ARE THE WORLD'S LANGUAGES CONSOLIDATING? THE DYNAMICS AND DISTRIBUTION OF LANGUAGE POPULATIONS*

Short Title: Are the World's Languages Consolidating?

David Clingingsmith

Scholars have conjectured that the return to speaking a language increases with the number of speakers. Long-run economic and political integration would accentuate this advantage, increasing the population share of the largest languages. I show that, to the contrary, language size and growth are uncorrelated except for very small languages (<35,000 speakers). I develop a model of local language coordination over a network. The steady-state distribution of language sizes follows a power law and precisely fits the empirical size distribution of languages with $\geq 35,000$ speakers. Simulations suggest the extinction of 40% of languages with <35,000 speakers within 100 years.

If as one people, speaking the same language, they have begun to [build this tower], then nothing they plan to do will be impossible for them. Come, let us go down and confuse their language so they will not understand each other.

—The Tower of Babel, Genesis 11:6 (New International Version)

Paulina Sugiarto's three children played together at a mall here the other day, chattering not in Indonesia's national language, but English... 'They know they're Indonesian,' Ms. Sugiarto, 34, said. 'They love Indonesia. They just can't speak Bahasa Indonesia. It's tragic.'

—*New York Times*, July 25, 2010

Why don't all people speak the same language? At least since the book of Genesis, scholars have puzzled over the diversity of spoken languages. In the story of the Tower of Babel, the descendants of Noah all speak a common language in the aftermath of the great flood. They decide to build a city with a great tower that stretches to heaven as an expression of their collective strength. Yahweh, the God of the Hebrew Bible, is displeased by this challenge and recognizes

*Corresponding author: David Clingingsmith, 11119 Bellflower Road, Room 281, Cleveland, Ohio, USA. Email: david.clingingsmith@case.edu. I thank Rachel Griffith (the editor) and two anonymous referees for many helpful suggestions. This paper also benefitted from conversations with Inge-Marie Eigsti, Silke Forbes, Scott Shane, Matt Sobel and Mark Votruba. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. The Ohio Supercomputer Center provided a grant of computing resources.

that a common language is essential to the endeavour. To blunt their power, Yahweh fragments the people into many groups each speaking a different language.

The ancient writers of this story clearly recognized that languages have increasing returns to scale. Differences in language create costly barriers to the flow of information and to economic exchange. The theory of network externalities in economics, while primarily concerned with technologies such as operating systems, media players, and telecommunications, also applies to languages (Katz and Shapiro, 1985, 1986; Lazear, 1999; Klemperer, 2008). Joining a network produces a positive externality grounded in a standardized way of exchanging information. The theory of network externalities argues that networks having increasing returns to size.

Increasing returns to size suggests the number of languages should be small. The rise of English as the world's *lingua franca* is often linked to the benefits of speaking a widely known language. English has 335 million native speakers and more than 1.2 billion second language speakers (Crystal, 2003). Since learning two languages is more costly than one, surely some children of those 1.2 billion non-native English speakers will learn English as their native language. English therefore ought to grow relative to those other languages. If this is true for English, why not for Mandarin, with its 178 million second language speakers, or Russian, with its 110 million? What about Spanish or the several large languages of India? This logic suggests increasing returns will lead to a consolidation of languages.

In this paper, I investigate to what extent the world's languages are consolidating. I build a theoretical model of language size in which agents interact over a locally connected network and can choose their language. The model equilibrium shows that increasing returns are compatible with a large number of languages. I develop empirical evidence on the growth and size of languages that fits the model's predictions well.¹ Only languages with less than 35,000 speakers appear to be consolidating at present. However, the model also predicts highly non-linear changes in the distribution as agents become more deeply interconnected.

Between 6,000 and 7,000 distinct languages are presently spoken as a mother tongue. Figure

¹Programs and datasets used in the analysis are available online at this journal's web site.

1 graphs the population size of 6,210 languages using data from the World Language Mapping System (WLMS, 2011). The size distribution is strongly right-skewed. Language sizes span nine orders of magnitude (panel a). While the median language has only 10,000 speakers, the 16 largest languages are spoken by fully half of the world population (panel b). The mean language size is 120 times larger than the median language.

Existing research on language consolidation proposes that increasing returns to language size favours large languages (Grin, 1992; de Swaan, 1993; Laitin, 1993; Church and King, 1993; Fishman, 1998; Van Parijs, 2000; Choi, 2002; Van Parijs, 2005; Wickstrom, 2005; Van Parijs, 2010). The main forces these scholars argue are opposing increasing returns are protection by the state, social prestige, and cultural vitality. They argue that since 1) state protection is difficult to acquire and 2) prestige and vitality are relatively weak forces, the few dozen largest languages have a very strong advantage over the rest. The overwhelming dominance of these few dozen we see in the size distribution of languages is taken as evidence of consolidation in action. Their general view is that over the next hundred years, only a few dozen large languages and those smaller languages that have official state protection will survive. Perhaps 350 languages currently fall into this group. However, the existing evidence offered for an ongoing consolidation is primarily based on cases, most compellingly the rise of large European languages such as French and English. The question of language consolidation is ripe for study using population-based statistical methods.

The theory of language consolidation has a simple implication that has yet to be tested directly: large languages must be growing more slowly than small languages. The first task of this paper is to assemble and analyse a new dataset of language population sizes and growth rates. National censuses are the best source for this data. I collected and tabulated census data from 15 countries covering 334 languages, some over multiple periods, for a total of 628 growth observations. This is a selected sample, so we need to use caution in interpreting the results, but it offers significant advantages over the existing evidence.

The new data shows that language population growth is actually *independent* of language size for languages with more than 35,000 speakers. There is no size trend in either the mean or variance

Figure 1: *Size Distribution of the World's Languages*

Notes: Language size is defined as the number of people who speak it as their first language or mother tongue. Panel (a): Kernel density estimate of the log population for 6,210 languages. Dashed lines show the $\mu \pm 2\sigma$ tails of a lognormal fit to the data. Bandwidth computed using Silverman's plug-in estimate (Cameron and Trivedi, 2005). Panel (b): Cumulative population by language rank. Data is from the World Language Mapping Survey (WLMS, 2011). Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

of growth rates above this level. Below 35,000 speakers, the growth rate is negatively correlated with size. OLS regressions show the relationship is robust to controlling for time-varying country characteristics. There are around 1,900 languages worldwide with more than 35,000 speakers, a

much larger viable set than many have feared.

When the size and growth rate of entities are independent, they are said to follow Gibrat's law, after the French economist who took note of the relationship in his study of inequality (Gibrat, 1931; Sutton, 1997). Gibrat's work, along with that of Zipf (1949) and Champernowne (1953), spawned a large literature on how the sizes of socio-economic aggregates are determined and what distributions they follow that I draw on in the present study. The most studied aggregates are firms, cities, and nations.²

The next task of the paper is to reconcile the evidence that languages follow Gibrat's law with the compelling idea that languages have increasing returns to size. How can increasing returns be consistent with an equilibrium in which language sizes range over four orders of magnitude? To answer this question, I build a model of language choice in which interaction agents benefit from having a common language. Interaction happens among agents who are connected in a network. Agents know the languages spoken by their potential interaction partners and can occasionally revise the single language they speak. The key insight that generates a diversity of spoken language despite increasing returns comes from thinking about how this network should be connected. Direct human interaction is primarily local. Even in the age of telecommunications, most of an individual's economic activity is conducted with other people who live relatively near them. Agents in the model play a coordination game with their local neighbours in geographic space. In equilibrium, any two agents who can reach each other by traversing the network will speak the same language. Agents whose parts of the network are isolated from each other will speak different languages. These isolated but internally connected parts of the network are called connected components. The size distribution of languages is the same as the size distribution of connected components of the network.

As a concrete example of such a language network, imagine two large, circular islands that have uniform, dense populations and radius r . The edges of the two islands are separated by d .

²Key papers include Axtell (2001), Luttmer (2007) and Saichev *et al.* (2009) on firms, Krugman (1996), Gabaix (1999), Eeckhout (2004), Gabaix and Ioannides (2004), Cordoba (2008), and Rybski *et al.* (2009) on cities, and Alesina and Spolaore (1997), Alesina and Wacziarg (1998), Alesina and Spolaore (2003), and González-Val and Sanso-Navarro (2009) on nations.

Suppose people are linked in the language network only with neighbours, defined as those who live within $1/4r$ of their location. Given a dense population, it is easy to see how, even though most of an island's population will not be directly connected to a given individual, any other individual on that island will be indirectly reachable from that individual by traversing the links of the network of neighbours. Suppose all individuals begin speaking their own language. The model predicts that in equilibrium all inhabitants of each island will share a single language. The situation between islands depends on how far apart they are. If $d < 1/4r$, contact between the shore dwellers on each island will link the networks, and the same language will be spoken on both islands. If they are further apart, there will be no contact and different languages will be spoken on each island.

Simulations of locally connected random networks reveal that the distribution of connected component sizes, and hence the distribution of language sizes predicted by the model, is highly skewed. Depending on the maximum distance between linked agents and how clustered agents are in space, mean sizes range from many to tens of times larger than the median. Increasing returns are thus theoretically consistent with a highly skewed distribution of language size when interaction is local. Consolidation does not follow inevitably from increasing returns because what matters to individuals is what their neighbours speak, not which language has the greatest return overall.

The double Pareto (DP) distribution provides the best fit to the simulated data among a set of right-skewed alternatives. The DP is defined for sizes above a lower threshold s_0 . On a plot of log size against log rank, the DP produces a piece-wise linear shape with slopes α and β on either side of a threshold s^* . The double Pareto is related to the simpler single-sloped Pareto or power law distribution most commonly fit in the size distribution literature. Figure 2 shows a rank-size plot for languages in the WLMS with more than 35,000 speakers. The language data shows an approximately linear shape from $10^{4.5}$ to $10^{7.5}$, or three orders of magnitude. Above this level, a steeper but still linear relationship appears to hold.

I fit the DP distribution to the WLMS data on language size using maximum likelihood. Since the DP is defined only for sizes larger than a lower threshold s_0 , I must specify a value of s_0 rather

Figure 2: *Rank-Size Distribution of Larger Languages*

Notes: Plot of rank and size for the 1,868 languages with population above 35,000. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text. And small amount of noise has been added to make the mass more evident.

than find it as a parameter. I use two methods for determining s_0 that both are in rough agreement. The first method finds the value of s_0 for which the DP is the best fit to the data by minimizing the Kolmogorov-Smirnov (KS) D -statistic. This gives me $\hat{s}_0=17,900$. The second method supposes that the data follows the lognormal distribution below s_0 and the DP distribution above s_0 . I obtain the estimate $\hat{s}_0=26,900$ by maximizing the likelihood of this composite distribution.

I settle on a conservative choice of $s_0=35,000$. The fit to the data is precise. A KS test fails to reject the null hypothesis that the data is drawn from the DP distribution with $p=0.95$. I estimate $\hat{\alpha}=0.342$ (s.e. 0.023) below a threshold of $\hat{s}^*=863,400$ (s.e. 17,390) and $\hat{\beta}=0.663$ (s.e. 0.026) above it.³ These scaling exponents are smaller than those typically found for the distribution of firm and city sizes, which are closer to 1. This implies that the distribution of languages is more

³Gomes *et al.* (1999) also found a two-regime scaling law using this data, though they did not statistically estimate the coefficients or thresholds. They report coefficients of 0.5 and 1.0 for the regimes, which it appears they eyeballed as there is no discussion of estimation. (Wichmann, 2005), perhaps unaware of this earlier work, claims that no scaling relationship exists, but seems to have considered only a single-regime law and did not consider the issue of thresholds.

extreme than firms and cities.

A lower threshold for equilibrium of approximately $s_0=35,000$ emerges from both 1) fitting the equilibrium size distribution from the model to a comprehensive dataset and 2) directly observing the regime in which growth and size are uncorrelated in the selected census data. One implication of this result is that we can think of 35,000 as a rough measure of the current minimum viable size of a language. I explore the implications of my results for language extinction through simulation. I apply the size-growth relationship estimated from census data to the current size distribution of languages. I use 100 speakers as my extinction cutoff. The simulation runs for 100 years. In the average run, 1,608 languages go extinct, which is about 26% of the current total. This is close to the estimates produced both by an intensive risk analysis conducted by UNESCO (Moseley, 2010) and an application of criteria used to assess animal endangerment to languages (Sutherland, 2003).

While all the empirical evidence points to a language size distribution in equilibrium above 35,000 speakers, the model simulations reveal an important non-linearity in how the size distribution changes with the size of agents' local neighbourhood. Over a narrow range of neighbourhood sizes, the population share of the largest language jumps from less than 10% to more than 90%. Network theory refers to this phenomenon as the emergence of a *giant component* and it is a feature of many network structures. It is also called a *phase transition*. The largest human language is Mandarin, with approximately 13% of the human population. The model suggests it is possible that the technological changes that increases the distance over which humans communicate and the number of connections they have could spark a phase transition in which Mandarin would rapidly come to dominate. The model excludes several factors, including political boundaries, bilingualism, ethnic preferences, and official language policy, that work against the phase transition.

The remainder of this paper is organized as follows. Section 1 produces some initial empirical estimates of language growth from census data and shows that Gibrat's law applies to languages. Section 2 provides background on how language evolution is related to local interaction. I develop the model in section 3 and derive the steady state size distribution. Section 4 fits the distribution to data on all languages and performs tests. Section 5 simulates the extinction of languages over the

next 100 years. Section 6 provides further discussion of the results and their implications for the evolution of language and language policy.

1 Gibrat’s Law for Languages

If languages are consolidating, then there must be a positive correlation between the population growth rates of languages and their size. This section presents evidence from population censuses that shows growth rates and sizes are in fact independent for all but the smallest languages. This relationship is known as Gibrat’s law. The evidence suggests that, over a large range, the size distribution of languages is in equilibrium.

I collected data on the sizes and growth rates of languages from national censuses for 14 countries.⁴ I believe this is the first such dataset to be assembled. National censuses are the only source of high-quality information about how language populations change, though only a fraction of censuses have asked comparable questions about language across multiple years. My principal sources are the IPUMS census microdata archive and the published volumes of the Census of India (India, 2008; Minnesota Population Center, 2011). The IPUMS contains usable data on language population for 14 countries. Each of these censuses contain variants of the question: ‘What was the language you first learned in childhood?’ My population estimates are based on this question. The Census of India is the only national census to collect information on language over a long span of time.

While I have the best data available, it is nevertheless a selected sample. We should proceed with this caveat in mind. A country is more likely to appear in my sample if 1) it has a census that releases microdata to IPUMS and 2) the languages spoken by the population are a matter of policy interest. Such countries will tend to be richer and more linguistically diverse, biasing down the growth rates among small languages. This leads to a bias in favour of finding a positive correlation.

The dataset contains 628 average annual growth observations for 334 languages. Population

⁴The countries are Bolivia, Canada, Mexico, Peru, Hungary, India, Indonesia, Mali, Morocco, Philippines, Romania, Senegal, South Africa, and Switzerland.

Figure 3: *Gibrat's Law for Languages*

Notes: The circles plot the average annual growth rate of a language across two census years against its population. Lines plot estimates of the conditional mean and standard deviation from local-linear kernel regressions along with 99% bootstrapped confidence intervals. The data come from population censuses for Bolivia, Cambodia, Canada, India, Mexico, Peru, Hungary, Indonesia, Mali, Morocco, Philippines, Romania, Senegal, South Africa, and Switzerland. (The Indian data is only available for languages with more than 10,000 speakers.) Base years range from 1970 to 2000: Population is adjusted to 2000 using country-specific growth rates calculated from the WDI database (World Bank, 2012).

estimates from the microdata samples employ the appropriate weights for each sample. Base years for the growth rates range from 1970 to 2000. Censuses in linguistically diverse countries, such as India and the Philippines, occasionally adjust the category labels used for different languages. For example, a language may be reported by its dialects in some years and not others, or a language with several names may be labelled differently depending on the year. In preparing the data, I check the coding of languages against the Ethnologue database to ensure consistency across census years (Lewis *et al.*, 2014). Ethnologue uses the standard ISO 639-6 language codes and reports relationships among languages, alternative names, and other useful information.

I first present plots of the data in Figure 3. I have adjusted the population sizes to the year 2000 using country-specific population growth rates from the WDI database (World Bank, 2012).

Growth rates and population size appear to be uncorrelated. The Spearman rank correlation of growth rate and size is a very small $\rho = -0.05$. I fail to reject the null hypothesis that growth rate and size are independently distributed with $p = 0.22$.

I compute an estimate of the mean growth rate conditional on size using a local-linear kernel regression (Cameron and Trivedi, 2005). The estimates are shown as a solid line. Dashed lines provide 99% confidence intervals computed using the wild bootstrap method (Davidson and Flachaire, 2008). The mean rises with population size until $\approx 10^{4.5}$, or about 35,000, and shows no trend after. I cannot reject a fixed mean growth rate for languages larger than 35,000. The data is sparse and confidence intervals wide for size below 10^4 because the Indian census only tabulated growth rates for languages above that size. There are also more outliers among small populations, some of which is likely due to sampling error.

1.1 Controlling for Country Characteristics

There are many economic and social processes that determine the growth rate of a population. The null correlation shown in Figure 3 could be misleading if demographic or economic factors related to language population growth are also related to the size of languages. To explore this possibility, I conduct a regression analysis of language population growth that includes country demographic and economic characteristics along with language population size. The results are shown in Table 1. The analysis shows that Gibrat's law still holds when we condition on these variables.

Column 1 shows the bivariate correlation between the natural log of language population and the language population growth rate. An increase of one in log language population, which corresponds to the level increasing by 2.7 times, leads to an increase in the language growth rate of 0.2 percentage points. In column 2, I capture the non-linearity revealed in Figure 3 by introducing a dummy for languages smaller than 35,000 and its interaction with log language population. As expected, this column shows that the positive correlation between log language population and the growth rate exists primarily for these smaller languages.

I introduce the natural log of country population and its growth rate into the regression in

columns 3 and 4. Population growth differs substantially across countries with different levels of economic development and at different stages of the demographic transition, and is an important potential confounding factor. Both of the added variables have strong positive associations with language growth. The coefficient on log language population falls in size and becomes indistinguishable from zero in the linear specification (column 3). Coefficients on language population variables in the non-linear specification do not change much with the country population controls (column 4).

I next add the natural log of GDP per capita in 2005 U.S. dollars and its growth rate in columns 5 and 6. These show no significant relationship to language population growth and their addition does not change any other coefficients. Finally, I add additional controls for the dependency ratio, life expectancy, urban share, and log population density in columns 7 and 8, also with no noteworthy changes on the coefficients of interest.

In summary, while added demographic and economic controls have sensible relationships with language population growth, they do not alter the conclusion that language size and growth are uncorrelated for languages with more than 35,000 speakers.

Table 1: *Gibrat's Law and Country Characteristics: Determinants of the Language Population Growth Rate*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ln Language Population	0.002** (0.00)	-0.001 (0.00)	0.001 (0.00)	-0.001 (0.00)	0.001* (0.00)	-0.001 (0.00)	0.001 (0.00)	-0.001 (0.00)
Ln Lang. Pop. $\times$ Language < 35,000		0.012*** (0.00)		0.011** (0.00)		0.011** (0.00)		0.012*** (0.00)
Language < 35,000		-0.114*** (0.04)		-0.103** (0.04)		-0.111** (0.04)		-0.113*** (0.04)
Ln Country Population			0.002*** (0.00)	0.002*** (0.00)	0.002** (0.00)	0.002*** (0.00)	0.001 (0.00)	0.003* (0.00)
Country Pop. Growth Rate			0.721*** (0.23)	0.661*** (0.23)	0.701*** (0.26)	0.718*** (0.26)	0.961*** (0.34)	0.633*** (0.32)
Ln GDP/Capita, 2005					-0.001 (0.00)	0.000 (0.00)	-0.005 (0.00)	-0.004 (0.00)
GDP/Capita Growth Rate					-0.034 (0.04)	-0.020 (0.04)	-0.043 (0.05)	-0.020 (0.05)
Constant	0.995*** (0.01)	1.025*** (0.01)	0.219 (0.23)	0.315 (0.23)	0.293 (0.27)	0.273 (0.27)	0.073 (0.32)	0.351 (0.31)
Additional Controls	N	N	N	N	N	N	Y	Y
Adjusted R-squared	0.014	0.071	0.054	0.099	0.060	0.113	0.057	0.111
Observations	635	635	635	635	612	612	612	612

Notes: Additional controls are the dependency ratio, life expectancy, urban share, and log population density. Robust standard errors in parentheses. Stars indicate coefficient statistically different from zero at the 10% (*), 5% (**), or 1% (***) levels.

2 Language Evolution and Local Interaction

When Gibrat's law holds, the size distribution of languages is in equilibrium. The next task is to reconcile this fact with the compelling idea that languages have increasing returns to scale just like other networked technologies. I develop a theory of language population dynamics that treats language choice as a coordination problem. I then fit the equilibrium size distribution predicted by the model to data on the actual size distribution of languages. Before proceeding to the model, I present some general facts about language that underlie several key assumptions.

A human language is a coding system that allows information to be transmitted between individuals. Two individuals who share the same code, or whose codes are sufficiently similar, can speak to and be understood by each other. No master copy of this code exists. This is an important way in which languages differ from the networked technologies more commonly studied by economists. Instead, a copy of the code resides in each individual and is transmitted from parent to child in early childhood. From this perspective, languages have deep parallels to genes. Unlike genes, however, individuals update their knowledge of a language during their lifetimes. They learn new words and adopt new grammatical constructs from the individuals with whom they speak. Such linguistic mutations will propagate across populations that are in regular contact. When there is a barrier to contact, differences in the linguistic code can accumulate between two populations. A new language arises when a substantial differentiation has taken place. What counts as substantial is a matter of judgement, but a bright line is usually drawn when separated populations can no longer communicate easily with one another.

Over the long span of human history, social contacts have been overwhelmingly local in nature. The vast majority of linguistic interaction is direct and involves people who live nearby, such as family members, neighbours, friends, work colleagues, and service workers. The rise of telecommunications, beginning in the 19th century, has made non-local indirect communication increasingly common, particularly in high-income countries where it is relatively cheap. Yet even interaction via telecommunications tends to be local relative to the human population as a whole.

There are few people whose economic lives transcend geographic space.

Barriers that constrain local direct contact include land features such as mountains, marshes, deserts, rivers, lakes, and oceans, as well as political boundaries. Languages tend to remain coherent within barriers and to drift apart across them. The important implication is that, given the drifting of language codes when there is no contact, speakers of a given language tend to be clustered together in space.

Figure 4 shows an example of the spatial clustering of languages around geographic barriers in the central Philippines. The Philippines is an island nation, so the primary barriers to interaction are the oceans between islands and the topography within islands. Different islands tend to speak different languages, though some languages span multiple islands that are close together. The islands shown in the figure are just south of the central island of Luzon, where Manila is located. The islands vary in topology. Mindoro and Panay are quite mountainous, while Bohol and Masbate are relatively flat. Cebu is hilly. Negros has a spine of mountains running northeast to southwest. Mindoro and Panay have a handful of small language areas, while Masbate, Cebu, and Bohol have only one or two. Negros divides into two language areas along its mountainous spine.

In the next section, I develop a model that shows how this spatial arrangement emerges from local interaction among individuals in contact.

3 A Theory of Language Population Dynamics

In this section I will develop a model of language population dynamics. The model shows how macro-level patterns of language size can emerge from local coordination. The model is grounded in two facts about language. The first is there are gains to sharing a common language with those with whom you might engage in economic activity. There is thus an incentive for economically integrated groups to coordinate on a common language. The second fact is that direct human interaction is primarily local in nature, as I argued in the previous section. Individuals will have a greater incentive to speak the language of those who are spatially proximate to them. The model explore the implications of these facts for the equilibrium size distribution of languages. I use

Figure 4: *Languages of the Central Philippines*

Notes: Each number indicates a distinct language. Names identify the major islands. Adapted from the Ethnologue's Map of the Southern Philippines (Lewis *et al.*, 2014). Used with permission.

Figure 5: *Simple Locally-Connected Network with Two Components*

Notes: The connection protocol specifies that two vertices closer than a fixed distance d are connected. No other vertices are connected.

concepts from evolutionary game theory and network theory in developing the model, particularly Young's model of local interactions (Young, 1998; Newman *et al.*, 2001; Jackson, 2008).

3.1 Overview of Model

I will begin by describing the model primitives and give an informal sketch of how equilibrium is reached. Then I will give a more formal treatment. The model describes the behaviour of a population of N agents. Each agent i in model has four primary properties: 1) it's location at a vertex or node of the graph Γ , 2) a language ℓ_i that it knows how to speak, 3) a set of neighbours N_i to whom it is directly linked by undirected edges,⁵ and 4) knowledge of the languages spoken by members of N_i . The agent periodically has an opportunity to change the language it speaks. All other properties are fixed.

Any two agents who can be reached by traversing edges through a series of one or more linked vertices are said to be connected. A connection protocol defines the method for determining which other vertices are in N_i . Depending on how the connection protocol is defined, all agents in the

⁵An undirected edge between vertices A and B means that communication can flow from both A to B and B to A. A directed edge allows only one of those possibilities.

model may be connected to each other, or agents may be divided into disjoint sets called connected components within which all are connected. The spatial aspect of the model is expressed in the connection protocol. For example, the vertices of Γ might have coordinates in two-dimensional space, with the probability of an edge between two vertices a decreasing function of the distance between them. In this case, most of a given agent's neighbours will be spatially proximate. Figure 5 shows a simple example in which the connection protocol specifies that only vertices closer than d are connected. There are two connected components in the graph with eight and four members each.

Agents earn payoffs from their language ability as follows. In each of an infinite number of periods, each agent i encounters one of its neighbours from N_i at random. If the agent and neighbour share a language in common, they engage in a mutually beneficial exchange and both earn positive payoffs. Otherwise they both earn zero. Agents know which languages are currently spoken by each of their neighbours, and occasionally get a chance to revise their own language without cost. Language choice is thus a coordination game among neighbours, and any given agent will maximize its expected payoffs by selecting the language most commonly used by its neighbours. When an agent changes its language, it alters the choice problem for all of its neighbours. Since agents within a connected component of the graph Γ are indirectly linked through overlapping sets of neighbours, the choice set of an agent A may be indirectly influenced over time by the decisions of agent B even though B is not a neighbour of A. I will show that if agents revise their choices with some probability of making a mistake, eventually all agents in each connected component will speak the same language. The population size distribution of languages will be identical to the size distribution of connected components.

As the number of connections in a graph increases, the size of the largest connected component often abruptly shifts from being relatively small to relatively big, radically changing the network structure. This is called a *phase transition* by analogy to the discontinuous effects a change in temperature can have on the state of matter. When the largest connected component is relatively big, it is called a *giant* component. I show that when Γ is close to this phase transition, the size

Table 2: *Payoff Matrix for G*

		ℓ_i				
		1	2	3	$\dots$	L
ℓ_j	1	1,1	0,0	0,0	$\dots$	0,0
	2	0,0	1,1	0,0	$\dots$	0,0
	3	0,0	0,0	1,1	$\dots$	0,0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	L	0,0	0,0	0,0	$\dots$	1,1

of connected components, and hence the size of languages, follows a Pareto distribution with an exponential cutoff. This result holds regardless of what connection protocol is used to define each agent's neighbours. When neighbours are defined according to spatial proximity, simulation results show that for a wide range of parameter values the double Pareto distribution provides the best fit to the distribution of language sizes.

3.2 Language Coordination over a Network

I begin the formal explication of the model in this section with the network defined by an arbitrary connection protocol. After developing some general results, I introduce an explicitly spatial model of neighbours using simulation in section 3.3.

Consider a coordination game played among the N members of a set of players $\mathcal{P}$. Each member of $\mathcal{P}$ is located at a vertex of the graph Γ . Each player i can speak a language $\ell_i \in \mathcal{L}$. Every player is connected to at least one other player by an undirected edge. The indicators e_{ij} tell us whether there is an edge connecting any two players i and j . If i and j are connected, $e_{ij} = 1$ otherwise $e_{ij} = 0$. The set of players to which player i is connected is called the neighbourhood of i and is denoted N_i . In other words, $j \in N_i \iff e_{ij} = 1$. The set of all edges between pairs of connected vertices $\{h, k\}$ is $\mathcal{E}$. There are $E = \sum_i \sum_j e_{ij}$ elements in $\mathcal{E}$.

Each period, every player is randomly paired with another member of N_i . The two neighbours play a coordination game G in which the actions taken are the player's respective languages. If both players have the same language, they each get a payoff of 1, otherwise they get a payoff of

0.⁶ Table 2 shows the payoff matrix. All of the diagonal entries in which players coordinate on the same action are pure Nash equilibria. None of the equilibria are risk dominant. If the payoffs to coordination differed by language, then the language associated with the largest diagonal entry would be a pure Nash equilibrium.

In population game theory, if player utility in a game can be rescaled by real numbers λ_i such that the change in scaled utility from a unilateral deviation is the same for all players, i.e. that $\lambda_i u_i(x'_i, x_{-i}) - \lambda_i u_i(x_i, x_{-i}) = \rho(x'_i, x_{-i}) - \rho(x_i, x_{-i})$, that game is called a potential game (Young, 1998; Sandholm, 2011). Games in which payoffs are symmetric for all players are belong to the class of potential games. For game G , I define the potential $\rho(\ell, N_i, t)$ of language choice ℓ_i in the neighbourhood of player i at time t as

$$\rho(\ell_i, N_i, t) = \sum_{j \in N_i} \mathbf{1}(\ell_i^t = \ell_j^t). \quad (1)$$

A particular language choice ℓ_i is a best response if it has the highest potential, which is the same as being the most widely spoken language in the neighbourhood. It is easy to show that if G is a potential game, then the overall game involving all players in $\mathcal{P}$ is also a potential game. Let the vector $\vec{\ell}$ indicate the current language choices of all members of $\mathcal{P}$ and the pair $\{h, k\}$ index an edge between players h and k . The potential function for the overall game is then

$$\rho^*(\vec{\ell}, t) = \frac{N}{E} \sum_{\{h, k\} \in \mathcal{E}} \mathbf{1}(\ell_h^t = \ell_k^t). \quad (2)$$

The potential function gives the expected total payoff to all players from any set of language choices $\vec{\ell}$. Note that this is just the fraction of edges that link two players who share a language multiplied by the number of players.

Players update their action choices according to a revision protocol. Each player possesses an alarm clock that rings to signal them to revise their action. The time between rings follows an

⁶I make payoffs the same for each language in $\mathcal{L}$ because linguists believe there are no intrinsic communication advantages of one language over another.

exponential distribution. At any particular time t , at most one player is engaging in revision. The revising player knows the actions most recently played by all of its neighbours $\{\ell_j^t\}_{j \in N_i}$. Since the revising player will encounter members of the neighbourhood randomly when pairing happens, the best-response language choice is to select the language most widely spoken among neighbours.

A deterministic best-response revision protocol would specify that the revising player would compute the potential $\rho(\ell_i, N_i, t)$ for each $\ell_i \in \mathcal{L}$, breaking ties by randomizing, and then select the language with the highest potential. While this seems like the most straightforward approach, it turns out that the equilibrium reached by this protocol is difficult to characterize because we cannot be assured that all parts of the strategy space are reached with positive probability.

The situation improves if we allow non-best-response choices to occur with a positive probability. Consider the log-linear response rule in which the probability of selecting a particular language z_i is

$$p(z, N_i, \beta) = \frac{e^{\beta \rho(z_i, N_i, t)}}{\sum_{z' \in \mathcal{L}} e^{\beta \rho(z', N_i, t)}}, \quad (3)$$

where $\beta \geq 0$. This probabilistic rule is known as a β -response function (Young, 1998). Languages with higher potential are more likely to be selected. The parameter β controls how sensitive the likelihood of a choice is to differences in potential. When β is large, the probability of choosing the language with highest potential approaches one and the probability of choosing non-best-responses approaches zero. When $\beta = 0$, all languages in $\mathcal{L}$ are chosen with equal probability. The parameter β thus captures how error-prone the players are in a fairly natural way.

Having set up the payoff matrix, potential function, and revision protocol, we can now use an existing result to characterize the equilibrium (Young, 1998, Theorem 6.1). Young shows that a symmetric potential game played on a finite graph using the β -response revision protocol leads to a unique stationary distribution in which the strategies played maximize the potential function for the overall game. The potential function $\rho^*(\vec{\ell}, t)$ is maximized when all players select the same languages as their neighbours. This implies that all players within each connected component of

the graph will coordinate on the same language. Since the payoff matrix states that no language is risk dominant, no language is favoured for coordination. If the graph has more than one connected component, then, the languages may differ between them.

If we suppose that at $t = 0$ the players' languages ℓ_i^0 are unique, then with a high probability the equilibrium size distribution of languages will be identical to the size distribution of connected components of the graph Γ .

RESULT 1. *The equilibrium size distribution of languages will be identical to the size distribution of the connected components of Γ .*

We thus investigate the distribution of language sizes, our ultimate object in building the model, through study of the distribution of connected components of Γ , which I will denote $H(s)$. I will develop an analytic result about the distribution of component size for an arbitrary connection protocol that holds in a special case. Later, I will use simulation of a specific local connection protocol to reach additional results.

The degree of a vertex is defined as the number of edges that connect to it. Many theoretical results about random graphs concern a special case called the Poisson random graph, in which edges are created between any two vertices i and j with independent probability p . Graphs with a spatial connection protocol are not of this type because the probability p_{ij} of an edge existing between i and j will typically be a decreasing function of the Euclidean distance between i and j rather than independent. This means that the many results developed for Poisson random graphs will not help here. I therefore draw on the theory of random graphs with arbitrary degree distributions developed in Newman *et al.* (2001) to further the analysis.

Define the probability distribution generating function for the vertex degree of graph Γ as

$$G(s) = \sum_{k=0}^{\infty} p_k s^k, \quad (4)$$

where p_k is the probability that a randomly selected vertex will have degree k , normalized so that $G(1) = 1$. $G(s)$ clearly depends on p_{ij} through p_k , but for the development that follows

we can leave the details unspecified. We are interested in the properties of the related generating function for the distribution of the sizes of components. To get there, we first define $L(s)$ as the size distribution of components reached by randomly selecting an *edge* from Γ and following it to one of its ends. We exclude the giant component, if there is one, from $L(s)$ to ensure that, with high probability, the component reached from the selected edge will not contain a closed loop. Under these fairly non-restrictive conditions, Newman *et al.* (2001) show that we can relate the size distribution $H(s)$ to the vertex degree generating function $G(s)$ using equations

$$L(s) = s \frac{G'(L(s))}{z} \quad \text{and} \quad (5)$$

$$H(s) = sG(L(s)), \quad (6)$$

where z is the average degree of the vertices of Γ . For any particular connection protocol, calculation of the p_k in principle allows us to use equations 5 and 6 to solve explicitly for $H(s)$. Newman *et al.* (2001) argue that this is usually impossible in practice. However, they use some asymptotic properties of generating functions to argue that, in the special case where Γ is close to the phase transition where a giant component emerges, the upper tail of component sizes approximately follows a Pareto distribution with an exponential cutoff for all connection protocols.

RESULT 2. *When Γ is near the phase transition when a giant component emerges, the upper tail of $H(s)$ approximately follows a Pareto distribution with an exponential cutoff as given in equation 7.*

$$PE(s) \sim s^{-1-\alpha} e^{-\lambda \frac{s}{s^*}}. \quad (7)$$

The probability of seeing a component larger than s approximately follows the Pareto distribution with exponent α when $s \ll s^*$. The exponential portion reduces the relative probability of very large components for any given value of α , where ‘very large’ is determined by the value of s^* . Several other phenomena that have power law size distributions are found exhibit this cutoff

behaviour (Clauset *et al.*, 2009).

3.3 *Simulating Language Coordination over a Locally Connected Network*

Results 1 and 2 hold no matter what connection protocol is used to form the network. However, our result on the size distribution only applies when the language network is close to the emergence of a giant component. We do not know whether the largest language, Mandarin, constitutes a giant component or not. I proceed with the study of the size distribution for the specific case in which connections are created based on the spatial proximity of agents. I use simulation to examine the size distributions of these locally connected networks.

Let the graph Γ_g be defined as follows. Each of N agents is located on a two-dimensional torus at randomly generated coordinates (x_i, y_i) . The Euclidean distance between any two individuals is given by $d(i, j)$. The probability that individuals i and j are connected is given by $p_{ij} = k(d(i, j))$. The importance of geographic proximity for connection means $k_d < 0$. Since we will be conducting simulations, I will work with the simple connection protocol given by

$$p_{ij} = \begin{cases} 1 & \text{if } d(i, j) \leq \tau \\ 0 & \text{if } d(i, j) > \tau. \end{cases} \quad (8)$$

Two agents are connected only if the distance between them is less than τ . We can use the p_{ij} to define the $N \times N$ adjacency matrix C with elements $c_{ij} = p_{ij}$. This matrix completely describes how the network is connected. Tarjan (1972) provides an algorithm to find the connected components of Γ_g using C . Once the connected components have been identified, it is easy to produce and analyse the empirical CDF of component sizes.

I show an example of a simulated locally connected language network in Figure 6. A population of 360 individuals were randomly assigned a location on a 30×30 torus. The torus is shown unwrapped on a plane.⁷ An individual's neighbours are those with $\tau < 1.7$, which are the eight

⁷The top and bottom edges and the left and right edges are in fact adjacent to each other. The torus can be recreated by rolling the grid up from top to bottom and then joining the ends.

Figure 6: *Example Language Network on a Torus*
with $\bar{x} = \bar{y} = 30$ and $\tau = 1.7$

Notes: Each square represents an individual and each color a connected component of the network, and therefore a equilibrium language. This network of 360 individuals has 33 languages.

squares immediately bordering their own. Application of Tarjan's algorithm reveals that there are 33 connected components in the graph, including singleton individuals who are unconnected by convention. The non-singleton components correspond to the distinct languages in the equilibrium of the model. Each individual is represented on the grid by the rank of his or her language in descending order. The size distribution of languages is heavily skewed: The mean language size is 11 while the median is only 4. Nearly a third speak the largest language, but there are nine languages with only one speaker.

Human populations tend to be spatially clustered on the earth's surface rather than randomly

distributed. Humans gravitate toward more productive ecosystems and away from deserts, high mountains, and cannot survive in the ocean. I take this spatial clustering as exogenous in the present model, and introduce it into the simulation through an additional parameter γ . This parameter ranges from zero to infinity and controls an algorithm that takes the N randomly assigned locations (x_i, y_i) on T as its input and clusters them (Lennartsson *et al.*, 2012). In the clustering algorithm, the locations are transformed into frequency space using a fast Fourier transformation and the resulting amplitudes are scaled by γ . The rescaled amplitudes are transformed back to two-dimensional space, and the N locations that have the strongest signal become the new set of coordinates. Values of γ greater than zero produce spatial clustering of the individuals.

Each simulation run is characterized by a choice of population size, torus size, neighbourhood radius, and clustering parameter: $\{N, (\bar{x}, \bar{y}), \tau, \gamma\}$. Since the size distribution of connected components is our object of interest and we expect this distribution to be heavy tailed, the population should be as large as is computationally practical. For any given torus size, the number of connected components falls as the population density rises. To be able to fit and test distributions with the simulation data, we need at least 100 or so components. On the other hand, the size of the adjacency matrix C is $N \times N$, which means that computation time rises quickly with N . With these concerns in mind, my simulations fix the torus size to 200×200 and the population to 10,000. I vary neighbourhood size τ between 1.5 and 2.3, which results in neighbourhood area between 0.02% to 0.04% of total area. The clustering parameter γ varies between 0 and 2. When $\gamma = 2$, the average individual has more than twice as many neighbours within a three-unit radius as when $\gamma = 0$.

I begin my analysis of the simulation data with a table of summary statistics in Table 3. The table shows average values for 250 simulation runs for each $\{\tau, \gamma\}$ combination. The median language has two speakers for most of the parameter values. The distribution of language sizes has a heavy right tail for all parameter values. The mean size is larger than the median everywhere and ranges from 4 to 174. In general, the mean size increases with both γ and τ . Intuitively, larger connected components can form when individuals are more spatially clustered and when

Table 3: *Simulation Summary Results by τ and γ*

Median Language Size					
γ	τ				
	1.5	1.7	1.9	2.1	2.3
0.0	2.0	2.0	2.0	2.0	1.2
0.5	2.0	2.0	2.0	2.0	1.6
1.0	2.0	2.0	2.0	2.0	2.0
1.5	2.0	2.0	2.0	1.9	2.0
2.0	1.8	1.8	1.8	1.8	2.0

Mean Language Size					
γ	τ				
	1.5	1.7	1.9	2.1	2.3
0.0	4	4	4	12	174
0.5	5	5	5	12	83
1.0	6	6	6	12	41
1.5	9	9	9	15	30
2.0	16	16	16	24	39

Population Share of Largest Language					
γ	τ				
	1.5	1.7	1.9	2.1	2.3
0.0	0.01	0.01	0.01	0.07	0.99
0.5	0.01	0.01	0.01	0.08	0.96
1.0	0.04	0.04	0.04	0.15	0.83
1.5	0.13	0.13	0.13	0.30	0.64
2.0	0.37	0.35	0.37	0.50	0.65

Notes: Each cell presents the average value of the indicated statistic for 250 simulations runs using the indicated values for τ and γ .

neighbourhoods are larger. The share of the population speaking the largest language varies from 1% to 99%, so the parameter space ranges across the emergence of a giant component.

Result 2 suggests that the size distribution of components will be approximately Pareto-exponential

Table 4: *Log Likelihood Ratios for Simulation Data: Double Pareto against other Long-Tailed Distributions*

γ		τ				
		1.5	1.7	1.9	2.1	2.3
0	<i>Pareto-Exponential</i>	7.08	6.68	10.42	2.30	0.72
	<i>Pareto</i>	21.96	21.48	25.32	5.22	0.72
	<i>Truncated Lognormal</i>	19.43	18.91	22.80	5.06	0.69
	<i>Truncated Weibull</i>	18.31	17.80	21.69	5.49	2.04
	<i>Truncated Exponential</i>	275.24	276.39	279.12	376.87	240.65
0.5		5.89	-1.54	4.01	1.88	0.43
		14.29	6.57	12.06	3.60	0.43
		13.65	5.98	11.50	3.57	2.04
		13.68	6.05	11.60	4.34	3.24
		356.25	351.75	358.45	412.78	321.71
1.0		-1.23	1.42	1.51	1.00	0.26
		0.20	2.88	3.02	1.27	0.26
		0.22	2.90	3.04	8.76	-0.05
		2.10	4.80	4.97	10.43	0.89
		509.88	509.84	508.03	504.57	370.18
1.5		0.51	0.61	0.54	0.92	-0.33
		0.57	0.66	0.62	0.94	-0.33
		-0.16	0.68	-0.05	0.09	-0.32
		2.91	3.72	3.00	2.40	0.81
		658.82	645.69	660.78	579.58	405.09
2.0		-1.20	0.61	0.67	1.01	0.12
		-1.20	0.62	0.67	1.01	0.12
		7.06	-0.50	-0.01	0.04	4.54
		10.06	2.56	3.15	2.06	5.63
		687.30	678.13	701.25	526.35	352.61

Notes: Each cell presents the average likelihood ratio of the double Pareto and indicated distribution for 250 simulations runs at the values for τ and γ .

near the emergence of a giant component. I fit this distribution to the simulation data along with five related distributions that also have heavy right tails. These are 1) the ordinary Pareto distribution, which is a pure power law and can be viewed as a special case of the Pareto-exponential with $\lambda = 0$, 2) the double Pareto, which has two power parameters α and β on either side of a threshold s^* , 3) the left-truncated lognormal distribution, 4) the left-truncated Weibull distribution,

Figure 7: *Emergence of the Giant Component as τ Grows*

Notes: Plot shows statistics about the size of the largest connected component over 50 simulation runs for the indicated values of τ . All runs set $\gamma = 0$. Dots show the mean size. Capped spikes show the 10th and 90th percentiles of size.

and 5) the left-truncated exponential distribution. I can then compare the values of the optimized likelihood functions using likelihood ratio tests. The distribution that has the highest likelihood is the best fit to the simulation data.

It turns out that the best-fitting distribution overall is the double Pareto (DP). For $\alpha < \beta$ and sizes above s_0 , the DP distribution is given by

$$DP(s) \sim \begin{cases} \left(\frac{s}{s^*}\right)^{-1-\alpha} & \text{for } s_0 \leq s < s^* \\ \left(\frac{s}{s^*}\right)^{-1-\beta} & \text{for } s \geq s^*. \end{cases} \quad (9)$$

The DP distribution is similar to the Pareto-exponential in that the parameter s^* indicates a ‘very large’ size above which, relative to the value of α , the relative probability of seeing very large components is smaller.

Table 4 shows the results of likelihood ratio tests comparing the DP to the other distributions. Positive values favour the DP over the alternative. The differences in fit is greatest when the

clustering parameter γ and the neighbourhood radius τ are small. When they are larger, it becomes more difficult to distinguish the DP from the Pareto-exponential, Pareto, and truncated lognormal distributions.

I explore the emergence of the giant component through additional simulations that use a more granular set of values for τ while keeping $\gamma = 0$. The gross structure of the language distribution is quite sensitive to small changes in parameters. Figure 7 shows that the average population share of the giant component, shown with dots, grows very rapidly between $\tau = 2.3$, when it comprises 20% of the vertices, and $\tau = 2.425$, when it comprises 80%. The size of the giant component is also much more varied during the phase transition. Capped spikes indicate the 10th and 90th percentiles of the distribution of giant component share in the simulation runs. This non-linear behaviour is of particular interest because real-world processes such as innovation in transportation and communication technology are arguably analogous to increasing the neighbourhood radius. Technological shocks may thus cause the distribution of language size change in highly non-linear way.

4 The Distribution of Language Population

I now fit the DP distribution to population data on all of the world’s languages from the WLMS (WLMS, 2011). This dataset has been introduced in Figures 1 and 2. The WLMS incorporates the language data from Ethnologue, including the population of speakers, and is collated and maintained by The Summer Institute of Linguistics. It is the most comprehensive resource of language population available.

The WLMS population estimates are drawn from censuses, the linguistics research literature, and submissions from field correspondents. The estimates are made at the country level. Since not all estimates are made at the same time, I adjust them to 2000 using country-specific growth rates computed from the World Development Indicators database (World Bank, 2012). To make the adjustment, I need to know the date a population estimate in the WLMS was made. This information exists for 89% of the country-language observations. I then aggregate the adjusted

population estimates to the language level. This procedure produces population estimates for 6,210 languages.

Figure 8 presents the full counter-cumulative distribution in panel (a) and a close up for languages with more than 10,000 speakers in panel (b). I have added a random jitter to the plots to make it easier to see where the mass of the distribution lies.

4.1 Fitting the Steady-State Distribution

The steady state of the model of language size is best described by the double Pareto distribution. The DP density function is given by

$$f(s; \alpha, \beta, s^*) = \begin{cases} C \left(\frac{s}{s^*} \right)^{-1-\alpha} & \text{for } s_0 \leq s < s^* \\ C \left(\frac{s}{s^*} \right)^{-1-\beta} & \text{for } s \geq s^*, \end{cases} \quad (10)$$

where

$$C = \frac{s_0^\alpha \alpha \beta}{s^* (s_0^\alpha (\alpha - \beta) + s^{*\alpha} \beta)}.$$

To fit the DP, we will use the log-likelihood function

$$\ln L = \sum_{i=1}^n \ln f(s_i; \alpha, \beta, s^*) \text{ for } s_i \geq s_0. \quad (11)$$

To fit equation 11 by maximum likelihood, we must first choose a value for s_0 . This is a problem faced whenever fitting a power law. Early work on power laws chose s_0 by inspecting a log-log plot of the empirical counter-cumulative distribution. The single Pareto distribution appears linear on this type of plot for $s > s_0$, while the double Pareto appears piece-wise linear. Panel (b) of Figure 8 shows this plot for larger languages. The plot looks slightly concave overall, while some segments appear to be linear.

Figure 8: *Counter-Cumulative Distribution of Language Population*

(a) Full Distribution

(b) Languages Larger than 10,000

Notes: Panel (a) contains the full distribution, $N=6,210$. Panel (b) shows languages with more than 10,000 speakers, $N=2,900$. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

The job of choosing s_0 for the DP distribution difficult to do by eye as it should fall below two linear sections with potentially different slopes. Further, an important question for the paper is precisely the range of languages for which the DP distribution might be consistent with equilibrium. I therefore apply two formal, principled approaches for choosing s_0 .

The first approach looks at Kolmogorov-Smirnov D -statistics for different values of s_0 . The D -statistic measures how well our empirical distribution fits the theoretical distribution given particular parameters. I estimate these parameters using maximum likelihood and then compute the statistic. This method is widely employed in the literature (e.g. Clauset *et al.*, 2009; Rybski *et al.*, 2009). We will find s_0 that provides the best possible fit using this method and also see what other values are plausible.

Let $F(s)$ be the empirical CDF and $G(s)$ be the theoretical CDF. We estimate the parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{s}^*$ by maximum likelihood and insert them into the theoretical CDF. The Kolmogorov-Smirnov D -statistic measures the maximum distance between the empirical and theoretical CDFs:

$$D = \max_{s \geq s_0} |F(s) - G(s)|. \quad (12)$$

The statistic $T = D\sqrt{N}$ allows us to test the null hypothesis that the empirical CDF follows the theoretical CDF.

I maximize equation 11 for a range of possible values of s_0 and compute both D and a p-value for the test statistic T . Figure 9 shows the results. The top panel shows that $s_0=17,900$ provides the best fit of the data to the DP distribution. It is a global minimum, though values of s_0 up to about 36,000 fit nearly as well. The bottom panel shows that we fail to reject the null hypothesis that the data follows the DP distribution for $s_0 > 14,000$ at the 10% level. The conclusion we draw here is that any value of s_0 above 14,000 is plausible, with values between 17,900 and 36,000 providing the best statistical fits.

The second method for choosing s_0 notes that empirical distribution follows a smooth curve for smaller values of s . By simply truncating the data, as we have done in the first method, I am

Figure 9: *Kolmogorov-Smirnov D Statistic and P Values*

Notes: Figure shows Kolmogorov-Smirnov D statistic and p Values for the empirical distribution against the Double Pareto. The top panel shows the D statistic from equation 12 computed for different values of s_0 . The bottom panel shows corresponding p values for the null hypothesis that the language population data follow the double Pareto distribution. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

in effect allowing that there may be as many parameters for the distribution below s_0 as there are observations. An alternative approach fits a likelihood that follows a right-truncated parametric distribution below s_0 and the DP above s_0 . The lognormal distribution seems an obvious choice for the left tail since it has a heavy right tail that is close to a power law in shape. Visual inspection of a kernel density estimate of log language size in Figure 1 suggests the lognormal is a reasonable fit. The LNDP has previously been fit to city data by Giesen *et al.* (2010).

Let p be the fraction of observations above s_0 and μ and σ be the mean and standard deviation of the right-truncated lognormal distribution. We can write a log-likelihood function for the entire distribution as

$$\ln L(\theta) = \sum_{s_i < s_0} \frac{(1-p)}{s\sigma} \frac{\phi(\frac{\ln s - \mu}{\sigma})}{\Phi(\frac{\ln s_0 - \mu}{\sigma})} + \sum_{s_0 \leq s_i < s^*} p C \left(\frac{s}{s^*} \right)^{-1-\alpha} + \sum_{s_i \geq s^*} p C \left(\frac{s}{s^*} \right)^{-1-\beta}, \quad (13)$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal density and distribution, respectively, and $\theta = (\alpha, \beta, \mu, \sigma, s_0, s^*)$. Maximizing this likelihood yields the estimate $\hat{s}_0=26,900$. This likelihood is somewhat difficult to estimate because 1) it has two discontinuities and 2) the lognormal and DP distributions have very similar shapes in the right tail. I used both differential evolution and mesh adaptive direct search, which do not require continuous or differentiable objective functions, to do the optimization and obtained similar results from both methods.⁸

Recall that I showed evidence in Section 1 that language population growth is consistent with Gibrat's law for languages larger than about 35,000. Below that size, there is weak evidence that average growth is lower. The values of s_0 suggested by the LNDP and the KS test are in broad agreement with this other evidence about where the language size distribution appears to be stable. I make a conservative choice of $s_0=35,000$ as my preferred estimate. Thirty percent of all languages are larger than 35,000.

Figure 10 shows the fitted distribution overlaid on a jittered plot of the data. The fit is very good. Table 5 presents the maximum likelihood estimates of the parameters for $s_0=35,000$. The

⁸These methods are both implemented in R. Differential evolution is available in the `DEoptim` package and mesh adaptive direct search is available in the `crs` package.

Figure 10: *Counter-Cumulative Distribution of Language Population with DP Fit*

Notes: The dots plot the empirical counter-cumulative distribution of languages with population above 35,000. $N=1,868$. The line shows the maximum likelihood fit of the double Pareto distribution. The parameters are those shown in Table 5. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

fitted distribution follows the exponent $\hat{\alpha}=0.342$ below the $\hat{s}^*=836,400$ threshold and $\hat{\beta}=0.663$ above the threshold. Using the KS test for goodness of fit, I fail to reject the null hypothesis that the empirical distribution follows the estimated theoretical distribution with $p=0.95$. Figure 11 shows the LNDP fit to the entire dataset.

The estimated power law exponents $\hat{\alpha}=0.342$ and $\hat{\beta}=0.663$ are small compared to those for cities or firms, which are typically close to one (Zipf, 1949; Axtell, 2001; Rozenfeld *et al.*, 2011). They also fall on the small side for other phenomena that exhibit power law behaviour (Clauset *et al.*, 2009; Rybski *et al.*, 2009). Smaller exponents correspond to a longer left tail. The 30 or so largest languages fall below the fitted line, meaning that they are smaller than we would expect if the languages followed the fitted theoretical distribution. This is not surprising as the theoretical distribution has a non-finite mean for $\beta < 1$, while the empirical data must have a finite mean. Under these conditions, the very largest observations should fall below the fitted line in expectation (Newman, 2005).

Table 5: *Double Pareto ML Estimates with $s_0 = 35,000$*

Parameter	Coefficient	Standard Error
α	0.342	0.023
β	0.663	0.026
s^*	836,353	17,390

Notes: N=1,868. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

4.2 Double and Single Pareto Fits Compared

Most of the empirical literature on size distributions fit the single Pareto distribution to the upper tail. In this section I provide a comparison of single and double Pareto fits to the data. I begin by computing the augmented likelihood LNSP and KS tests through simple modification to the procedure in section 4.1. In this case, the two methods of estimating s_0 give very different results. Maximum likelihood estimation of the LNDP produces $\hat{s}_0 = 3.231 \times 10^7$. The value of s_0 that minimizes the difference D between the theoretical and empirical distribution is 534,564. The latter value corresponds well to the estimated s_* for the DP distribution. The sharply different estimates are not surprising when we consider that the lognormal and SP distributions are very close in shape in the upper tail.

Figure 12 plots fitted Pareto distributions with the two values of s_0 . I estimate $\hat{\alpha}=0.645$ with $s_0=534,564$ and $\hat{\alpha}=1.281$ with $s_0 = 3.23 \times 10^7$. The augmented likelihood places just 37 languages above s_0 . All of these lie below the fitted CDF for both the DP distribution and the Pareto fitted with the D-statistic-minimizing s_0 . Most of the observations determining the fitted parameters of the DP distribution are fit to the lognormal part of the LNSP.

It is difficult to compare the fit of the estimated SP and DP distributions as they cover different subsets of the overall data. However, a likelihood ratio test will show whether the LNDP or LNSP fits better. Since the distributions are nested, the LNDP, with its additional parameter, will necessarily fit at least as well as the LNSP. Let L_N be the null likelihood, the LNSP, and L_A

Figure 11: *Lognormal-double Pareto Fit to Language Populations*

Notes: The dots plot the empirical counter-cumulative distribution of languages. $N=6,210$. The line shows the maximum likelihood fit of the lognormal-double Pareto distribution. Data are from the WLMS. Population data at the country level is adjusted to year 2000 and aggregated as described in the text.

be the alternative likelihood, i.e. the LNDP. I compute the test statistic T is using the estimated likelihoods

$$T = -2 \log \left(\frac{\widehat{L}_N}{\widehat{L}_A} \right). \quad (14)$$

T approximately follows a χ^2 distribution with one degree of freedom (Clauset *et al.*, 2009). I compute $T = 20.74$, and reject the null distribution in favour of the alternative with $p < 0.001$. This is quite a large difference in ratios, meaning that the additional parameter makes a substantial improvement in the fit.

Figure 12: *Counter-Cumulative Distribution of Language Population with Pareto Fits*

Notes: The dots plot the empirical counter-cumulative distribution of languages with population above 534, 564. $N=509$. A jitter has been added for readability. The lines show maximum likelihood fits of the Pareto distribution with $s_0 = 534, 564$ and $s_0 = 3.231 \times 10^7$. The fitted values of $\hat{\alpha}$ are 0.645 and 1.281, respectively. Data are from the WLMS. Population data is adjusted to year 2000 as described in the text.

5 Language Extinction

Scholars and policy makers have been concerned since at least the early 1990s about the extinction risk faced by small languages. As I discussed in the introduction, the existing literature on language consolidation suggests that this risk is faced by the great majority of languages.

The minimum viable size for a language depends on circumstances such as geographic isolation, the extent of bilingualism, state support, domains in which the language is used, properties of alternative languages, and the robustness of intergenerational transmission. A wide range of thresholds have been proposed for a rule of thumb below which a language should be considered in danger. Some scholars have suggested 1,000,000 (de Swaan, 1991; Fishman, 1998; Graddol, 2004). Others have suggested 10,000 (Hale, 1992; Krauss, 1992; Crystal, 2000).

Growing alarm at the potential loss in cultural diversity from language extinction prompted UNESCO to conduct a thorough and careful risk assessment for all languages (Moseley, 2010).

The assessment evaluates nine factors and results in a judgement that the language is safe, vulnerable, or endangered. In addition to the number of speakers, the assessment includes to what degree the language is being learned by children, social domains in which the language is used and not used, government policy, and community attitudes.

Figure 13 shows a kernel density estimate of the size distribution of the 1,649 languages UNESCO judges to currently be endangered. These languages have a mean size of 26,800 speakers, a median size of 609, and a 90th percentile of 27,000. UNESCO includes seven languages with more than a million speakers on the list. These are mostly minor European languages, such as Lombard, Emiliano-Romagnolo, and Yiddish.

Figure 13: *Distribution of Endangered Languages by Log Population Size*

Notes: Data are from UNESCO. Kernel density estimate of the size distribution for 1,649 languages considered endangered. Bandwidth computed using Silverman's plug-in estimate (Cameron and Trivedi, 2005).

Most UNESCO endangered languages are below the threshold at which Gibrat's law holds (section 1) and below the lower bound of the DP fit. These estimates were produced using a completely different procedure, and it is reassuring that they are in broad agreement about the sizes of languages that are threatened with decline.

What does my analysis suggest about how many languages will go extinct? To explore this

question, I simulated 100 years of evolution by applying the size-growth rate relationship estimated in section 1 to the full size distribution of the world’s languages. I took the relationship as fixed over the simulation, and did not allow speciation. I estimated the conditional mean and variance of annual growth at 10,000 log size bins from a local-linear regression. In each simulation year, each language is matched to the appropriate size bin, and a growth rate is drawn from a normal distribution with the estimated mean and variance of growth associated with that bin. Languages that fall below 100 speakers are considered to be extinct and are not included in the next year.

Averaging over 1,000 simulation runs, 1,608 languages, or 26% of the current total, have gone extinct. This estimate is quite similar to the 1,649 languages UNESCO considers to be endangered. If we suppose a reasonable measure for endangerment is a 25% probability of going extinct in 100 years, I find 1,957 languages to be endangered. They have a median current size of 729 speakers, which is close to the UNESCO median of 609. The main difference between the UNESCO estimate and the one produced here is that, by considering factors other than size, UNESCO includes more large and fewer small languages on its list.

As another check, I follow Sutherland (2003) in applying criteria used by the IUCN (International Union for the Conservation of Nature) to assess the endangerment of biological species. I find that for 1,645 languages, or 26% of the current total, the average population across simulation runs declines by 80% or more. This corresponds to the IUCN ‘critically endangered’ classification. According to Sutherland, languages are much more endangered than birds or mammals, of which 3.2% and 6.0% of species are critically endangered.

6 Discussion

The main empirical results I have established are that 1) language growth rates follow Gibrat’s law in a selected sample of languages, and 2) language sizes follow the DP distribution in the upper 30% tail for the entire universe of languages. I have linked these findings through a model of language choice. The DP distribution emerges as the equilibrium when coordination about language occurs via a locally connected network. The precise fit of the distribution for the *universe*

of languages larger than 35,000 supports the evidence of Gibrat's law on the *selected* sample of languages.

My analysis provides an account of the dynamics of language populations that fits the aggregate data better than the existing models. It suggests that forces that would drive a divergence in language size operate on a smaller scale than previously believed. The evidence suggests humans will still speak many thousands of languages a hundred years from now.

The biological constraints on language learning mean that changes in mother tongue necessarily happen on a generational time scale. It is possible that major innovations in communication of the late 20th century, such as the internet and widely disseminated broadcast media, will fundamentally alter the relationship between language size and growth. Networks are known to undergo phase transitions in which the size of the largest component, which for human languages is Mandarin, very rapidly expands its share of the population. We should not forget, however, that economic and cultural integration has been under way for a long time, and much of that integration is reflected in the data I have presented.

Weatherhead School of Management

Case Western Reserve University

References

- Alesina, A. and Spolaore, E. (1997). 'On the number and size of nations', *The Quarterly Journal of Economics*, vol. 112(4), pp. 1027–1056.
- Alesina, A. and Spolaore, E. (2003). *The Size of Nations*, Cambridge, MA: MIT Press.
- Alesina, A. and Wacziarg, R. (1998). 'Openness, country size and government', *Journal of Public Economics*, vol. 69(3), pp. 305–321.
- Axtell, R.L. (2001). 'Zipf Distribution of U.S. Firm Sizes.', *Science*, vol. 293(5536), pp. 1818–20.
- Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press, 16 edn.
- Champernowne, D. (1953). 'A Model of Income Distribution', *The Economic Journal*, vol. 63(250), pp. 318–351.
- Choi, E. (2002). 'Trade and the Adoption of a Universal Language', *International Review of Economics & Finance*, vol. 11(3), pp. 265–275.
- Church, J. and King, I. (1993). 'Bilingualism and Network Externalities', *Canadian Journal of Economics*, vol. 26(2), pp. 337–345.
- Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). 'Power-Law Distributions in Empirical Data', *SIAM Review*, vol. 51(4), pp. 661–703.
- Cordoba, J. (2008). 'On the Distribution of City Sizes', *Journal of Urban Economics*, vol. 63(1), pp. 177–197.
- Crystal, D. (2000). *Language Death*, Cambridge: Cambridge University Press.
- Crystal, D. (2003). *English as a Global Language*, Cambridge: Cambridge University Press.

- Davidson, R. and Flachaire, E. (2008). 'The Wild Bootstrap, Tamed at Last', *Journal of Econometrics*, vol. 146(1), pp. 162–169.
- de Swaan, A. (1991). 'Notes on the Emerging Global Language System: Regional, National and Supranational', *Media, Culture & Society*, vol. 13(3), pp. 309–323.
- de Swaan, A. (1993). 'The Evolving European Language System: A Theory of Communication Potential and Language Competition', *International Political Science Review*, vol. 14(3), pp. 241–255.
- Eeckhout, J. (2004). 'Gibrat's Law for (All) Cities', *American Economic Review*, vol. 94(5), pp. 1429–1451.
- Fishman, J. (1998). 'The New Linguistic Order', *Foreign Policy*, vol. 113(113), pp. 26–32.
- Gabaix, X. (1999). 'Zipf's Law for Cities: An Explanation', *Quarterly Journal of Economics*, vol. 114(3), pp. 739–767.
- Gabaix, X. and Ioannides, Y. (2004). 'The Evolution of City Size Distributions', *Handbook of Regional and Urban Economics*, vol. 4(617), pp. 2341–2378.
- Gibrat, R. (1931). *Les Inegalites Economiques: Applications aux Inegalites des Richesses, a la Concentration des Entreprises, aux Populations des Villes, aux Statistiques des Familles, etc., d'une Loi Nouvelle, la Loi de l'Effet Proportionnel*, Paris: Librairie du Recueil Sirey.
- Giesen, K., Zimmermann, A. and Suedekum, J. (2010). 'The Size Distribution Across All Cities Double Pareto Lognormal Strikes', *Journal of Urban Economics*, vol. 68(2), pp. 129–137.
- Gomes, M., Vasconcelos, G., Tsang, I. and Tsang, I. (1999). 'Scaling Relations for Diversity of Languages', *Physica A: Statistical Mechanics and its Applications*, vol. 271(3-4), pp. 489–495.
- González-Val, R. and Sanso-Navarro, M. (2009). 'Gibrats Law for Countries', *Journal of Population Economics*, vol. 23(4), pp. 1371–1389.

- Graddol, D. (2004). 'The Future of Language.', *Science*, vol. 303(5662), pp. 1329–31.
- Grin, F. (1992). 'Towards a Threshold Theory of Minority Language Survival', *Kyklos*, vol. 45(4), pp. 69–97.
- Hale, K. (1992). 'On Endangered Languages and the Safeguarding of Diversity', *Language*, vol. 68(1), pp. 1–3.
- India (2008). *Language: India, States and Union Territories (Table C-16)*, New Delhi: Office of the Registrar General, India.
- Jackson, M.O. (2008). *Social and Economic Networks*, Princeton, N.J.: Princeton University Press.
- Katz, M. and Shapiro, C. (1985). 'Network Externalities, Competition, and Compatibility', *The American Economic Review*, vol. 75(3), pp. 424–440.
- Katz, M. and Shapiro, C. (1986). 'Technology Adoption in the Presence of Network Externalities', *The Journal of Political Economy*, vol. 94(4), pp. 822–841.
- Klemperer, P. (2008). 'Network Goods (Theory)', in (S. Durlauf and L. Blume, eds.), *New Palgrave Dictionary of Economics and the Law, Second Edition*, Basingstoke: Palgrave Macmillan, http://www.dictionaryofeconomics.com/article?id=pde2008_N000138. Last Accessed: December 8, 2014.
- Krauss, M. (1992). 'The World's Languages in Crisis', *Language*, vol. 68(1), pp. 4–10.
- Krugman, P. (1996). *The Self-Organizing Economy*, Oxford: Blackwell.
- Laitin, D.D. (1993). 'The Game Theory of Language Regimes', *International Political Science Review*, vol. 14(3), pp. 227–239.
- Lazear, E. (1999). 'Culture and Language', *Journal of Political Economy*, vol. 107(S6), pp. S95–S126.

- Lennartsson, J., Hkansson, N., Wennergren, U. and Jonsson, A. (2012). ‘SpecNet: A spatial network algorithm that generates a wide range of specific structures.’, *PloS One*, vol. 7(8), pp. e42679–e42679.
- M. P. Lewis, G. F. Simons and C. D. Fennig, eds. (2014). *Ethnologue: Languages of the World*, SIL International, 17 edn.
- Luttmer, E. (2007). ‘Selection, Growth, and the Size Distribution of Firms’, *The Quarterly Journal of Economics*, vol. 122(3), pp. 1103–1144.
- Minnesota Population Center (2011). *Integrated Public Use Microdata Series, International: Version 6.1 [Machine-Readable Database]*, Minneapolis: University of Minnesota.
- C. Moseley, ed. (2010). *Atlas of the Worlds Languages in Danger*, 3rd edn., Paris: UNESCO Publishing, update of July 9, 2012.
- Newman, M. (2005). ‘Power Laws, Pareto distributions and Zipf’s law’, *Contemporary Physics*, vol. 46(5), pp. 323–351.
- Newman, M.E.J., Strogatz, S.H. and Watts, D.J. (2001). ‘Random graphs with arbitrary degree distributions and their applications’, *Physical Review E*, vol. 64(026118), pp. 1–17.
- Rozenfeld, H., Rybski, D., Gabaix, X. and Maske, H. (2011). ‘The Area and Population of Cities: New Insights from a Different Perspective on Cities’, *American Economic Review*, vol. 101(August), pp. 2205–2225.
- Rybski, D., Buldyrev, S.V., Havlin, S., Liljeros, F. and Makse, H.a. (2009). ‘Scaling Laws of Human Interaction Activity.’, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106(31), pp. 12640–5.
- Saichev, A., Malevergne, Y. and Sornette, D. (2009). *Zipf’s Law for Firms: Relevance of Birth and Death Processes*, Springer-Verlag.

- Sandholm, W.H. (2011). *Population Games and Evolutionary Dynamics*, Cambridge, Mass.: MIT Press.
- Sutherland, W.J. (2003). 'Parallel Extinction Risk and Global Distribution of Languages and Species.', *Nature*, vol. 423(6937), pp. 276–9.
- Sutton, J. (1997). 'Gibrat's Legacy', *Journal of Economic Literature*, vol. 35(1), pp. 40–59.
- Tarjan, R. (1972). 'Depth-first search and linear graph algorithms', *SIAM Journal on Computing*, vol. 1(2), pp. 146–160.
- Van Parijs, P. (2000). 'The Ground Floor of the World: On the Socio-Economic Consequences of Linguistic Globalization', *International Political Science Review*, vol. 21(2), pp. 217–233.
- Van Parijs, P. (2005). 'Europes Three Language Problems', in (D. Castiglione and C. Longman, eds.), *The Challenge of Multilingualism in Law and Politics*, Oxford: Hart Publishing.
- Van Parijs, P. (2010). 'Linguistic justice and the territorial imperative', *Critical Review of International Social and Political Philosophy*, vol. 13(1), pp. 181–202.
- Wichmann, S. (2005). 'On the Power-Law Distribution of Language Family Sizes', *Journal of Linguistics*, vol. 41(1), pp. 117–131, ISSN 0022-2267.
- Wickstrom, B.A. (2005). 'Can Bilingualism be Dynamically Stable?: A Simple Model of Language Choice', *Rationality and Society*, vol. 17(1), pp. 81–115.
- WLMS (2011). *World Language Mapping System, version 16*, Colorado Springs: Global Mapping International.
- World Bank (2012). *World Development Indicators*, Washington D.C.: World Bank, update of July 9, 2012.
- Young, H.P. (1998). *Individual Strategy and Social Structure*, Princeton, N.J.: Princeton University Press.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley Press.