"What is Your Primary Language?": Spatial Considerations of Primary Language Identification in a Multilingual Rural Region

Abstract

In a multilingual environment, individuals routinely use three or more languages and may identify a primary language among them. The identification of this language has profound implications in multilingual societies yet has not received adequate attention in the scholarly literature, let alone from a spatial perspective.

In this study, we aim to evaluate the impacts of place on primary language identification in a rural region of Cameroon. The places examined include the residence of multilingual individuals and the birthplace of their family members. The specific objectives are twofold: (1) we predict the primary language identified by a multilingual individual based on these two types of place and (2) we then evaluate the impact of these places and associated sociolinguistic characteristics of the individuals on the identification. To support these research goals, we leverage spatial-sociolinguistic data in the region and a random forest classification model.

Results show that the two types of place and an individual's sociolinguistic characteristics predict primary languages to a high level of accuracy. Among the factors, the two types of place contribute the most—across multiple perspectives—to the identification of primary languages.

This work contributes to and broadens the current discussion of the connection between space and language. The knowledge gained is valuable for understanding and maintaining the linguistic ecology in small-scale rural societies where language endangerment is a prominent concern.

key words: primary language, place, space-language connection, multilingualism, small-scale rural society.

Introduction

In a multilingual environment, individuals routinely use three or more languages, a phenomenon referred to as multilingualism (Kemp 2009). Multilingual individuals will typically identify one of the languages as their *primary languag*e, and this will often be the one that is most frequently used. The identification of primary language is a reflection of how individuals perceive their linguistic identity, community belonging, and social capital. It has multifaceted linguistic, social, and cultural implications, especially in small-scale rural multilingual societies, although these implications are significantly underexplored (S. Bordia and P. Bordia 2015; Singer and Harris 2016; Chernela 2018; Di Carlo 2018). Although empirically observable, the concept of primary language has not received adequate attention in the scholarly literature.

The identification of a primary language is closely related to multilingual acquisition because the primary language is part of individuals' multilingual profiles (Dressler 2014; Ellis 2016). The profile here refers to the composition of linguistic repertoires, i.e., the set of languages that a multilingual individual knows. There has been active discussion of multilingual acquisition in the linguistic literature, including characteristics of multilingual profiles and considerations behind language choice (Hua and Wei 2005; Paradis 2007; Canagarajah and Wurr 2011; Lüpke and Storch 2013; Paradowski and Bator 2018). Primary languages have been an integral part of the discussion, but they have never been a focus, and the discussion as a whole is rarely from a spatial perspective.

Spatial considerations are as essential as sociolinguistic considerations to multilingual acquisition but are often seen in linguistic studies as a backdrop within which multilingual acquisition takes place (Whiteley 1974; Van der Merwe 1993; Blommaert, Collins, and Slembrouck 2005; Veselinova and Booza 2009; Di Carlo and Pizziolo 2012). The expression of the backdrop tends to be simplified (Ambrose and Williams 1991; Breton 1993; Williams 1996; Di Carlo 2022). For example, a multilingual region is often presented as a homogenous area on a map, as found, for instance, in standard linguistic reference sources such as Ethnologue (Eberhard, Simons, and Fennig 2023).

More recent studies have substantially deepened the role space plays in multilingualism, such as how multiple languages interact in intricate and dynamic spatial processes (Prochazka and Vogl 2017; Ranacher, Van Gijn, and Derungs 2017; Hiippala et al. 2020; Paul 2020; Ranacher et al. 2021; Väisänen et al. 2022). These studies showcase the rich context that

spatial considerations can bring to multilingualism and put space at the forefront of multilingualism research.

Various factors have been found to drive multilingual acquisition, including social, cultural, historical, political, economic, religious, and environmental considerations (Hua and Wei 2005; Paul 2020; Canagarajah and Wurr 2011; Nicholls, Eadie, and Reilly 2011; Westergaard 2021). Among them, community influence and kinship influence are two widely acknowledged considerations (Grenier 1984; Chiswick, Lee, and Miller 2005; Kharkhurin 2008; Paradis 2007; Skutnabb-Kangas et al. eds 2009; Chevalier 2012; Paradowski and Bator 2018).

Multilingual acquisition is essentially a spatial experience fused with social, cultural, linguistic, and environmental meanings. Two types of place, the residence of multilingual individuals and the birthplace of their family members, are central to the influences of community and kinship, respectively. In the community of residence where multilingual individuals live, the sociolinguistic environment affects their multilingual profile (Gumperz 1964). Their "residence languages" are typically acquired through daily activities with their friends, neighbors, and other people with whom individuals interact. The influence of residence place often leads to a high level of acquisition and frequent use of the acquired languages (Hoffmann and Ytsma eds 2004; Hua and Wei 2005; Paradis 2007).

On the other hand, languages acquired by family members at their birthplace also play an irreplaceable role (Barron-Hauwaert 2003; Braun and Cline 2010). Close family members, such as parents, will have acquired "birthplace languages" early in their life that will often be

among their most used languages (These are the "residence languages" of family members in their early life). Multilingual individuals acquire family members' birthplace languages primarily at home. The influences from parents are large in quantity and high in quality and frequency (Barron-Hauwaert 2003; Braun and Cline 2010; Unsworth 2016; Diskin 2020). Spouses, also important family members, affect individuals' multilingual profiles as well, although only later in their life (Chiswick, Lee, and Miller 2005).

Although the importance of place is clear in conceptual terms, it is unclear how place affects how multilingual individuals identify their primary languages and, specifically, what role the two types of place play in the choice. Answering these questions is a key first step in understanding the spatial roots of primary language identification, among many underexplored spatial considerations and implications.

Primary language identification is of special interest in the context of "small-scale multilingualism," an understudied phenomenon widely observed across continents (François 2012; Lüpke 2016; Singer and Harris 2016; Cobbinah et al. 2017; Chernela 2018; Di Carlo, Good, and Ojong Diba 2019). The term refers to observed patterns of multilingualism in small-scale societies, where individuals primarily interact with other individuals whom they know (Reyes-Garcia et al. 2017). Compared to many bilingual (or multilingual) studies that are conducted in urban areas and in the context of immigration (Henrich, Heine, and Norenzayan 2010), the linguistic landscape in small-scale rural societies has distinct characteristics.

Individuals in these societies routinely use a large number of local languages, and their multilingual profiles differ from each other because they typically emerge out of their own

distinctive sociolinguistic trajectories. A key factor that contributes to this diversity is the fact that individuals consider the local languages to be more or less equal in status, with none of them being more socially prestigious or having any overall higher value than any other (Lüpke 2016). Furthermore, intermarriage is common among the communities, and it is a major mechanism for a local language to be used in multiple communities, even those where another language predominates. These characteristics enrich the diversity of both residence languages and birthplace languages and add complexity to the identification of primary language (Lüpke and Storch 2013; Esene Agwara 2020).

To fill the research gaps discussed above, we aim to evaluate the impacts of place on primary language identification in a rural region of Cameroon, whose societies have been characterized by small-scale multilingualism. The objectives of this study are twofold: (1) to predict the primary language of multilingual individuals based on the two types of place and their associated sociolinguistic characteristics and (2) to evaluate the impact of the places and characteristics. To support these research goals, spatial-sociolinguistic data of multilingual individuals in the region are analyzed using a random forest classification model.

Results of this work provide a baseline understanding of primary languages, the spatial-linguistic roots behind their identification, and interactions between language and space. The examination of linguistic identity at the individual level, in particular, advances the systematic exploration of this domain since previous work has examined it primarily at the community level (Sutton 1997; Slotta 2015). These efforts may stimulate further discussion of the space-language connection in the subdiscipline of linguistic geography. In addition, this study can

inform policies in language preservation for many small-scale multilingual societies whose endangered linguistic ecologies are an immediate and prominent concern (Pakendorf, Dobrushina, and Khanina 2021).

Study Area

The study area, Lower Fungom, is a rural region of Cameroon (Figure 1). Geographically, Lower Fungom is part of Sub-Saharan Africa, covering approximately 240 square kilometers in a hilly area with rugged terrain (Di Carlo and Pizziolo 2012; Di Carlo and Good 2014). The region is dominated by a tropical monsoon climate, forest-savanna mosaic vegetation, and well-distributed rivers and streams.

[Figure 1. The study area of Lower Fungom.]

There are thirteen village-chiefdoms in the region — i.e., polities headed by a chief and which were traditionally politically independent. No official census for Lower Fungom is available. The village population size is estimated by field observations and other sources (Appendix A). Lower Fungom has long been known to be culturally distinctive and characterized by both an exceptional level of linguistic diversity across the region and a high degree of individual-level multilingualism (Warnier 1980).

There are two different characterizations of the region's linguistic diversity. The scholarly linguistic classification categorizes the speech varieties of the region's thirteen villages into eight distinct languages. It also recognizes that, even when two villages are classified as

speaking the same language, each village is associated with a distinctive variety (i.e., its own dialect). Many of the languages and varieties are endangered (Good et al. 2011; Di Carlo 2018).

By contrast, local residents treat each of the thirteen village-chiefdoms of the region as speaking its own "language," summing to thirteen local languages. Language names are often given that are the same as the village names among residents. We adopt the perception of local residents in this study as it is directly relevant to our goal to assess the impact of place on primary language and because it also reflects how residents understand their own multilingual profiles. Also used in Lower Fungom are two additional varieties associated with the village of Mekaf and its associated hamlet of Small Mekaf, both near Lower Fungom in the neighboring region. These 15 languages are analyzed as the primary languages in this study.

The highly multilingual individuals of this region use five languages on average, with the extreme being nineteen languages (Table 1). The languages include the local residence languages, birthplace languages, and local languages from other villages inside or outside Lower Fungom (Di Carlo and Pizziolo 2012; Esene Agwara 2013; Di Carlo 2015).

Table 1. Languages used in Lower Fungom villages according to local residents' perceptions. The village's own local language is listed first for each village.

Village	**Languages used by residents**
Abar	Abar, Aghem, Ajumbu, Biya, Buu, Fang, Fungom, Koshin, Kung, Missong, Mmen, Mufu, Mundabli, Munken, Naki-Mashi, Naki-Mekaf, Ngun, Weh
Ajumbu	Ajumbu, Fungom, Kung, Missong, Mmen, Naki-Mekaf
Biya	Biya, Abar, Naki-Mekaf, Ngun

Buu	Buu, Abar, Biya, Fang, Koshin, Kung, Mbuk, Missong, Mufu, Mundabli, Munggaka, Munken, Ngun
Fang	Fang, Abar, Ajumbu, Bum, Buu, Kom, Koshin, Kung, Munggaka
Koshin	Koshin, Abar, Fang, Naki-Mashi, Naki-Mekaf, Small Mekaf
Kung	Kung, Fang, Naki-Mekaf, Naki-Nkang, Small Mekaf
Mashi	Naki-Mashi, Koshin, Kung, Naki-Mekaf, Small Mekaf
Missong	Missong, Abar, Adjume, Aghem, Ajumbu, Biya, Bororo, Bum, Buu, Dumbu, Fang, Fungom, Hausa, Isu, Jukun, Kom, Koshin, Kung, Mmen, Modele, Mufu, Mundabli, Munggaka, Munken, Naki-Mashi, Naki-Mekaf, Ncane, Ngun, Weh
Mufu	Mufu, Abar, Buu, Missong, Mundabli, Munggaka, Naki-Mashi
Mundabli	Mundabli, Abar, Buu, Fulani, Missong, Mufu, Naki-Mashi
Munken	Munken, Abar, Aghem, Bafut, Bali, Bambili, Bambui, Biya, Bum, Buu, Fungom, Isu, Jukun, Kom, Kung, Lamnso', Mankon, Mendakwe, Missong, Mmen, Mufu, Mundabli, Munggaka, Naki-Mashi, Ngun, Nkambe, Nkwen, Oku, Weh
Ngun	Ngun, Abar, Biya, Naki-Mekaf

Data

Data description

The data used in this study are obtained from a sociolinguistic survey conducted in Lower Fungom between 2012 and 2015 (the survey was approved by the Internal Review Board of the institutions that conducted the survey). More than 200 residents across the 13 villages participated in the survey. Among them, 176 residents have complete information for this study and are included in the analysis.

We extract two sets of spatial-sociolinguistic data from the survey with respect to the 176 residents and their family members, respectively (Table 2). To distinguish the two datasets,

the residents are referred to as “multilingual individuals” and their family members as “family members of multilingual individuals” in the subsequent discussion.

The multilingual individual dataset includes their village of residence, the primary language identified by them, and their gender. The number of speakers of the 15 primary languages among the multilingual individuals is listed in Appendix B. The family member dataset includes their village of birth, the languages they use, and their gender. The birth villages include the 13 villages inside Lower Fungom for family members born in the region and additional villages for those born outside the region. The family member dataset is divided into paternal, maternal, and spousal sub-datasets to explore whether the father, mother, and spouse have different impacts on primary language identification.

Of the datasets, the village of residence of multilingual individuals and the village of birth of their family members are used to represent the two types of place discussed above. Languages used by family members are used to represent the impact of diverse languages. Gender is used to represent the impact of spouses, given the common intermarriages among communities. The primary languages identified by the multilingual individuals are used to validate the prediction of their primary languages.

Table 2. The datasets of multilingual individuals and their family members.

Set	Data	Description
Multilingual individual dataset	Residence	The village of residenc of a multilingual individual

		Primary language	The primary language of a multilingual individual
		Gender	The gender of a multilingual individual
Family member dataset	Paternal dataset	Paternal birthplace	Village of birth of an individual's father
		Paternal languages	All languages used by an individual's father
	Maternal dataset	Maternal birthplace	Village of birth an individual's mother
		Maternal languages	All languages used by an individual's mother
	Spousal dataset	Spousal birthplace	Village of birth an individual's spouse
		Spousal languages	All languages used by an individual's spouse
		Spousal gender	The gender of an individual's spouse

Data preprocessing

To prepare for the random forest analysis, we derive two sets of features from the two datasets (Table 3). One set of features corresponds to the multilingual individual dataset. The village of residence of the multilingual individuals is expressed as a nominal feature, and the gender of the multilingual individuals is expressed as a binary feature where "1" indicates male and "0" indicates female. The second set of features corresponds to the family member dataset. The birthplace of the family member is expressed as a nominal feature. The languages

they use are expressed as a series of binary features, one feature for each language used by each family member, where "1" indicates the use of the language and "0" otherwise. The gender of the family members is expressed as a binary feature where "1" indicates male and "0" indicates female. The features in this set are divided into paternal, maternal, and spousal subsets. All features used in this study, totaling 124, are listed in Appendix C.

We enter the two sets of features into the random forest model to predict the primary languages. We then use the primary languages identified by the multilingual individuals to validate the prediction. Lastly, the contributions of spatial-linguistic features are evaluated.

Table 3. Features used in the random forest model.

Feature Group		**Feature Name**	**Number of Features**	**Format**
Multilingual individual features		Residence	1	Nominal
		Gender	1	Binary
Family member features	Paternal features	Paternal birthplace	1	Nominal
		Paternal languages	48	Binary
	Maternal features	Maternal birthplace	1	Nominal
		Maternal languages	40	Binary
	Spousal features	Spousal birthplace	1	Nominal
		Spousal languages	30	Binary
		Spousal gender	1	Binary
Total			124	

Methodology

Random forest classification model

The random forest model is a machine learning method to predict observations into classes (Breiman 2001). The model constructs an ensemble of decision trees to predict the class based on features, where the features represent the driving principles governing the classes. The prediction process includes a training phase and a testing phase. The training phase captures the relationship between the features and the classes, while the testing phase predicts the class for observations using the relationships obtained during the training phase. During training, each tree uses a series of randomly selected features to predict the class. The process iterates until the results converge when the predicted classes closely approximate the observed classes. The majority vote from the entire ensemble of tree predictions is used as the final prediction (Breiman et al. 1984; Breiman 2001; Bishop 2006; Gislason, Benediktsson, and Sveinsson 2006).

The model is deemed advantageous for this study due to a number of considerations. First, the model is well known for its tolerance of small sample sizes (Qi 2012; Luan et al. 2020), which suits this study given the limited amount of data available on primary languages. In addition, the model can process a large number of features, which is the case for this study because the diverse languages used by the family members result in a large number of features (Table 3). Further, the model identifies the contribution of each of the features and helps differentiate their role, especially in cases where there is a large number of them (Izmirlian 2004; Tagliamonte and Baayen 2012; Fassnacht et al. 2014; Jiao et al. 2020).

Model implementation

We implement the model with a number of critical parameters. These include the division of training set vs. testing set, the number of decision trees, and the maximum number of features for a tree. First, for the division between the training and testing sets, we adopt the common 80% versus 20% rule (Breiman 2001; Gholamy, Kreinovich, and Kosheleva 2018). For both sets, data are stratified according to the number of speakers of the 15 primary languages (Appendix B).

Second, for the number of decision trees, we select 45 based on two rounds of experiments designed to maximize the prediction accuracy while minimizing underfitting and overfitting. The first round of experiments uses the number of trees ranging from 10 to 100, with an increment of 10. The minimum value of 10 is used to accommodate the small size of the dataset, and the maximum value of 100 is used based on the typical choice discussed in the literature (Latief et al. 2019; Wibowo et al. 2020). The model performs best when the number of trees is 40 and 50. Subsequently, the second round of experiments focuses on a narrower range between 40 and 50 trees with a refined increment of 1. The best result is observed with a setting of 45 trees.

For the third parameter, the maximum number of features, we set the number as 11 after experimenting with three options recommended in the literature (Kyriakides and Margaritis 2019). The first option uses the total number of features, which is 124 features in this case, as the maximum for a tree. The second option uses the logarithm base 2 of the total number of

features for a tree, which is seven in this case. The last option uses the square root of the total number of features, which is 11 features in this case, and this option delivers the best results. The other two options either reduce the intended randomness of the ensemble design (the first option) or render a number too small to support the prediction (the second option). The best model performance during the experiments and selection process is confirmed using leave-one-out cross validation.

Model evaluation

We assess prediction accuracy using a confusion matrix and F1 score. The confusion matrix reports the overall accuracy, commission errors, and omission errors of prediction. The overall accuracy is expressed as the percentage of correctly predicted primary language classes over all observed language classes. The commission error is the percentage of observations committed to incorrect language classes. The omission error is the percentage of predictions omitted from correct language classes. Both errors range between 0 and 1, and a lower value indicates a better prediction. The F1 score integrates precision and recall, expressed as the harmonic mean of the two (Powers 2011). The precision accounts for the number of correct predictions among all predictions, while the recall accounts for the number of observations correctly predicted among all observations. The F1 score ranges from 0 to 1, with 1 meaning a perfect prediction and 0 an ineffective prediction. We use both the confusion matrix and F1 score as they evaluate the model performance from different angles.

To evaluate the impact of features on primary language identification, the model quantifies their contributions by the Gini importance index embedded in the random forest model. For a given feature, its contribution is expressed as the decrease in Gini impurity if the feature is removed from the model (Strobl, Malley, and Tutz 2009). The values of all feature contributions sum to 1 (100 percent), and the features are ranked according to their contribution values. In this study, given the total of 124 features, the average value of feature contribution is 0.008, which serves as a relative reference point when evaluating the contribution of the 124 features.

To compare the results of the random forest model with a statistically oriented analytic, we use a multinomial logistic regression analytic to predict the primary languages. The regression evaluates the contribution of features through a statistical relationship between them and the primary languages. The 15 classes of primary languages are the dependent variable for the regression, and the 124 features serve as independent variables. The results of multinomial logistic regression are also assessed using a confusion matrix and F1 score.

Results and Discussion

Primary languages can be predicted

For our first research objective, predicting primary languages, the overall accuracy achieved by the random forest model is 88.9%, and the F1 score is 0.894. These results provide evidence that the primary language can be predicted at the individual level with satisfactory accuracy, even though the random forest model has only rarely been used in sociolinguistic research. The impact of two types of place and their associated sociolinguistic characteristics are able to

capture the essence of primary language identification within a linguistically complex environment.

The commission errors and omission errors are associated with different primary language classes (Table 4A, Appendix D, Table D.1). Among the cases of misclassification of commission, Biya as a primary language is misclassified as Munken, which is the predominant local language of the nearest village to Biya. Residents of Biya interact with those in nearby Munken more frequently than those in other villages. Furthermore, the intermarriage rate between Biya and Munken is relatively high compared to the others. The spatial and social overlap of Biya and Munken speakers explains the misclassifications. Other misclassifications of commission are due to similar reasons. For the omission errors, the misclassifications are attributed to two situations. The majority of them are due to spatial and social overlap between pairs of nearby villages. The remaining omission error involves Mekaf, a village outside Lower Fungom. The use of Mekaf is scattered in the region due primarily to intermarriage, making this primary language less predictable.

The interactions between places contribute to the identification of primary language. Such interaction is important in its own right as it extends the impact of place beyond the place itself. This importance is seldom discussed in multilingualism studies and the general linguistic literature.

Table 4. Commission error and omission error of the confusion matrix, and precision and recall of the F1 score for primary language prediction by (A) the random forest model where the overall accuracy is 89%

Primary language	Commission error	Omission error	Precision	Recall
Abar	0.0	0.2	1.0	0.8
Ajumbu	0.0	0.0	1.0	1.0
Biya	0.5	0.0	0.5	1.0
Buu	0.5	0.0	0.5	1.0
Fang	0.0	0.0	1.0	1.0
Koshin	0.0	0.0	1.0	1.0
Kung	0.33	0.0	0.67	1.0
Missong	0.0	0.0	1.0	1.0
Mufu	0.0	0.5	1.0	0.5
Mundabli	0.17	0.0	0.83	1.0
Munken	0.0	0.25	1.0	0.75
Mashi	0.0	0.0	1.0	1.0
Mekaf	0.0	1.0	1.0	0.0
Ngun	0.0	0.0	0.0	0.0
Small Mekaf	0.0	0.0	1.0	1.0

(B) the multinomial regression model where the overall accuracy is 78%

Primary language	Commission error	Omission error	Precision	Recall
Abar	0.25	0.4	0.75	0.6
Ajumbu	0.0	0.0	1.0	1.0
Biya	1.0	1.0	0.0	0.0

Buu	0.0	1.0	1.0	0.0
Fang	0.0	0.0	1.0	1.0
Koshin	0.0	0.0	1.0	1.0
Kung	0.0	0.0	1.0	1.0
Missong	0.2	0.0	0.8	1.0
Mufu	0.5	0.5	0.5	0.5
Mundabli	0.2	0.33	0.8	0.67
Munken	0.25	0.25	0.75	0.75
Mashi	0.0	0.0	1.0	1.0
Mekaf	0.0	0.0	1.0	1.0
Ngun	0.0	0.0	0.0	0.0
SmallMekaf	0.0	0.0	1.0	1.0

Results of the multinomial logistic regression model achieve an overall accuracy of 77.8% and an F1 score of 0.789. The regression predicts most primary languages correctly yet yields twice as many misclassifications as the random forest model (Table 4B, Appendix D, Table D.2). The ensemble design of the random forest makes full use of all information in the data, especially advantageous for datasets of limited observations. Moreover, its use of majority voting prevents undue effects from possible outliers (Huang et al. 2022). For the logistic regression, the large number of classes of dependent variables and the vast number of

independent variables compared to the number of observations involved may affect the prediction (Augustin, Cummins, and French 2001). Because the random forest model achieves better results than the logistic regression model, we focus on the results of the random forest model in the rest of the paper.

Place matters

Regarding our second objective, namely to evaluate the impact of places and associated characteristics on primary language, the contribution of the 124 features is reported in Appendix C. Three groups of features emerge based on their contribution values when referencing the average (0.008) and natural breaks in the values (Table 5). Group 1 includes the top four ranked features whose contribution values are considerably higher than the average. Group 2 includes 33 features (ranked from the fifth to the thirty-seventh) whose contribution values are above the average but notably below the features in Group 1. Group 3 includes the remaining 87 features that ranked thirty-eighth to the hundred-and-twenty-fourth, whose contribution values are below the average.

Table 5. Three groups of features

Group	**Number of features**	**Range of contribution**	**Features**	**Top features in the group**
1	4	0.1232 – 0.0539	4 place-centric features	Paternal birthplace Residence Maternal birthplace Spousal birthplace

2	33	0.0288 – 0.0084	10 out of 48 paternal languages; 10 out of 40 maternal languages; 11 out of 30 spousal languages; Gender; Spousal gender	Whether father uses Koshin Whether mother uses Ajumbu Whether mother uses Missong Whether mother uses Naki-Mashi Whether spouse uses Mundabli
3	87	0.0079 – 0	The remaining Paternal, Maternal, and Spousal languages	

The top-ranked group (Group 1) exclusively includes place-centric features: paternal birthplace (contribution 0.1232), multilingual individual's residence (contribution 0.0909), maternal birthplace (contribution 0.0818), and spousal birthplace (contribution 0.0539). Of the four features, the paternal, maternal, and spousal birthplaces (ranked first, third, and fourth, respectively) are discussed together as they represent the impact of family members' birthplaces. We then discuss the residence of multilingual individuals (ranked second).

Of the three family member features, the father's birthplace plays the most critical role in primary language identification. The majority of the societies of Lower Fungom are patrilineal in terms of social and cultural heritage and hierarchy (Di Carlo 2018). Multilingual individuals may prefer or feel compelled to use their father's birthplace language at home and in the extended family. In Lower Fungom, families with a common male ancestor often live next to each other in a village. This facilitates frequent use of the father's birthplace language when communicating with relatives nearby. These internal spatial arrangements add another dimension to the impact of place, and this dimension is seldom examined in the current literature.

The importance of the mother's birthplace is prominent and ranked third in the top group. Similar to the father's birthplace language, the mother's birthplace language is frequently used at home. Multilingual individuals typically want to maintain social ties with their mother's family. This is an incentive for them to acquire and use the mother's birthplace language as needed. Yet, given the patrilineal structure of societies in Lower Fungom, it is understandable that the mother's birthplace contributes less than the father's.

The spouse's birthplace, ranked the last in the top group, is found to have noticeable, but clearly lower, importance than the parents' birthplace for primary language identification. Due to the linguistic diversity of the region, the spouses' birthplace languages are often different from the residence language of the multilingual individuals. The spouse brings their birthplace language later into the multilingual individual's life, and the impact on the individual's primary language is not as clear as that of parental influence, except for the difference by gender.

Typically, women move to live in their husband's villages after marriage. That is, in addition to patrilineality, the region is also dominated by patrilocality, another spatial notion. In this study, approximately half of the multilingual individuals are male, and the other half are female. If multilingual individuals are male, they may keep their primary languages from before marriage while understanding the primary language of their spouse (the wife). Alternatively, if multilingual individuals are female, when they relocate to the residence of their spouse (the husband), they tend to accommodate the primary language of their spouse

by learning it and using it daily (Esene Agwara 2020). It is not a surprise that the impact of spouses is identifiable, while not as strong as that of the parents.

Besides the birthplace of family members, the residence of multilingual individuals is also in the top group (contribution 0.0909). This type of place has its own spatial characteristics, which collectively increase the likelihood of using the residence languages over the family member's birthplace languages.

Houses in villages are densely located near each other, and people interact frequently with neighbors. The residence languages are commonly used in communication, to an extent where their impact may compete with the impact of the birthplace languages used in the extended family. The internal spatial arrangements, characteristic of this society, stand out as a critical dimension of spatial impact.

Collective characteristics of residence place also exert their impact. Ritual events held in a village are exclusively performed in the language associated with it, i.e., the residence language. The institutionalization of these languages directly or indirectly reinforces their use in daily interactions (Di Carlo 2018). In addition, the localist attitude that a village has towards its own language might motivate residents to identify it as their primary language to achieve a sense of social acceptance and security. Given these circumstances, it is not surprising that the language of one's village of residence may be a primary language instead of their parents' birthplace languages.

Group 2 is the distant second in the ranking, where all contribution values of the features (0.0289–0.0084) are lower than the top group (0.1232–0.0539) but above the average of 0.008. Of the 33 features in this group, 31 are birthplace languages of family members, and two are gender features, one for multilingual individuals and the other for their spouses. This group explicitly highlights the prominent role of place from the lens of birthplace languages. Gender features are recognized in this group because they are related to the impact of the spouse's birthplace language.

Group 3 includes 88 features, and their contribution values are below the average. Within the group, the majority of features are languages used by family members other than their birthplace languages. Most of these languages either originate from outside Lower Fungom or have a low presence in the daily life of multilingual individuals.

In sum, these results substantiate the role of place with respect to an individual's identification of their primary language. The impact is present across multiple dimensions, including the characteristics of a village, the spatial arrangements within a village, and spatial interactions between villages. This results in clear connections between place and primary language identification.

On the one hand, the focus on primary languages and analysis of their spatial-sociolinguistic roots establish a baseline level of knowledge for the field. On the other hand, this knowledge enriches our understanding of spatial concepts such as location, locale, place identity, and sense of place (Shaw and Sui 2020).

Conclusions

This study explores the understudied topic of primary language identification through a spatial lens. The two types of place (residence of multilingual individuals, birthplace of family members) and their associated sociolinguistic characteristics are effective in predicting primary languages to a high level of accuracy. Different features make distinct contributions to the identification, where the two types of place contribute the most, followed by the birthplace languages of family members. These results substantiate the impact of place and contribute to current discussions regarding the connection between space and language.

This study also provides a new perspective to explore small-scale multilingualism, a complex phenomenon that deserves greater attention. Although this study focuses on a specific area, multilingualism is commonplace in both urban and rural contexts across the world, and it is an important topic for both the humanities and social sciences. The approach in this study has the potential to be applied to other similar contexts. Of note here is that the results of this study are based on sociolinguistic survey techniques that can be replicated in other contexts relatively straightforwardly, thus allowing work of this kind to be readily extended to other parts of Sub-Saharan Africa and globally.

From a spatial perspective, research questions that arise from this work include understanding the role of spatial and social interactions in maintaining the linguistic ecology of small-scale societies. Addressing these issues will likely provide insights into the maintenance and revitalization of endangered languages and endangered patterns of multilingualism.

If a dataset with considerably more observations is available, we might be able to observe additional spatial-linguistic patterns and gain greater insights into space-language dynamics. However, it is not achievable in this study because of the extreme physical and social challenges in collecting a large dataset in the study area due to ongoing conflict. These limitations warrant future exploration of small-scale multilingualism in other rural regions. We expect that studies in similar contexts will broaden the findings in this study now that the validity of the approach has been established.

Acknowledgments

[Removed for anonymity.]

References

Ambrose, J. E., and Williams, C. H. 1991. Language made visible: Representation in Geolinguistics. In *Linguistic minorities, society, and territory*, ed. by C. H. Williams, 298–314. Clevedon: Multilingual Matters.

Augustin, N. H., R. P. Cummins, and D. D. French. 2001. Exploring spatial vegetation dynamics using logistic regression and a multinomial logit model. *Journal of Applied Ecology* 38 (5):991–1006.

Barron-Hauwaert, S. 2003. A study of children growing up with three Languages. In *The multilingual mind: Issues discussed by, for, and about people living with many languages*, ed. T. Tokuhama-Espinosa, 129–50. Westport, CT: Praeger Publishers.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. New York: Springer.

Bordia, S., and P. Bordia. 2015. Employee's willingness to adopt a foreign functional language in multilingual organizations: The role of linguistic identity. *Journal of International Business Studies* 46 (4):415–28.

Blommaert, J., J. Collins, and S. Slembrouck. 2005. Spaces of multilingualism. *Language & Communication* 25 (3):197–216.

Braun, A., and T. Cline. 2010. Trilingual families in mainly monolingual societies: Working towards a typology. *International Journal of Multilingualism* 7 (2):110–27.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth.

Breton, R. J. 1993. "Easy geolinguistics" and cartographers. In *The cartographic representation of linguistic data (Discussion Papers in Geolinguistics Nos. 19–21): Selected papers from a Geolinguistic Seminar (Le Paily, France, September 10–13, 1992)*, ed. Y. Peeters and C. H. Williams, 47–9. Stoke-on-Trent: Staffordshire Polytechnic.

Canagarajah, A. S., and A. J. Wurr. 2011. Multilingual communication and language acquisition: New research directions. *The Reading Matrix* 11:1–15.

Chernela, J. 2018. Language in an ontological register: Embodied speech in the Northwest Amazon of Colombia and Brazil. *Language & Communication* 63:23–32.

Chevalier, S. 2012. Active trilingualism in early childhood: The motivating role of caregivers in interaction. *International Journal of Multilingualism* 9 (4):437–54.

Chiswick, B. R., Y. L. Lee, and P. W. Miller. 2005. Family matters: The role of the family in immigrant's destination language acquisition. *Journal of Population Economics* 18 (4):631–47.

Cobbinah, A., A. Hantgan, F. Lüpke, and R. Watson. 2017. Carrefour des langues, carrefour des paradigmes. In *Espaces, mobilités et éducationn plurilingues: Éclairages d'Afrique ou d'ailleurs*, ed. M. Auzanneau, M. Bento, and M. Leclère, 79–97. Paris: Édition des Archives Contemporaines.

Di Carlo, P. 2015. Multilingualism, solidarity, and magic: New perspectives on traditional language ideologies in the Cameroonian Grassfields. In *Plurilinguismo, sintassi: Atti del XLVI congresso internazionale di studi della Società di Linguistica Italiana (SLI) – Siena, 27–29*

settembre 2012, ed. S. Casini, C. Bruno, F. Gallina, and R. Siebetcheu, 287–302. Rome: Bulzoni.

Di Carlo, P. 2018. Towards an understanding of African endogenous multilingualism: Ethnography, language ideologies, and the supernatural. *International Journal of the Sociology of Language* 254:139–63.

Di Carlo, P. 2022. The geographic sides of small-scale multilingualism: New challenges in linguistic cartography. In *New directions in linguistic geography: Exploring articulations of space*, ed. by G. Niedt, 49–85. Singapore: Palgrave Macmillan.

Di Carlo, P., and J. Good. 2017. The vitality and diversity of multilingual repertoires: Commentary on Mufwene. *Language* 93 (4):e254–e262.

Di Carlo, P., and G. Pizziolo. 2012. Spatial reasoning and GIS in linguistic prehistory: Two case studies from Lower Fungom (Northwest Cameroon). *Language Dynamics and Change* 2 (2):150–83.

Di Carlo, P., J. Good, and R. A. Ojong Diba. 2019. Multilingualism in rural Africa. *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.227.

Diskin, C. 2020. New speakers in the Irish context: Heritage language maintenance among multilingual migrants in Dublin, Ireland. *Frontiers in Education* 4:1–7.

Dressler, R. 2014. Exploring linguistic identity in young multilingual learners. *TESL Canada Journal* 32 (1):42–52.

Eberhard, D. M., G. F. Simons, and C. D. Fennig (eds.). 2022. *Ethnologue: Languages of the world* (twenty-fifth edition). Dallas, TX: SIL International. http://www.ethnologue.com.

Ellis, E. M. 2016. “I may be a native speaker but I’m not monolingual”: Reimagining all teacher’s linguistic identities in TESOL. *TESOL Quarterly* 50 (3):597–630.

Esene Agwara, A. D. 2013. Multilingualism in Lower Fungom: Analyses from an ethnographically-oriented sociolinguistic survey. MA thesis, University of Buea.

Esene Agwara, A. D. 2020. What an ethnographically-informed questionnaire can contribute to the understanding of traditional multilingualism research: Lessons from Lower Fungom. In *African multilingualisms: Rural linguistic and cultural diversity*, ed. P. Di Carlo and J. Good, 183–205. Lanham: Lexington Books.

Fassnacht, F. E., F. Hartig, H. Latifi, C. Berger, J. Hernández, P. Corvalán, and B. Koch. 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment* 154:102–14.

François, A. 2012. The dynamics of linguistic diversity: Egalitarian multilingualism and power imbalance among northern Vanuatu languages. *International Journal of the Sociology of Language* 214:85–110.

Gholamy, A., V. Kreinovich, and O. Kosheleva. 2018. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics* 11 (2):105–11.

Gislason, P. O., J. A. Benediktsson, and J. R. Sveinsson. 2006. Random forests for land cover classification. *Pattern Recognition Letters* 27 (4):294–300.

Good, J., J. Lovegren, J. P. Mve, C. Nganguep Tchiemouo, R. Voll, and P. Dicarlo. 2011. The languages of the Lower Fungom region of Cameroon: Grammatical overview. *Africana Linguistica* 17 (1):101–64.

Grenier, G. 1984. Shifts to English as usual language by Americans of Spanish mother tongue. *Social Science Quarterly* 65 (2): 537–50.

Gumperz, J. J. 1964. Linguistic and social interaction in two communities. *American Anthropologist* 66:137–53.

Henrich, J., Heine, S. J., and Norenzayan, A. 2010. Most people are not WEIRD. *Nature* 466 (7302):29.

Hiippala, T., T. L. A. Väisänen, T. Toivonen, and O. Järv. 2020. Mapping the languages of Twitter in Finland: Richness and diversity in space and time. *Neuphilologische Mitteilungen* 121 (1):12–44.

Hoffmann, C., and J. Ytsma, eds. 2004. *Trilingualism in family, school, and community*. Clevedon: Multilingual Matters.

Hua, Z. and L. Wei. 2005. Bi- and multilingual language acquisition. In *Clinical sociolinguistics*, ed. M. Ball, 165–79. Oxford: Blackwell.

Huang, X., J. Lu, S. Gao, S. Wang, Z. Liu, and H. We. 2022. Staying at home is a privilege: Evidence from fine-grained mobile phone location data in the United States during the COVID-19 pandemic. *Annals of the American Association of Geographers* 112 (1):286–305.

Izmirlian, G. 2004. Application of the random forest classification algorithm to a SELDITOF proteomics study in the setting of a cancer prevention trial. *Annals of the New York Academy of Sciences* 1020 (1):154–74.

Jiao, S., Y. Gao, J. Feng, T. Lei, and X. Yuan. 2020. Does deep learning always outperform simple linear regression in optical imaging? *Optics Express* 28 (3): 3717–31.

Kemp, C. 2009. Defining multilingualism. In *The exploration of multilingualism*, ed. L. Aronin and B. Hufeisen, 11–26. Amsterdam: John Benjamins.

Kharkhurin, A. V. 2008. The effect of linguistic proficiency, age of second language acquisition, and length of exposure to a new cultural environment on bilinguals' divergent thinking. *Bilingualism: Language and Cognition* 11 (2):225–43.

Kyriakides, G., and K. G. Margaritis. 2019. *Hands-on ensemble learning with Python: Build highly optimized ensemble machine learning models using scikit-learn and Keras*. Birmingham: Packt Publishing.

Latief, M.A., T. Siswantining, A. Bustamam, and D. Sarwinda, D. 2019. A comparative performance evaluation of random forest feature selection on classification of hepatocellular carcinoma gene expression data. *Proceedings of the 3rd International Conference on Informatics and Computational Sciences (ICICoS 2019)*:1–6.

Luan, J., Zhang, C., Xu, B., Xue, Y., and Ren, Y. 2020. The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research* 227:105534.

Lüpke, F., and A. Storch. 2013. *Repertoires and choices in African Languages*. Berlin: De Gruyter Mouton.

Lüpke, F. 2016. Uncovering small-scale multilingualism. *Critical Multilingualism Studies* 4 (2):35–74.

Nicholls, R. J., P. A. Eadie, and S. Reilly. 2011. Monolingual versus multilingual acquisition of English morphology: What can we expect at age 3? *International Journal of Language & Communication Disorders* 46 (4):449–63.

Pakendorf, B., N. Dobrushina, and O. Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25 (4):835–59.

Paradis, J. 2008. Early bilingual and multilingual acquisition. *Handbook of multilingualism and multilingual communication*, ed. P. Auer and L. Wei, 15–44. Berlin: Mouton de Gruyter.

Paradowski, M. B., and A. Bator. 2018. Perceived effectiveness of language acquisition in the process of multilingual upbringing by parents of different nationalities. *International Journal of Bilingual Education and Bilingualism* 21 (6):647–65.

Paul, B. K. 2020. Identifying and analyzing the dominant languages in small island developing states. *The Professional Geographer* 72 (1):121–30.

Powers, D. M. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2 (1): 37–63.

Prochazka, K., and Vogl, G. 2017. Quantifying the driving factors for language shift in a bilingual region. *Proceedings of the National Academy of Sciences* 114 (17):4365–69.

Qi, Y. 2012. Random forest for bioinformatics. In *Ensemble machine learning: Methods and applications*, ed. by C. Zhang and Y. Ma, 307–23. New York: Springer.

Ranacher, P., N. Neureiter, R. Van Gijn, B. Sonnenhauser, A. Escher, R. Weibel, P. Muysken, and B. Bickel. 2021. Contact-tracing in cultural evolution: A Bayesian mixture model to detect geographic areas of language contact. *Journal of The Royal Society Interface* 18 (181):20201031.

Ranacher, P., R. Van Gijn, and C. Derungs. 2017. Identifying probable pathways of language diffusion in South America. Paper presented at AGILE conference 2017, Wageningen, May 9–May 12.

Reyes-García, V., D. Zurro, J. Caro and M. Madella 2017. Small-scale societies and environmental transformations: Coevolutionary dynamics. *Ecology and Society* 22 (1). https://doi.org/10.5751/ES-09066-220115.fcl

Shaw, S. L., and D. Sui. 2020. Understanding the new human dynamics in smart spaces and places: Toward a spatial framework. *Annals of the American Association of Geographers* 110 (2):339–48.

Singer, R., and S. Harris. 2016. What practices and ideologies support small-scale multilingualism? A case study of Warruwi Community, northern Australia. *International Journal of the Sociology of Language* 241:163–208.

Skutnabb-Kangas, T., R. Phillipson, A. K. Mohanty, and Panda, M. eds. 2009. *Social justice through multilingual education*. Clevedon: Multilingual Matters.

Slotta, J. 2012. Dialect, trope, and enregisterment in a Melanesian speech community. *Language & Communication* 32 (1):1–13.

Strobl, C., J. Malley, and G. Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14 (4):323–48.

Sutton, P. 1997. Materialism, sacred myth and pluralism: Competing theories of the origin of Australian languages. In *Scholar and sceptic: Australian Aboriginal studies in Honour of L. R. Hiatt*, edited by F. Merlan, J. Morton, A. Rumsey (eds.), 211–42. Canberra: Aboriginal Studies Press.

Tagliamonte, S. A., and R. H. Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24 (2):135–78.

Unsworth, S. 2016. Quantity and quality of language input in bilingual language development. In *Bilingualism across the lifespan: Factors moderating language proficiency*, ed. E. Nicoladis and S. Montanari, 103–22. Berlin: De Gruyter Mouton.

Väisänen, T., O. Järv, T. Toivonen, and T. Hiippala. 2022. Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems* 97:101857.

Van der Merwe, I. 1993. The urban geolinguistics of Cape Town. *Geo Journal* 31 (4):409–17.

Veselinova, L. N., & Booza, J. C. 2009. Studying the multilingual city: A GIS-based approach. *Journal of Multilingual and Multicultural Development* 30 (2):145–65. https://doi.org/10.1080/01434630802582476.

Warnier, J.-P. 1980. Des précurseurs de l'École Berlitz: Le multilinguisme dans les Grassfields du Cameroun au 19e siècle. In *L'expansion bantoue: Actes du colloque international du CNRS, Viviers (France) 4–16 avril 1977: Volume III*, ed. by L. Bouquiaux, 827–44. Paris: SELAF.

Westergaard, M. 2021. Microvariation in multilingual situations: The importance of property-by-property acquisition. *Second Language Research* 37 (3):379–407.

Whiteley, W. H. 1974. Some patterns of language use in the rural areas of Kenya. In *Language in Kenya*, ed. by W. H. Whiteley, 319–50. Oxford: OUP.

Williams, C. H. 1996. Geography and contact linguistics. In *Contact linguistics: An international handbook of contemporary research*, ed. H. Goebl, P. H. Nelde, Z. Stary, and W. Wolck, 63–75. New York: Walter de Gruyter.

Wibowo, A., F. A. Nugroho, E. A. Sarwoko, and I. M. A. Setiawan. 2020. Classification of headache disorder using random forest algorithm. *Proceedings of the 4th International Conference on Informatics and Computational Sciences (ICICoS 2020)*:1-5.

Appendix A. Population of Lower Fungom villages by category.

>1,000: Fang; Koshin

500-1,000: Abar; Munken; Kung; Missong

100-500: Mundabli; Ajumbu; Mashi; Ngun; Buu; Biya; Mufu

Appendix B. Number of speakers of the 15 primary languages among multilingual individuals.

Abar	Ajumbu	Biya	Buu	Fang	Kung	Koshin	Mashi	Mekaf	Missong	Mufu	Mundabli	Munken	Ngun	SmallMekaf
20	12	5	5	14	10	14	8	5	21	11	26	17	2	4

Appendix C. Contributions of features.

Rank	Feature	Contribution
Group 1		
1	Paternal birthplace	0.1232
2	Multilingual individual residence	0.0909
3	Maternal birthplace	0.0818
4	Spousal birthplace	0.0539
Group 2		
5	Whether father uses Koshin	0.0288
6	Whether mother uses Ajumbu	0.0240
7	Whether mother uses Missong	0.0228
8	Whether mother uses Naki-Mashi	0.0221
9	Whether spouse uses Mundabli	0.0195
10	Whether father uses Missong	0.0185
11	Whether father uses Fang	0.0178
12	Whether mother uses Kung	0.0178
13	Whether father uses Ajumbu	0.0177
14	Whether mother uses Abar	0.0176
15	Whether mother uses Munken	0.0158

16	Whether father uses Kung	0.0157
17	Whether mother uses Pidgin	0.0156
18	Whether father uses Abar	0.0153
19	Whether mother uses Fang	0.0153
20	Whether spouse uses Ajumbu	0.0152
21	Whether father uses Naki-Mashi	0.0150
22	Whether spouse uses Koshin	0.0145
23	Whether spouse uses Naki-Mashi	0.0140
24	Whether spouse uses Mufu	0.0129
25	Whether father uses Munken	0.0127
26	Whether spouse uses Kung	0.0127
27	Whether spouse uses Abar	0.0112
28	Gender of the spouse	0.0104
29	Whether mother uses Koshin	0.0102
30	Whether father uses Biya	0.0100
31	Gender	0.0098
32	Whether mother uses Mufu	0.0095
33	Whether father uses Mundabli	0.0095
34	Whether father uses Pidgin	0.0091

35	Whether spouse uses Biya	0.0088
36	Whether spouse uses English	0.0086
37	Whether spouse uses Missong	0.0084
Group 3		
38	Whether spouse uses Mekaf	0.0079
39	Whether spouse uses Buu	0.0078
40	Whether spouse uses Pidgin	0.0078
41	Whether father uses Buu	0.0075
42	Whether father uses Kom	0.0074
43	Whether father uses Mekaf	0.0068
44	Whether mother uses Biya	0.0061
45	Whether mother uses Ngun	0.0058
46	Whether mother uses Buu	0.0052
47	Whether spouse uses Ngun	0.0049
48	Whether spouse uses Munken	0.0047
49	Whether mother uses Mundabli	0.0046
50	Whether mother uses Mekaf	0.0045
51	Whether spouse uses Fang	0.0044
52	Whether father uses Mungbam	0.0041

53	Whether spouse uses Mungbam	0.0040
54	Whether father uses Mufu	0.0035
55	Whether father uses Small-Mekaf	0.0035
56	Whether spouse uses Munggaka	0.0035
57	Whether mother uses Small-Mekaf	0.0034
58	Whether mother uses Isu	0.0033
59	Whether spouse uses Small-Mekaf	0.0032
60	Whether father uses Isu	0.0029
61	Whether mother uses Weh	0.0027
62	Whether spouse uses Jukun	0.0025
63	Whether father uses Bum	0.0024
64	Whether spouse uses Aghem	0.0024
65	Whether spouse uses Isu	0.0023
66	Whether father uses Mmen	0.0021
67	Whether spouse uses French	0.0021
68	Whether mother uses Mungbam	0.0019
69	Whether mother uses Mmen	0.0019
70	Whether father uses Aghem	0.0018
71	Whether spouse uses Mmen	0.0018

72	Whether spouse uses Weh	0.0017
73	Whether father uses Mumfu	0.0016
74	Whether mother uses Fungom	0.0013
75	Whether father uses Bambui	0.0012
76	Whether father uses Bali	0.0010
77	Whether mother uses Aghem	0.0010
78	Whether father uses Dumbu	0.0009
79	Whether mother uses Kom	0.0009
80	Whether father uses Fungom	0.0009
81	Whether mother uses Munggaka	0.0008
82	Whether father uses Jukun	0.0008
83	Whether father uses Ncane	0.0008
84	Whether spouse uses Dumbu	0.0008
85	Whether spouse uses Kom	0.0007
86	Whether spouse uses Bali	0.0007
87	Whether father uses English	0.0007
88	Whether spouse uses Hausa	0.0006
89	Whether mother uses Mumfu	0.0006
90	Whether mother uses Nsungli	0.0006

91	Whether father uses Nunggaka	0.0005
92	Whether father uses Lamnso	0.0005
93	Whether father uses Zhoa	0.0005
94	Whether mother uses Zhoa	0.0004
95	Whether mother uses Nchanti	0.0004
96	Whether father uses Misaje	0.0004
97	Whether mother uses Bum	0.0003
98	Whether mother uses Jukun	0.0003
99	Whether spouse uses Fungom	0.0003
100	Whether father uses Weh	0.0002
101	Whether mother uses French	0.0002
102	Whether father uses Limbum	0.0002
103	Whether father uses Banso	0.0002
104	Whether father uses Hausa	0.0002
105	Whether father uses Oku	0.0001
106	Whether mother uses Naki	0.0001
107	Whether father uses Bakweri	0
108	Whether father uses French	0
109	Whether father uses Nkwen	0

110	Whether father uses Naki	0
111	Whether father uses German	0
112	Whether father uses Nsungli	0
113	Whether father uses Nchanti	0
114	Whether father uses Njikum	0
115	Whether father uses FulaniAku	0
116	Whether father uses Kumfutu	0
117	Whether mother uses English	0
118	Whether mother uses Hausa	0
119	Whether mother uses Bali	0
120	Whether mother uses Nkwen	0
121	Whether mother uses Tsa	0
122	Whether mother uses Adjume	0
123	Whether mother uses Mgbeuh	0
124	Whether mother uses Bafmen	0

Appendix D. Confusion matrices.

Table D.1 Confusion matrix of primary language prediction by the random forest model.

Observed \ Predicted	Abar	Ajumbu	Biya	Buu	Fang	Koshin	Kung	Missong	Mufu	Mundabli	Munken	Mashi	Mekaf	Ngun	Small Mekaf
Abar	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Ajumbu	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Biya	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Buu	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Fang	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Koshin	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
Kung	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
Missong	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
Mufu	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Mundabli	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
Munken	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0
Mashi	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Mekaf	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Ngun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Small Mekaf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table D.2 Confusion matrix of primary language prediction by the multinomial logistic regression model.

Observed \ Predicted	Abar	Ajumbu	Biya	Buu	Fang	Koshin	Kung	Missong	Mufu	Mundabli	Munken	Mashi	Mekaf	Ngun	Small Mekaf
Abar	3	0	0	0	0	0	0	0	0	1	0	0	0	1	0
Ajumbu	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Biya	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Buu	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Fang	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Koshin	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
Kung	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
Missong	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Mufu	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Mundabli	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0
Munken	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0
Mashi	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Mekaf	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Ngun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Small Mekaf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1