Ukhwejo (ukhw1241). In Lika, generalized animacy-based agreement is obligatory on subject agreement on verbs, which only differentiates between animate and inanimate subjects, while gender agreement on adnominal modifiers is syntactic (Augustin 2010). In Mpiemo and Bibaka Ukhwejo, generalized animacy-based agreement optionally runs through the entire agreement system. Animate nouns may trigger gender 1/2 agreement whereas inanimate nouns trigger agreement in gender 7/8 (Mpiemo, Thornell 2010) or 7 (Bibaka Ukhwejo, Thornell 2012). This can happen with all agreement targets, irrespectively of the lexical gender of a noun. Syntactic agreement is still in use, particularly among the older genera-

tions of speakers. Di Garbo & Verkerk (2022) argue that the agreement systems attested in these three languages offer crucial empirical evidence for how the erosion of gender marking may possibly be connected with the expansion of animacy-based agreement. When, as happens in Lika, Mpiemo, and Bibaka Ukhwejo, semantic agreement is generalized to inanimate nouns, this leads to a situation in which at least some agreement targets (such as the verb in Lika) or, optionally, all of them (as in Mpiemo and Bibaka Ukhwejo) only encode two gender values, the animate and inanimate. If this polarization of agreement around animacy contrasts is generalized to all existing agreement targets, and becomes obligatory, a gender systems of the type found in Bila, with only two genders, animate and inanimate, may emerge.

In Di Garbo & Verkerk's (2022) sample, complete gender loss is attested in Homa (homa1239), Komo (komo1260), Kituba (two varieties, one spoken in the Kwilu-Kwango regions of the DRC [kitu1246] and one in the Lower Congo region of the Republic of Congo [kitu1245]), Polri (pomo1271), and Yansi (yans1239). Homa is a now extinct language and was once spoken in South Sudan. We know next to nothing about it, except that it may retain some fossilized forms of prefixal animacy-based marking on some attributive adjectives (Santandrea 1963). Komo, spoken in the DRC, still has fossilized prefixes on nouns, but no gender agreement. Thomas (1994) notes the different forms these fossilized nominal prefixes have, and hypothesizes that Komo may have borrowed extensively from other Bantu languages with more conservative systems. The two varieties of Kituba are creoles whose system of nominal prefixation is mostly intact, but which do not have any gender agreement left (Mfoutou 2009; Mufwene 1997; Buchanan 1996–1997; Stucky 1978). In Polri, spoken in Cameroon and Congo, and Yansi, spoken in the Kwilu and Mai-Ndombe provinces of the DRC, relics of the now lost gender system are left on nouns, whereas patterns of agreement only express number distinctions (Wega 2012; Mufwene 2006).

Although they exhibit substantial variation in their respective patterns of gender marking, the languages of Di Garbo & Verkerk's (2022) sample cluster into four main typological profiles, as shown in Fig. 1: (1) languages with only syntactic and no animacy-based agreement (in black, 121 languages); (2) languages with both syntactic and animacy-based agreement (in blue, 40 languages); (3) languages with only animacy-based agreement (in orange, 11 languages); and (4) languages devoid of gender (in green, 6 languages).[7] An over-

7 These counts do not include Shiwa, a language of Gabon whose diverging gender system does not seem to directly relate to the rise and spread of animacy-based agreement (Ollomo Ella 2013).

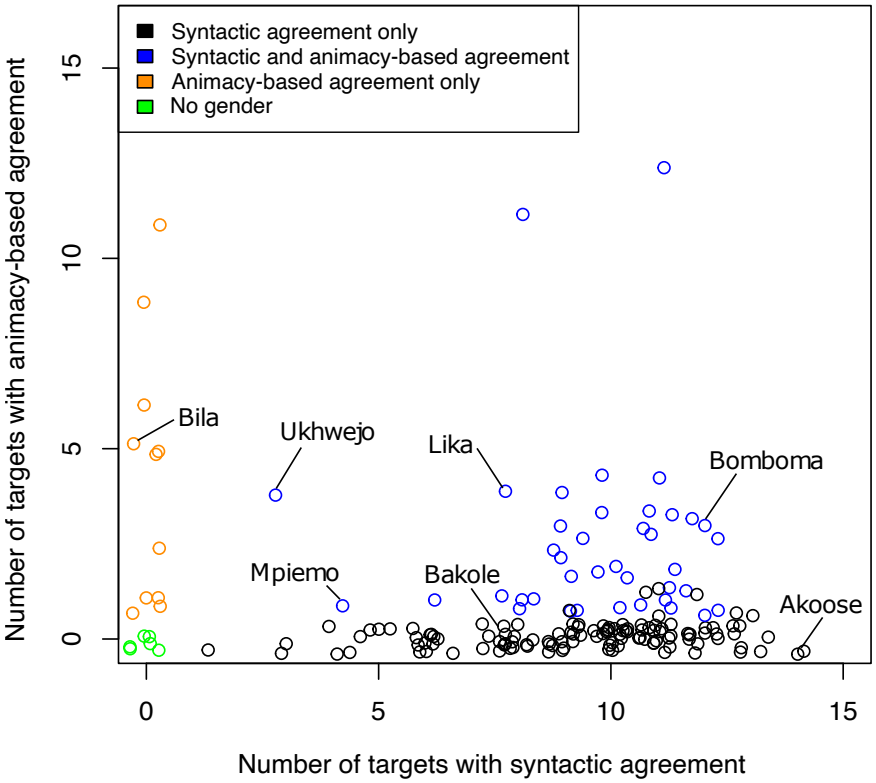

FIGURE 1 Plot of the 179 northwestern Bantu languages in Di Garbo & Verkerk (2022), based on how many of the fifteen agreement targets in each language show animacy-based agreement and how many show syntactic agreement. Points have been jittered so they do not overlap

view of the distribution of syntactic and animacy-based agreement in all the languages of the sample is included as Supplementary Information 1, "Distribution of syntactic and animacy-based agreement per language and across target types."

Di Garbo & Verkerk (2022) make several points regarding diachronic change and the geographical distribution of the four types identified above, which we cannot examine in detail here. Figure 2 illustrates some of these tendencies, by representing the languages and types of gender systems under study in the context of surrounding languages. For instance, the figure shows that there are clear clusters of languages with only syntactic agreement (in purple) and with both syntactic and animacy-based agreement (in dark blue). Languages with only animacy-based gender (in orange) or no gender (in green) are scattered, but primarily found along the northern edge of the reach of the Bantu fam-

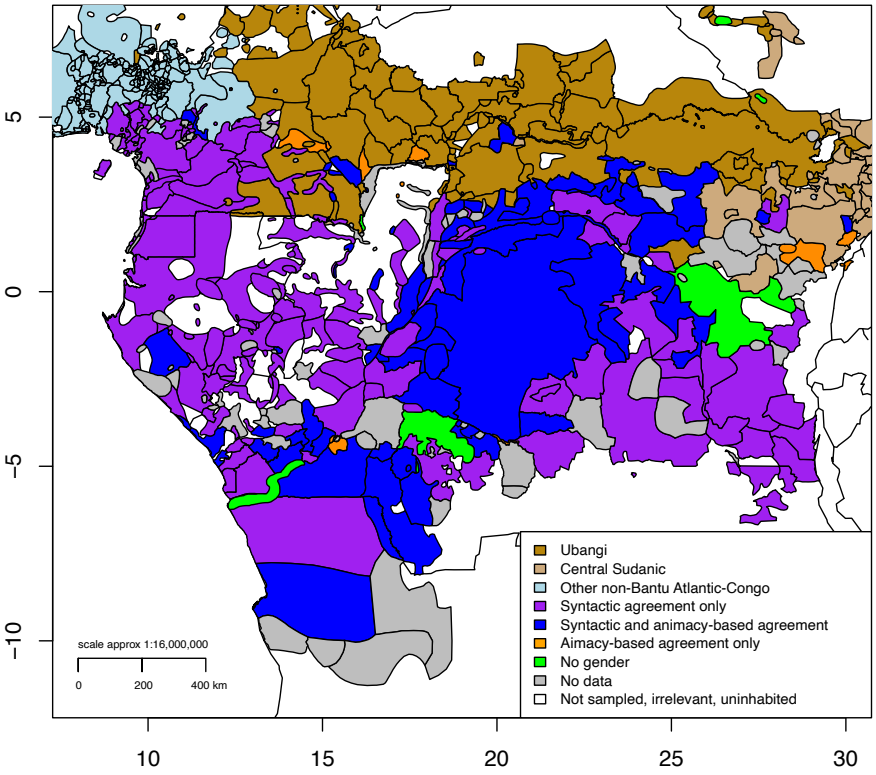

FIGURE 2 Polygon map representing the distribution of the four-way typology among northwestern Bantu languages described in the text in the context of neighboring languages and their genealogical affiliations. Language polygons are taken from the World Language Mapping System (version 17, Global Mapping International, 2015, http://worldgeodatasets.com). White areas represent uninhabited land or bodies of water; towards the north, other Atlantic-Congo, Central Sudanic, and Ubangi languages are not drawn.

ily, as also noticed by Maho (1999). The most northwestern corner of the area investigated contains zone A and B languages, which are closest to the root of the Bantu family tree (Grollemund et al. 2015). Since these are almost all languages with only syntactic agreement (in purple in Fig. 2), Di Garbo & Verkerk (2022) infer that Proto-Bantu probably also had only syntactic agreement, an assumption we also make here.[8]

8 Proto-Bantu reconstructions going back to Meeussen (1967) would align with the conservative assumptions we make here. While Meeussen's Proto-Bantu reconstructions in the domain of verbal inflections (subject and object prefixes excluded) are currently being questioned by specialists in the field, those pertaining to noun classes and agreement markers, as

To summarize, Bantu gender systems are most often described as complex and conservative nominal classification systems, with several gender distinctions, pervasive patterns of syntactic agreement, and pockets of animacy-based agreement mostly localized in the eastern coastal regions (Contini-Morava 2008). However, recent studies suggest that many NWB languages also show pervasive animacy-based agreement and at least seventeen of them display highly eroded systems of gender marking, which only distinguish between animate and inanimate nouns, or have no gender at all. Building upon the data set and findings of Di Garbo & Verkerk (2022), in this paper we test the hypothesis that these pervasive patterns of restructuring occurred in languages characterized by a history of intense and prolonged contact with non-Bantu populations and/or in creolized Bantu varieties.

3 Methods

3.1 *Variable design: Identifying sociohistorical and environmental correlates of gender restructuring*

Based on existing literature (see Section 2.1), we can posit two potential factors that might have impacted patterns of gender marking in the NWB context: (1) contact with non-Bantu languages without gender or with non-Bantu-like gender systems (see Nichols [1992, 2003] as well as Di Garbo [2020] on the influence of neighboring languages on the retention or loss of gender systems) and (2) massive non-native language learning resulting in pidginization and creolization (McWhorter 2001, Trudgill 1999). In this section, we summarize how these factors relate to the seventeen most-eroded gender systems of our 179 NWB language sample. As mentioned in Section 2.3, of these seventeen languages, eleven have solely animacy-based gender while six lack gender altogether. Our aim in this section is to use insights from sociolinguistic and ethnographic resources on the languages and population histories of these communities as a basis for variable design.

The seventeen languages are given here grouped by their closest genealogical relations (according to Glottolog; Hammarström et al. 2018):[9]

– Ababuan (Central-Western Bantu): Bera (bera1259), Amba (amba1263), Bila (bila1255), and Komo (komo1260) (Komoic); Homa (homa1239), Bodo

well as to nominal and verbal derivation, are generally considered to be more solidly established and less in need of revision (Bostoen 2019: 322; Van de Velde 2019).

9 Glottolog's classifications are based on family-specific literature, for these groups; see McMaster (1988), Bastin et al. (1999: 204–205), Grollemund et al. (2018) and Pacchiarotti et al. (2019).

(bodo1272), and Kari (kari1306) (Ngbele-Ngenda); and Beeke (beek1238) (Bali-Beeke)
- Likouala-Sangha (Central-Western Bantu): Pande (pand1264) and Mbati (mbat1248)
- West-Coastal Bantu (Central-Western Bantu): Yansi (yans1239) and Nzadi (nzad1234)
- Makaa-Kako (Bantu A–B10–B20–B30): Polri (pomo1271) and Kako (kako 1242)
- Two creoles: Lingala (ling1263) and Kituba (two varieties, kitu1245 and kitu1246)

For each of these languages, we consulted grammars and ethnographies, as well as sociolinguistic overviews and survey reports, in order to gather information on the (historical) sociolinguistics of the speaker population.[10] We also gathered data, when there was any to be found, on other remarkable grammatical features of the languages as well as on sources of lexical borrowing. A summary of this literature review is included in Appendix C.

The Ababuan group (McMaster 1988) contains, as far as we found, only languages with solely animacy-based gender or no gender. In the other groups, we have pairs of closely related languages (Pande and Mbati, Polri and Kako) or singleton languages (Yansi, Nzadi) that have closely related sister languages that do display syntactic gender agreement (and sometimes animacy-based agreement). The majority of languages with solely animacy-based gender or no gender in our sample (twelve out of seventeen) are located in the northern Bantu borderlands. These are the Ababuan, Likouala-Sangha, and Makaa-Kako groups listed above. The location of these twelve languages is displayed in Fig. 3, together with the locations of the most relevant Ubangi and Central Sudanic languages and of several Pygmy populations taken from Bahuchet (2012).

Data on the social history of these twelve northern borderlands languages, and those in the east of the sampled area, are not always clear nor extensive. However, where information is available, it points towards contact with various non-Bantu populations both in the ethnographic sense and also in the

10 We are aware that existing ethnographies and sociolinguistic overviews often rely on oral traditions collected during colonial times. As problematic as they can be, these accounts are often the only resources available about the social history of many speech communities of Africa. Such a substantial lack of data justifies, in our opinion, using these sources as a basis for building at least a tentative understanding of the history of the populations we study in this paper.

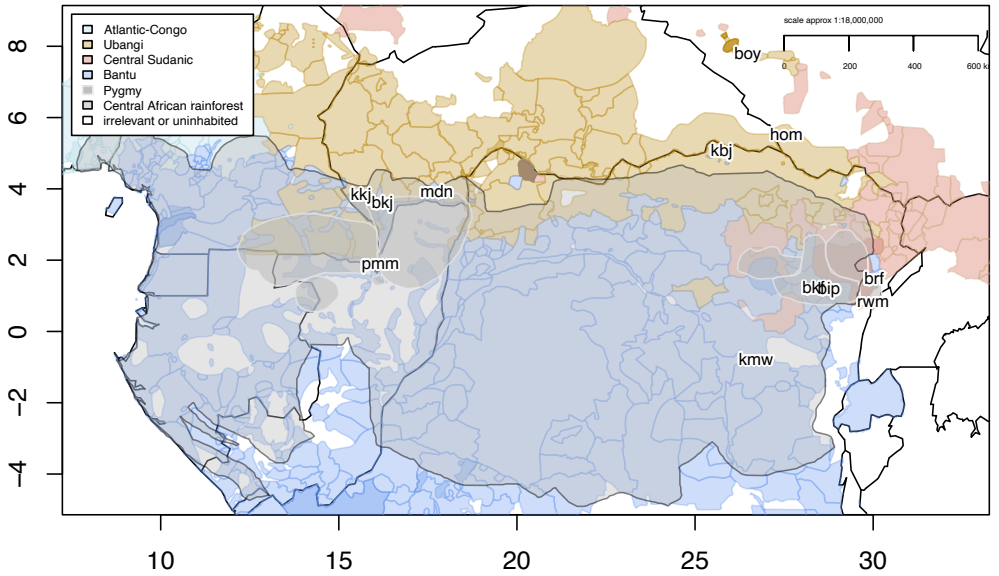

FIGURE 3 Location of the twelve languages with only animacy-based gender marking or no gender at all in relation to surrounding non-Bantu languages, and the Central African rainforest. Language polygons are taken from the World Language Mapping System (version 17, Global Mapping International, 2015, http://worldgeodatasets.com). White areas represent uninhabited land or bodies of water; towards the north, other Central Sudanic and Ubangi languages are not drawn. Locations of Pygmy populations are taken from Bahuchet (2012). The rainforest polygon was taken from Grollemund et al. (2015). The darker shade of some language polygons can be ignored.

linguistic sense (loanwords, unusual features that have likely been borrowed). These contacts feature Ubangi and Central Sudanic speakers as well as Pygmy populations.

Ubangi speech communities came into contact with Bantu speakers when the former migrated from the north as a result of Nilo-Saharan migrations even further north (Bostoen & Donzo 2013: 440, and references therein). Historical records point to a history of prolonged contact, characterized by cultural and linguistic assimilation and influence in both directions, also involving bidirectional language shift (Bostoen & Donzo 2013: 440; Vansina 1990: 66). The Kari (Bantu) speakers, for example, are described as being assimilated into the Zande-speaking community (Ubangi) which suggests long-standing bilingualism in Zande. The prestige of Zande as the majority language of the area may have impacted Kari in a substantial way, perhaps even determining the conditions for Kari to fall out of use altogether.[11] Ubangi languages either lack gender

11 About Mbati, Bouquiaux & Thomas (1994) write to the point: "La question se pose de savoir

or, if they do have gender systems, these may differ from those typical of Bantu languages. For instance, in several Ubangi languages, gender marking tends to be suffixal rather than prefixal (Boyd 1989), and in Zande, gender marking is only pronominal and features a combination of sex-based and non-sex-based distinctions (Corbett 1991). A few Ubangi languages exhibit an interesting combination of syntactic and animacy-based agreement, which is partly reminiscent of what we also encounter in some NWB languages (Fedden & Corbett 2017: 32–37; Fiedler et al. 2021 on Mba).

To the east of these northern borderlands, western Bantu farmers came in contact with Central Sudanic speakers, who were traditionally herders or cereal farmers (Vansina 1990). Vansina (1990: 66–67) reports on the dramatic impact that the expansion of speakers of the Mamwu-Balese subgroup had on western Bantu-speaking communities such as the Bira and Komo, which are both part of our sample and exhibit eroded gender systems. These newly arriving Central Sudanic populations imposed their social organization on the area, but, in the end, "fused with the original farmers" (Vansina 1990: 66). Central Sudanic languages typically lack grammatical gender, and compared to other Nilo-Saharan groupings, they tend to display reduced systems of nominal number marking (Dimmendaal 2000). While, in these languages, nominal morphology is generally suffixal, instances of prefixation also occur, which, in some cases, can be explained as borrowings from neighboring Bantu languages (Dimmendaal 2000: 237; Blench 2018). Finally, the languages originally spoken by the Pygmy populations before they shifted to Bantu, Ubangi, and/or Central Sudanic languages are unfortunately unknown (Bostoen & Gunnink forthcoming 2022; Klieman 2003).

Güldemann (2018: 506–507) and Bostoen & Gunnink (forthcoming 2022) suggest that some of the diverging structural features attested in the northern Bantu borderlands arose as the result of substrate effects from the native languages of the Pygmy groups. Güldemann (2018: 506–507) describes how in the central part of the Macro-Sudan belt and the Bantu spread zone, lower altitude and denser forestation are more prone to pattern with restructured gender systems than higher altitude and savanna-like environments. This typological

s' il s'agit d'une langue bantoue qui a perdu ses principales caractéristiques au contact de ses voisines oubanguiennes (mais pourquoi n'est-ce pas le cas pour le ngando et l'aka?) ou s'il ne s'agit pas tout bonnement d'une langue bantoue empruntée par des Oubanguiens." (It is an open question whether we are dealing with a Bantu language that has lost its main characteristics because of contact with Ubangi neighbors (but why this is not the case for Ngando and Aka?) or whether we are dealing with a Bantu language simply borrowed by Ubangi populations.)

distribution could be connected with the history of early Bantu settlements in this area. The first populations arriving in these areas were not familiar with lowlands rainforest environments. Adapting to this unfamiliar environment required investment in contact with the local forest-based communities (see also Klieman 2003). This scenario must have favored intimate contact between communities, which ultimately led the autochthonous communities to shift to the languages of the newcomers, and created the right context for shift-induced changes to occur. Restructuring of gender marking could be counted as one of these substrate effects. The retention of traditional gender systems in other rainforest Bantu languages does not contradict this hypothesis as speakers of the more conservative languages probably arrived in these areas much later, when familiarity with the environment had already been acquired, and the contact dynamics between rainforest populations were already more asymmetrical and less tight. Considering the genealogical groups listed above, it seems that Ababuan, where we find only languages with heavily restructured gender systems, may fit the Pygmy substrate scenario, although we should note that many of the sources mentioned in Appendix C also remark on contact with Central Sudanic. For the other borderland languages (Pande, Mbati, Polri, and Kako), a more localized scenario involving contact with Ubangi speakers seems more fitting.

Based on this overview, we can distinguish at least two layers of contact-induced cultural adaptations. The first layer concerns interactions between Bantu farmers, newcomers in the area, and local forest-specialists (Pygmies). The second layer concerns interactions between NWB speakers and Ubangi and Central Sudanic populations, who also expanded into the Central African rainforest. As the current discussion and the data in Appendix C show, both layers of contact had profound effects on the linguistic profile of these areas, either in the form of rapid and abrupt language shifts or in the form of prolonged bilingualism and reciprocal assimilation. We can speculate that rapid and abrupt language shift may explain the radically restructured gender systems or the complete loss of gender which we find attested in these twelve northern Bantu borderlands languages.[12] Conversely, prolonged bilingualism

12 The area where these radically restructured gender systems are attested largely coincides with the northern Congo Basin region, as defined by Seidensticker et al. (2021) in a recent study of population collapse in the Congo rainforest. The study reports a lower incidence of population collapse for this region (see Fig. S5 in their Supplementary Materials), which may be suggestive of prolonged cohabitation between Bantu and non-Bantu populations. This may in turn explain why more pervasive instances of contact-induced restructuring

could account for the coexistence of syntactic and animacy-based agreement, which is observed both among NWB and Ubangi languages of the area.

As mentioned above, gender restructuring may also result from pidginization and creolization. In our sample, this is clearly associated with Kituba and (Kinshasa) Lingala. As described in Appendix C, the process of pidginization at the origin of these languages in both cases precedes the arrival of the European colonizers and stems from trade practices between local populations around the Congo. What is remarkable in both cases is that while the restructuring of gender marking around animacy-based distinctions is pervasive in the agreement domain, to the effect that no agreement is left in Kituba and only animacy-based agreement is found in Kinshasa Lingala, class prefixation on nouns tends to be maintained. Maho (1999: 140) argues that this type of development is typical for Bantu languages of wider communication.

Finally, it is worth noting that the relationship between contact and gender restructuring is not one-to-one. While northern Bantu borderlands languages that have both syntactic and animacy-based agreement (e.g., Pagibete, spoken on the border of Northern Ngbandi) or only syntactic agreement (e.g., Mpongmpong, spoken slightly south of Polri and Kako, and surrounded by Baka and Ubangi languages (Bouquiaux & Thomas 1994)) can perhaps be explained as later arrivals to the rainforest, there are also instances of restructuring where no clear evidence for language contact exists. This is the case for Yansi (no gender)[13] and Nzadi (only animacy-based gender, with a human/nonhuman contrast). These two areally close languages are spoken in the southwesternmost part of Fig. 3, on the southern border of the rainforest, far away from attested non-Bantu populations. To the best of our knowledge, there are no clear-cut sociolinguistic clues as to why these languages exhibit heavily restructured profiles from a general Bantu perspective (both Yansi and Nzadi have other peculiar characteristics beyond the gender domain; see Pacchiarotti & Bostoen 2021). While we cannot exclude the possibility of internal change, Bostoen & Gunnink (forthcoming 2022) mention an unknown substrate as a potential cause.

in the gender domain are attested in languages spoken in this area. We thank Brigitte Pakendorf for this suggestion.

13 Yansi must be characterized by large dialectal variation, as there are sources that describe it as preserving a fairly traditional gender system, with some animacy-based agreement (Mayanga 1985), but also sources which describe it as devoid of any productive pattern of gender agreement, with retention of the nominal prefixes and number-based object indexation on the verb (Maho 1999; Mufwene 1994, 2006). It is the latter variety which we chose to consider in this study.

While the overview presented in Appendix C and summarized in the current section is limited in scope, the results of this survey point towards contact with various non-Bantu languages, at different points in time, and to the impact of large numbers of second language learners as relevant factors for the emergence of animacy-based gender and the complete loss of gender. We translate these observations into a set of sociolinguistic and geographical variables, which we aim to model alongside our typological variables, and which we operationalize as follows:

– **Sociogeographic variables: Ancestor in rainforest; current rainforest overlap; border with Ubangi/Central Sudanic**. We introduce the first two variables here in order to test whether past or current presence in the rainforest predicts a higher incidence of restructuring in gender marking. They are meant to test the hypothesis that restructuring in gender marking is a substrate effect, directly connected with contact between Bantu populations and autochthonous rainforest-specialist populations. The third variable is meant to complement the rainforest variables, and to directly test whether proximity with Ubangi- and/or Central Sudanic-speaking populations predicts a higher incidence of restructuring in the languages of our sample.
– **Geographic variables: Latitude; longitude.**[14] Since radically restructured languages tend to be located in the northern Bantu borderlands, the geographical position of languages as captured by these two variables may be an important factor in explaining their distribution. Beyond these languages, the presence of animacy-based agreement also has a distinct geographic signature (see Fig. 2).
– **Demographic variable: Number of L1 speakers**. As per the discussion of Lingala and Kituba above, this variable is relevant to capture processes of language change that might take place in languages of wider communication. By including this variable, we also expand on previous research which looked at whether the sheer number of gender distinctions correlates with community size (see, e.g., Sinnemäki & Di Garbo 2018).

We believe that, taken together, these variables can provide a representative picture of the NWB language ecology as depicted by the sources that we con-

14 We included these as fixed effects on the premise that decisions on what should be fixed and what should be random effects are erratic (see Bolker 2018 and references therein). However, there are several other options to incorporate control for spatial autocorrelation, including the use of the spatial conditional autoregressive term in the package brms of the statistical program R and including latitude and longitude as smoothed interaction terms in generalized additive models (Wood 2017).

sulted (see Section 5.1 for discussion). While these variables are neither exact nor exhaustive, they appear to shed light on aspects of the interaction between gender systems and the local sociogeographic environment, which we want to test further with the support of quantitative methodologies and based on the full data set of 179 NWB languages. Choosing these variables does not exclude the possibility that other processes, which these variables do not capture, may influence the distribution of types of gender system that we observe in our sample. Given this proviso, we factor in these variables in our quantitative analyses, which we introduce in the next section.

3.2 *Variable implementation: Quantifying sociohistorical and environmental correlates of gender restructuring*

We investigate the hypothesis that a relationship exists between the type of gender system a NWB language has and the sociohistorical and environmental factors identified in Section 3.1. To assess this relationship, we use regression analyses, where the response variable (the type of gender system) is assessed in terms of the sociolinguistic and sociogeographic variables identified in Section 3.1. Our response variable—that is, the types of gender systems that we observe in our sample—can be operationalized in three different ways in our models:[15]

1. As a four-way typology, which we presented in Section 2.3. Languages are classified as having either (1) only syntactic agreement; (2) both syntactic and animacy-based agreement; (3) only animacy-based agreement; or (4) no gender. This translates into a four-way multinomial measure, or three binary measures if the first group is taken as the reference group.
2. As a simplified binary typology. This is obtained by dropping the distinction between types (2), (3), and (4) as given above, that is, by devising a binary measure which contrasts languages with only syntactic agreement with languages that have any other type of gender system (i.e., languages that have various amounts of animacy-based agreement or no gender)
3. As the counts of the number of targets that receive syntactic agreement and of the number of targets that receive animacy-based agreement, that is, as two count measures.

For regression modeling, we exclusively use two of these response measures: the simplified typology (binary measure; only syntactic agreement vs. any other type of gender system) and the counts of how many targets receive syntac-

15 Additional possibilities exist, some of which we also explored. However, we do not discuss them here for the sake of brevity (see Supplementary Information 2 for an overview).

tic agreement and how many receive animacy-based agreement. The four-way typology given in Section 2.3 classifies the languages of the data set into four groups, some of which (those containing languages with only animacy-based gender or no gender) are very small. We attempted to fit such a response variable at earlier stages of this project, and faced issues related to statistical power because of the skewed distribution of languages across groups (see Supplementary Information 2, Section 7, for a discussion of statistical power). From a typological point of view, choosing where to draw the line in the range from languages with a small amount of animacy-based marking through to those with no gender agreement at all is somewhat artificial, as differences between these types can sometimes be minimal (see Section 2.3). For instance, languages with no syntactic agreement and where animacy-based marking occurs on a single target only minimally differ from languages with no gender agreement. Similarly, languages with both syntactic and animacy-based agreement, where the latter is potentially available for all agreement domains, are in effect not so different from languages with only animacy-based gender. In view of this, we think that if any binary distinction is to be made in our typology, it should be between languages that only have syntactic agreement and all other languages (i.e., languages with any amount of animacy-based agreement or no gender marking at all). We do use the four-way typology for plotting in Sections 2.3 and 4 as we find it more fine-grained and insightful.[16]

The data on the sociogeographic variables were collected from the following sources:

– Ancestor in rainforest and current location in rainforest. The data on whether the ancestor language was located in the rainforest are based on the reconstructed route of the Bantu speakers by Grollemund et al. (2015: Fig. 3).[17] Following Grollemund et al.'s analysis, most of the NWB languages in the current sample have an ancestor in the rainforest; only languages in Guthrie zones A30, B60–B70, D20–D50, and H10–H20 do not. The data on current location in the rainforest are based on the ratio of the overlap between each language area and the Central Africa rainforest (ranging from 0 if the language is spoken entirely outside the rainforest to 1 if spoken

16 We suspect that some languages which we currently classify as only having syntactic agreement might also have instances of animacy-based agreement (marginal or otherwise), as this is considered to be a phenomenon that tends to be underreported in grammars (see also Maho 1999). Hence we feel that the binary typology is most conservative.

17 We have chosen this data set because it is the only currently available quantitative data set on the route of the Bantu speakers.

entirely inside the rainforest). The polygons used for these calculations were taken from Grollemund et al. (2015) and the World Language Mapping System.[18]
- Latitude and longitude. These data are almost entirely taken from Glottolog (Hammarström et al. 2018). When data were missing from Glottolog, information was taken from primary sources.
- Border with Ubangi and/or Central Sudanic. The language polygons from the World Language Mapping System and/or identified from the literature were used to identify speaker communities that are in close contact (i.e., touching borders) with Ubangi- and/or Central Sudanic-speaking communities. Data on where these communities are located were taken from *Ethnologue* (Lewis et al. 2016).
- Number of L1 speakers. These data are taken from *Ethnologue* (Lewis et al. 2016).[19]

Given that we excluded Shiwa (see footnote 7) and two other languages due to missing data on population size (see footnote 19), the final data set contains 176 languages. The relationship between the type of gender system and demographic and geographic factors was assessed using several different models. As we deal with a mixture of data types (binary, continuous, counts), we primarily constructed generalized linear mixed effects models (GLMMs) using the package brms (Bürkner 2017) in R (R Core Team 2017). This package allows the user to fit Bayesian multilevel models in R using the probabilistic programming language STAN (Carpenter et al. 2017). The case for using mixed models in language typology has been made by Cysouw (2010) and Coupe (2018) and is gaining ground (Sinnemäki & Di Garbo 2018; Schmidtke-Bode & Levshina 2018; Schmidtke-Bode 2019; among others). Our immediate reason for choosing Bayesian GLMMs over frequentist statistics was that Bayesian packages in R

18 World Language Mapping System, version 17, Global Mapping International, 2015, http://worldgeodatasets.com. For languages that are not featured in *Ethnologue* and the World Language Mapping System, we identified locations of speaker communities from the literature and created corresponding polygons.

19 Five languages in our sample are not featured in Lewis et al. (2016): Bafoto (bafo1235), Lwel (lwel1234), Mpama (mpam1239), Nzadi (nzad1234), and Zamba (zamb1245). We searched the literature for the number of L1 speakers for these languages, and found the following: For Mpama, there are no sources on the number of speakers. Bafoto is listed as extinct by Motingea Mangulu (2001–2002: 151) and, correspondingly, we set the number of speakers at 0. Lwel is estimated to have 35,000 speakers in 1976 (Khang Levy 1979: xi). de Boeck (1948: 862) writes that there are 10,000 Mpama speakers. Nzadi is said to have several thousand speakers (Crane et al. 2011), but since this is not an exact number, we did not use it. Therefore, Nzadi and Mpama are excluded from the quantitative analyses.

(brms [Bürkner 2017] and MCMCglmm [Hadfield 2010]) allow for a wider range of models, including a mixture of categorical and continuous fixed effects, as well as random effects that control for genealogical relations by using the full structure of a phylogenetic tree (set). Additional reasons for opting for a Bayesian framework are discussed by McElreath (2020). Other recent typological studies that opt for a Bayesian approach are Erben Johansson et al. (2020), Urban & Moran (2021), and Guzmán Naranjo & Becker (2021).

Since our sample exclusively consists of closely related languages, we employ statistical methods that can account for historical relationships between these languages. We do this by using (a) grouping variables, that is, classifications in terms of big(gish) genealogical language groups; and (b) phylogenetic trees. The grouping variables capture the relatedness of the languages of our sample at 3000–5000 years in the past (Grollemund et al. 2015). Phylogenetic trees capture the entire pedigree of the sampled languages, as they consist of hierarchically nested groups of closely related languages.

The grouping variable that we used for the analyses reported in the main text is the classification of Narrow Bantu by Glottolog (Hammarström et al. 2018).[20] It contains six groups: Ababuan (22 languages in our sample), Bantu A–B10–B20–B30 (70), Central-Western Bantu (130), East Bantu (12), Lebonya (6), and Mbam-Bube (15).[21] Our sample does not include all Central-Western and all East Bantu languages, as many of these are spoken in different Guthrie zones. We included only languages from the NWB area, from Guthrie zones A, B, C, D, and H (see Section 2.3). For more information regarding the use of Glottolog as the main grouping variable and for the phylogenetic trees used for additional analyses, see Supplementary Information 2, Section 3.

20 Glottolog's classifications (Hammarström et al. 2018) are based on existing literature. In the case of Narrow Bantu, the reference literature is Bostoen & Gregoire (2007), which in turn is based on Bastin et al. (1999).

21 Bube is a stand-alone genealogical grouping in Glottolog (Hammarström et al. 2018), containing a single language. As this is not a workable grouping in a GLMM, we add Bube to Mbam, which has been proposed as the closest relative of Bube.

4 Results: Modeling the sociogeographic correlates of restructuring and erosion in NWB gender systems

We constructed various generalized linear mixed effects models (GLMMs) using the package brms (Bürkner 2017). For reasons of space, only the most relevant aspects of the analyses are discussed here. This means that we focus on identifying those variables that are significant predictors of the different types of gender systems we identified. We also discuss the robustness of our findings in terms of details that we feel are relevant for the current study as well as for sociolinguistic typology at large. These discussions focus on different ways to control for genealogical dependencies, the influence of languages with the most marked gender systems on the results, and using random slopes. Details on the analyses that are not discussed in the text, including full output summaries, are given in Supplementary Information 2. The data set and code to run the various models is available as online Supplementary Information 3.

We first illustrate the demographic and geographic variables as well as the grouping variables in Figs. 4 and 5. Figure 4 contains simple descriptive plots of the nonlinguistic variables crossed with the four-way typology described in Section 2.3. The variables "ancestor in rainforest" and "border with Ubangi/Central Sudanic" are binomial, that is, represent a two-way option between whether each language had an ancestor in the Central African rainforest or not and whether each language shares a border with Ubangi or Central Sudanic languages or not. Most languages had an ancestor in the rainforest; there are no languages with no gender system or only animacy-based gender marking that did not have an ancestor in the rainforest. There is a much smaller proportion of languages with only syntactic agreement among languages that border Ubangi or Central Sudanic languages than among languages that do not border these groups. The remaining sub-figures of Fig. 4 deal with continuous variables, shown using box plots rather than bar charts. The languages with the four different types of gender system differ in their mapping onto these four nonlinguistic variables, with the differences between types being larger for longitude and current forest overlap, and smaller for number of L1 speakers and latitude. Figure 5 gives an overview of the Glottolog grouping variables crossed with: the four-way typology, introduced in Section 2.3; sharing a border with Ubangi or Central Sudanic; and the number of L1 speakers. In the Glottolog groupings, "Central-Western" languages form the biggest and most diverse group. Other groups, such as Mbam-Bube, are less varied in that they do not contain languages of one or more types, and they do not include languages on the Ubangi or Central Sudanic border. The difference in the number of L1 speakers across Glottolog groupings is minimal.

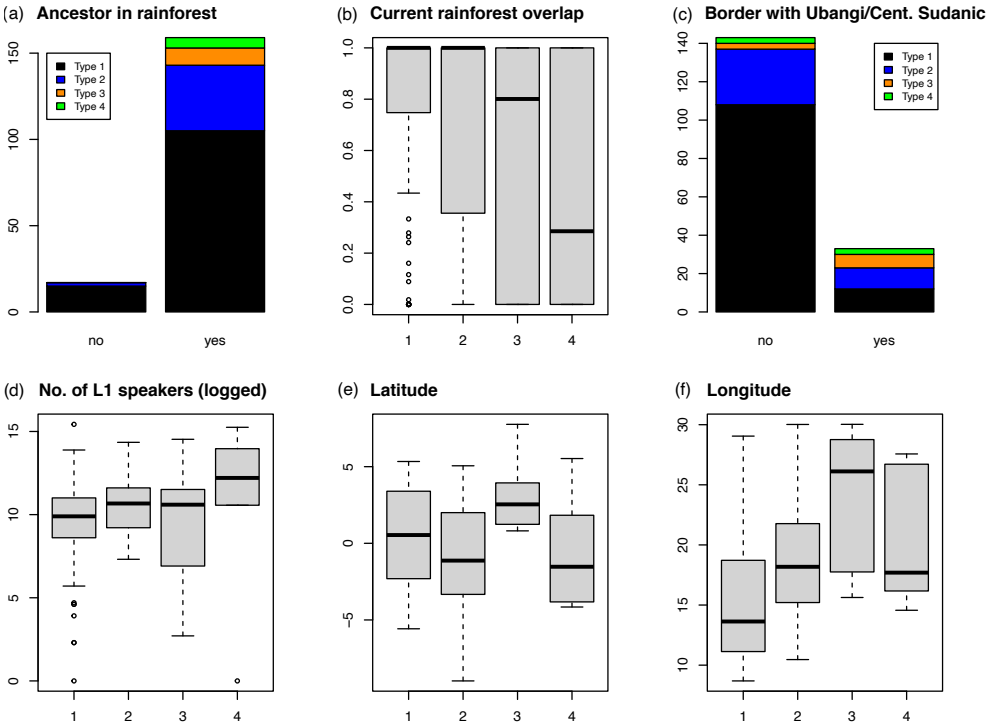

FIGURE 4 Types of NWB gender systems across six demographic and geographic variables. Types of gender system are: (1) only syntactic agreement and no animacy-based agreement; (2) both syntactic and animacy-based agreement; (3) only animacy-based agreement and no syntactic agreement; or (4) no gender. (a), the number of languages with/without an ancestor in the rainforest, by type of gender system. (b), the ratio of the area where the language is spoken inside the rainforest to the total area where the language is spoken, by type of gender system. (c), the languages with/without a border with Ubangi and/or Central Sudanic, by type of gender system. (d), the natural logarithm of the number of L1 speakers of each language by type of gender system. (e) and (f) display respectively the latitude and longitude of each language taken from Glottolog, by type of gender system.

We start the discussion of our results with reports on a set of three GLMMS that explain diversity in gender systems in terms of the six sociogeographic and demographic variables. The first model ("bin_Glot_int") uses the binary typology as response variable, contrasting languages with only syntactic agreement with languages exhibiting any other type of gender agreement or no gender at all. The second ("syn_counts_Glot_int") and third ("ani_counts_Glot_int") model use the counts measures, which detail respectively how many targets receive syntactic agreement and how many receive animacy-based agreement. These models factor in genealogical relations through random intercepts using the Glottolog groupings (for more details, see Supplementary Information 2,

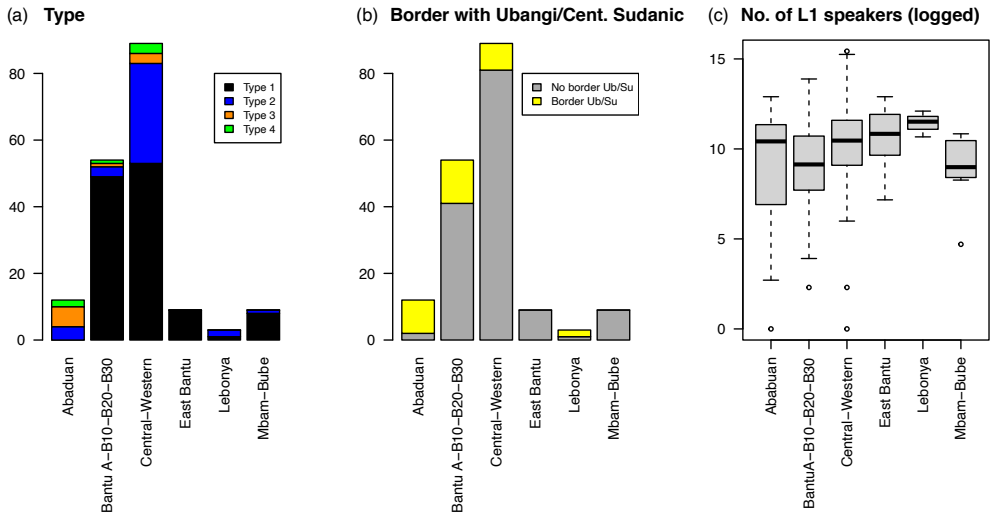

FIGURE 5 (a), number of languages in the main clades of Narrow Bantu (as found on Glottolog) by type of gender system; types of gender system are: (1) only syntactic agreement and no animacy-based agreement; (2) both syntactic and animacy-based agreement; (3) only animacy-based agreement; or (4) no gender. (b), number of languages in the main clades of Narrow Bantu by whether they border Ubangi or Central Sudanic languages. (c), languages plotted by natural logarithm of number of speakers, across the different clades of Narrow Bantu.

Section 3). In all of the models we report on, including those discussed in Supplementary Information 2, random intercepts (always included) and random slopes (included in models discussed in Supplementary Information 2, Section 7) are relevant model components. Hence we do not discuss their specifics extensively. Relevance is assessed by the 95% credibility interval, which should not include zero, however, the strength of the effect may be gauged by how far the distribution is removed from zero.

In "bin_Glot_int," the binary typology (response variable) is modeled in terms of all six predictor variables, and a random intercept using the Glottolog groupings is used to control for genealogical non-independence. An overview of the relevance of the predictors is given in Fig. 6; asterisks (*) preceding predictor names in that figure indicate relevant contributions of that predictor to explaining the response measure. Given the binary nature of the response variable, this analysis is structured in the same way as a simple logistic regression analysis. Since the reference category is having only syntactic agreement (coded as 0), the coefficients are telling us something about the probability of having animacy-based gender marking in any form or no gender marking at all (coded as 1), in terms of the predictor values. There are two relevant independent measures, sharing a border with Ubangi or Central Sudanic and current

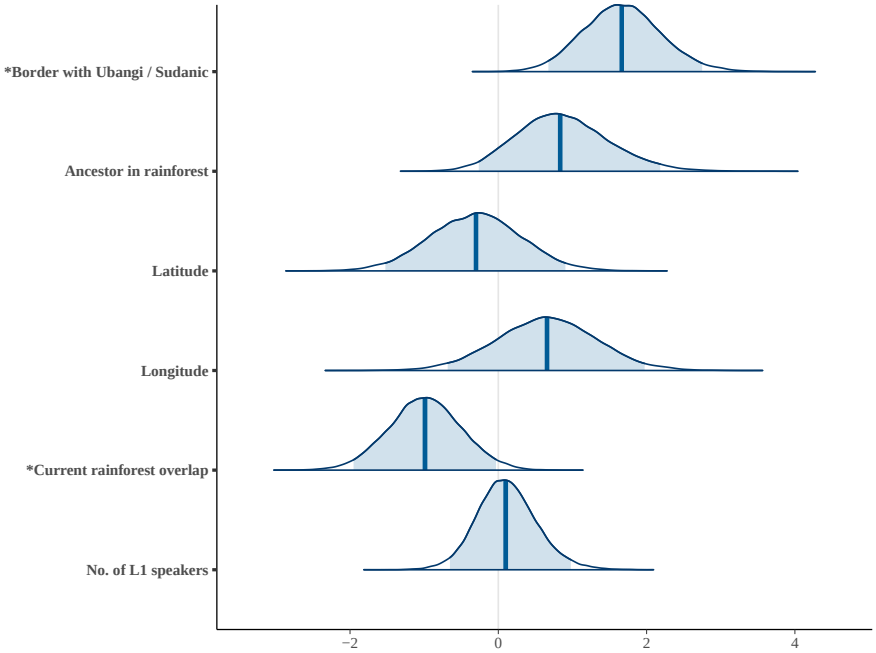

FIGURE 6 Density plots of the posterior distribution of the fixed effects (*b* coefficients) of the binary typology model "bin_Glot_int." The model includes the six predictors and random intercepts using Glottolog groupings.

rainforest overlap. Figure 6 indicates that languages that display any type of animacy-based marking and/or no gender (i.e., languages deviating from the reference category) are characterized by sharing a border with languages from the Ubangi or Central Sudanic families, and by a lower current rainforest overlap.[22]

We can contrast the results on the binary typology as response variable depicted in Fig. 6 with those of the models that take the target counts as dependent measures. We first address the model taking the number of targets that agree syntactically as the response variable; a visual representation is included in Fig. 7. In this model, we find that the number of L1 speakers and sharing a

22 The random intercept that is estimated in this model using the Glottolog groupings is a relevant component of the analysis. The estimated deviation between the average gender type of each Glottolog grouping and the overall average is 2.53, standard deviation 1.32, 95% credibility interval 0.80 to 5.83 (excludes zero). The six Glottolog groupings behave mostly similar except for Ababuan, which has the only significantly diverging random intercept (estimate 3.29, 95% credibility interval 0.18 to 8.78, excludes zero).

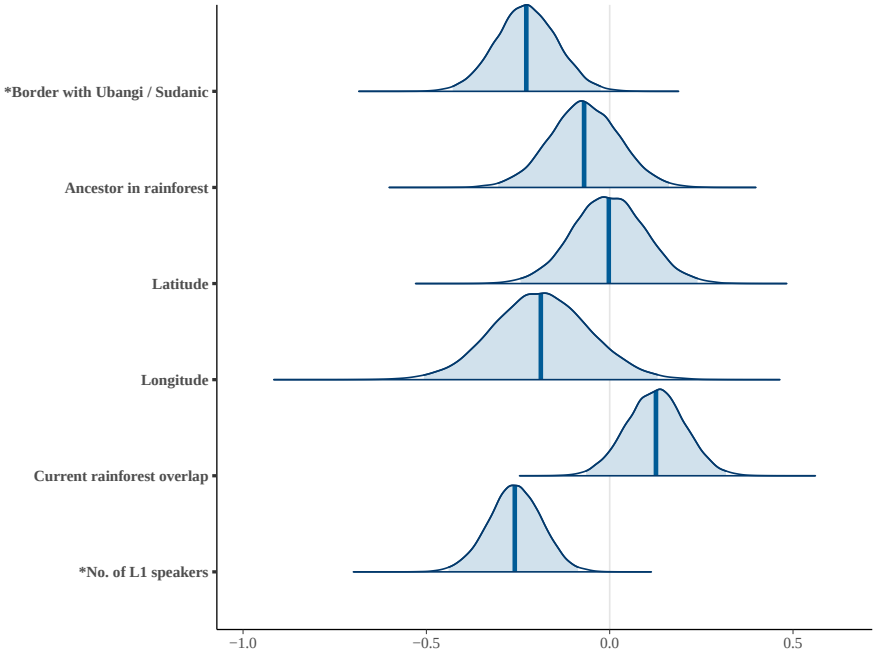

FIGURE 7 Density plots of the posterior distribution of the fixed effects (*b* coefficients) of the number of syntactic agreement target counts model "syn_counts_Glot_int." The model includes the six predictors and random intercepts using Glottolog groupings.

border with Ubangi or Central Sudanic are both relevant negative predictors. This suggests that languages with a larger number of targets that agree syntactically have lower numbers of L1 speakers and less contact with the Ubangi and Central Sudanic language families.[23]

The results of the model "ani_counts_Glot_int," taking the number of targets that receive animacy-based agreement as the response variable, are displayed in Fig. 8. Sharing a border with Ubangi or Central Sudanic and longitude are relevant positive predictors. Languages with a larger number of targets that are marked for animacy-based contrasts tend to share a border with Ubangi or Central Sudanic and to have higher longitudes (i.e., a more easterly posi-

23 Again, the random intercept is a relevant component of the analysis. The estimated deviation between the average gender type of each Glottolog grouping and the overall average is 0.98, standard deviation 0.43, 95% credibility interval 0.46 to 2.08 (excludes zero). Again, Ababuan is the outlier with the only significantly diverging random intercept (estimate −1.23, 95% credibility interval −2.17 to −0.34, excludes zero).

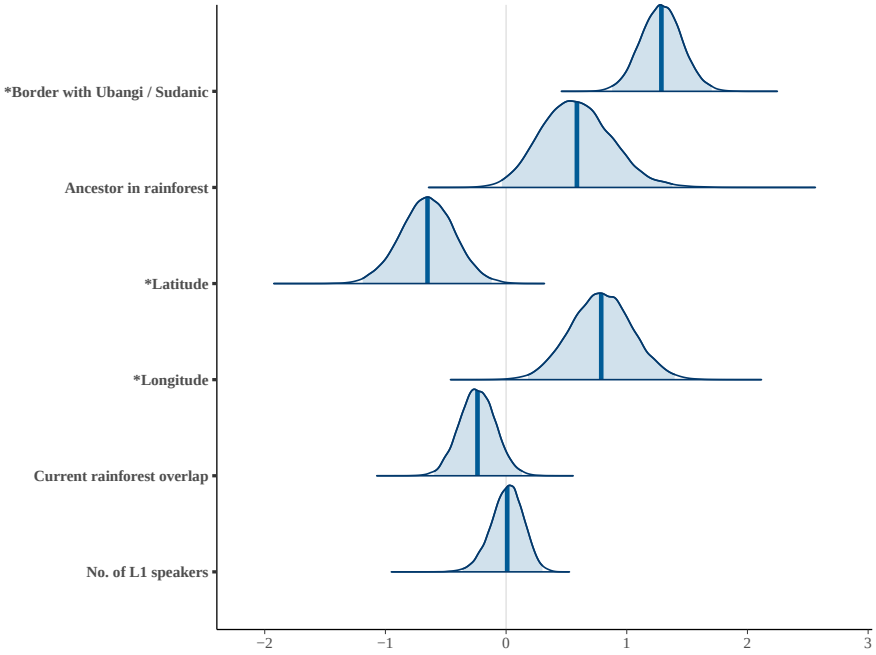

FIGURE 8 Density plots of the posterior distribution of the fixed effects (*b* coefficients) of the number of animacy-based target counts model "ani_counts_Glot_int." The model includes the six predictors and random intercepts using Glottolog groupings.

tion); languages with a smaller number of targets marked for animacy-based contrasts are associated with lower latitudes (i.e, a more southerly position). The effect of bordering Ubangi or Central Sudanic is much stronger than those of latitude and longitude (as it is further away from 0). Additionally, there is an almost significant effect for ancestor in rainforest (95% credibility interval −0.03 to 1.38).[24]

These three models point to five relevant effects, which we further discuss using a set of analyses and figures:

– **Sharing a border with Ubangi or Central Sudanic** (in all three models). Languages sharing a border with Ubangi or Central Sudanic are more likely to display animacy-based marking or have no gender; they have fewer targets

24 Again, the random intercept is a relevant component of the analysis. The estimated deviation between the average gender type of each Glottolog grouping and the overall average is 1.46, standard deviation 0.82, 95% credibility interval 0.49 to 3.52 (excludes zero). None of the Glottolog groups have significantly diverging estimates.

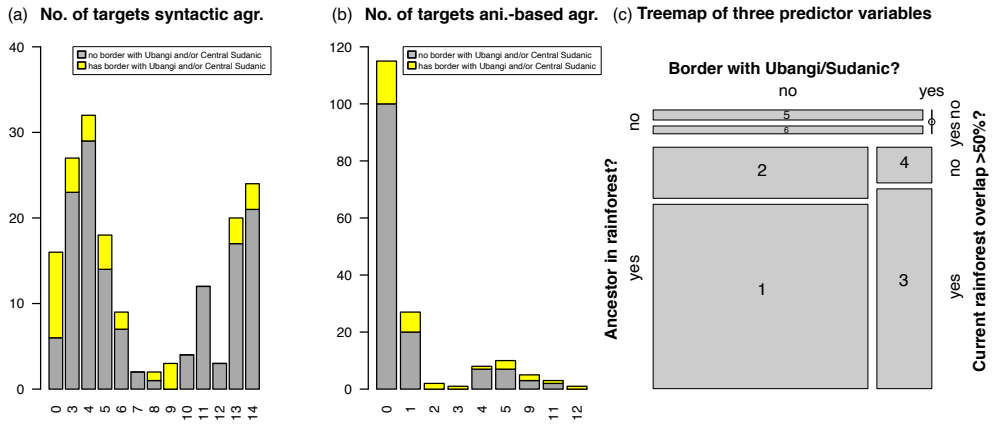

FIGURE 9 Three plots regarding the interaction between "sharing a border with Ubangi/Central Sudanic" and two response variables (a and b), and two other predictor variables (c). Plot (c) is to be read as a 3×2 table; cell size corresponds to group size. In plot (c), *Cell 1*, languages with an ancestor in the rainforest; no border with Ubangi and/or Central Sudanic; and current rainforest overlap > 50%. *Cell 2*, languages with an ancestor in the rainforest; no border with Ubangi and/or Central Sudanic; no current rainforest overlap > 50%. *Cell 3*, languages with an ancestor in the rainforest; a border with Ubangi and/or Central Sudanic; and current rainforest overlap > 50%. *Cell 4*, languages with an ancestor in the rainforest; a border with Ubangi and/or Central Sudanic; no current rainforest overlap > 50%. *Cell 5*, languages with no ancestor in the rainforest; no border with Ubangi and/or Central Sudanic; no current rainforest overlap > 50%. *Cell 6*, languages with no ancestor in the rainforest; no border with Ubangi and/or Central Sudanic; current rainforest overlap > 50%. The crossed o sign in the top right indicates that there are no languages that have a border with Ubangi and/or Sudanic that do not have an ancestor in the rainforest.

agreeing syntactically, and a larger number of targets that display animacy-based agreement. Figure 9 displays the interaction between the binary predictor of sharing a border with Ubangi or Central Sudanic and the number of targets that agree syntactically (left plot) and those that receive animacy-based marking (central plot), as well as two other sociogeographic predictors, having an ancestor in the rainforest and current rainforest overlap (made binary by taking an overlap greater than 0.5 as present within the rainforest, and less than 0.5 as not present, right plot). It shows that more than half of the languages (10 out of 16) which have no targets agreeing syntactically share a border with Ubangi or Central Sudanic (left). Out of all languages sharing a border with Ubangi or Central Sudanic, more than half (18 out of 33) have animacy-based agreement on at least one target (center). When comparing the three variables relating to the Central African rainforest related variables (right), there is a big overlap between having an ancestor in the rainforest and the binary version of being currently present in the rain-

forest, but no big overlap with the variable of sharing a border with Ubangi or Central Sudanic.

– **Number of L1 speakers** (only in "syn_counts_Glot_int"). Languages with a higher number of L1 speakers have a smaller number of targets exhibiting syntactic agreement. Figure 10 illustrates the interaction between the number of targets that agree syntactically and the number of L1 speakers. The main cause for this interaction seems to be languages with no syntactic agreement, which have higher numbers of L1 speakers than those with three to six (or even more) targets of syntactic agreement. However, these large languages with no syntactic agreement are a minority in our sample. We thus implement a control strategy for these outliers, by excluding fifteen languages with more than 400,000 speakers from the models (Supplementary Information 2, Section 4). When we do this, the opposite effect emerges: in this subset of languages, a higher incidence of syntactic agreement is associated with larger population sizes. Since our sample contains only a few "big" languages and many (very) small languages, these results may suggest that the population size measure does not fit particularly well. Its effect on the patterning of the data is observable only above a certain cutoff point, which covers a rather small portion of the data set.
– **Current rainforest overlap** (only in "bin_Glot_int"). Languages with mild or pervasive animacy-based agreement or which do not have gender at all are likely to have a lower current rainforest overlap. This is perhaps an unexpected finding, as we hypothesized a relationship between animacy-based restructuring of gender marking and the Central African rainforest (see Section 3.1). However, we can observe in Figs. 2 and 3 that many of these languages are indeed spoken outside this area.
– **Latitude** (only in "ani_counts_Glot_int"). In general, languages with a lower number of targets that are marked for animacy-based agreement are associated with lower latitudes, that is, a more southerly position, which is what we also observe in Fig. 11, right plot, not counting the languages devoid of any animacy-based gender marking.
– **Longitude** (only in "ani_counts_Glot_int"). Languages with a higher number of targets that are marked for animacy-based agreement are associated with higher longitudes, that is, a more easterly position in the area where NWB languages are spoken. For an illustration, see Fig. 11, left plot.

To summarize, while sharing borders with Ubangi or Central Sudanic languages and present location in or outside the rainforest are relevant in predicting the distribution of our binary types (languages with only syntactic agreement vs. languages with any type of animacy-based marking or no gender marking at all), five predictors out of the six seem to be relevant in accounting for the distri-

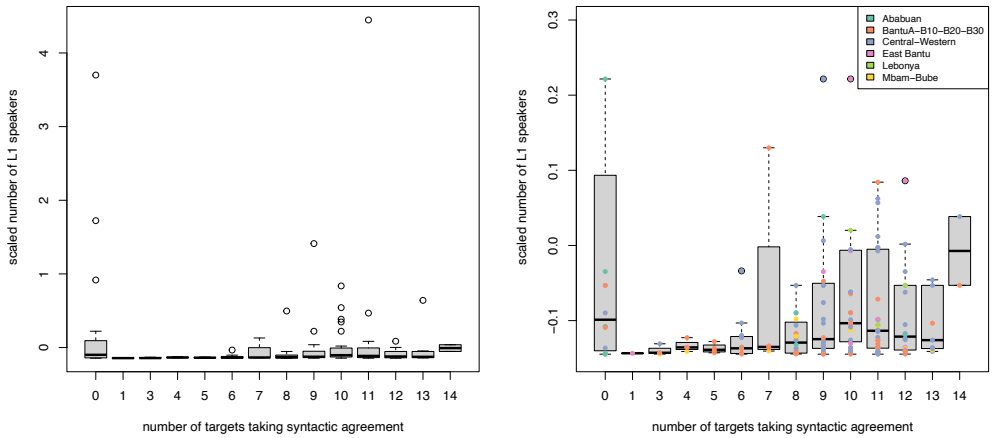

FIGURE 10 Interaction between the number of targets that receive syntactic agreement in each language and the scaled number of L1 speakers, including outliers (left) and zoomed in (right). The number of L1 speakers was scaled using the methodology described by Gelman & Hill (2007: 56–57) and Gelman et al. (2008: 1380) such that the transformed measure has mean 0 and standard deviation 0.5, see Supplementary Information 2. Glottolog groupings are shown using color labels in the right plot.

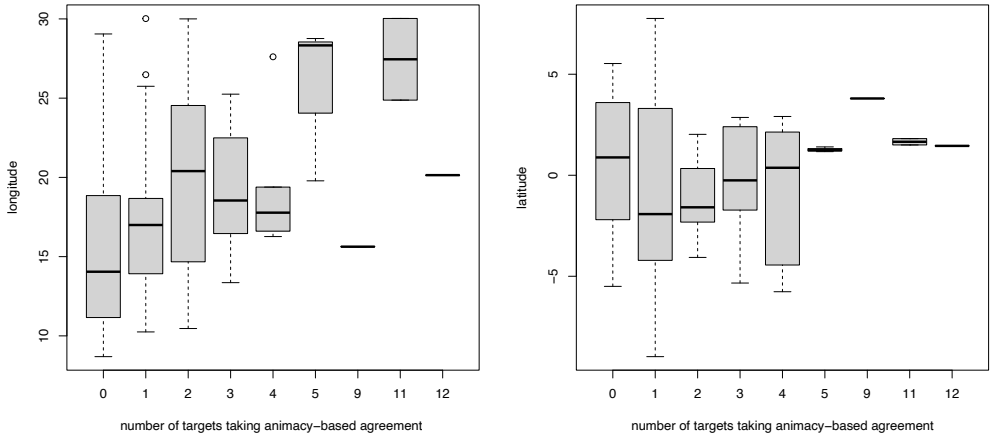

FIGURE 11 Interaction between the number of targets that receive animacy-based gender in each language and the longitude (left) and latitude (right) where the language is spoken.

bution of languages with lower or higher degrees of syntactic agreement versus languages with lower or higher degrees of animacy-based agreement. Having an ancestor language spoken in the rainforest does not play a role in any of the proposed models (except marginally in the model "ani_counts_Glot_int"). This result can probably be explained by the fact that the large majority of the sampled languages did have an ancestor in the rainforest, thus making the data set very homogeneous with respect to this feature.

The robustness of the findings presented above can be addressed from several angles, based on the information provided by additional analyses reported in Supplementary Information 2. We would like to comment on the following: whether the same effects are found when different ways are used to account for genealogical relatedness; the extent to which the effects listed above are dependent on the sampling of languages with only animacy-based gender or no gender; and whether the effects listed above remain relevant when including random slopes.

First, in Supplementary Information 2, Section 3, we show that when using additional analyses with different genealogical controls, only the number of L1 speakers and sharing a border with Ubangi/Central Sudanic are robust, as these variables appear to be relevant across different analyses. Given the exploratory nature of our analyses, we feel it is a moot point to discuss whether some methods are better suited than others here. Rather, we feel it is important to note that some effects are more robust than others given different genealogical controls.

Second, given that languages with only animacy-based gender or no gender represent a rather marked type in the landscape of the gender systems attested in our sample, we ran our models once again while excluding the seventeen languages displaying highly eroded systems. Figure 4 in Supplementary Information 2, Section 5, is a comparison between the results of the analyses on this sample lacking these seventeen languages and the results on the full set of languages presented above in Figs. 6, 7, and 8. In the binary typology model, the effect of current rainforest overlap disappears. In the model of the number of targets that receive syntactic agreement, both effects from sharing a border with Ubangi or Central Sudanic and the number of L1 speakers disappear, and instead there are effects of current rainforest overlap and latitude. In the model of the number of targets that receive animacy-based agreement, the effect of longitude disappears. Hence, in our models, the contribution of these seventeen languages is especially striking when considering the variables relating to the number of targets that display syntactic agreement and the number of L1 speakers. This is in line with the observation we made above concerning the relationship between absence of syntactic agreement and higher numbers of L1 speakers, and the fact that the languages with more than 400,000 speakers are outliers in this data set. When the languages with high numbers of speakers are excluded from the analyses, the effect of this variable also disappears.

Third, in Supplementary Information 2, Section 7, we compare models with random slopes for those predictors that were reported to be relevant above, adding them on a one-by-one basis. Random slopes are additional controls for genealogical relatedness such that the predictor can have a different relationship to the response variable depending on the Glottolog groupings. Random

slopes were added to the random intercepts that were already included in the models discussed so far, for each relevant predictor in an individual model. All effects reported above disappear when random slopes for the specific predictors are added. We believe that this is primarily due to lack of power, and an extensive discussion regarding power is included in Supplementary Information 2, Section 7. In this context, it is worth noting that the models including full phylogenetic trees (that is, maximum clade credibility trees or tree sets), which we may regard as a midway solution between a random intercept model and a random intercept + slope model when controlling for genealogical relatedness, do show effects for number of L1 speakers and sharing a border with Ubangi or Central Sudanic (see Supplementary Information 2, Section 3).

5 Discussion

One of the aims of this paper is to demonstrate the methodological advantages of new approaches to sociolinguistic typological variable design by studying the sociolinguistic typology of NWB gender systems. In this section, we discuss our methods and findings and also outline prospects for future research in the field.

5.1 *Discussion of methods and findings*

This paper explored the hypothesis that gender systems tend to be eroded or even break down under the pressure of language contact (Trudgill 1999; Dahl 2019). NWB languages form an excellent test bed for this hypothesis. This area features languages with traditional gender systems, where only syntactic agreement occurs, and languages with different types of restructured or eroded gender systems, which display various forms of animacy-based agreement, or have no gender at all. We identified three potential factors that may have impacted NWB gender systems: substrate influence from languages that were originally spoken by native Pygmy peoples, who later switched to Bantu (or Ubangi or Central Sudanic) languages as these spread (Klieman 2003; Güldemann 2018; Bostoen & Gunnink forthcoming 2022); contact with Ubangi or Central Sudanic languages without (Bantu-like) gender systems (Vansina 1990; Bouquiaux & Thomas 1994; Wega 2012); and creolization of trade languages which speakers with different language backgrounds came to learn (Fehderau 1966; Samarin 1991; Bokamba 2009; Meeuwis 2013). The first scenario may have applied to the Ababuan languages; the second scenario probably accounts for Polri, Kako, Pande, and Mbati but also for Bodo, Homa, and Kari, and for all northern borderlands Bantu languages to some extent; and the third scenario

applies to Kituba and (Kinshasa) Lingala, as well as, presumably, to other Bantu lingua francas that we have not sampled here.

In Section 4, we modeled the three above-mentioned language contact scenarios using six predictors: number of L1 speakers, latitude, longitude, having an ancestor in the Central African rainforest, current location with respect to the Central African rainforest, and proximity to the Ubangi and Central Sudanic language families. We found that five of these six variables are relevant predictors for one or more of the three main models presented (in Figs. 6, 7, and 8), with two of them, the number of L1 speakers and sharing a border with Ubangi or Central Sudanic, as the most robust. The number of L1 speakers and sharing a border with Ubangi or Central Sudanic were found to be relevant predictors across a variety of methods to control for genealogical relatedness (see Supplementary Information 2, Section 3). None of the reported effects "survives" a more rigorous control for genealogical relatedness, the use of random slopes (Supplementary Information 2, Section 7).[25]

Since the paper deals with exploratory rather than confirmatory analyses, it is useful to consider what insights these findings may offer to more confirmatory studies of the long-debated idea that gender systems become eroded or break down under language contact. We do this by further commenting on the nature of the typological measurements of gender systems that we use in our models, the effect of demographic variables, and the role that heavily restructured languages play in the shaping of our results.

First, the three response variables (the binary typology and the target counts) are inter-correlated, and we should consider them as a relevant set of measures, rather than trying to pick the "best" one. Contrasting languages that have only syntactic agreement with languages that exhibit any kind of animacy-based restructuring in the binary typology, we find that the latter are more likely to share a border with Ubangi or Central Sudanic, as well as to have lower current overlap with the rainforest. This matches what we find when using target counts as the response variables. Languages with animacy-based agreement or no gender at all tend to have fewer or no targets for syntactic agreement and more targets for animacy-based agreement. The number of targets that receive animacy-based agreement can be explained by a set of geographical predictors. This suggests that animacy-based agreement shows the greatest sensitivity to contact effects, either as a result of language shift or con-

25 Note that we have not applied corrections for multiple testing such as the Bonferroni correction, which would have been relevant if this paper were more confirmatory in nature. In addition, we cannot exclude the possibility that one or more of our predictors is dependent on hidden or latent variables that are not included in this study.

vergence with neighboring languages. Our study clearly illustrates that using different types of typological measures of complex morphosyntactic phenomena, such as gender marking, has important consequences for the output of sociolinguistic typological studies.

As far as the effect of demographic variables is concerned, languages devoid of syntactic agreement have higher numbers of L1 speakers than most languages that mark syntactic agreement on any number of targets. However, these "big" languages are clearly outliers in the data set, as most languages of the sample have moderate to small L1 populations, and, as discussed above, the effect of population size disappears (or even changes direction) when these big outliers are excluded. One explanation could be that while the number of L1 speakers (discussed in Section 4 and in Supplementary Information 2, Section 4) might be a relevant factor for worldwide sociolinguistic typological investigations (Lupyan & Dale 2010; Bentz & Winter 2013; Sinnemäki & Di Garbo 2018), this may not necessarily be the case for family- or area-based studies (in Africa), where the demographic profiles of languages may not be normally distributed. In our data set, L1 speaker populations are skewed in both directions. We have many languages spoken by (very) small communities, and a small set of languages that are spoken by (several) millions of speakers. The effect of number of L1 speakers on the type of gender system is different in these subsets of languages, and we would argue that this is a by-product of the family-based approach that we took. We would also suggest that more relevant and fine-tuned variables than a count of L1 speakers can be constructed in order to capture contact dynamics related to incidence of contact and adult second language learning.[26]

Furthermore, the models excluding languages with only animacy-based gender or no gender (Supplementary Information 2, Section 5) suggest that, overall, the effects that we find significantly rely on this subset of most restructured languages. These observations may suggest that such languages should be actually excluded from future work on the typology and evolution of Bantu gender systems, because they are so markedly different from the rest of the family. We would strongly disagree with this idea for a number of reasons. First, several of the sociogeographic effects remain relevant in the abovementioned models

26 Incidentally, a recent study by Dobrushina & Moroz (2021) finds that speakers of smaller languages tend to be more multilingual than speakers of languages with large populations. The effect is most visible with populations up to 120,000 speakers, and is no longer observable in populations above 400,000 speakers. These results align with our findings in suggesting that the predictive power of population size on language-related facts may not be linear/continuous.

even excluding the languages with only animacy-based gender or no gender, which suggests that these factors are also important for explaining the distribution of languages with more traditional gender systems. Second, in those areas where we find several languages with only animacy-based gender or no gender, there are also several languages that we could not include in the sample due to lack of data (see Fig. 2). We cannot ignore that more instances of heavily restructured gender systems could be found among these. Languages with heavily restructured gender should ultimately not be seen as marginal outliers, or at least not in this part of the Bantu-speaking world. Most importantly, irrespectively of how many they are, these languages are of crucial importance to understand how gender systems change, both inside and outside NWB.

5.2 *Prospects for future research*

This paper is the first large-scale comparative study to demonstrate a palpable link between the shape and fabric of gender systems and a number of demographic and geographic factors. However, we feel that this is not the end of the story but rather the beginning. The next steps to be taken are those where we measure more exactly the impact that population movements within the Bantu family and contact with non-Bantu populations had on the evolution of Bantu gender systems. We can imagine the following options, all of which (unfortunately) require detailed research that falls outside the scope of this paper:

- Comparative studies of contact with Ubangi and Central Sudanic. This could be implemented through qualitative studies of language contact dynamics on the languages of our sample that border with Ubangi or Central Sudanic languages. Information regarding the gender systems of these languages, if they have any, would be vital to find out how NWB gender systems changed under "pressure" from non-Bantu-like gender systems and nominal morphosyntax. Such a qualitative study would also inspire further quantitative analyses of gender systems, as well as further relevant sociolinguistic and sociogeographic measures.
- Measures of the amount of time spent inside the rainforest. This could be achieved by looking at branch lengths in the trees of Grollemund et al. (2015), Koile et al. (in review), and future phylogenetic investigations of the Bantu language family. A variable of this kind could help to better disentangle the difference between older and recent waves of expansion in the rainforest, which, as argued by Güldemann (2018), is crucial to motivate the presence of both languages with only syntactic agreement and with restructured gender systems in present-day rainforest areas.
- Measures of sociolinguistic dynamics at the macro- and micro-level, including demography, processes of pidginization or creolization, information on

register usage, passive understanding versus active usage, age at which a language was learned, and so on. This is a goal which is not achievable based on presently available sociolinguistic data, but ongoing research in the field of sociolinguistic typology (Di Garbo et al. 2021) makes it possible to believe that further advancement in the field could be achieved in the near future.

– Measures of the sociopolitical complexity of speech communities, for instance along the lines of the variable "Jurisdictional hierarchy beyond local community" in D-Place (Kirby et al. 2016). See also David Gil's ideas (in the LingTyp discussion list, October 2018) about a scale of language-related sociopolitical complexity ranging from national languages, through local varieties of national languages and local languages historically part of a larger political entity, to local languages only recently part of a larger political entity. Such a measure could capture aspects of language use whenever more detailed sociolinguistic information is not available.
– Comparison of linguistic data on the distribution of restructured gender systems with molecular anthropological data. While such analyses are not immediately pursuable given the current level of genetic sampling in the relevant areas of interest (Pakendorf et al. 2011: 59), future advances in the field could make this possible.

A valid alternative to the steps above would be to compare the NWB data with better documented contact scenarios within the Bantu-speaking world. Probably the best candidate for such a comparative endeavor would be the southern Bantu languages, whose history of contact with non-Bantu, "Khoisan"-speaking populations is well documented both from a linguistic and a molecular anthropological point of view, and with the added advantage that the relevant contact languages are still spoken in the area and relatively well described. While extensive work has been done on Khoisan influence on southern Bantu at the phonetic level (Pakendorf et al. 2017), and the population dynamics in the context of which these patterns of language convergence took place have also been reconstructed with the support of molecular anthropological data, not much work has been done in the domain of nominal morphosyntax, let alone gender marking. Nevertheless, existing research points to the possibility that in the southern Bantu languages, contact-induced change in nominal morphosyntax goes in the direction of the weakening, and sometimes loss, of the locative and evaluative (diminutive and augmentative) genders, and, in parallel, of the grammaticalization of suffixes which express these functions in alignment with the marking strategies that are found in neighboring Khoisan languages (Creissels 1999; Güldemann 1999). Whether, in combination with these developments, any form of animacy-based agreement also occurs, and

can be motivated as the result of contact with neighboring Khoisan languages, is, at present, unexplored territory.

In addition to areally restricted comparisons, a pan-Bantu study encompassing all sub-areas of the family would be the most natural long-term development of this work, especially as we suspect that animacy-based agreement is more widespread throughout the family than previously thought (Wald 1975; Maho 1999: 135). Through a comprehensive study of gender systems across the entire Bantu family, data on synchronic distributions could be used to test transition probabilities between types of gender systems and patterns of gender agreement. Their interaction with relevant language-external factors beyond those considered for the NWB area could also be studied.

6 Concluding remarks

We have shown how the general crosslinguistic and NWB-specific tendency for gender systems to be restructured around semantically transparent animacy contrasts may tie in with the history of the NWB populations. Ethnographic accounts pointed to relevant measures of population contact dynamics, which were then tested using GLMMs. We found that environmental variables related to geography and demography are relevant predictors of the distribution of restructuring in NWB gender systems. We conclude that hypotheses about linguistic adaptation and the interaction between language structures and the environment are best tested through integrated research approaches, where quantitative analyses generalizing over large sets of data are as much as possible based upon qualitative data, which zoom in to the specifics of individual languages and speech communities. We also demonstrated that an analysis of patterns of gender marking which focuses both on the *locus* of marking, that is, the word classes that carry gender inflections, and the *type* of marking, that is, the distinction between, on the one hand, syntactic agreement, and, on the other hand, semantic, in our case animacy-based, agreement, is more revealing of the evolutionary dynamics of gender systems than analyses based on sheer number of gender distinctions, such as those conducted in previous studies (Lupyan & Dale 2010; Blasi et al. 2017; Sinnemäki & Di Garbo 2018; Dahl 2019). A more nuanced and ecologically informed approach to variable coding like the one adopted here for grammatical gender is thus key to test hypotheses on linguistic adaptation in individual grammatical domains, language families, and areas.

Supplementary materials

This paper is accompanied by three pieces of Supplementary Information that can be found here: https://doi.org/10.6084/m9.figshare.17261090:

- Supplementary Information 1: A figure entitled "Distribution of syntactic and animacy-based agreement per language and across target types."
- Supplementary Information 2: A PDF document with details on the analyses.
- Supplementary Information 3: All relevant code and data, including code for making the figures.

Acknowledgments

This paper is the result of a research collaboration begun in May 2016, during the first Quantitative Methods Spring School, organized by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History (Jena). We are very thankful to the spring school's directors, Russell Gray and Fiona Jordan, and to all participants for support and inspiration. Ongoing analyses and preliminary results of this research have been presented at various conferences, workshops, and other academic venues. We are particularly thankful to the audiences of the workshop on "Language shift and substratum interference in (pre)history" (Jena 2017), the SLE conferences of 2017 (Zürich) and 2018 (Tallinn), and the African Linguistics research seminar of the Humboldt University (Berlin, 2017). We wish to thank Ines Fiedler, Tom Güldemann, Brigitte Pakendorf, and Bernhard Wälchli for feedback and stimulating discussions during the course of this work. For help with data collection we are thankful to Harald Hammarström. For input on the statistical analyses we thank Simon Branford, Dan Dediu, Cara Evans, Russell Gray, Simon Greenhill, Ezequiel Koile, Catherine Sheard, and Kaius Sinnemäki. We thank the reviewers for valuable feedback and Tim Curnow for excellent copy editing. Di Garbo has received funding for this research from the Wenner-Gren Foundations (Sweden) and partly from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 805371). The usual disclaimers apply.

References

Asobo, Irene Swiri. 1989. *The Noun Class System of Kɔlɛ*. MA thesis, Université de Yaoundé I.

Augustin, MaryAnne. 2010. *Selected features of syntax and information structure in Lika (Bantu D.20)*. MA thesis, Graduate Institute of Applied Linguistics.

Bahuchet, Serge. 2012. Changing language, remaining pygmy. *Human Biology* 84. 11–43.

Bastin, Yvonne, A. Coupez & M. Mann. 1999. *Continuity and divergence in the Bantu languages: Perspectives from a lexicostatistic study*. Tervuren: Royal Museum for Central Africa.

Bentz, Christian & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3. 1–27.

Blasi, Damian E., Susanne Maria Michaelis & Martin Haspelmath. 2017. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour* 1(10). 723–729. doi:10.1038/s41562-017-0192-4.

Blench, Roger. 2018. Core and peripheral noun morphology in Central Sudanic languages. Paper presented at the 13th Nilo-Saharan Conference, University of Addis Ababa, May 6, 2017.

de Boeck, L.B. 1948. La classification des langues en Afrique. *Bulletin des séances de l'Institut Royal Colonial Belge* 19. 846–873.

Bokamba, Eyamba. 2009. The spread of Lingala as a lingua franca in the Congo Basin. In Fiona McLaughlin (ed.), *The languages of urban Africa*, 50–70. London: Continuum.

Bolker, Ben. 2018. *GLMM FAQ*. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random. Accessed on November 12, 2021.

Boone, Douglas W. & Kenneth S. Olson. 1995. Bua bloc survey report. Bunia (Zaïre): SIL Eastern Zaire.

Bostoen, Koen. 2019. Reconstructing Proto-Bantu. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu languages: Second edition*, 308–334. London: Routledge.

Bostoen, Koen & Jean-Pierre Donzo. 2013. Bantu-Ubangi language contact and the origin of labial-velar stops in Lingombe (Bantu, C41, DRC). *Diachronica* 30. 435–468.

Bostoen, Koen & Claire Gregoire. 2007. La question Bantoue: bilan et perspectives. *Mémoires de la Société de linguistique de Paris* 15. 73–91.

Bostoen, Koen & Hilde Gunnink. forthcoming 2022. The impact of autochthonous languages on Bantu language variation: A comparative view on Southern and Central Africa. In Salikoko Mufwene & Anna Maria Escobar (eds.), *Cambridge handbook of language contact*, Cambridge: Cambridge University Press.

Bouquiaux, Luc & Jacqueline M.C. Thomas. 1994. Quelque problèmes comparatifs de langues bantoues C10 des confins oubanguiens: Le case du mbati, du ngando e de

l'aka. In Geider, Thomas and Kastenholz, Raimund (eds.), *Sprachen und Sprachzeugnisse in Afrika: eine Sammlung philologischer Beiträge Wilhelm J.G. Möhlig zum 60. Geburtstag zugeeignet*, 87–106. Köln: Cologne: Rüdiger Köppe Verlag., 87–106. Cologne: Rüdiger Köppe.

Boyd, Raymond. 1989. Adamawa-Ubangi. In John Bendor-Samuel & Rhonda L. Hartell (eds.), *The Niger-Congo Languages*, 178–215. Lanham, MD: University Press of America.

Bruel, Georges. 1910–1911. Les populations de la Moyenne Sanga. *Revue d'ethnographie et de sociologie* 1–2. 3–32, 111–125.

Buchanan, Deborah L. 1996–1997. The Munukutuba noun class system. *Journal of West African languages* 26(2). 71–86.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28. doi:10.18637/jss.v080.i01.

Burnham, Philip, Elisabeth Copet-Rougier & Philip Noss. 1986. Gbaya et Mkako: Contribution ethno-linguistique à l'histoire de l'Est-Cameroun. *Paideuma* 87–128.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). doi:10.18637/jss.v076.i01.

Contini-Morava, Ellen. 2000. Noun class as number in Swahili. In Hellen Contini-Morava & Yishai Tobin (eds.), *Between grammar and lexicon*, 1–30. Amsterdam: John Benjamins.

Contini-Morava, Ellen. 2008. Human relationship terms, discourse prominence, and asymmetrical animacy in Swahili. *Journal of African Languages and Linguistics* 29. 127–171.

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2013. Systems of gender assignment. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/32. Accessed February 14, 2014.

Coupé, Christophe. 2018. Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale and shape. *Frontiers in Psychology* 9. 513. doi:10.3389/fpsyg.2018.00513.

Crane, Thera Marie, Larry M. Hyman & Simon Nsielanga Tukumu. 2011. *A grammar of Nzadi [B865]: A Bantu language of the Democratic Republic of Congo*. Berkeley: University of California Press.

Creissels, Denis. 1999. Origin et évolution des diminutifs et augmentatifs dans quelques langues africaines. *Silexicales* 2. 29–35.

Cysouw, Michael. 2005. Quantitative methods in typology. In Reinhard Köhler, Gabriel

Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics: An international handbook*, 554–578. Berlin: Mouton de Gruyter.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14. 221–234. doi:10.1515/lity.2010.010.

Dahl, Östen. 2000. Elementary gender distinctions. In Barbara Unterbeck, Matti Rissanen, Terttu Nevalaien & Mirja Saari (eds.), *Gender in grammar and cognition*, 577–593. Berlin: Mouton de Gruyter.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Dahl, Östen. 2019. Gender: Exoteric or esoteric? In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity*, vol. 1: General issues and specific studies, 53–61. Berlin: Language Science Press.

Denny, Peter & A. Chet Creider. 1976. The semantics of noun classes in Proto-Bantu. *Studies in African Linguistics* 7. 1–30.

Di Garbo, Francesca. 2020. The complexity of grammatical gender and language ecology. In Peter Arkadiev & Francesco Gardani (eds.), *The complexities of morphology*, 193–229. Oxford: Oxford University Press. doi:10.1093/oso/9780198861287.003.0008.

Di Garbo, Francesca, Eri Kashima, Ricardo Napoleão De Souza & Kaius Sinnmäki. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Silvia Ballarè & Guglielmo Inglese (eds.), *Tipologia e sociolinguistica: Verso un approccio integrato allo studio della variazione – Atti del workshop della società linguistica italiana 20 settembre 2020* 5, 143–176. Officinaventuno. doi:10.17469/O2105SLI000005.

Di Garbo, Francesca & Annemarie Verkerk. 2022. A typology of northwestern Bantu gender systems. To appear in *Linguistics*.

Dijkmans, Joseph J.M. 1936. De Akare. In *Exspectatio gentium: Compte rendu de la XIIIe semaine de missiologie de Louvain (1935)* (Museum Lessianum; section missiologique 24), 116–136. Bruxelles: L'Édition Universelle.

Dimmendaal, Gerrit. 2000. Number marking and noun categorization in Nilo-Saharan languages. *Anthropological Linguistics* 42. 214–261.

Dobrushina, Nina & George Moroz. 2021. The speakers of minority languages are more multilingual. *International Journal of Bilingualism* 25. 921–938. doi:10.1177/13670069211023150.

Eberhard, David M., Gary F. Simons, & Charles D. Fennig (eds.). 2019. *Ethnologue: Languages of the world, twenty-second edition*. Dallas, TX: SIL International. Online version: http://www.ethnologue.com.

Erben Johansson, Niklas, Andrey Anikin, Gerd Carling & Arthur Holmer. 2020. The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology* 24(2). 253–310. doi:10.1515/lingty- 2020-2034.

Faraclas, Nicholas. 1986. Cross River as a model for the evolution of Benue-Congo nominal class/concord systems. *Studies in African Linguistics* 17(1). 39–54.

Fedden, Sebastian & Greville G. Corbett. 2017. Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: A journal of general linguistics* 2(1). 34. doi:10.5334/gjgl.177.

Fehderau, W. Harold. 1966. *The origin and development of Kituba* (*lingua franca Kikongo*). PhD Thesis, Ithaca: Cornell University.

Fiedler, Ines, Tom Güldemann & Benedikt Winkhart. 2021. The two concurrent gender systems of Mba. *STUF—Language Typology and Universals* 74(2). 303–325. doi: 10.1515/stuf-2021-1034.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau & Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4). 1360–1383.

van Geluwe, H. 1956. *Les Bira et les peuplades limitrophes* (Ethnographic Survey of Africa: Central Africa 2). London: International African Institute.

van Geluwe, H. 1960. *Les Bali et les peuplades apparentées* (*ndaka-mbo-beke-lika-budunyari*) (Ethnographic Survey of Africa: Central Africa, Belgian Congo 5). London: International African Institute.

Good, Jeff. 2012. How to become a "Kwa" noun. *Morphology* 22. 293–335.

Grollemund, Rebecca, Simon Brandford, Koen Bostoen, Andrew Meade, Chris Venditti & Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Science of the United States of America* 112. 13296–13301.

Grollemund, Rebecca, Jean-Marie Hombert & Simon Branford. 2018. A phylogenetic study of North-Western Bantu and South Bantoid languages. In Rajend Meshtrie & David Bradley (eds.), *The dynamics of language: Plenary and focus lectures from the 20th International Congress of Linguists, Cape Town, July 2018*, 118–132. Cape Town: UCT Press.

Güldemann, Tom. 1999. Head-initial meets head-final: Nominal suffixes in eastern and southern Bantu from a historical perspective. *Studies in African Linguistics* 29. 49–91.

Güldemann, Tom. 2018. Areal linguistics beyond contact, and linguistic areas of Afrabia. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 448–545. Berlin: Mouton de Gruyter.

Güldemann, Tom & Ines Fiedler. 2019. Niger-Congo "noun classes" conflate gender with deriflection. In Francesca Di Garbo, Bruno Olsson & Bernard Wälchli (eds.), *Grammatical gender and linguistic complexity*, vol. 1: General issues and specific studies, 95–145. Berlin: Language Science Press.

Guthrie, Malcolm & Archibald Norman Tucker. 1956. *Linguistic survey of the northern*

Bantu borderland, vol. 1. London: Oxford University Press & International African Institute (IAI).

Guthrie, Malcom. 1967–1971. *Comparative Bantu: An introduction to the comparative linguistics and prehistory of the Bantu languages*. Farnborough: Gregg International Publishers.

Guzmán Naranjo, Matías & Laura Becker. 2021. Statistical bias control in typology. *Linguistic Typology*. doi: 10.1515/lingty-2021-0002.

Hadfield, Jarrod. 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software* 33. 1–22.

Hammarström, Harald, Robert Forkel & Martin Haspelmath (eds.). 2018. *Glottolog 3.3*. Jena: Max Planck Institute for the Science of Human History. doi:10.5281/zenodo.1321024. http://glottolog.org. Accessed on May 11, 2017.

Harvey, Tammie K. 1997. *The Bali of Northeastern Congo-Kinshasa: Uncovering the history of a people shrouded by the Ituri rain forest*. PhD dissertation, University of Texas at Arlington.

Hedinger, Robert. 2008. *A grammar of Akoose: A Northwest Bantu language* (SIL International and The University of Texas at Arlington Publications in Linguistics 143). Dallas: SIL International.

Igartua, Iván & Ekaitz Santazilia. 2018. How animacy and natural gender constrain morphological complexity: Evidence from diachrony. *Open Linguistics* 4(1). 438–452.

Jacquot, André & Irvine Richardson. 1956. Report of the western team: Atlantic coast to Oubangui. In Malcolm Guthrie & Archibald Norman Tucker (eds.), *Linguistic survey of the northern Bantu borderland*, vol. 1, 9–62. London: Oxford University Press & International African Institute (IAI).

Joset, Paul E. 1952. Les baamba et les babwizi du Congo Belge et de l'Uganda Protectorate. *Anthropos* 47. 369–387, 909–946.

Katamba, Francis. 2003. Bantu nominal morphology. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 103–120. London: Routledge.

Kempe, Vera & Patricia J. Brooks. 2018. Linking adult second language learning and diachronic change: A cautionary note. *Frontiers in Psychology* 9. 480. doi:10.3389/fpsyg.2018.00480.

Khang Levy, Nyi-m'shum. 1979. *Elements de grammaire morphologique de la langue Lwel*. MA thesis, Lubumbashi: Université Nationale du Zaïre (UNAZA) dissertation.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale & Michael C. Gavin. 2016. D-place: A global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11(7). e0158391.

Klieman, Kairn A. 2003. *The Pygmies Were Our Compass: Bantu and Batwa in the History of West Central Africa, Early Times to c. 1900 C.E.* Portsmouth, NH: Heinemann.

Koile, Ezequiel, Simon J. Greenhill, Damian E. Blasi, Remco Bouckaert, and Russell D. Gray. (in review) Phylogeographic analysis of the Bantu language expansion supports a rainforest route.

Kutsch Lojenga, Constance. 2003. Bila (D 32). In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 450–474. London: Routledge.

Lewis, M.P., Gary F. Simons & Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the world, nineteenth edition*. Dallas: SIL International. http://www.ethnologue.com.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5(1). 1–10.

Maho, Jouni. 1999. *A comparative study of Bantu noun classes*. Göteborg: Acta universitatis gothoburgensia.

Maho, Jouni. 2003. A classification of the Bantu languages: An update of Guthrie's referential system. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 639–651. London: Routledge.

Maho, Jouni. 2009. New updated Guthrie list online. https://brill.com/fileasset/downloads_products/35125_Bantu-New-updated-Guthrie-List.pdf.

Marchese, Lynell. 1988. Nouns classes and agreement systems in Kru: A historical approach. In Michael Barlow & Charles A. Ferguson (eds.), *Agreement in natural language: Approaches, theories, descriptions*, 323–341. Stanford: Center for the Study of Language and Information.

Mayanga, Tayeye. 1985. *Grammaire yansi* (*Rép. du Zaïre*). Zaire: CEEBA, Bandundu.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and STAN* (*2nd edition*). Chapman and Hall/CRC Press.

McMaster, Mary Allen. 1988. *Patterns of interaction: A comparative ethnolinguistic perspective on the Uele region of Zaïre ca. 500 A.D.to 1900 A.D.* PhD dissertation, University of California at Los Angeles.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5. 125–166.

McWhorter, John. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. New York: Oxford University Press.

Meeussen, Achille E. 1967. Bantu grammatical reconstructions. *Africana Linguistica* 3. 79–121.

Meeuwis, Michael. 2013. Lingala. In Susanne Michaelis, Philipe Maurer, Martin Haspelmath & Magnus Huber (eds.), *The survey of pidgin and creole languages*, vol. 3, Contact languages based on languages from Africa, Asia, Australia and the Americas, 25–33. Oxford: Oxford University Press.

Mertens, Joseph. 1935–1939. *Les Ba Dzing de la Kamtsha*. Brussels: Campenhout.

Mfoutou, Jean-Alexis. 2009. *Grammaire et lexique Munukutuba: Congo-Brazzaville, République démocratique du Congo, Angola*. Paris: L'Harmattan.

Motingea Mangulu, Andre. 2001–2002. Situation actuelle des parlers minoritaires au

Nord-Ouest de la République Démocratique du Congo. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* 41. 147–154.

Mufwene, Salikoko S. 1994. Restructuring, feature selection and markedness: From Kimanyanga to Kituba. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 20(supplement). 67–90.

Mufwene, Salikoko S. 1997. Kitúba. In Sarah Grey Thomason (ed.), *Contact languages: A wider perspective* (Creole language library 17), 173–208. Amsterdam & Philadelphia: John Benjamins. doi:10.1075/cll.17.09muf.

Mufwene, Salikoko S. 2006. How Bantu is Kiyansi? In F.K. Erhard Voeltz (ed.), *Studies in African linguistic typology*, 327–335. Amsterdam: John Benjamins. doi:10.1075/tsl.64.18muf.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna. 2003. Diversity and stability in language. In Brian Joseph & Richard Janda (eds.), *The handbook of historical linguistics*, 283–310. Oxford: Blackwell.

Nurse, Derek & Gérard Philippson. 2003. Introduction. In Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 1–12. London: Routledge.

Ollomo Ella, Régis. 2013. *Description linguistique du shiwa, langue bantu du Gabon: Phonologie, morphologie, syntaxe, lexique*. Université de la Sorbonne Nouvelle (Paris 3).

Ouzilleau, F. 1910–1911. Notes sur les langues des Pygmyées de la Sanga: Suivies de vocabulaires. *Revue d'ethnographie et de sociologie* 2. 75–92.

Pacchiarotti, Sara & Koen Bostoen. 2021. Final vowel loss in Upper Kasai Bantu (DRC) as a contact-induced change. *Journal of Language Contact* 14. 438–476. doi: 10.1163/19552629-14020007.

Pacchiarotti, Sara, Natalia Chousou-Polydouri & Koen Bostoen. 2019. Untangling the West-Coastal Bantu mess: Identification, geography and phylogeny of the Bantu B50–80 languages. *Africana Linguistica* 21. 87–162.

Pakendorf, Brigitte, Koen Bostoen & Cesare de Filippo. 2011. Molecular perspectives on the Bantu expansion. *Language Dynamics and Change* 1. 50–88.

Pakendorf, Brigitte, Hilde Gunnink, Bonny Sands & Koen Bostoen. 2017. Prehistoric Bantu-Khoisan language contact: A cross-disciplinary approach. *Language Dynamics and Change* 7. 1–46.

Philippson, Gérard & Rebecca Grollemund. 2019. Classifying Bantu languages. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu languages: Second edition*, 335–354. New York: Routledge Taylor & Francis Group.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/.

Richardson, Irvine. 1957. *Linguistic survey of the Northern Bantu borderland*, vol. 2. Oxford: Oxford University Press.

Samarin, William J. 1991. The Origins of Kituba and Lingala. *Journal of African Languages and Linguistics* 12(1). 47–78. doi:10.1515/jall.1991.12.1.47.

Santandrea, Stefano. 1948. Little known tribes of the Bahr El Ghazal. *Sudan Notes and Records* 29. 78–106.

Santandrea, Stefano. 1963. Short notes on the Bɔdɔ, Huma and Kare languages. *Sudan Notes and Records* 44. 82–99.

Schmidtke-Bode, Karsten. 2019. Attractor States and Diachronic Change in Hawkins's "Processing Typology". In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*, 123–148. Berlin: Language Science Press.

Schmidtke-Bode, Karsten & Natalia Levshina. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In Ilja A. Seržant & Alena Witzlack-Makarevich (eds.), *Diachrony of differential argument marking*, 509–537. Berlin: Language Science Press.

Seidensticker, Dirk, Wannes Hubau, Dirk Verschuren, Cesar Fortes-Lima, Pierre de Maret, Carina M. Schlebusch & Koen Bostoen. 2021. Population collapse in Congo rainforest from 400 CE urges reassessment of the Bantu Expansion. *Science Advances* 7(7). doi:10.1126/sciadv.abd8352.

Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. doi:10.3389/fpsyg.2018.01141.

Smith-Stark, Cedric. 1974. The plurality split. *Chicago Linguistic Society* 10. 657–671.

Stucky, Suzanne U. 1978. How a noun system may be lost: Evidence from Kituba (lingua franca Kikongo). *Studies in the Linguistic Sciences* 8(1). 216–233.

Thomas, John Paul. 1994. Bantu noun-class reflexes in Komo. *Africana Linguistica* 142. 177–195.

Thornell, Christina. 2010. Morphology of plant names in the Mpiemo language. In Karsten Legère, Christina Thornell, Bernd Heine & Whilhelm J.G. Möhlig (eds.), *Bantu languages: Analyses, description and theory*, 249–270. Cologne: Rüdiger Köppe Verlag.

Thornell, Christina. 2012. Simplification of the nominal class system in Central African Bantu Bendo [ɓendɔ]. Paper presented at the 7th World Congress of African Linguistics (WOCAL), Buea, Cameroon, August 20–24, 2012.

Toronzoni, Ngama-Nzombio. 2004. *Esquisse du Bomboma* [*langue Bantu de zone C*]. Porto: Centre de Recherche en Pédagogie Appliquée.

Trudgill, Peter. 1999. Language contact and the function of linguistic gender. *Poznań Studies in Contemporary Linguistics* 35. 133–152.

Urban, Matthias & Steven Moran. 2021. Altitude and the distributional typology of language structure: Ejectives and beyond. *PLOS ONE* 16(2). e0245522. doi:10.1371/journal.pone.024552.

Vansina, Jan. 1966. *Introduction à l'éthnographie du congo* (Éditions Universitaires du Congo). Kinshasa: Université Lovanium.

Vansina, Jan. 1990. *Paths in the rainforests: Toward a history of political tradition in equatorial Africa*. London: University of Wisconsin Press.

Van de Velde, Mark. 2019. Nominal morphology and syntax. In Koen Bostoen & Mark van de Velde (eds.), *The Bantu languages: Second edition*, 237–269. New York: Routledge.

Vihman, Virve-Anneli, Diane Nelson & Simon Kirby. 2018. Animacy distinctions arise from iterated learning. *Open Linguistics* 4(1). 552–565. doi:10.1515/opli-2018-0027.

Vorbichler, Anton. 1963. Zu dem Problem der Klasseneinteilung in Lebendiges und Lebloses in den Pygmäen- und Waldnegerdialekten des Ituri, Congo. In *Festschrift Paul Schebesta zum 75. Geburtstag, gewidmet von Mitbrüdern, Freuden und Schülern* Studia Instituti Anthropos 18, 23–34. St. Augustin: Anthropos-Inst.

Wälchli, Bernhard & Francesca Di Garbo. 2019. The dynamics of gender complexity. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity II: World-wide comparative studies*, 201–364. Berlin: Language Science Press.

Wald, Benji V. 1975. Animate concord in northeast coastal Bantu: Its linguistic and social implications as a case of grammatical convergence. *Studies in African Linguistics* 6. 267–314.

Wayland, E.J. 1929. Notes on the Baamba. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 59. 517–524.

Wega, Simeu Abraham. 2012. *Grammaire descriptive du pɔ́lrɔ̀: Eléments de phonologie, de morphologie et de syntaxe*. PhD dissertation, Université de Yaounde I, Department de langues africaines et linguistique.

Winter, Edward H. 1953. *Bwamba: A structural analysis of a patrilineal society*. PhD dissertation, Harvard University.

de Wit-Hasselaar, Alida. 1995. *Ndaka / Mbo / Beeke survey report*. Bunia: SIL.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC Press 2nd edn.

Appendix A: Coding model for data on gender marking

The coding model used in this paper is the same as the one followed by Di Garbo & Verkerk (2022: Appendix 1). All languages included in the study were coded based on this model.

A.1 *Gender marking on nouns*

– How many singular noun class forms?
– How many plural noun class forms?
– How many number-invariant noun class forms?
– How many singular/plural pairings of noun class forms?

A.2 *Gender marking on agreement targets*

– How many distinguishable singular agreement classes?
– How many distinguishable plural agreement classes?
– How many number-invariant agreement classes?
– How many paired singular/plural agreement classes?

A.3 *What are the word classes that carry syntactic agreement?*

The coding for these variables is a "yes/no/unknown" type of coding. Except for the variable "Other," which is listed at the end, variable names are ordered alphabetically.

– **Attributive adjectives**: adnominal modifiers encoding property words.
– **Copula-like constructions**: constructions expressing nominal and/or locative predications.
– **Demonstrative modifiers**: adnominal modifiers indicating different degrees of spatial distance from the speaker and/or the listener.
– **Demonstrative pronouns**: pronominal expressions indicating different degrees of spatial distance from the speaker and/or the listener.
– **Genitives/connectives**: markers that are used to introduce nominal possessors. In Bantu languages, they generally consist of the stem *a* preceded by a pronominal prefix, which agrees in gender with the possessor. They are also used to encode adjectival types of meanings with modifying nouns encoding properties and/or entities.
– **Independent third person pronouns**: anaphoric pronouns corresponding to 'he/she/it' in English.
– **Numerals**: adnominal modifiers encoding cardinal numbers. In Bantu languages, ordinal numbers also agree in gender, but they are expressed through genitive constructions with cardinal numbers as modifiers (thus gender agreement is marked on the genitive relator rather than on the numeral as such).

– **Quantifiers**: adnominal modifiers encoding quantity expressions such as, for instance, 'some', 'all', 'many'.
– **Possessive pronouns**: pronominal prefixes agreeing in gender with the possessee.
– **Predicative adjectives**: property words used predicatively, i.e., taking TAM inflections.
– **Question words**: selective interrogative such as 'how many?' and 'which?', as well as interrogative pronouns ('who?', 'what?')
– **Reflexive pronouns**: reflexives in Bantu are usually invariable prefixes, which are part of the set of inflectional markers that a verb can take. However, in some cases, there can be reflexive intensifiers, which are independent words that can sometimes take pronominal markers in agreement with the gender of the noun. This is what we target in our coding.
– **Relative pronouns/Relative constructions**: independent pronominal expressions functionally comparable to the English 'who/whom/which/that'. Constructions encoding relative clauses, which do not fall under "relative pronouns."
– **Verbs**: lexemes for the encoding of prototypical predicative expressions (actions, states).
– **Other targets and/or domains of gender marking**: here we include anything that cannot be captured by the features listed above.

A.4 *What are the word classes that carry animacy-based agreement?*

The coding for these variables is a "yes/no/unknown" type of coding. Except for the variable "Other," which is listed at the end, variable names are ordered alphabetically.

– **Attributive adjectives**: adnominal modifiers encoding property words.
– **Copula-like constructions**: constructions expressing nominal and/or locative predications.
– **Demonstrative modifiers**: adnominal modifiers indicating different degrees of spatial distance from the speaker and/or the listener.
– **Demonstrative pronouns**: pronominal expressions indicating different degrees of spatial distance from the speaker and/or the listener.
– **Genitives/connectives**: markers that are used to introduce nominal possessors. In Bantu languages, they generally consist of the stem *a* preceded by a pronominal prefix, which agrees in gender with the possessor. They are also used to encode adjectival types of meanings with modifying nouns encoding properties and/or entities.
– **Independent third person pronouns**: anaphoric pronouns corresponding to 'he/she/it' in English.

- **Numerals:** adnominal modifiers encoding cardinal numbers. In Bantu languages, ordinal numbers also agree in gender, but they are expressed through genitive constructions with cardinal numbers as modifiers (thus gender agreement is marked on the genitive relator rather than on the numeral as such).
- **Quantifiers:** adnominal modifiers encoding quantity expressions such as, for instance, 'some', 'all', 'many'.
- **Possessive pronouns:** pronominal prefixes agreeing in gender with the possessee.
- **Predicative adjectives:** property words used predicatively, i.e., taking TAM inflections.
- **Question words:** selective interrogative such as 'how many?' and 'which?', as well as interrogative pronouns ('who?', 'what?')
- **Reflexive pronouns**: reflexives in Bantu are usually invariable prefixes, which are part of the set of inflectional markers that a verb can take. However, in some cases, there can be reflexive intensifiers, which are independent words that can sometimes take pronominal markers in agreement with the gender of the noun. This is what we target in our coding.
- **Relative pronouns/Relative constructions:** independent pronominal expressions functionally comparable to the English 'who/whom/which/that'. Constructions encoding relative clauses, which do not fall under "relative pronouns."
- **Verbs:** lexemes for the encoding of prototypical predicative expressions (actions, states).
- **Other targets and/or domains of gender marking:** here we include anything that cannot be captured by the features listed above.

A.5 *Additional questions*

- Is animacy-based agreement obligatory outside the NP?
- Is animacy-based agreement obligatory everywhere?
- Does agreement only signal number?
- Do noun class forms only mark number?
- Do noun class forms only mark animacy?
- Do noun class forms mark animacy and number?
- Is there extra marking of animacy on nouns (e.g., animacy markers are juxtaposed to the nominal gender markers)?
- Is there extra marking of plurality on nouns (e.g., in addition to their nominal gender markers, nouns take an additional plural marker which is gender-invariant)?
- Is there extra marking of animacy and number on nouns (e.g., animacy/number markers are juxtaposed to the nominal gender markers)?

– Notes (this is a free text variable where the coder can write in any additional remark on the language which is being described).

Appendix B: The languages of the sample

Languages are classified into types based on the classification introduced in Section 2.3.

Name	Isocode	Glottocode	Guthrie	Type
Akoose	bss	akoo1248	A15	1
Akwa	akw	akwa1248	C22	1
Amba (Uganda)	rwm	amba1263	D22	3
Babango	bbm	baba1263	C441	1
Bafaw-Balong	bwt	bafa1247	A141	1
Bafia	ksf	bafi1243	A53	1
Bafoto	–	bafo1235	C611	1
Bakaka	bqz	baka1273	A15	1
Bakole	kme	bako1250	A231	1
Baloi	biz	balo1261	C31	1
Bamwe	bmg	bamw1238	C412	1
Bangala	bxg	bang1353	C30A	1
Bangi	bni	bang1354	C32	2
Bangubangu	bnx	bang1350	D27	1
Bankon	abb	bank1256	A42	1
Basa (Cameroon)	bas	basa1284	A43a	1
Bassossi	bsi	bass1260	A15	1
Batanga	bnm	bata1285	A32	1
Bebele	beb	bebe1248	A73a	1
Beeke	bkf	beek1238	D335	3
Beembe	beq	beem1239	H11	1
Bekwil	bkw	bekw1242	A85b	1
Bembe	bmb	bemb1255	D54	1
Bera	brf	bera1259	D32	3
Bila	bip	bila1255	D311	3
Bodo (Central African Republic)	boy	bodo1272	D308	3
Boko (Democratic Republic of Congo)	bkp	boko1263	C16	1
Bolia	bli	boli1255	C35	2

(*cont.*)

Name	Isocode	Glottocode	Guthrie	Type
Boloki	bkt	bolo1262	C36e	2
Boma	boh	boma1246	B82	1
Bomboma	bws	bomb1262	C411	2
Bomitaba	zmx	bomi1238	C14	1
Bongili	bui	bong1284	C15	2
Bube	bvb	bube1242	A31	1
Bubi	buw	bubi1250	B305	1
Budu	buu	budu1250	D32	1
Budza	bja	budz1238	C37	1
Bulu (Cameroon)	bum	bulu1251	A74	1
Bushoong	buf	bush1247	C83	1
Buyu	byi	buyu1239	D55	1
Bwa	bww	bwaa1238	C44	2
Bwela	bwl	bwel1238	C42	2
Dengese	dez	deng1250	C81	1
Dibole	bvx	dibo1245	C101	1
Ding	diz	ding1239	B86	2
Duala	dua	dual1243	A24	1
Duma	dma	duma1253	B51	1
Eton (Cameroon)	eto	eton1253	A71	1
Ewondo	ewo	ewon1239	A72	1
Fang (Equatorial Guinea)	fan	fang1246	A75	1
Gyele	gyi	gyel1242	A801	1
Holoholo	hoo	holo1240	D28	1
Homa	hom	homa1239	D304	4
Hungana	hum	hung1278	H42	1
Isu (Fako Division)	szv	isuf1235	A23	1
Kélé	keb	kele1257	B22	1
Kaamba	xku	kaam1238	H112A	2
Kako	kkj	kako1242	A93	3
Kande	kbs	kand1300	B32	1
Kari (Democratic Republic of Congo)	kbj	kari1306	D301	3
Kele (Democratic Republic of Congo)	khy	kele1255	C55	2
Kimbundu	kmb	kimb1241	H21	2
Kituba (Congo)	mkw	kitu1245	H10A	4
Kituba (Democratic Republic of Congo)	ktu	kitu1246	H10A	4

(*cont.*)

Name	Isocode	Glottocode	Guthrie	Type
Kol (Cameroon)	biw	kolc1235	A832	1
Komo (Democratic Republic of Congo)	kmw	komo1260	D23	4
Koongo	kng	koon1244	H14	1
Koonzime	ozm	koon1245	A842	1
Kota (Gabon)	koq	kota1274	B25	1
Koyo	koh	koyo1242	C24	1
Kwakum	kwu	kwak1266	A91	1
Kwasio	nmg	kwas1243	A81	1
Laari	ldi	laar1238	H16f	2
Lefa	lfa	lefa1242	A51	2
Lega-Mwenga	lgm	lega1250	D25	1
Lega-Shabunda	lea	lega1249	D251	1
Lengola	lej	leng1258	D12	2
Libinza	liz	libi1244	C321	2
Ligenza	lgz	lige1238	C414	2
Lika	lik	lika1243	D201	2
Likila	lie	liki1240	C31	1
Likwala	kwc	likw1239	C26	1
Lingala (Bokamba)	lin	ling1263	C30b	3
Lobala	loq	loba1239	C16	1
Lombo	loo	lomb1260	C54	1
Lumbu	lup	lumb1249	B44	1
Lusengo	lse	luse1252	C36	1
Lwel	–	lwel1234	B85	1
Mabaale	mmz	maba1270	C31	1
Mahongwe	mhb	maho1248	B252	1
Makaa	mcp	maka1304	A83	1
Mbala	mdp	mbal1257	H41	1
Mbangwe	zmn	mban1268	B23	1
Mbati	mdn	mbat1248	C13	3
Mbere	mdt	mber1257	B61	1
Mbesa	zms	mbes1238	C51	1
Mbo (Cameroon)	mbo	mboc1235	A15	1
Mboko	mdu	mbok1243	C21	1
Mbole	mdq	mbol1247	D11	1
Mbosi	mdw	mbos1242	C25	1

(*cont.*)

Name	Isocode	Glottocode	Guthrie	Type
Mbule	mlb	mbul1262	A623	1
Mituku	zmq	mitu1240	D13	1
Mmaala	mmu	mmaa1238	A62	1
Moi (Congo)	mow	moic1236	C32	2
Mokpwe	bri	mokp1239	A21	1
Mongo (Atlantic-Congo)	lol	mong1338	C61	2
Mpama	–	mpam1239	C323	1
Mpiemo	mcx	mpie1238	A86c	2
Mpongmpong	mgg	mpon1254	A86	1
Mpuono	zmp	mpuo1241	B84	1
Myene	mye	myen1241	B11	1
Ndambomo	nxo	ndam1254	B204	1
Ndasa sud	nda	ndas1238	B201	1
Ndobo	ndw	ndob1238	C31	1
Ndumu	nmd	ndum1239	B63	2
Ngando (Democratic Republic of Congo)	nxd	ngan1302	C63	2
Ngelima	agh	ngel1238	C45	2
Ngom nord	nra	ngom1270	B22	1
Ngombe (Democratic Republic of Congo)	ngc	ngom1268	C41	2
Ngongo (Democratic Republic of Congo)	noq	ngon1267	H31	1
Ngungwel	ngz	ngun1272	B72a	1
Njebi	nzb	njeb1242	B52	1
Njyem	njy	njye1238	A84	1
Nkongho	nkc	nkon1247	A151	1
Nomaande	lem	noma1260	A46	1
Ntomba	nto	ntom1248	C35	2
Nubaca	baf	nuba1241	A621	1
Nugunu (Cameroon)	yas	nugu1242	A622	1
Nyali	nlj	nyal1250	D33	2
Nyanga	nyj	nyan1304	D43	1
Nyokon	nvo	nyok1243	A45	1
Nzadi	–	nzad1234	B85	3
Ombamba	mbm	omba1241	B62	1
Ombo	oml	ombo1238	C76	1
Oroko	bdu	orok1266	A101	1
Pagibete	pae	pagi1243	C401	2

(*cont.*)

Name	Isocode	Glottocode	Guthrie	Type
Pande	bkj	pand1264	C12	3
Pinji	pic	pinj1243	B304	1
Polri	pmm	pomo1271	A92	4
Punu	puu	punu1239	B43	1
Sakata	skt	saka1287	C34	2
Sake	sak	sake1247	B251	1
San Salvador Kongo	kwy	sans1272	H16a	2
Sangu (Gabon)	snq	sang1333	B42	1
Seki	syi	seki1238	B21	1
Sengele	szg	seng1278	C33	2
Shiwa	–	shiw1234	A803	NA
Sighu	sxe	sigh1238	B202	1
Simba	sbw	simb1254	B302	1
Sira	swj	sira1266	B41	2
So (Cameroon)	sox	soca1235	A82	1
So (Democratic Republic of Congo)	soc	sode1235	C52	1
Songomeno	soe	song1305	C82	2
Songoora	sod	song1300	D24	1
Suku	sub	suku1259	H32	2
Suundi	sdj	suun1239	H131	2
Teke-Ebo	ebo	teke1278	B74b	1
Teke-Fuumu	ifm	teke1274	B77b	1
Teke-Tege	teg	teke1275	B71	1
Tembo (Motembo)	tmv	temb1272	C37	2
Tetela	tll	tete1250	C71	1
Tibea	ngy	tibe1274	A54	1
Tiene	tii	tien1242	B81	1
Tsaangi	tsa	tsaa1242	B53	1
Tsogo	tsv	tsog1243	B31	1
Tuki	bag	tuki1240	A601	2
Tunen	tvu	tune1261	A44	1
Ukhwejo	ukh	ukhw1241	A802	2
Vili	vif	vili1238	H12	2
Viya	gev	eviy1235	B301	1
Wumbvu	wum	wumb1242	B24	1
Yaka (Congo)	iyx	yaka1274	B73	1

(*cont.*)

Name	Isocode	Glottocode	Guthrie	Type
Yaka (Democratic Republic of Congo)	yaf	yaka1269	H31	2
Yansi	yns	yans1239	B85	4
Yasa	yko	yasa1242	A33a	1
Yela	yel	yela1238	C74	1
Yombe	yom	yomb1244	H16c	1
Zamba	–	zamb1245	C16	2
Zimba	zmb	zimb1251	D26	1

Appendix C: The sociolinguistic typology of languages with radically restructured gender systems

ISO	Glottocode	Name	Group	History
brf	bera1259	Bera	Komo-Bila	Homeland: Nile basin. Settlement in current location around 1800. Possibly the first Bantu arriving in this area
rwm	amba1263	Amba (Uganda)	Komo-Bila	Origin: probably the Bira community
bip	bila1255	Bila	Komo-Bila	no info
kmw	komo1260	Komo (DRC)	Komo-Bila	Homeland: Nile basin. At current place of settlement, fighting against local populations first and Arabic settlers later
hom	homa1239	Homa	Kari	No known L1 speakers. Last speaker died by 1975. The Huma used to live round about Mt. Bangenze from where they fled to the hills neighboring Tombora
boy	bodo1272	Bodo (Central African Republic)	Kari	Recently, from Mafaya (a locality near the sources of the Kuru River, within the former Bahr el Ghazal province), to the present location, Deim Zubeir
kbj	kari1306	Kari (DRC)	Kari	Living at their current location for a long time, but not the first/original inhabitants. Being assimilated into the Azande society

Contact	Borrowed lexical material	Other features	Sources
Pygmies; Hima (Nkore-Kiga, East Bantu); Lese (Central Sudanic); Ngwana (?); Sudanese in general	no info	no info	van Geluwe (1956)
Pygmies; Toro (Niger-Congo, Plateau) and Nyoro (East Bantu)	no info	no info	Joset (1952); Wayland (1929); Winter (1953)
Pygmies; other languages, unclear which	no info	Nine-vowel system with ATR-harmony, something also found in several Central-Sudanic languages	Harvey (1997); Kutsch Lojenga (2003)
Exogamy was very rigorous in the past but on the way to being less strictly respected. Unclear with which languages there was contact with (Vansina 1990: 66–67 mentions that contact with expanding Central Sudanic populations had dramatic consequences on the language)	A large number of lexical items from Lingala. Borrowing in the nominal domain along with semi-productive nominal prefixes; sources hypothesize that Komo has undergone extensive creolization	Apart from the heavily eroded noun class system, Komo still retains typical features of Bantu grammar	Thomas (1994); van Geluwe (1956)
no info	Some Zande loans	no info	Eberhard et al. (2019); Santandrea (1948, 1963)
Supposedly, the community ended up in Mafaya during early Bantu migrations and underwent influences from Sudanic there. Source mentions that informations are more conversant in Zande than in their native tongue	Some, from Kresh (Kresh-Aja); Ndogo (Ubangi); Zande (Ubangi)	Some constructions match those found in Zande, such as male/female gender of animals, treatment of attributes, demonstrative acting as a neuter pronoun	Santandrea (1948, 1963)
Certainly with Zande (Ubangi) speakers, previously probably with other communities	Depending on source, either majority of 60 basic vocabulary items is not Bantu (Dijkmans 1936), or clearly a Bantu language (Santandrea 1948). Quinary numerals reminiscent of Bwa (Bantu, Ababuan). Some Zande (Ubangi) loans	no info	Dijkmans (1936); Santandrea (1948, 1963)

(*cont.*)

ISO	Glottocode	Name	Group	History
bkf	beek1238	Beeke	Bali-Beeke	At current location since about 1960. Earlier history unknown, except that origin is the Ituri river. Various sources give irreconcilable accounts of closest genealogical affiliations
bkj	pand1264	Pande	Pande-Mbati	Supposedly, the Pande migrated into the Sanga basin folowing an east-to-west trajectory, coming from the edges of the Ubangi
mdn	mbat1248	Mbati	Pande-Mbati	The Mbati are said to have crossed the Ubangi as invaders and settled in their present location between 1750 and 1850 (Bouquiaux & Thomas 1994) or, alternatively, less than a century ago (Richardson 1957)
pmm	pomo1271	Polri	Polri-Kako	Different accounts; they agree that the Pol come from the east. One source indicates this migration took place in the 20th century
kkj	kako1242	Kako	Polri-Kako	Homeland: somewhere in the Batouri and Mambere valleys (Central African Republic). Migration waves supposedly started off from three distinguished macro-groups whose dispersal followed the rivers. The current place of settlement is a convergence area for migration waves coming from different directions; these created numerous conflicts in the area

Contact	Borrowed lexical material	Other features	Sources
(Sua-Mbuti) Pygmies; Ndaka (Bantu, Ababuan)	Yes, from Ndaka	no info	de Wit-Hasselaar (1995); Vorbichler (1963); van Geluwe (1960: 11)
The indigenous populations of Bakota, Bagandu, the Mandinga of Mbaere, and Bollemba speak pretty much the same language variety, which the Pande call Lindzali/Linzeli. This common linguistic identity, paired with some shared traditions and beliefs, indicates common historical origins	Yes, from Yangere (Ubangi)	no info	Bruel (1910–1911); Ouzilleau (1910–1911); Richardson (1957)
Yes, with Ngando and Bakota (Bantu), as well as with Ngbaka and Monzombo (Ubangi). This contact goes back at least three centuries.	Yes, from Ali (Gbaya, Niger-Congo); Sango (Ubangi); general Ubangi	There are only traces left of the connective, the tonal system is very unusual	Bouquiaux & Thomas (1994); Richardson (1957)
Yes, both with Bantu (Makaa, Njem, Mpiemo) and non-Bantu languages (Gbaya-Mandja group (Niger-Congo), Mbum (Ubangian), Mbonga (Jarawan)). Also with Baka Pygmies	The vocabulary is largely Bantu, with affinities with the Makaa-Njem complex in particular	Wega (2012) describes Pomo as a transition language, with features that are clearly Bantu and other features that are clearly non-Bantu	Bruel (1910–1911); Wega (2012); Burnham et al. (1986); Jacquot & Richardson (1956)
Coming into east Cameroon, the Kako (and the Gbaya) came into contact with the Mpiemo, Maka, Pol, Kwakum (all Bantu), Mbum (Ubangian), Mbonga (Jarawan), and the Baka Pygmies. Contacts with Muslim communities from 1840 onwards. Genealogical relationships with neighboring languages unclear	The vocabulary is largely Bantu, with strong affinities with the Makaa-Njem cluster. Many Gbaya (Niger-Congo) words are found in Kako. Borrowing from Pol/Pomo, Kwakum (both Bantu)	no info	Burnham et al. (1986); Guthrie & Tucker (1956)

(*cont.*)

ISO	Glottocode	Name	Group	History
yns	yans1239	Yansi	Yansi-Nzadi	Mertens (1935–1939) claims that the Yansi group came into the forest from the savanna in the south.
No ISO code	nzad1234	Nzadi	Yansi-Nzadi	Oral tradition regarding settlement reports on three prominent facts: fights with Sengele Pygmies, settlement at Kwamuntu village, migration across the Kasai River
lin	ling1263	Lingala (Kinshasa)	creole	Kinshasa Lingala is the descendent of the Bangala pidgin, originally spoken in the Bangala state post and spread northeast later on. This variety escaped the standardization imposed by the Scheutists, which later gave rise to Makanza Lingala. Bangala is the descendant of Bobangi, a riverine trade language spoken in the western part of the Congo River. The Europeans started using Bobangi as a medium of communication, which led to pidginization and fostered substantial influences from European and West African languages. This variety was later imposed as the language of communication at the Bangala post.

Contact	Borrowed lexical material	Other features	Sources
Two possible scenarios: (1) Relative isolation in the transition zone between the equatorial rainforest and southern savanna (Vansina 1966). (2) Substrate influence from non-Bantu languages spoken by autochthonous hunter-gatherer groups, which came in contact with the Bantu in the vicinity of their homelands, after break up of Proto-West-Coastal Bantu (Bostoen & Gunnink forthcoming 2022)	no info	Atypical features in the domain of phonology, morphology, and syntax	Vansina (1966); Bostoen & Gunnink (forthcoming 2022)
B80 languages are not well studied. They have been in contact with each other as well as with other Bantu languages from which they borrowed words. Areal convergence in other domains as well. Borrowing from contact Bantu languages is mentioned but unclear which	no info	Out of the B80 languages, Nzadi is the language that has undergone most reduction: shortened words, loss in the domain of derivational morphology, largely isolating syntax. A simplified Bantu language rather than a language that has developed West-African-Benue-Congo features	Crane et al. (2011)
Definitely a contact variety	European and West African influences	no info	Bokamba (2009); Meeuwis (2013)

(*cont.*)

ISO	Glottocode	Name	Group	History
ktu	kitu1246	Kituba (DRC)	creole	Three accounts: Fehderau (1966) posits that Kituba descends from a Kikongo Pidgin, developed in the area of the Manianga market before the arrival of the European traders, and became the trade language of the Lower Congo region. The process can be traced back to the 15th century. The Belgians widely promoted its use. The creolization process started around 1940, especially in urban areas. Samarin (1991) argues that the earliest evidence of the ancestor of Kituba goes back to 1905, there is no evidence of a Kikongo Pidgin before the 1890s. A third explanation is that Kituba originated through contact between the Bakongo and foreign (West African?) workers.
mkw	kitu1245	Kituba (Congo)	creole	Same as for the DRC variety

Contact	Borrowed lexical material	Other features	Sources
Kituba came into use as a lingua franca and is still largely used as a lingua franca today. The relatively small number of L1 speakers suggests that Kituba is a pidgin which has only recently undergone creolization. Contact between Lower and Upper Congo river people prompted Lingala influences into Kituba. Contact with Lingala deeply influenced Kituba, both lexically and grammatically	Portuguese via Kikongo; Lingala; French (especially starting from the creolization phase)	no info	Fehderau (1966); Samarin (1991)
Same as for the DRC variety	Same as for the DRC variety	no info	Fehderau (1966); Samarin (1991)