STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 71–87, DOI 10.21307/stattrans-2019-035
Submitted – 14.05.2019; Paper ready for publication – 01.10.2019

MODELLING LANGUAGE EXTINCTION USING SUSCEPTIBLE-INFECTIOUS-REMOVED (SIR) MODEL

N. A. Ikoba[1], E. T. Jolayemi[2]

ABSTRACT

The study presents a stochastic epidemic model applied to the model of indigenous language extinction. The Susceptible-Infectious-Removed (SIR) categorization of an endemic disease has been reformulated to capture the dynamics of indigenous language decline, based on the assumption of non-homogeneous mixing. The time in which an indigenous language is expected to be extinct was derived using a modified SIR model with the population segmented into several sub-communities of small sizes representing family units. The data obtained from the 2016 indigenous language survey conducted in several parts of Nigeria and from the 2013 Nigeria Demographic Health Survey (NDHS) were used to estimate the key parameters of the model for Nigeria's several indigenous languages. The parameters of interest included the basic reproduction number, the threshold of endemicity, and the time in which a language is expected to be extinct, starting from the endemic level. On the basis of the time in which a language is expected to be extinct, several of the surveyed languages appeared to be in a precarious condition, while others seemed virile, thanks to a high language transfer quotient within families.

Key words: language extinction, stochastic epidemic model; non-homogeneous mixing, quasi-stationary distribution, time in which a language is expected to be extinct.

1. Introduction

The world today is littered with thousands of languages and several hundreds have been documented to have become extinct (Crystal, 2000). Of the known 7,102 living languages, 22% of them have been categorized as 'in trouble', 13% dying, while there is a loss rate of about 6 languages per year (Lewis *et al.*, 2015).

It is claimed that more than 400 Nigerian languages are endangered and there is a declining level of transfer of indigenous language ability to the younger generation (Ohiri-Aniche, 2014). It was therefore projected that many Nigerian languages will become extinct by 2084 (Ohiri-Aniche, 2014).

A lot of languages globally are tethering on the brink of extinction (Nuwer, 2014). Over the past century alone, it is estimated that around 400 languages -

[1] Department of Statistics, University of Ilorin, Ilorin, Nigeria. E-mail: ikoba.na@unilorin.edu.ng.
[2] Department of Statistics, University of Ilorin, Ilorin, Nigeria.

about one every three months - have gone extinct, and it is estimated that 50% of the world's remaining 6,500 languages will be gone by the end of the 21st century (Nuwer, 2014).

Languages usually reach the point of crisis after being displaced by a socially, politically and economically dominant one. Sometimes, especially in immigrant communities, parents will decide not to teach their children their heritage language, perceiving it as a potential hindrance to their success in life (Nuwer, 2014).

The problem of declining use of indigenous languages at the expense of global languages like English and French has become much pronounced in previously colonized countries in the developing world.

While there is a common view that many indigenous languages are dying, the depth of the problem is yet to be sufficiently captured, as there has been minimal use of scientific approaches to reflect the gravity of the problem especially among indigenous languages of sub-Saharan Africa. Thus, the desire to provide an analytical model that sufficiently captures the language decline in a society over an extended period of time is the main motivation for this research.

A research that motivated the pursuit of language modelling using epidemic models is the work of Daley and Kendall (1964) which modelled the spread of rumours using the basic SIR epidemic model.

The main focus in the adaptation of stochastic epidemic models to study the decline of indigenous languages is to derive conditions under which the indigenous language can be propagated without the threat of extinction. In other words, it is of interest to estimate thresholds above which such languages will continue to survive with minimal fluctuations around the threshold.

The SIR model was originally used to model the dynamics of an epidemic within a population consisting of susceptible individuals (S), infectious individuals (I) and those who have recovered or are removed (R) and could no longer contribute to the spread of the epidemic. There are numerous versions of the SIR epidemic model to capture different disease dynamics for both closed populations and populations that incorporate demographic turnover.

Under the context of indigenous language decline in a community, using the principle of the SIR epidemic model, the susceptible group is viewed as all children born into the community; the infectious group corresponds to children who later acquire indigenous language ability; while the removed group contains those who exit the community either through death or migration.

There are several useful similarities between the study of epidemics and language decline, with seemingly opposite objectives. While under the epidemic context, the goal of modelling is to ensure the termination of the spread of the disease in the population, in the context of language modelling, the goal is to ensure continuous propagation of the language to enhance its survival in the population. It is noted that this is a novel application of stochastic epidemic models to capture indigenous language dynamics. A search of the literature reveals that there are seemingly no papers applying epidemic models in the area of indigenous language extinction. A few of the relevant researches germane to this study are now presented.

Trapman and Bootsma (2009) established a relation between the spread of infectious diseases and the dynamics of the M/G/1 queueing system with processor sharing. They showed that the number of infectious individuals in a standard SIR epidemic model at the moment of first detection of the epidemic was geometrically distributed. They derived the distribution of the number of infectious individuals at the moment of first detection in a broad class of epidemics in large populations.

An inherent relationship exists between infectious diseases and human populations, as reflected by Dobson and Carper (1996). The relationship can be extended to that of languages and human populations.

Allen and Burgin (2000) compared the dynamics of some deterministic and stochastic discrete-time epidemic models with fixed and varying population. In some cases, the time to extinction was very long, and in such cases, if the probability distribution is conditioned on non-extinction, then for values of the basic reproduction number $R_0 > 1$, there exists a quasi-stationary distribution whose mean agrees with the deterministic endemic equilibrium. The expected duration of the epidemic was also investigated numerically.

Ball and Lyne (2000) analyzed the spread of an SIR epidemic in a closed, finite population using a household model. Both local and global infection was possible within the population. We are interested in such household models that could be modified to fit the goals of language modelling.

Nasell (2002) studied several stochastic models with demography for various endemic infections. Approximations of the quasi-stationary distributions and expected time to extinction were derived for the SI, SIS, SIR and SIRS models. The approximations were valid for sufficiently large population sizes, and in comparison with corresponding deterministic models, the stochastic models provided realistic parameter estimates.

Nasell (2005) examined quasi-stationarity and time to extinction for the classic endemic model, with focus restricted to the transition region in the parameter space where the quasi-stationary distribution is non-normal. An approximation of the marginal distribution of the infected individuals in quasi-stationarity was proposed. Simulation results showed that the analytical approximations performed reasonably.

Verdasca et. al (2005) studied the effect of spatial correlations on the spread of infectious diseases using a stochastic SIR model on complex networks. Heavy stochastic fluctuations tend to limit the utility of deterministic models under such circumstances.

Lindholm and Britton (2007) studied an SIR model with the population subdivided into k sub-communities of equal sizes n, assumed large. Lindholm and Britton (2007) model provides a useful platform for the adaptation of epidemic modelling to language extinction. The general SIR model with homogeneous mixing is called SIR-HM and the SIR model with heterogeneous mixing in which the population is divided into sub-communities is called SIR-SC (Lindholm and Britton, 2007).

Burr and Chowell (2008) analyzed SEIR-type models under the assumption of non-homogeneous mixing. Their goal was to evaluate possible retrospective signatures of non-homogeneous mixing behaviour. From simulated outbreaks, it was concluded that such signatures can detect non-SEIR-type behaviour in some of the social structures considered.

Schwartz et. al (2009) investigated random extinction processes in a class of epidemic models, examining the rate of disease extinction as a function of disease spread. It was shown that the effective entropic barrier for extinction in a SIS model displays scaling with the distance to the bifurcation point, with an unusual critical component. Analytical results were compared with numerical simulations and were found to be good.

Billings et. al (2013) considered the effect of randomly distributed intervention as disease control on large finite populations, and showed how intervention control modulates the expected time to extinction, which in turn was a function of population size and rate of infection spread.

In section two, the relevant SIR model is conceptualized to capture the language decline dynamics. The model analysis is also presented, and estimates of the quasi-stationary distribution, threshold of endemicity and expected time to extinction are obtained. The model parameter estimation from survey and historical data is also presented in section 3. The results from an indigenous language survey carried out in some parts of Nigeria in order to determine the extinction status of some Nigerian languages are presented and discussed in section 3. Finally, our conclusions are presented in section 4.

2. SIR model with demography for language extinction

The main interest motivating the adaptation of the relevant SIR model to study indigenous language extinction is the very close relationship between epidemics and languages with regards to extinction. When demographic turnover and heterogeneities in the population are incorporated into SIR models for endemic diseases with infectious periods that are of the same order as the lifetime distribution of individuals, it is seen that such scenario will sufficiently capture the indigenous language decline dynamics in a population.

2.1. Description of the SIR model with demography and heterogeneous mixing

One of the fundamental assumptions of the standard SIR epidemic model is that the population is a closed one, that is, there is no demographic turnover (births, deaths, migrations) in the population. For the SIR model with demography, this assumption is relaxed in that there can be births of susceptible individuals into the population and also deaths or migration out of the population. This relaxed assumption therefore introduces population dynamics as a component of the evolution of the epidemic.

The SIR model with demography provides a useful adaptation of endemic diseases modelling to capture the decline of languages with some modifications. Andersson and Britton (2000) provided an excellent description of the SIR model with demography, stressing its appropriateness in modelling endemic diseases, which have a long infectious period in relation to the lifetime of the individual. They noted that when modelling the spread of an endemic disease in a very large population, demographic turnover cannot be ignored. A disease is called endemic if it is able to persist in a population for a long time, without the need of introducing new infectious individuals from some external population (Lindholm, 2007). When the susceptible population in a large community is augmented fast enough through births and/ or migrations, the epidemic tends to persist for a very long time even without the introduction of new infectious individuals into the population.

For the language scenario, the assumption is imposed that the infectious period and the residual lifetime of the individual coincide. In fact, the overall lifetime of an individual is the sum of the language acquisition (latency) period and the infectious period. The system is viewed at specified epochs.

The focus in the adaptation of stochastic epidemic models to study the decline of indigenous languages is to derive conditions under which the indigenous language can be propagated without the threat of extinction. In other words, interest is centred in achieving those thresholds above which such languages will continue to survive with minimal fluctuations around the threshold.

According to Lindholm and Britton (2007), for diseases in which the infectious period is of the same order with the lifetime distribution, one can assume that when an individual recovers from the infection, this individual will likely be removed due to death within a relatively short time period. This is a fundamental backbone of the language adaptation of stochastic epidemic models, as it fairly approximates the scenario in the language setting, where it is assumed that the infectious period and the residual lifetime of the individual are identically distributed.

The relevant SIR model with demography is now conceptualized in the indigenous language extinction approach.

The population in the community contains k families labelled 1,2,...,k. Let n_i be the size of the i^{th} family, then the size of the entire population is

$$N = \sum_{i=1}^{k} n_i$$

Individuals are born into the population at a constant rate μk and each of them is assumed to have an exponentially distributed lifetime with intensity θ.

Initially, it is assumed that there are zero susceptible and k indigenous language-speaking individuals in the population, denoting the initial language-speaking population in the community before the advent of external influences like colonization. A given indigenous language speaker stays ‘infectious’ for a time period that is exponentially distributed with intensity ν (unless he dies of other causes before the end of that period). During that time, the infective in the j^{th} family makes contact with children in his family at rate β/n_j. With probability p, a contacted child imbibes the language.

All random variables and counting processes (which are Poisson processes due to the exponential lifetimes assumption) are assumed to be mutually independent. Some of the relevant variables are next defined:

X(t)= Number of susceptibles at time t; Y(t)= Number of language speakers at time t.

$\{X(t), Y(t), t \geq 0\}$ is a Markov process with transition rates:

$$\left\{\begin{array}{lll} From & to & at\ rate \\ (s_j, i_j) & (s_j+1, i_j) & \mu n_j \\ (s_j, i_j) & (s_j-1, i_j) & \theta s_j \\ (s_j, i_j) & (s_j-1, i_j+1) & (\beta/n(1+\varepsilon(k-1)))s_j(i_j+\varepsilon\sum_{u\neq j} i_u) \\ (s_j, i_j) & (s_j, i_j-1) & (\theta+\nu)i_j \end{array}\right. \tag{2.1}$$

We are mainly interested in the susceptible (S) and language-speaking (I) population, as the recovered/ removed population is assumed to have no further influence on the indigenous language dynamics.

Non-homogeneous mixing models such as those for outbreaks in social networks are often believed to provide better predictions of the benefits of the various mitigation strategies such as isolation or vaccination (Burr and Chowell, 2008). The homogeneous mixing assumption is often an unrealistic one, but due to the tractability of their analysis, such models provide useful insights on the disease dynamics. Social structure is a type of non-homogeneous mixing, such as most individuals having preferential contact with work or family members compared to the general population.

The application of the SIR model with demography and non-homogeneous mixing to the language scenario imply that an infectious individual can mainly infect his offspring. There is a sort of family-based transmission of language ability from parent to offspring. The dynamics of the spread of the language becomes more intricate with the assumption of non-homogeneous mixing. If the population is divided into families, some families may not have any indigenous language speaker (language-free), while most others will contain a mixture of language speakers and non-speakers. As the population evolves in time, the tendency is to have an increasing number of non-language speakers, mainly the young ones born into the population but unable to imbibe their indigenous language.

It is of interest to study situations where k is large in relation to the $n_j's$. This appropriately captures the scenario in communities where there are smaller family sizes in relation to the number of families in the community. It is also assumed that the contact rates within families are the same.

Define the parameter

$\varepsilon =$ proportion of an individual's contacts that are with other families.

$\varepsilon = 0$ implies that the families are isolated and there is no transmission between families. $\varepsilon = 1$ implies a single large community in which there is interaction among individuals in the population irrespective of their family.

The overall infectious pressure in the entire population is kept constant regardless of the value of ε.

The standard SIR model assumes homogeneous mixing in the population, therefore results from the analysis of the standard SIR model are not applicable when the assumption of homogeneous mixing is relaxed. An appropriate analysis of the SIR model with the population divided into sub-communities has been given by Lindholm and Britton (2007). For the case with sub-communities, an infected individual in the j^{th} family makes contacts with any given individual within its own family at rate β'/n_j and at rate $\varepsilon\beta'/n$ with a given individual in any of the k-1 surrounding sub-communities, where n is taken as the mean family size in the population.

Drawing from the principles of Lindholm and Britton (2007), the probability that a contact is within the j^{th} family is

$$\Pr(within\ contact) = \frac{1}{1+\varepsilon(k-1)}$$

and the basic reproduction number, R_0 is given as

$$R_0 = \frac{1}{(\theta+\nu)}\left(\frac{n\beta'}{n} + \frac{(k-1)n\beta'}{n}\right) = \frac{\beta'}{(\theta+\nu)}\big(1+\varepsilon(k-1)\big) \tag{2.2}$$

Define $\alpha = (\theta+\nu)/\theta$ as the ratio of mean lifetime to mean duration of language ability and $\beta = \beta'(1+\varepsilon(k-1))$, then the basic reproduction number can be expressed as

$$R_0 = \frac{\beta}{\theta\alpha} \tag{2.3}$$

which is of the same form as the SIR model with homogeneous mixing.

It is noted that the basic reproduction number, R_0 is an increasing function of ε. In fact, it can be seen from equation (2.2) that $R_0(1) = kR_0(0)$ and all other values of $R_0(\varepsilon)$ are bounded in the interval $(R_0(0), R_0(1))$. Hence, as the value of ε increases, the basic reproduction number also increases, with the additional property that $R_0(1)$ is an integer multiple of $R_0(0)$.

2.2. The quasi-stationary distribution

The quasi-stationary distribution is important when modelling endemic diseases, since interest is on the behaviour of the epidemic until it goes extinct. Similar reasoning also show that the quasi-stationary distribution in the language context is also of relevance, as a language that has progressed to the stage of quasi-stationarity has become endangered and is most likely to become extinct in the absence of any external revitalization measures. The quasi-stationary distribution of the population is defined as the conditional distribution that the process has not been absorbed after a long period of time. The endemic level can be thought of as the mean of this distribution, which the process fluctuates around (Lindholm and Britton, 2007).

Let $Q = \{q_{x,y}\}$ denote the quasi-stationary distribution.

$$q_{x,y} = \lim_{t\to\infty} Pr\{X(t) = x, Y(t) = y | Y(t) > 0\}$$

If in addition, the infectious period is exponentially distributed, then by reason of the memoryless property of the exponential distribution,

$$Pr(T_Q > t+s | T_Q > t, \big(X(0), Y(0)\big) \sim\ Q) = Pr(T_Q > s | \big(X(0), Y(0)\big) \sim\ Q)$$

where T_Q is the time to extinction of the language given that the process is in the quasi-stationary state.

A relevant proposition is next presented (Lindholm and Britton, 2007):

Proposition 1: The time to extinction given that the process is started in the quasi-stationary distribution T_Q, is exponentially distributed with mean

$$E(T_Q) = \tau = \frac{1}{\theta\, \alpha\, q_{.,1}} \tag{2.4}$$

where

$$q_{.,1} = \sum_x q_{x,1} \tag{2.5}$$

is the marginal probability that 1 indigenous language speaker is left in the population.

Lindholm and Britton (2007) derived an approximation of $q_{.,1}$, and thus T_Q using diffusion approximation.

Drawing from the central limit theorem, and with a population of size n, then approximations for the mean number of indigenous language speakers, μ_y and the standard deviation, σ_y are given as (Lindholm and Britton, 2007)

$\mu_y = n\frac{R_0-1}{\alpha R_0}$ and $\sigma_Y = \frac{\sqrt{n}}{R_0}\surd(R_0 - 1 + R_0^2/\alpha)$, and when $\alpha >> R_0$

$\sigma_Y \approx \sqrt{n(R_0-1)}/R_0$ and

$$q_{.,1} \approx \frac{R_0}{\sqrt{2\pi n(R_0-1)}} \exp\left(\frac{-n\,(R_0-1)}{2\alpha^2}\right) \tag{2.6}$$

The quasi-stationary distribution obtained by Lindholm and Britton (2007), as presented in equation (2.6) would not be appropriate in the case where the lifetime of the individual and the infectious period are of the same order (α small). This is the nature of the indigenous language dynamics, hence we propose an alternative approximation for the quasi-stationary distribution as

$$q_{.,1} \approx \frac{R_0}{\sqrt{2\pi n(R_0-1+R_0^2/\alpha)}} \exp\left(\frac{-n\,(R_0-1)^2}{2\alpha^2(R_0-1+R_0^2/\alpha)}\right) \tag{2.7}$$

which is better suited for the language characterization due to the fact that the residual lifetime and the period language ability are of the same order.
It is noted that equation (2.7) reduces to equation (2.6) when $\alpha >> R_0$.

2.3. Expected time to extinction of the language

As established in Andersson and Britton (2000), the expected time to extinction for the homogeneous mixing case of the SIR model, using the normal approximation (equation 2.6) yields

$$\tau \approx \frac{\sqrt{2\pi n(R_0-1)}}{\theta\alpha R_0} \exp\left(\frac{n\,(R_0-1)}{2\alpha^2}\right) \tag{2.8}$$

When the life length is long in relation to the length of the infectious period, the above approximation produces too wide estimates when the sub-community sizes are only moderately large (Lindholm and Britton, 2007). A better approximation, well suited for the language scenario with smaller family sizes, could be obtained using equation (2.7) instead:

$$\tau \approx \frac{\sqrt{2\pi n(R_0-1)}}{\theta\alpha R_0}\exp\left(\frac{n\,(R_0-1)^2}{2\alpha^2\left(R_0-1+\frac{R_0^2}{\alpha}\right)}\right) \tag{2.9}$$

Under the scenario of long life length in relation to the length of the infectious period, Nasell (2005) proposed that the quasi-stationary distribution of the number of infected individuals could be approximated with a geometric distribution with $p = 1/\mu_Y$. If $Y \sim Geo(p)$, then $E(Y) = 1/p = \mu_Y$

When the quasi-stationary distribution of Y is approximated as such, then

$$\tau_n \approx \frac{n(R_0-1)}{\theta\alpha^2 R_0} \tag{2.10}$$

The expected time to extinction when all the k families are started at the endemic level, where there is a proportion ε of contacts between families, $\tau(\varepsilon)$ provides a credible challenge in obtaining the estimates of the mean time to extinction τ that is dependent on ε. In actual fact, $\tau(\varepsilon) = \tau(\varepsilon, n, k, \theta, \alpha, R_0)$.

When $\varepsilon = 0$, all k families are isolated and independent, and starting at the endemic level of infection, the expected time until one of the k families recovers is τ_n/k, due to independence and that the expected duration of an epidemic within a family is exponentially distributed with mean τ_n. Due to the assumption that the language-free states are absorbing states of the process, when $\varepsilon = 0$, the expected time until one of the k-1 remaining families recovers is $\tau_n/(k-1)$. Repeating the argument yields

$$\tau(0) = \tau_n \sum_{i=1}^{k} \frac{1}{i} \tag{2.11}$$

As the number of families, k increases, the sum $\sum_{i=1}^{k}\frac{1}{i}$ approaches a finite value, its limiting value. Hence, the time to extinction will increase with greater number of families even with zero interaction between families.

On the other hand, when $\varepsilon = 1$, all k families behave as one large community of size $\sum_{j=1}^{k} n_j$, and we can use the SIR-HM approximation of τ_n with n replaced by $\sum_{j=1}^{k} n_j$ (Lindholm and Britton, 2007).

$$\tau(1) = \tau_n$$

If the $n_j's$, the size of the families are small, Lindholm and Britton (2007) suggested that the geometric approximation of τ_n may be used to approximate $\tau(0)$ and $\tau(1)$, to yield

$$\frac{\tau(0)}{\tau(1)} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}$$

which is less than 1 for $k > 1$, and this implies that $\tau(0) < \tau(1)$, hence $\tau(\varepsilon)$ is an increasing function.

For large sizes of the families, Lindholm and Britton (2007) also suggested that the truncated normal approximation of τ_n should be used. Furthermore, for population sizes such that $\mu_Y/\sigma_Y > 3$ and with $\alpha >> R_0$, τ_n is approximated by equation (2.8), and $\frac{\tau(0)}{\tau(1)}$ is also smaller than 1 for sufficiently large n.

As established by Andersson and Britton (2000), the stationary points of the process are the language-free state (1,0) and

$$(\hat{x}, \hat{y}) = \left(\frac{\nu+\theta}{\beta}, \frac{\theta}{\nu+\theta}\left(1 - \frac{\nu+\theta}{\beta}\right)\right) = \left(\frac{1}{R_0}, \frac{R_0 - 1}{\alpha R_0}\right) \tag{2.12}$$

It should be noted that the language-free state (1,0) is stable for $R_0 < 1$ and unstable for $R_0 > 1$. The point $(\hat{x}, \hat{y})$ is stable for $R_0 > 1$ (and is otherwise negative).

If $R_0 < 1$, the language is predicted to die out fairly quickly in the community, and if $R_0 > 1$ then it will rise towards a positive level called the endemic level.

The basic reproduction number, R_0 works as a threshold determining the dynamics of the indigenous language, and is dimensionless. If $R_0 \leq 1$, the language will go extinct rather quickly. Otherwise the language has a positive probability to persist in the population over a long period of time.

Using principles from the Central Limit Theorem, let us assume that $R_0 > 1$ and suppose that the process is started close to the endemic level $(n\hat{x}, n\hat{y})$. The process is positively recurrent and it will become absorbed into the set of language-free states $\{(i, 0), i \geq 0\}$ in finite time. Prior to absorption, small fluctuations may be observed around the endemic level and the nature of these fluctuations can be examined. Define for $t \geq 0$

$$\left(\tilde{X}_{t_i}, \tilde{Y}_{t_i}\right) = \sqrt{n}(\bar{X}_{t_i} - x(t_i), \bar{Y}_{t_i} - y(t_i))$$

Andersson and Britton (2000) established that $\left(\tilde{X}_{t_i}, \tilde{Y}_{t_i}\right)$ converges weakly on compact time intervals to a Gaussian process $(\tilde{X}, \tilde{Y})$ with mean vector $\underline{0}$ and covariance matrix Σ.

The time to extinction of the indigenous language, T_Q is defined as (Andersson and Britton, 2000)

$$T_Q = \inf\{i \geq 0 : Y(t_i) = 0\}$$

T_Q is finite for any fixed population size if $R_0 > 1$.

According to Andersson and Britton (2000), it is a classical (and very difficult) problem to obtain the estimates of T_Q. Not even the expected value $E(T_Q)$ is easily estimated. One approach is to let the population size to become very large and regard language extinction as the result of a large deviation from a high endemic level. In this regard, asymptotic approximations of $E(T_Q)$ are available. There is also an heuristic derivation of an approximate expression for $E(T_Q)$ (Nasell, 1999), noting that the coefficient of variation of the number of infectious individuals in endemicity is given by

$$\frac{\sqrt{n\hat{\Sigma}_{22}}}{n\hat{y}} = \frac{\sqrt{n\left(R_0 - 1 + \frac{R_0^2}{\alpha}\right)}}{R_0}\frac{\alpha R_0}{n(R_0 - 1)} \approx \frac{\alpha}{\sqrt{n(R_0 - 1)}} \tag{2.13}$$

The last approximation is appropriate when $\alpha >> R_0^2$.

For a real-life disease in a community with heterogeneous mixing, extinction may be caused by a normal fluctuation from a not so high endemic level or by a large deviation from a high level. In other words, in the absence of a catastrophe that may wipe out the disease and the population, the extinction of the disease follows a gradual process.

Simulation results from Andersson and Britton (2000) also show that, for realistic parameter values, the Nasell (1999) formula gives a much better approximation to the observed time to extinction than does the formula derived by van Herwaarden and Grasman (1995).

The distribution of T_Q depends on the parameters $R_0, \alpha,$ and θ as well as the size of the population. The extinction times will increase as the basic reproduction number increases. In fact, a population that sufficiently reproduces itself will have a high value of R_0, which in turn, will produce higher extinction times.

2.4. Threshold of endemicity of the language

At endemicity, the system fluctuates minimally around the endemic level. For the process $\{X(t), Y(t), t \geq 0\}$, we are interested in the proportion of indigenous language speakers at time t. We introduce the random variable

$$Z(t) = \frac{Y(t)}{X(t) + Y(t)} = \frac{Y(t)}{N(t)}$$

At time t=0, it is assumed that there are no susceptible individuals and m language-speaking individuals, that is, the initial population had indigenous language ability. This implies that $Z(0) = 1$. The value of the random variable $Z(t)$ at the endemic level, ($\hat{z}(t)$) is the threshold beyond which the language becomes endangered, called the *threshold of endemicity*. Once this threshold is reached, there is a high probability that the language will become extinct in the population in finite time, in the absence of any external intervention.

The threshold of endemicity is given as

$$\hat{z} = \frac{\hat{y}}{\hat{x} + \hat{y}} = \frac{R_0 - 1}{\alpha + R_0 - 1} \tag{2.14}$$

Once a language gets to the endemic level, then the possibility of extinction from that point onwards can be very high. At the endemic level, there are small fluctuations before extinction. The endemic level is a somewhat stationary position of the process.

For language survivability, there should only be a small proportion of families in the language-free state.

2.5. Parameter estimation

In order to estimate the parameters of the SIR model of language extinction for a specified indigenous language, input parameters like the birth rate, life expectancy, language transfer rate, mean language-speaking period, mean family sizes, etc, have to be extracted from historical data or via a survey of the language. These parameters are the inputs needed to obtain the estimates derived in the previous section.

Using these input parameters, the basic reproduction number is computed, the expected time to extinction is also obtained, as well as all the other relevant metrics to ascertain the virility of the indigenous language, as described in the previous section.

3. Results and discussions

An indigenous language survey was conducted in some cities in Nigeria in 2016 and language use data was extracted for some tribes over two generations. The questionnaire was constructed such that a respondent could provide demographic information and details on the language ability of his siblings and children, if any. In addition, data on the fertility profile of Nigerian women and average life expectancy of Nigerians, obtained from the 2013 Nigeria Demographic Health Survey (NDHS) (NPC, 2014) were used as inputs to estimate the model parameters for the surveyed languages.

The surveyed languages were Yoruba, Igbo, Bini, Urhobo, Esan and Isoko. Apart from the major languages of Yoruba and Igbo, the other languages emanate mainly from Edo and Delta states in the Niger Delta region of Nigeria.

The questionnaire contained 19 brief questions with the goal of eliciting information on the basic demographic characteristics of the respondent, indigenous language use ability of the respondent’s sibling as well as the respondent’s children, if any. The questionnaire also contained questions relating to the possible reasons for lack of intergenerational transfer of language ability from parent to children. A total of 607 respondents provided language use data on themselves, their parents, siblings and their children, if any. Hence, the data contained information on over 5,000 persons across three generations.

Data from the indigenous language survey questionnaire were processed via the Statistical Package for the Social Sciences (SPSS) and R computing software to yield the relevant estimates of the conceptualized model parameters.

Table 1 is a summary of the estimates of the model input parameters for the group of parents and non-parents (first and second generation, respectively) surveyed in the various languages. Table 2 captures the estimates of the quasi-stationary distribution, time to extinction, and threshold of endemicity, for the two generations. The estimate of the mean lifetime was 54 years (NPC, 2014), and this was assumed to be the same for all the languages.

Table 1. Estimates of the model input parameters (mean family size n, birth rate μ, indigenous language transfer rate β, basic reproduction number R_0, mean duration of the language transmission period $1/\nu$, and the ratio mean lifetime to the mean period of language ability α) for the first (1st) and second (2nd) generations respectively

Language	**n**		$\boldsymbol{\mu}$		$\boldsymbol{\beta}$		$\boldsymbol{R_0(0.001)}$		$\boldsymbol{R_0}$		$\boldsymbol{1/\nu}$	$\boldsymbol{\alpha}$
	1st	**2nd**	**1st**	**2nd**	**1st**	**2nd**	**1st**	**2nd**	**1st**	**2nd**		
Igbo	8	5	5.97	2.82	5.61	1.80	4.22	1.35	3.83	1.23	25	3.16
Yoruba	8	5	5.84	3.04	5.39	1.44	4.00	1.80	3.64	1.64	26	3.08
Urhobo	9	5	6.27	3.04	4.43	1.44	3.33	1.08	3.03	0.98	25	3.16
Esan	9	6	6.20	3.78	5.57	2.61	4.08	1.91	3.71	1.74	27	3.00
Bini	9	5	7.09	3.00	6.46	1.50	4.68	1.09	4.25	1.00	28	2.93
Isoko	10	5	7.21	2.25	5.90	1.06	4.49	0.81	4.08	0.73	24	3.25
Others	10	5	7.11	3.17	6.29	1.83	4.61	1.34	4.19	1.22	27	3.00

Table 2. Estimates of the marginal quasi-stationary distribution ($q_{.,1}$), time to extinction (τ) and threshold of endemicity ($\hat{z}$) for the languages surveyed for the first (1st) and second (2nd) generations respectively

Language	$\boldsymbol{q_{.,1}}$		$\boldsymbol{\tau_n}$		$\boldsymbol{\tau(0)}$		$\boldsymbol{\tau}$		$(\hat{z})$	
	1st	**2nd**	**1st**	**2nd**	**1st**	**2nd**	**1st**	**2nd**	**1st**	**2nd**
Igbo	0.1286	0.2558	32.0	5.06	165.82	26.25	132.84	66.8	0.47	0.07
Yoruba	0.1275	0.2213	33.1	11.13	171.67	57.74	137.65	79.3	0.46	0.17
Urhobo	0.1245	0.3280	32.61	-	169.15	-	137.26	52.1	0.39	-
Esan	0.1104	0.1930	39.44	15.31	204.61	79.42	163.02	93.2	0.47	0.20
Bini	0.1022	0.3053	43.33	0	224.78	0	180.34	60.4	0.53	0
Isoko	0.1040	-	38.59	-	200.20	-	159.84	-	0.49	-
Others	0.0941	0.2524	45.68	5.41	236.96	28.06	191.34	71.3	0.52	0.07

The mean family sizes n, ranged from 5 to 10. In order to ensure that all the time variables were in the same unit of time, the corresponding annual birth rate based on the generational birth rates were computed and used for the subsequent calculations. The mean infectious period $1/\nu$ was taken as the difference between the mean lifetime and the mean age at marriage. It varied across the languages surveyed.

Since the number of households in Enumeration Areas (EAs), the basic sampling unit in Nigerian population censuses, ranged between 80 and 100, we chose a value of k=100, which is a reasonable value for the number of households in a typical community.

For the choice of minimal interaction between families, we chose a value of $\varepsilon = 0.001$, which is smaller than Lindholm and Britton (2007) choice of 1/365=0.0027 and yields $\Pr(within\ contact) = 0.91$, in comparison with Lindholm and Britton (2007) choice which yields $\Pr(within\ contact) = 0.79$.

The choice of the value of ε, the proportion of an individual's language interaction with other families was motivated by the desire to keep the probability of inter-family interaction below 0.1. That is, we desired that the probability of within family interaction should be around 0.9, or that 90% of the language interactions among children should be within their family, enforcing the assumption of intergenerational transfer of language ability from parent to children.

The basic reproduction number R_0 when $\varepsilon = 0.001$ was slightly higher than the corresponding value when $\varepsilon = 0$ for all the tribes. The values of R_0 were between 3.03 and 4.68 for the first generation data and between 0.73 and 1.80 for the second generation data.

The values of α, the ratio of life's length to the length of the infectious period, were between 2.93 and 3.25 among the tribes. We observed minimal variation in the values of α, pointing to the fact that similar conditions affect individuals in the population irrespective of their language.

The marginal distribution of 1 infectious individual in the family at quasi-stationarity, $q_{.,1}$, had 0.13 as its largest values in the first generation (corresponding to the Igbo language) and 0.33 for the Urhobo language in the second generation.

Due to the generational decline in the values of R_0, the expected time to extinction for the families, τ_n also exhibited declines between the two generations. Similarly, the same scenario is replicated for $\tau(0)$ and τ.

The threshold of endemicity $\hat{z}$ also showed sometimes sharp decline between both generations across the surveyed languages. However, as should be expected for languages on the lower fringes of the extinction scale, the declines were steeper. The Urhobo and Bini languages had the lowest threshold of endemicity.Higher values of $\hat{z}$ imply that the process will remain in endemicity for a longer time before it is absorbed, while lower values imply that the process will only spend little time until absorption.

Once the process gets to endemicity/ quasi-stationarity, then in finite time, in the absence of any external intervention, the process will be absorbed even with $R_0 > 1$.

For the language to remain virile, it should be far away from the quasi-stationary level and programmes and policies put in place to ensure steady growth of the language-speaking population.

The total fertility rate, which is the expected number of children an average Nigerian woman would have born at the end of the childbearing cycle (15–49 years) was about 5.5 (NPC, 2014). This value was not too far off from our survey data for all the tribes.

4. Conclusion

The SIR model of language extinction provides a useful approach to ascertaining the status of any language globally, with the use of historical or survey data. This conceptualization therefore provides a tool that can be deployed to document the status of the world's numerous indigenous languages.

On the basis of the expected time to extinction, conditioned on quasi-stationarity, several of the surveyed Nigerian languages were seen to be in precarious conditions, while a few others are seemingly virile based on a high language transfer quotient within families. While it is difficult to correctly predict the time to extinction due to the limited nature of the survey, the rapid decline in indigenous language ability among younger Nigerians point to the possibility of extinction as the older generation exit the population.

The threshold of endemicity, which reflects the proportion of the language-speaking population in stationarity, also point to decline and threat of extinction in relation to the younger generation across the surveyed Nigerian languages.

The goal of indigenous language stakeholders is to enhance the level of use of the indigenous language in both private and public domains. In essence, language practitioners will desire the value of ε to be very close to 1. That will signal the virility of any indigenous language

In the presence of historical data, the model could provide a good perspective of the indigenous language dynamics vis a vis demographic, economic and social changes. In such situations, the progression of any indigenous language could be tracked effectively, so that when the proportion of speakers fall below the allowed threshold, necessary short-term and long-term interventions could be made by governments and other stakeholders.

Areas of possible extension of the model framework are given below:

i. The model could be modified to accommodate greater heterogeneities (age, sex, location, etc), as there could be certain segments of the population with varying indigenous language transmission rates and residual lifetimes;

ii. a simulation scheme could be conceptualized and developed to sufficiently capture the dynamics of any language decline, using initial parameter estimates obtained from survey or historical data;

iii. the value of the global infection rate can be modelled as a function of k, the number of families in the population, as the present model produces estimates of R_0 that tend to be far higher as k increases;

iv. better approximations for the expected time to extinction could be designed when
there are smaller family sizes; and

v. models that adequately accommodate the transient behaviour of language decline over times could be developed to study the indigenous languages.

REFERENCES

ALLEN, L., BURGIN, A. M., (2000). Comparison of Deterministic and Stochastic SIS and SIR Models in Discrete Time. Mathematical Biosciences, 163, pp. 1–33.

ANDERSSON, H., BRITTON, T., (2000). Stochastic Epidemic Models and Their Statistical Analysis, Lecture Notes, Department of Mathematics, Stockholm University, Sweden.

BALL, F. G., LYNE, O. D., (2000). Stochastic Multitype SIR Epidemics among a Population Partitioned into Households, Adv. in App. Prob., 33, pp. 99–123.

BILLINGS, L., MIER-Y-TERAN-ROMERO, L., LINDLEY, B., SCHWARTZ, I., (2013). Intervention-Based Stochastic Disease Eradication, PLoS ONE 8(8): e70211.

BURR, T. L., CHOWELL, G., (2008). Signatures of Non-Homogeneous Mixing in Disease Outbreaks. Mathematical and Computer Modelling, 48, pp. 122–140.

CRYSTAL, D., (2000). Language Death, New York: Cambridge University Press.

DALEY, D. J., KENDALL, D. G., (1964). Epidemics and Rumours. Nature, 204 (4963), 1118.

DOBSON, A. P., CARPER, E. R., (1996). Infectious Diseases and Human Population History, BioScience, 46(2), pp. 115–126.

HAGENAARS, T. J., DONNELY, C. A., FERGUSON, N. M., (2004). Spatial Heterogeneity and the Persistence of Infectious Diseases. J. Theo. Bio., 203, pp. 33–50.

LEWIS, M. PAUL, GARY F., SIMONS, CHARLES, D. FENNIG, (eds.)., (2015). Ethnologue:
Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. www.ethnologue.com.

LINDHOLM, M., (2007). Stochastic epidemic models for endemic diseases: the effect of population heterogeneities. Research Report 2007:10, Licentiate thesis, Department of Mathematics, Stockholm University, Sweden.

LINDHOLM, M., BRITTON, T., (2007). Endemic persistence or disease extinction: the effect of separation into families, Theo. Pop. Bio., 72, pp. 253–263.

NASELL, I., (1999). On the time to extinction in recurrent epidemics. Journal of the Royal Statistical Society, B 61, pp. 309–330.

NASELL, I., (2002). Stochastic Models of Some Endemic Infections. Mathematical Biosciences, 179, pp. 1–19

NASELL, I., (2005). A new look at the critical community size for childhood infections. Theoretical Population Biology, 67, pp. 203–216.

National Population Commission, (2014). Nigeria Demographic and Health Survey 2013.

Published by National Population Commission (Nigeria) and ICF International, www.npc.gov.ng.

NUWER, R., (2014, May 6). Why We Must Save Dying Languages. Retrieved from www.bbc.co.uk.

OHIRI-ANICHE, C., (2014). More than 400 Nigerian Indigenous Languages Endangered,
Vanguard Newspaper, Nigeria, www.vanguardngr.com, 26-04-14.

SCHWARTZ, I. B., BILLINGS, L., DYKMAN, M., LANDSMAN, A., (2009). Predicting Extinction Rates in Stochastic Epidemic Models. Journal of Statistical Mechanics:
Theory and Experiment, www.stacks.iop.org/JSTAT/2009/P01005 .

VERDASCA, J. A., TELO DA GAMA, M. M., NUNES, A., BERNADINO, N. R., PACHECO, J. M., GOMES, M. C., (2005). Recurrent Epidemics in Small World Networks, J. Theor. Biol., 233(4), pp. 553–561.