University of Pennsylvania Working Papers in Linguistics

Volume 22
Issue 2 *Selected Papers from New Ways of Analyzing Variation (NWAV 44)*

Article 2

12-1-2016

Methods for Modeling Social Factors in Language Shift

Maya R. Abtahian

Abigail C. Cohn

Thomas Pepinsky

Methods for Modeling Social Factors in Language Shift

Abstract
In this paper we expand our understanding of language endangerment by shifting the focus from small language communities to minority language communities with speaker populations in the millions. We argue for a methodological shift toward examining language shift scenarios more broadly and quantitatively for two main reasons: 1) it is becoming increasingly clear that a large speaker population does not protect against language shift (Anderbeck 2013); 2) we need to make a distinction between the symptoms and the causes of language shift, where factors such as a dwindling number of child speakers should be seen as symptoms of language shift that are caused by other factors (Himmelmann 2010). In this paper we use Indonesia as a case study and analyze a sample of the 2010 census. We treat language choice as a sociolinguistic variable and analyze the correlation between six social factors and language choice (local languages vs. the national language, Indonesian). These results provide a starting point for creating more comprehensive models of the sociolinguistics of language shift.

Methods for Modeling Social Factors in Language Shift

Maya Ravindranath Abtahian, Abigail C. Cohn and Thomas Pepinsky*

1 Introduction

Language shift is the process by which the dominant language of a speech community is replaced by a new primary language of communication. Although stable multilingualism may be "part of the social fabric of everyday life for hundreds of millions of people the world over" (Sankoff 2001:638), stable multilingualism in many cases gives way to shift toward a few dominant world languages and a corresponding decrease in the world's language diversity (Krauss 1992), at least partly as a result of macro language policies that have marginalized minority languages (cf. Tollefson 1991) and implicitly or explicitly favored monolingualism. Many of these cases involve shift toward global languages of wider communication (LWC; Fishman et al. 1977) such as English and French, but a number of cases which warrant investigation involve shift toward what we might call "local LWCs" such as Hindi in India, Hausa in Northern Nigeria, and, in the case we discuss here, Indonesian in Indonesia.

In this paper we examine shift away from the local languages of Indonesia toward Indonesian (Bahasa Indonesia), using census data to analyze the effect of six social factors on shift away from local languages toward Indonesian. Although the dialogue on language endangerment worldwide has largely focused on languages with small speaker populations (Bradley and Bradley 2002, Florey 2010, Krauss 1992), we see a need for a methodological shift toward examining language shift scenarios more broadly and quantitatively for two main reasons. First, it is becoming increasingly clear that a large speaker population does not protect against language shift (Anderbeck 2013, Ravindranath and Cohn 2014). Second, we need to distinguish between the symptoms and the causes of language shift, where factors such as a dwindling number of child speakers should be seen as symptoms of language shift that are caused by other factors (Himmelmann 2010). In this paper we expand our understanding of endangerment by taking a quantitative approach to examining language shift in larger language communities with speaker populations in the millions.

2 Language shift in Indonesia

We chose Indonesia for this project partly because its large population and enormous linguistic diversity offer rich opportunities for multivariate analyses of language shift scenarios. Indonesia is one of the most multilingual nations in the world. Indonesian is the national language of Indonesia and the primary language of instruction in Indonesian schools. Its institution as a unifying, 'national' language is generally dated to the 1928 *Sumpah Pemuda* 'Youth Pledge'; since independence, Indonesian has been the sole official language, and the language of government, law, and education (Sneddon 2003). It is spoken alongside over seven hundred languages of Indonesia as part of a complex linguistic landscape that we do not address this paper (although we will describe how we mitigate this complexity in a variety of ways in our study). Two dimensions of this linguistic complexity that are relevant to this particular study are i) the fact that Indonesian is closely related to and in some cases mutually intelligible with some of these languages that are also Malayic; and ii) that there is a distinction made by most speakers between Indonesian that is *bahasa resmi* (formal) versus *bahasa sehari-hari* (colloquial; note that the latter term may also be used by some speakers for local languages, as Zentz 2014 points out for Javanese).

From previous studies we know that since Independence, knowledge and use of Indonesian is

*The authors would like to thank J. Joseph Errington, Daniel Kaufman, Naomi Nagy, John Wolff and the audience at NWAV 44 in Toronto for their helpful comments on versions of this paper. We would also like to thank the Fulbright Foundation for funding Cohn's research in Indonesia, Atma Jaya Catholic University for hosting Cohn during her recent sabbatical, and both the Mario Einaudi Center for International Studies at Cornell University and the Center for the Humanities at the University of New Hampshire for funding portions of this research.

increasing in the population at large. Figure 1 (from Musgrave 2014, based on discussion in Steinhauer 1994) demonstrates the precipitous increase in the proportion of the population claiming knowledge of Indonesian between the years 1971 on the left and 1990 on the right, based on three censuses. By 1990, more than 90% of respondents in the 10–49 year old age group claimed knowledge of Indonesian, where only two decades before that percentage was 65%. Moreover, there is some evidence that the shift to increased use of the national language appears to be happening at the expense of local languages. As Adelaar (2010) writes, "In spite of their large speech communities, the Javanese, Sundanese and Madurese languages are losing some of their domains of usage to Indonesian and are not always passed on to the next generation" (2010:25). Margaret Florey (2005) also points out that "restricting the definition of 'endangered language' to those languages with small speaker populations disguises the extent of the problem" (2005:59). It is this scenario that we are examining in a variety of ways in our ongoing project on language maintenance and shift in Indonesia. Our assumption is that as Indonesian displaces local languages, these languages face risks of endangerment, despite the fact that many of them currently have speaker populations in the millions.

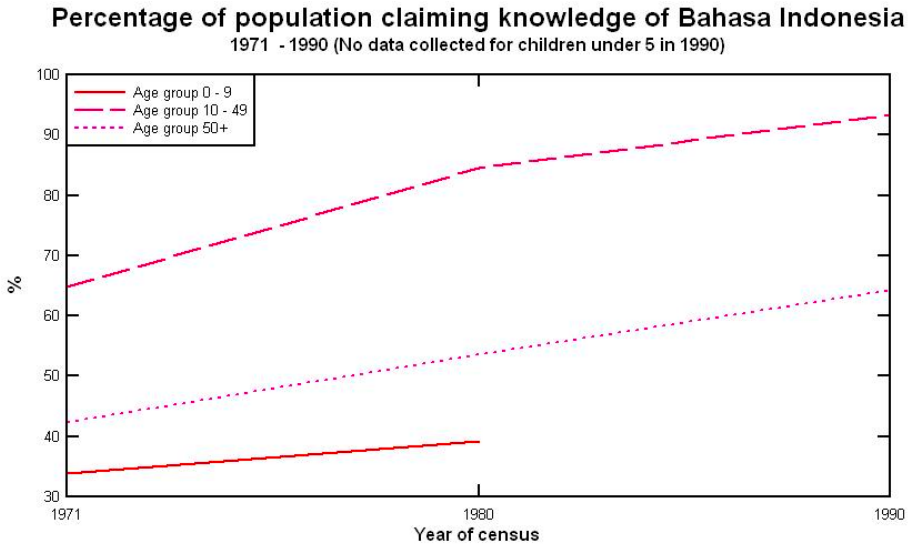

Figure 1: Increase in reported knowledge of Indonesian (Musgrave 2014, based on Steinhauer 1994).

2.1 Large language communities

Most of the language endangerment literature has focused on the examination of language shift scenarios at a very local level (Bradley and Bradley 2002, Dorian 1989, Florey 2010, Seliger and Vago 1991). Although a small speaker population is often equated with language endangerment, we see a small speaker population as a symptom, not a cause, of language death, and in fact size and vitality are not always closely correlated (Ravindranath and Cohn 2014). Rather, we argue that two approaches to the sociolinguistic study of language shift are necessary for understanding the process of shift. On the one hand, we need community-level approaches to understand the mechanics of language shift, that is, to examine more closely why speakers are making the language choices that they are. Yet we also need an approach that focuses on the language community with a more detailed quantitative analysis of many individual speakers in order to better understand the social factors that correlate with language shift. We feel that this type approach gives us a baseline that can help us go back to community level studies.

The approach to language shift that we take here follows from Himmelmann's (2010) description of a language endangerment scenario. As he writes, "it is rarely the case that one or two or three causes or factors lead to language endangerment. Instead, language endangerment results from the specific and complex constellation of a variety of such factors… an endangerment scenario" (2010:46). One useful benefit of studying larger languages is that their large and diverse

populations afford us the opportunity to undertake multivariate analyses of language endangerment scenarios that can consider several risk factors at the same time. The "big languages" that we consider in this study are all languages with speaker populations of at least one million.

3 Methods

We use a 1% sample of the 2010 Indonesian census, available through the Integrated Public Use Microdata Series (Minnesota Population Center 2014). The sub-sample of the census data that we are considering consists of the ten areas of Indonesia where a non-Malayic language with at least 1 million speakers is spoken. We have limited our sample in this way in order to avoid ambiguity due to dialect continua and labeling (with the varieties of Malay that are close to Indonesian). Our dependent variable is the answer to the question: What language does (RESPONDENT) use daily at home? ("*Apakah bahasa sehari-hari yang digunakan (NAMA) di rumah?*").

These language communities are the largest (non-Malayic) language communities in Indonesia. Each one is spoken in a province that has at least one major urban center, which allows us to compare urban vs. rural speakers. The language communities are geographically spread around the Indonesian archipelago, with communities in Java, Bali, Lombok, Sulawesi and Sumatra. This geographic spread also allows us to compare communities in the economically and politically central inner islands (Java, Bali and Lombok) to communities in two outer islands (Sulawesi and Sumatra). All of the language communities are listed in Table 1, along with their populations and EGID (Expanded Graded Intergenerational Disruption) Score (Lewis and Simons 2010), as cited in the Ethnologue (Lewis, Simons and Fennig 2016). These scores are listed in order to demonstrate how few of these languages are currently considered threatened by the wider linguistics community.

Language	**Province**	**Speaker population**	**EGIDS**
Javanese	Central, East Java	84.3 mil	2 Provincial
Sundanese	W. Java	34 mil	5 Developing
Madurese	Madura, E. Java	6.7 mil	5 Developing
Batak	N. Sumatra	5.5 mil	5 /6a Vigorous
Buginese	S. Sulawesi	5 mil	3 Wider Communication
Acehnese	Aceh	3.5 mil	5 Developing
Balinese	Bali	3.3 mil	5 Developing
Makassarese	S. Sulawesi	3.3 mil	6b Threatened
Sasak	Lombok	2.1 mil	5 Developing
Gorontalo	Gorontalo, Sulawesi	1 mil	6b Threatened

Table 1: Regional languages with over 1 million speakers, excluding Malayic varieties (Lewis and Simons 2010).

We can express the probability that a survey respondent speaks Indonesian using a logistic regression model, with the following formula:

$$\Pr(\text{Indonesian} = 1) = \frac{1}{1 + e^{-(a+bX)}}$$

Here, a and b are parameters to be estimated, and X is a vector of predictor variables that we hypothesize might affect language choice. The nonlinear functional form constrains the probability that a respondent reports speaking Indonesian to lie between 0 and 1, as is appropriate when modeling binary responses as probabilities.

We considered a number of independent social factors. First, we considered the predictive value of parent's first language on the first language of their children, as a measure of the success of intergenerational transmission (Fishman 1991). Then we looked at 6 primary social factors: age, urbanization, development index, education, religion, and gender, some of which we expected to be generalizable to other language shift scenarios and others that we specifically wanted to inves-

tigate in Indonesia. The choice of these social factors to investigate comes from previous work at the speech community level on the sociolinguistics of language shift and endangerment (e.g., Dorian 1981, Gal 1978), particularly in the Indonesian context (Errington 1998, Florey 2010, Kurniasih 2006, Setiawan 2012, Smith-Hefner 2009). There are of course complex relationships between many of these factors, and there are some important factors that this method leaves out, such as attitudes toward the different varieties and language use in different domains. These factors are actually the focus of other work that we are doing on this topic using a more local survey and interview data.

We find that all of these coefficients are significant at the .001 level. Outputs of the regression are not linguistically meaningful quantities. However, we can take the results of these and present them in a visual approach that gives the reader the power to assess the magnitude and direction of the effect, as presented in the following figures. We do this by transforming the logistic regression results into predicted probabilities for theoretically meaningful combinations of the predictor variables, and plotted them alongside one another.

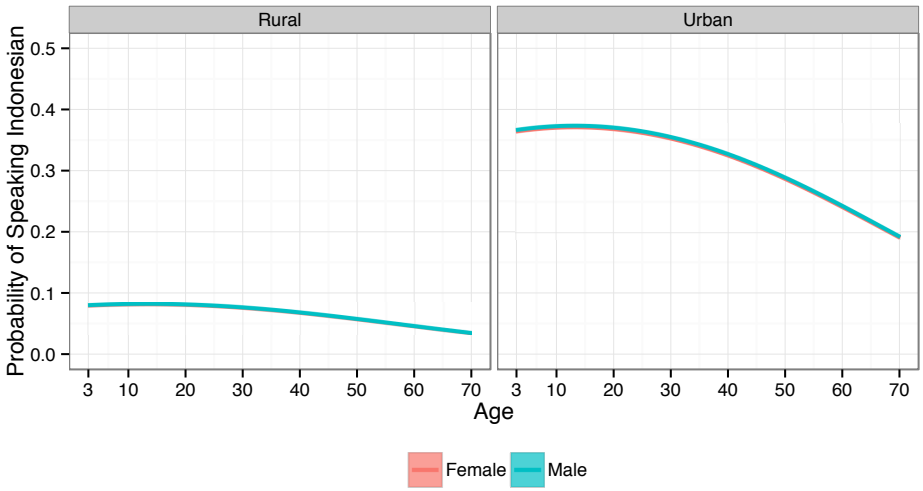

Figure 2: Baseline model for shift.

Figure 2 demonstrates our baseline model for shift. The Y-axis contains the predicted probability that a respondent speaks Indonesian for particular combinations of independent variables; in this figure age, gender, and urbanization, which a step-up/step-down logistic regression procedure always selects for inclusion. From this basic figure we can learn a number of things, including the following:

(1) Urban men and women are far more likely to speak Indonesian than rural men and women.
(2) Even at the maximum, 10- and 20-year olds still have less than a 40% chance of speaking Indonesian at home.
(3) The two lines representing predicted probabilities for women and men overlap nearly completely for both urban and rural respondents. Women and men do not differ much in their probability of speaking Indonesian (the difference is clearly miniscule compared to the magnitude of the other factors, although the regression shows it to be significant).

Our methodological approach involves a tradeoff between the rich qualitative insights that small-scale studies provide, and the panoramic overview of entire linguistic communities that statistical analysis of census data provide. Neither type of analysis is "correct"; in fact, both are required to understand both language shift from the micro- to the macro-level. However, quantitative analyses enable us to ask many different questions using the same dataset. It is unrealistic for small-scale qualitative studies to compare speakers across multiple demographic categories at once: urban and rural, high and low income, across the age range, for different ethnic groups. In

considering multiple factors at once, moreover, we are able to compare the *relative* strength of different factors (the effects of gender versus those of economic development, for example).

Another benefit of our quantitative approach using survey data is that it is cumulative: our findings are easily replicable using publicly available data. Future researchers may exploit census microdata from other countries to compare language shifts in other multilingual contexts. Or, they may probe the Indonesian data still further, examining other demographic factors that we have not considered here, or looking for higher order interactions and nonlinear effects across predictors.

4 Findings

4.1 Age

The examination of age allows us to measure the success of intergenerational transmission, and our data seems to demonstrate ongoing shift away from local languages and toward Indonesian. Moreover there is a lot we can learn from macro studies that is more nuanced than what you might expect, and this is demonstrated in our examination of age. For instance, we find some interesting non-linear effects by age in our Indonesian census data that are only likely to be apparent in a large sample size such as this. In Figure 3, for instance, we show that having one parent who speaks Indonesian has an effect on how likely the child is to speak Indonesian.

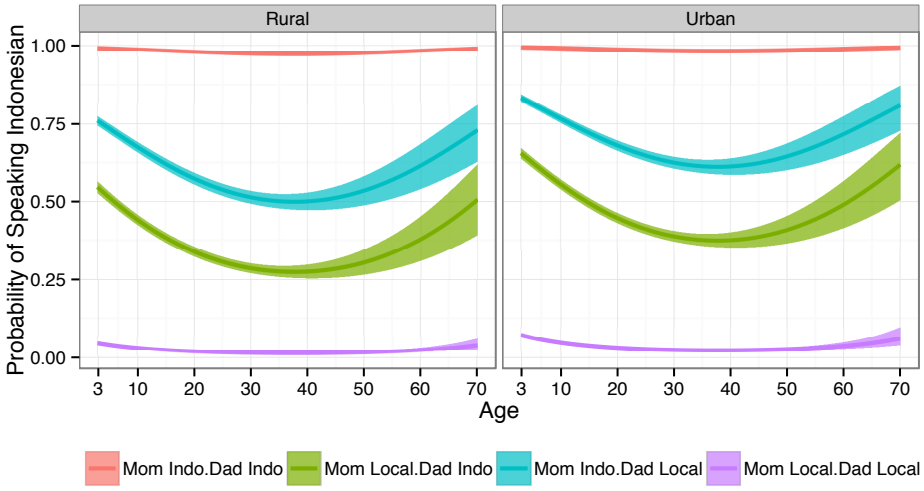

Figure 3: Age and parents' language.

A more nuanced result is that we show that the effect of having one parent who speaks Indonesian is actually higher among the oldest Indonesians than it is for middle-aged Indonesians. We hypothesize that this might reflect the residue of the uneven spread of Indonesian as a national language: those families where at least one parent spoke Indonesian four decades ago were probably particularly cosmopolitan or nationally-oriented at a time when this was still relatively rare. This finding of course would only come out with a large enough sample size to show these differences – we could not have gained this knowledge with a small sample.

4.2 Gender

Given previous work in Indonesia, we had reason to believe that gender should be a significant factor in language choice in Indonesia. Smith-Hefner's (2009) work in Central Java, for instance, shows her female university respondents were more likely than her male university respondents to report using Indonesian rather than Javanese, and in particular were more likely to report that they planned to use Indonesian with their children. The women that she interviewed reported preferring Indonesian over Javanese because it was a more egalitarian language. Interestingly however we find little effect of gender across a larger sample of Indonesians. How to reconcile these differ-

ences? Although community level studies do show an effect of gender, these effects may not be a result of gender per se, but of gender differences in social networks, access to education, and exposure to Indonesian. It may be, for example, that only within some middle class groups with higher educational aspirations do we find an effect of gender on language use. Nevertheless, by comparing the effects of gender to those of other demographic variables, we find that the gender differences that we do find in a large dataset pale in comparison to the effect of age or urban/rural.

4.3 Other social demographic factors

Three of our social factors we expect to be generalizable even beyond the Indonesian context, and these are urbanization, education, and development. To measure development, we created an index of socioeconomic development, which we term a "development index," that captures salient features of respondents' material conditions. The development index is an additive index of eight factors included in the census, including whether the respondent i) owns their own home; ii) has electricity; iii) has running water; iv) has a sewer; v) has a flush toilet; vi) has something other than a dirt floor (e.g., cement, wood, etc.); vii) owns a (non-mobile) phone; viii) owns a mobile phone. This index is not a proxy for class, income, or social status, but it does allow us to array respondents along a continuum of material "development" using indicators that are valid in the Indonesian context. The index runs from 0 to 8, with 8 being the highest level of material development.

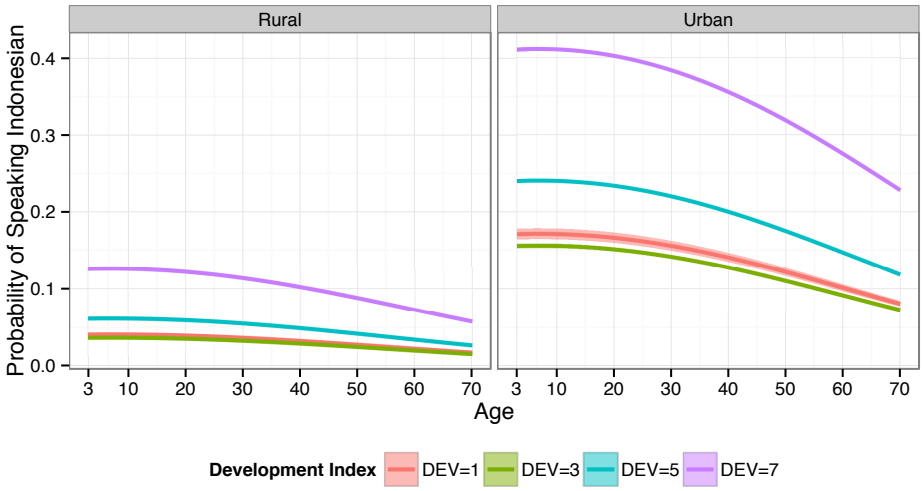

Figure 4: Language shift according to development index.

We show here Levels 1, 3, 5, and 7, where the predictive effects of development on Indonesian language use can be clearly seen (Figure 4). Moreover, while higher scores on the development index predict more Indonesian language use in both urban and rural areas, the substantive effect is much higher in urban areas and the effect of age is also greater in urban areas than rural.

Education (Figure 5) shows similar results. This is largely expected at the lower levels; since Indonesian is the language of national education, speakers will naturally have more exposure to, and be expected to use Indonesian more, the more education they have. However this effect holds even beyond primary and secondary education. Note that a 30 year old urban Indonesian with a postgraduate education has a 65% chance of reporting speaking Indonesian, compared to 28% for an urban speaker of the same age with a junior high education. We expect that there is an intersection here with both class and the perceived social, economic and cultural values of using Indonesian. Both development and education are effects that we expect are generalizable to other contexts globally.

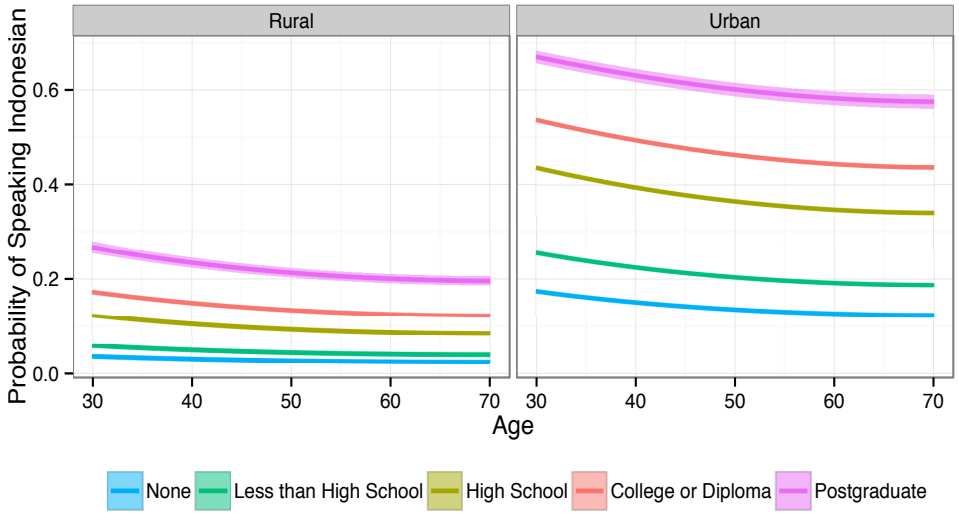

Figure 5: Language shift according to education level.

4.4 Ethnic group

Ethnicity, and correspondingly ethnolinguistic group, is an important part of group identity in Indonesia, as it is in many multi-ethnic and multi-linguistic societies. One of the interesting aspects of using Indonesia as a case study is that, with so many groups varying in size (and many ethnolinguistic groups with populations in the millions), we can examine ethnic group as a social factor independent of our other social factors.

Each one of our ten language community groups corresponds with one major province that has one major urban area, and each one of these is also associated with one major ethnolinguistic group. The division of our sample into "inner" island groups (in Java, Bali, and Lombok) and "outer" island groups (in Sulawesi and Sumatra) also becomes important in this analysis. First of all, we find that for the most part our results are similar across the ten different ethnic groups that we consider, which gives us some reason to suspect that these results may also be more generalizable to similar contexts in other countries. However, we find some interesting effects when we compare inner and outer island groups.

Figure 6 displays our baseline analysis, divided by ethnic group, and now looking at just urbanization and age. The groups are laid out by size of population, left to right and top to bottom. In every case we find that urban residents are more likely to speak Indonesian than rural residents. But more importantly, we find dramatic differences in the overall probability of speaking Indonesian by ethnic group, and we hypothesize that this is influenced by a few factors. First we see a general trend by size, where the languages with larger populations are more likely to be maintained. However that alone does not account for the ethnic group differences we see. Javanese, Sundanese, Madurese, Balinese and Sasak are all spoken in the inner islands. All of these communities have both more maintenance of the local language and less difference between urban and rural speakers. In contrast the languages spoken on the outer islands are uniformly further progressed in shift away from local languages. Four out of five of these still show a distinction between urban and rural speakers as well.

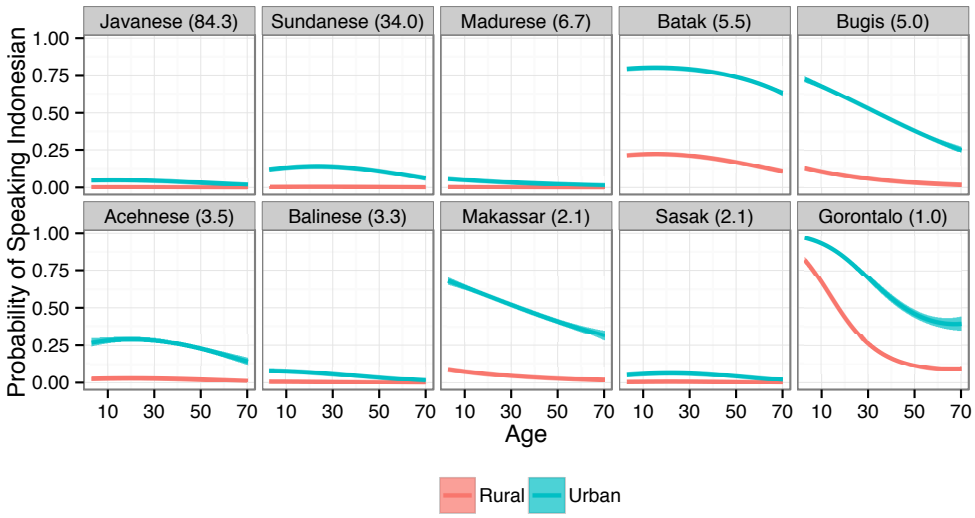

Figure 6: Ethnic group analysis.

The one exception to this is Gorontalo, the language in our analysis that is in the most vulnerable situation. We estimate that there is 75% chance that a ten-year old, rural, ethnic Gorontalo respondent living in Gorontalo province speaks Indonesian. We suspect that the vulnerability of Gorontalo may be related to factors beyond size or inner vs. outer island. One factor may be that it is the only one of these languages that is not written; but probably more important in our opinion is the point that Anderbeck (2015) makes, that since Gorontalo split from North Sulawesi province in 2000, many Gorontalos no longer feel the need to assert their distinct identity by using their language.

4.5 Religion

The last social factor that we consider is religion. Indonesia is a multi-religious state, and all Indonesians are required by law to be affiliated with one of six religions: Islam, Hinduism, Buddhism, Protestantism, Catholicism or Confucianism. Islam is the majority religion, with 87% of our sample coded as Muslims. Prior work suggests that religion is an important social factor in language choice in Indonesia, albeit one that is determined by other local factors.

In Figure 7 we have divided our sample into 20 groups by ethnicity and urban vs. rural. We find very strong evidence of differences in Indonesian language use by religion, and as with the other factors these are especially pronounced in urban areas. However we also find an interesting effect when we compare the ten different ethnic groups within our sample. When we examine religion by ethnic group, we see that *the predictive effect of religion depends on whether the religion is that ethnic group's majority religion*. Among Muslim-majority ethnic groups, non-Muslims are more likely to speak Indonesian, and when the majority ethnic group is not Muslim, Muslims are more likely to speak Indonesian. For example, Sundanese are a Muslim-majority ethnic group; among Sundanese, Muslims are more likely to speak the Sundanese, while minority Buddhists and Christians are more likely to use Indonesian. In contrast, if we look at Batak, where Christians form a small majority, Muslims are more likely to speak Indonesian. The same is true for Bali, where Hindus form the majority, and Muslims are more likely to use Indonesian.

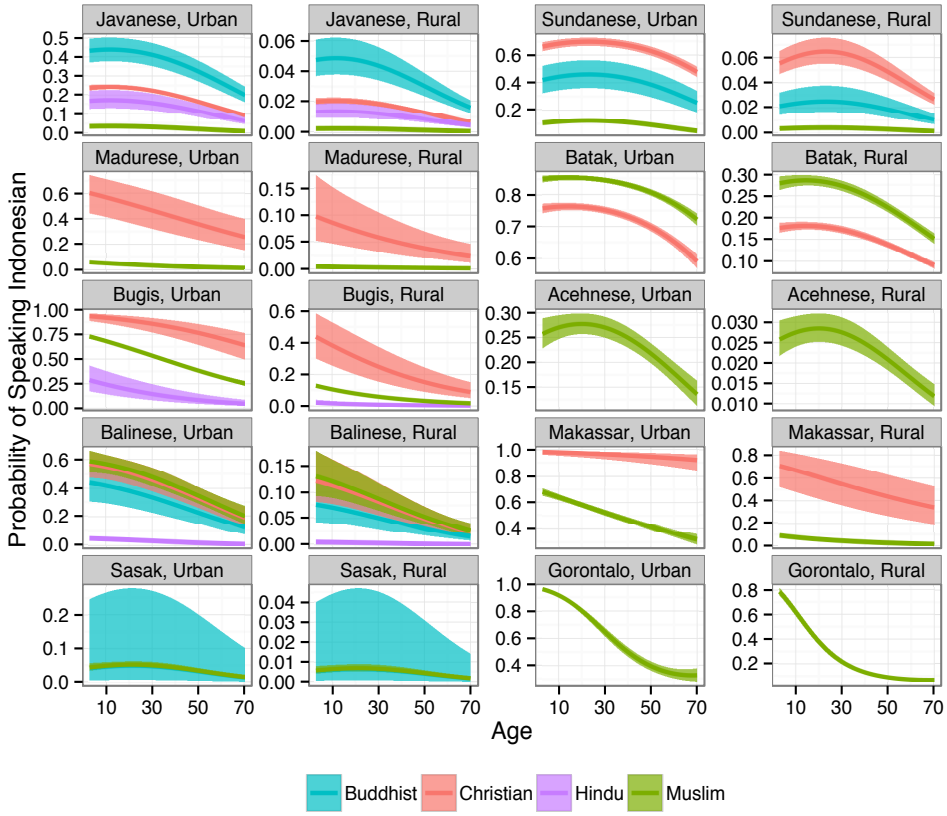

Figure 7: Religion by ethnic group.

5 Conclusion

As it becomes increasingly clear that a small population is a symptom of language shift and not a cause of language shift, and that even large languages experience language shift, multivariate quantitative studies of big languages undergoing shift allow us to create a better picture of language shift scenarios. That is, we can better model what it means to be female + rural + Muslim + Sundanese and what that means for language choice. We can also model other combinations of factors, and with great statistical precision. This is more difficult in a small or mid-level study, where either you mainly have rural Muslim Sundanese speakers, or you have only 3-5 speakers with any combination of social factors. Quantitative studies such as ours, in turn, may then provide a baseline for community level studies of shift, where local social factors, domains of use, and language attitudes may be more closely examined and triangulated by observations of speakers' actual language use. In our future work (see also Abtahian, Cohn and Pepinsky, to appear), we will continue to refine methods for multivariate analysis of language shift scenarios, and begin to narrow our lens on the different groups in Indonesia through surveys and interviews.

References

Abtahian, Maya, Abigail C. Cohn and Thomas Pepinsky. To appear. Modeling social factors in language shift. *International Journal of the Sociology of Language.*

Anderbeck, Karl. 2012. Portraits of Indonesian Language Vitality. Ms. URL https://sites.google.com/site/nusantaralanguagevitality/.

Fishman, Joshua, Robert Cooper and Andrew Conrad, eds. 1977. *The Spread of English.* Rowley, MA: Newbury House.

Florey, Margaret. 2005. Language shift and endangerment. In *The Austronesian Languages of Asia and Madagascar,* ed. A. Adelaar and N. Himmelmann, 43–64. New York: Routledge.

Himmelmann, Nikolaus P. 2010. Language endangerment scenarios: A case study from northern Central Sulawesi. In *Endangered Languages of Austronesia,* ed. M. J. Florey, 45–72. Oxford: Oxford University Press.

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68:4–10.

Kurniasih, Yacinta. 2006. Gender, class and language preference: A case study in Yogyakarta. In *Selected Papers from the 2005 Conference of the Australian Linguistic Society*, ed. K. Allan, 1-25.

Lewis, Paul M. and Gary F. Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique* 55:103–20.

Lewis, M. Paul, Gary F. Simons and Charles D. Fennig, eds. 2016. *Ethnologue: Languages of the World, Nineteenth Edition*. Dallas, Texas: SIL International. URL www.ethnologue.com. Date accessed: October 2015.

Minnesota Population Center. 2014. Integrated Public Use Microdata Series, International: Version 6.3. Minneapolis: University of Minnesota. URL https://www.ipums.org. Date accessed: October 2015.

Musgrave, Simon. 2014. Language shift and language maintenance in Indonesia. *Language, Education and Nation-building*, 87-105. Palgrave Macmillan UK.

Ravindranath, Maya and Abigail C. Cohn. 2014. Can a language with millions of speakers be endangered? *Journal of Southeast Asian Linguistics* 7:64–75.

Sankoff, Gillian. 2001. Linguistic outcomes of language contact. In *Handbook of Sociolinguistics,* ed. P. Trudgill, J. Chambers and N. Schilling-Estes, 638–668. Oxford: Basil Blackwell.

Setiawan, Slamat. 2012. *Children's language in a bilingual community in East Java*. Doctoral Dissertation, The University of Western Australia, Perth.

Smith-Hefner, Nancy. 2009. Language shift, gender, and ideologies of modernity in Central Java, Indonesia. *Journal of Linguistic Anthropology* 19:57–77.

Sneddon, James N. 2003. *The Indonesian Language: Its History and Role in Modern Society.* Sydney: University of NSW Press.

Steinhauer, Hein. 1994. The Indonesian language situation and linguistics: Prospects and possibilities. In *Bijdragen tot de Taal-, Land- en Volkenkunde, 150 Volumes of Bijdragen: A Backward Glimpse and a Forward Glimpse 150*, 755–784.

Tollefson, James. 1991. *Planning Language, Planning Inequality*. New York: Longman.

Zentz, Lauren. "Love" the local, "use" the national, "study" the foreign: Shifting Javanese language ecologies in (post-)modernity, postcoloniality and globalization. *Journal of Linguistic Anthropology* 24:339–359.

Maya R. Abtahian
Department of Linguistics
University of Rochester
Rochester, NY 14627
maya.r.abtahian@rochester.edu

Abigail C. Cohn
Department of Linguistics
Cornell University
Ithaca, NY 14853
acc4@cornell.edu

Thomas Pepinsky
Department of Government
Cornell University
Ithaca, NY 14853
pepinsky@cornell.edu