MODELING COMPETITIVE EVOLUTION OF MULTIPLE LANGUAGES

Zejie Zhou, Boleslaw K. Szymanski, and Jianxi Gao
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180, USA
gaoj8@rpi.edu

July 17, 2019

ABSTRACT

Increasing evidence demonstrates that in many places language coexistence has become ubiquitous and essential for supporting language and cultural diversity and associated with its financial and economic benefits. The competitive evolution among multiple languages determines the evolution outcome, either coexistence, decline, or extinction. Here, we extend the Abrams-Strogatz model of language competition to multiple languages and then validate it by analyzing the behavioral transitions of language usage over the recent several decades in Singapore and Hong Kong. In each case, we estimate from data the model parameters that measure each language utility for its speakers and the strength of two biases, the majority preference for their language, and the minority aversion to it. The values of these two biases decide which language is the fastest growing in the competition and what would be the stable state of the system. We also study the system convergence time to stable states and discover the existence of tipping points with multiple attractors. Moreover, the critical slowdown of convergence to the stable fractions of language users appears near and peaks at the tipping points, signaling when the system approaches them. Our analysis furthers our understanding of multiple language evolution and the role of tipping points in behavioral transitions. These insights may help to protect languages from extinction and retain the language and cultural diversity.

Keywords Behavioral transition · Language competition · Data analysis

1 Introduction

Language is for its speaker an essential component of their culture with great importance also for business and economic activities, especially those involving international knowledge transfer [1], interdisciplinary research [2], or international management processes [3]. The dynamic of the language competition has attracted considerable attention in the past decades, resulting in the development of the mathematical models for competition between two languages [4], language acquisition, variation across languages [5], and dynamics of language norm changes [6]. Language competitive dynamics fuel the changes in fractions of speakers of various languages [7], and their collisions [8], blending, and evolution. The final outcome of such evolution can be dominance of one language over the others, the extinction of dominated languages [9], language coexistence, or unification of close languages into one. This process is affected by both internal and external factors. Internal factors represent inherent characteristics of languages, such as lexical and phonological factors [10]. External factors account for social, political and economic influences, such as "The Speak Good English Movement" in Singapore [11] and implementation of standardized Mandarin in China [12]. Both factors influence how people choose their languages and indirectly determine the fraction of speakers of languages, leading to equilibrium with different fractions of speakers for those languages. Languages under the multiple language competitive dynamics may be in one of three different states: (i) dominate state, i.e., entire population only speak this language; (ii) coexistence state, i.e., there are positive fractions of the population using this language; (iii) extinction state, no one speaks this language.
Language dynamics is important for understanding the connections between languages [13] which is related to the language competition and language learning [5], and second language acquisition [14] that supports the existence of bilingual interactive activation [15]. Furthermore, analysis and modeling of language competition can also be extended to social sciences such as population interactions [16], formation of collective opinions [17], cultural evolution [18], increases language and cultural diversity [19], and opinion dynamics [20].

In the past decade, a variety of models have been developed to understand the competitive dynamics of languages. Most of the attention has been paid to the coexistence of two languages [16, 7, 21] and the corresponding bilingualism [22]. One research in this area is finding factors that affect the evolution of competing languages, such as social interaction networks [7], or microscopic competition between two languages [21]. Another research area is modeling abstract competition factors by using real data to estimate the model parameters for transforming language from old to new forms [23] or estimating all parameters in language competition [24]. Moreover, some researches focus on combining the area of language dynamics with the area of statistical physics [20, 25] or applying to them statistical laws to describe word uses [26]. These previous works develop approaches to modeling, analyzing, and even quantifying the competition and evolution of languages, enabling us to theoretically construct and simulate dynamics of two languages. In [27], the authors extend the Abrams-Strogatz model to the competition among multiple languages, but their model becomes very complicated with large number of languages. Still, this multiple language competition model has not been validated on real-word data. Thus, we ask the critical questions: What are the critical parameters that determine the existence and values of tipping points of language coexistence and extinction in the community? And is there a system metric that can indicate the approaching extinction of a language?

Here, we answer the questions raised above by using real-world language evolution data from Singapore [28] and Hong Kong [29, 30, 31] to find optimal parameters and validate the extended Abrams-Strogatz model [27]. The model parameters found in this process drive the utilities of competing languages, the strength of majority preference for the most popular language and minority aversion to this language and measure this utility’s impact on language evolution. Currently, all the modeled languages coexist. We investigate the behavioral transitions of the languages under the perturbations of each parameter. We find that the language with the highest language utility tends to grow faster and eventually gain the largest fraction of speakers. Moreover, when the majority preference is small than a certain critical value, the popular languages may lose its leading position during the evolution process. Finally, when the minority aversion is not sufficiently strong, the languages with small (including the language with the smallest) initial fractions of speakers may gain the largest fraction of speakers. From the above analysis, we obtain the complete phase diagram for each community, showing the evolution of each language in each dataset and the relation between transitions and parameters. Secondly, we analyze the relation between convergence time and state of competing languages, and show that the competition arises to the highest level when the language dominance switches from one language to another. Finally, we illustrate individual and combined effects of two language biases, the majority preference and the minority aversion, by simulating how languages with the largest initial fractions of speakers are affected.

2 RESULT

2.1 DATA AND MODELING

To model the real world language competitions, we use dataset of languages used in Singapore (the whole country), languages used in the Chinese community of Singapore, languages used in Indian Community of Singapore, and languages used in Hong Kong. We consider speakers of one language in our dataset as people who consider this language as their primary language.

**Singapore.** In the 1950s, dialects such as Hokkien were the most widely spoken language in Singapore. In the 1957 census, about 1.8% people mainly spoke English, about 0.1% people mainly spoke Chinese Mandarin. However, after the implementation of a series of policies from the 1950s to present, the proportions of speakers of different languages in Singapore considerably changed. Until 2010, English and Chinese Mandarin became the most spoken languages with speakers proportion of 32.3%, 35.6%, respectively in Singapore’s entire country [28].

**Hong Kong.** We use language data collected between 1949 to 2016 in Hong Kong [30, 31, 29]. The number of people who mainly speak English increased in these 67 years, surpassing the number of people who mainly speak Hakka, Hoklo, or Sze Yap. We do not include Cantonese speakers in our dataset since the number of people who use Cantonese as their common language is much larger than the speaker population in our dataset. We normalize all languages in our dataset before fitting them into our model.
We employ the extended Abrams-Strogatz model [27] to test the competition among multiple languages,

\[
\frac{dx_i}{dt} = \sum_{j=1, j \neq i}^{n} x_j P_{ij} - x_i \sum_{j=1, j \neq i}^{n} P_{ij}
\]

(1)

where \(x_i\) is the fraction of the population speaking language \(i\), and \(P_{ij}\) represents the transition rate from language \(j\) to language \(i\).

\[
P_{ij} = s_i x_i^\beta (1 - x_j)^{\alpha - \beta}
\]

(2)

where \(\beta \geq 0\) and \(\alpha - \beta \geq 0\) represent the strength of the majority preference and the minority aversion, respectively. \(s_i > 0\) is the utility of language \(i\), and \(\sum_{i=1}^{n} s_i = 1\).

We utilize numerical simulation to compute the parameters in our model. The following equation calculates the difference:

\[
D = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} (x_{i,j} - x'_{i,j})^2}
\]

(3)

where \(i\) is the index of languages varying from 1 to \(n\) which is the number of total time steps, \(j\) is the index of languages varying from 1 to \(m\), which is the number of languages. \(x_{i,j}\) is the rational value of fraction of language \(j\) users at time step \(i\), \(x'_{i,j}\) is the theoretical value of a fraction of language \(j\) users in time step \(i\).

To fit this model to our language data, we set a range for parameter majority preference \(\beta\) and minority aversion \(\alpha - \beta\), and then, iterating over this range of majority preference and minority aversion, we evaluate possible values for language utilities \(s_1, s_2, \ldots, s_m\). Given \(\beta\) and \(\alpha - \beta\) to minimize the difference \(D\) between real language fractions and theoretical language fractions we calculate from Eq. [1]. Then we repeat narrowing range of parameters and doing grid–search for language utilities, \(\beta\), and \(\alpha - \beta\) to keep increasing the precision of our theoretical model. We end the repetition until our theoretical model is close enough to the real world language data. Note that since parameter \(s_i\), which represents state utility in extended Abrams-Strogatz model, acts as a useful or advantageous factor or feature of state \(i\), in our language competition model, we change the definition of \(s_i\) to the utility of language \(i\).

By fitting the language data into extended Abrams-Strogatz model, we compare the real evolution processes in different districts and their corresponding simulation results in Fig. [1]. It is clear that the model successfully represents the behavior of the data, correctly showing the evolution of each language. In Singapore, the whole country dataset, Dialect starts with about 0.975 fractions of speakers and is surpassed by Mandarin in 1996, by English in 1997. As for Chinese community of Singapore dataset, the trend of speakers in a different language is similar to that of speakers in Singapore whole country dataset where English and Mandarin gradually replace dialects: it starts with about 0.766 fractions of speakers and is surpassed by Mandarin in 1994, by English in 2001. In the Indian community of Singapore dataset, Tamil started with the most substantial fraction (about 0.613) of speakers but continually lost its speakers and eventually was exceeded by English in 2003. The commonality of these three datasets is the increase of English speakers, which might be caused by the increasing number of English medium schools [28] in this period. The increasing usage of Mandarin in Singapore the whole country dataset and Chinese community dataset might benefit from “Speak Mandarin Campaign” implemented in 1979 in Singapore [32]. In Hong Kong dataset, Sze Yap, which is a Chinese vernacular in Hong Kong, owns the most substantial fraction (about 0.578) at the beginning, but is gradually replaced by Mandarin in 1994, by English in 2001. In the Indian community of Singapore dataset, Mandarin starts with about 0.975 fractions of speakers and is surpassed by Mandarin in 1996, by English in 1997. As for Chinese dialects, Dialect starts with the most substantial fraction (about 0.613) of speakers but continually loses its speakers and eventually was exceeded by English in 2003.

2.2 STATE DEFINITION

We will refer to the language with the largest number of speakers as the most popular, but if this language drives the competing languages to extinction, we will refer to is as dominant.

Here we need to define each state in the evolution of language competition. In our simulations, two distinct states are defined as: “coexistence state” and “dominance state”. “Coexistence state” arises if at least two languages survive but in this state, an extinction of some languages is still possible. In contrast, in “dominance state”, the survival of one language leads to the extinction of all others. The “coexistence state” and “dominance state” are illustrated in Fig. [2] where the left column shows example of “coexistence state” with its title showing the language with largest fraction of speakers, while the right column displays examples of “dominance state”. Note that the “coexistence state” and “dominance state” here all refer to the state after fractions of speakers for all languages converge to their steady states.

As shown in Fig. [2], during language competition, Dialect’s fraction of speakers drops dramatically, while English’s fraction of speakers increases. Eventually, none of the competing languages disappear, with English being the most
When speakers drops significantly. Eventually, Mandarin becomes dominant when its utility exceeds 0.36. English is dominant. Then, for exceeds 0.36, it becomes dominant. Similarly, in Fig. 3C, in the beginning, when Mandarin utility reaches "coexistence state" in which all competing languages’ fractions change dramatically. When the utility of Dialect "dominance state" as well (language utility of Dialect), error is . (C) Language data of Indian community in Singapore between 1957 and 2010. Parameters fit to dynamic model (, , English: , Dialect: , Mandarin: , error is . (D) Language data of Hong Kong between 1949 and 2016. Parameters fit to dynamic model (, , English: , Tamil: , Malay: , error is . Real data are shown by dots, theoretical data are shown by solid lines, real and theoretical data are all normalized. X represents the fraction of language, represents the time point. (A) Language data of the whole country in Singapore between 1957 and 2010. Parameters fit to dynamic model (, 0.824, English: , Dialect: , Mandarin: , error is . (B) Language data of Chinese community in Singapore between 1957 and 2010. Parameters fit to dynamic model (, 0.987, English: , Dialect: , Mandarin: , error is . (C) Language data of Indian community in Singapore between 1957 and 2010. Parameters fit to dynamic model (, , English: , Tamil: , Malay: , error is . (D) Language data of Hong Kong between 1949 and 2016. Parameters fit to dynamic model (, , English: , Hakka: , Hoklo: , Sze Yap: , error is . Figure 1: The language evolution over time. Real data are shown by dots, theoretical data are shown by solid lines, real and theoretical data are all normalized. X represents the fraction of language, represents the time point. (A) Language data of the whole country in Singapore between 1957 and 2010. Parameters fit to dynamic model (, 0.726, English: , Dialect: , Mandarin: , error is . (B) Language data of Chinese community in Singapore between 1957 and 2010. Parameters fit to dynamic model (, 0.998, English: , Dialect: , Mandarin: , error is . (C) Language data of Indian community in Singapore between 1957 and 2010. Parameters fit to dynamic model (, , English: , Tamil: , Malay: , error is . (D) Language data of Hong Kong between 1949 and 2016. Parameters fit to dynamic model (, , English: , Hakka: , Hoklo: , Sze Yap: , error is .

popular language. In contrast, English is shown to be the only survivor after the language competition in Fig. 2B, because English gains a substantial fraction of speakers while other languages lose almost all theirs. In Fig. 2C, Dialect loses a large number of speakers in the competition, causing its competitors to gain more speakers, but it still holds the most speakers and coexists with other competing languages. Dialect surpasses English and becomes dominant in Fig. 2D, where fractions of speakers of languages other than Dialect all reduce to a shallow level. Similarly in Fig. 2E, Mandarin owns the largest fraction of speakers and stays in "coexistence state," while in Fig. 2F, Mandarin leads to the extinction of all other languages. As we took the Singapore dataset as an example to define different states, we are now able to describe future simulations in a more straightforward way.

### 2.3 State Utility $s_i$

In the extended Abrams-Strogatz model, $s_i$ represents the utility of language $i$. Accordingly, in our language dataset, each language $i$ has its own $s_i$ representing its utility. We analyze the relation between $s_i$ and the competition between language $i$ and other languages when $s$ is in the range $[0, 0.6]$. Since the total of all language utilities is by definition 1, in each simulation, with the increasing of the utility of one language, increasing utility of one language decreases utilities of other languages proportionally to their current utility. In the Singapore whole country dataset (Fig. 2A-2C), we set majority preference $\beta = 0.726$ and minority aversion $\alpha - \beta = 0.283$. For Fig. 2A, when $s_i \in [0, 0.32]$, languages are in "dominance state" with Mandarin being dominant. As $s_i$ further increases, languages come into "coexistence state" for a short while, and the fraction of English speakers surpasses the fraction of Mandarin speakers for $s_i > 0.34$. When $s_i > 0.36$, English becomes dominant and drives others to extinction. In Fig. 2B, language competition starts in "dominance state" as well (language utility of Dialect $s_j \in [0, 0.24]$), with Mandarin being dominant. Then, the system reaches "coexistence state" in which all competing languages’ fractions change dramatically. When the utility of Dialect exceeds 0.36, it becomes dominant. Similarly, in Fig. 2C, in the beginning, when Mandarin utility $s_k \in [0, 0.32]$, English is dominant. Then, for $s_k \in [0.34, 0.36]$, the system is in "coexistence state" and the fraction of English speakers drops significantly. Eventually, Mandarin becomes dominant when its utility exceeds 0.36.
Figure 2: We take Singapore dataset to illustrate "coexistence state" and "dominance state" of a language. "X" represents the fractions of languages and $\tau$ represents time steps. (A) After $\tau = 2500$, competing languages stay in coexistence state with English acquiring the largest fraction of speakers. (B) English is dominant since from $\tau = 2100$, the fractions of all other languages decrease to an extremely low level. (C) Dialect gets the largest fraction of speakers without leading to the extinction of Dialect or Mandarin. (D) Dialect’s fraction of speakers increase to an extremely high level, leading to the extinction of other languages. (E) Mandarin’s fraction of speakers grows and surpass English and Dialect. Eventually, all languages stay in coexistence state and Mandarin owns the largest fraction of speakers. (F) As Mandarin increases its number of speakers significantly, other competing languages go extinct.
Figure 3: Fraction of language speakers in relation to this language utility, where $X^*$ represents the fraction of a language users at the steady state of competition, and $s$ stands for language utility. (A) Fraction of language speakers in relation to the English utility in the Singapore whole country dataset. (B) Fraction of language speakers in relation to Dialect utility in the same dataset. (C) Fraction of language speakers in relation to Mandarin utility in the same dataset. (D) Fraction of language speakers in relation to English utility in Chinese community in Singapore. (E) Fraction of language speakers in relation to Dialect utility in Chinese community in Singapore. (F) Fraction of language speakers in relation to Mandarin utility in Chinese community in Singapore. (G) Fraction of language speakers in relation to English utility in Indian community in Singapore. (H) Fraction of language speakers in relation to Tamil utility in Indian community. (I) Fraction of language speakers in relation to Malay utility in Indian community. (J) Fraction of language speakers in relation to English utility in Hong Kong. (K) Fraction of language speakers in relation to Hakka utility in Hong Kong. (L) Fraction of language speakers in relation to Hoklo utility in Hong Kong.
In Chinese community dataset (Fig. 3 DEF), we set majority preference \( \beta = 0.63 \) and minority aversion \( \alpha - \beta = 0.36 \). In Fig. 3D, Mandarin is dominant when \( s_i \in [0, 0.3] \). Then the system reaches "coexistence state" when \( s_i \) increases. With further increase of \( s_i \), Dialect becomes dominant. In Fig. 3E, when language utility of Dialect \( s_j \in [0, 0.2] \), Mandarin is dominant. When \( s_j \in [0.22, 0.4] \), the system is in "coexistence state". When \( s_j > 0.4 \), Dialect becomes dominant. In Fig. 3F, when language utility of Mandarin \( s_k \in [0, 0.18] \), English is dominant. When \( s_k \in [0.2, 0.4] \), the system is in "coexistence state". When \( s_k > 0.4 \), Mandarin becomes dominant.

In Indian community dataset (Fig. 3 GHI), we set majority preference \( \beta = 0.21 \) and minority aversion \( \alpha - \beta = 0.82 \). In Fig. 3G, when language utility of English \( s_j \in [0, 0.32] \), Tamil, which is the language with the largest utility, is dominant. Then the system reaches "coexistence state" as the utility of English further increases. When \( s_i \) exceeds 0.44, English starts to be dominant. In Fig. 3H, when language utility of Tamil \( s_j \in [0, 0.33] \), English is dominant. When \( s_j \in [0.34, 0.45] \), the system is in "coexistence state". When \( s_j > 0.45 \), Tamil becomes dominant. In Fig. 3I, English is dominant for the initial range of \( s_k \), then its fraction of speakers decrease significantly, causing other languages to grow as the system enters "coexistence state." When \( s_k > 0.43 \), Malay is dominant.

In Hong Kong dataset (Fig. 3 KL), we set majority preference \( \beta = 0.987 \) and minority aversion \( \alpha - \beta = 0.095 \). In Fig. 3J, when language attraction of English \( s_i \in [0, 0.28] \), Hoklo is dominant. Then Hoklo and English’s fractions of speakers change dramatically, leading to the dominance of English. In Fig. 3K, when language attraction of Hakka \( s_j \in [0, 0.27] \), English is dominant. Then, as language utility of Hakka exceeds 0.27, Hakka becomes dominant. Similarly, in Fig. 3L, in the beginning, English is dominant, but for \( s_k > 0.28 \), Hoklo becomes dominant.

From these simulations, we find that when the language utility \( s_i \) of language \( i \) is relatively small, one of the other languages, which is usually the language with the highest language utility, tends to be dominant. As \( s_i \) increases, languages might come into "coexistence state," which acts as a transition period for language \( i \) to become dominant. Moreover, when \( s_i \) is large enough, language \( i \) becomes dominant.

### 2.4 CONVERGENCE TIME

It is notably hard to predict the critical transition from one state to another because the state of the system may show little change before the tipping point [33]. Critical slowdown [34] defined in statistical physics is an indicator for early warning signals with applications to many fields, ranging from the economy [35] to ecology [36]. Here we employ critical slowdown as early warning signals for the behavioral transition in the language competition. Fig. 4 shows the convergence time of different datasets under different language competitions, where the \( x \)-axis represents the initial fraction of one language, the left \( y \)-axis represents the equilibrium fraction of each language, and the right \( y \)-axis represents the convergence time \( \tau \). In Fig. 4A, \( \tau \) reaches a peak when the initial fraction of Dialect increases to 0.56 and Dialect replaces English as dominant language. Similarly, at \( \tau \) in Fig. 4B, when the system transitions from "dominance state" to "coexistence state", \( \tau \) reaches its peak. In Hong Kong datasets, we find similar outcome. The convergence pattern observed in Fig. 4C is similar to the one seen in Fig. 4A, since the peak of \( \tau \) happens when the dominance switches from one language to another. As for Fig. 4D, the peak of \( \tau \) happens when the system transitions from "coexistence state" to "dominance state," which is exactly opposite to the transition in Fig. 4A, yet they show similar patterns of convergence time.

From these simulations, the peak of convergence time happens when the state transition happens of either switching the dominant language or moving from "coexistence state" to "dominance state," or vice versa. Such transitions can be caused by the comparable competing strength of different languages. The convergence time enables us to identify the "tipping point" of the system parameter values giving the control system enough time to prevent unwanted transition.

### 2.5 SENSITIVITY TO MAJORITY PREFERENCE AND MINORITY AVERSION

Here, we focus on how majority preference and minority aversion can affect language competitions. For each dataset, we set all parameters and languages’ initial fractions of speakers to the values used in modeling and data section except for majority preference \( \beta \) and minority aversion \( \alpha - \beta \).

In Fig. 5A, we have \( \beta = 0.726 \) and \( \alpha - \beta \) ranging from 0 to 1 with step 0.01. In the range of \( \alpha - \beta \in [0, 0.31] \), Mandarin has the largest fraction of speakers, and all languages are in "coexistence state". In the range of \( \alpha - \beta \in [0.32, 0.37] \), Mandarin is dominant, causing the extinction of Dialect and English. When the minority aversion start to exceed 0.38, then Dialect, which is the language with the largest initial fraction in this dataset, starts to be dominant. In Fig. 5C, in the range of \( \alpha - \beta \in [0, 0.4] \), the system stays in "coexistence state" and English is dominant. When \( \alpha - \beta \) further increases, English still is dominant, but as the minority aversion becomes large enough, language with the largest initial fraction (Dialect) of speakers starts to be dominant. In Fig. 5E, the system starts in "coexistence state" and English, which is the language with the largest language utility (\( s = 0.4 \)), is dominant in the range of \( \alpha - \beta \in [0, 0.96] \). Then,
when $\alpha - \beta \in [0.98, 1.02]$, English is again dominant. Tamil, the language with the largest initial fraction of speakers, is dominant when its minority aversion is larger than 1.02. In Fig. 5G, when $\alpha - \beta \in [0, 0.07]$, English, which is the language with the largest language utility ($s = 0.297$), is most popular, but the system is in "coexistence state". When $\alpha - \beta \in [0.08, 0.56]$, English is again dominant. When $\alpha - \beta \in [0.57, 0.71]$, Hakka and English compete with each other for dominance, with Hakka is dominant when $\alpha - \beta \in [0.75, 0.8]$ and English is dominant when $\alpha - \beta \in [0.66, 0.71]$. This may be caused by the two languages’ similar level of competitiveness determined by both initial fraction and utility. However, when minority aversion is high enough (larger than 0.71), Sze Yap, which is the language with the highest initial fraction, becomes dominant.

From these simulations, we conclude that low value of minority aversion favors the growth of languages with small initial fraction, and usually the language with the largest language utility $s$ among them, can gain the largest fraction of speakers. In contrast, high minority aversion favors the growth of languages with the largest initial fraction of speakers. In Fig. 5B we set the minority aversion $\alpha - \beta = 0.28$, and vary the majority preference $\beta$ from [0, 1]. Similar to language patterns in Fig. 5A, languages start in "coexistence state", with Mandarin owning the largest fraction of speakers when $\beta \in [0, 0.74]$. When the majority preference increases to ($\beta \in [0.75, 0.8]$), Mandarin becomes dominant, even though it had the smallest initial fraction of speakers. As the majority preference further increases and until it reaches 0.82, Dialect, the language with the largest initial fraction, starts to be dominant.

In Fig. 5D, when $\alpha - \beta = 0.36$, the system stays in "coexistence state", with Mandarin being dominant in the range of $\beta \in [0, 0.66]$. Then as $\beta$ increases to ($\beta \in [0.67, 0.83]$), Mandarin is dominant for a short range of $\beta$ because for $\beta > 0.83$, Dialect, the language with the smallest initial fraction, is dominant. In Fig. 5E, at the beginning with ($\beta \in [0, 0.3]$), English, which is the language with the largest language utility, has the largest fraction of speakers and all languages are in "coexistence state." For $\beta \in [0.31, 0.34]$, English is dominant, but for larger $\beta$, Tamil becomes dominant. Similarly, as shown in Fig. 5H, the system transitions from "coexistence state" to "dominance state", and the language with the largest initial fraction becomes dominant when majority preference is high enough; in this case, for $\beta \in [0.76, 0.96]$ English is dominant, while for $\beta \in [0.98, 1.16]$, Hakka takes over this role. Finally, when $\beta > 1.16$, Sze Yap, the language with the largest initial fraction, is dominant.

From these four simulations, we find that when majority preference is small, languages with small initial fraction (usually the language with the largest language utility) might own the largest fraction of speakers, and becomes dominant. However, when majority preference is high enough, the language with the largest fraction of speakers usually becomes dominant, driving other language to extinction.
Figure 5: The change of equilibrium point when $\alpha - \beta$ or $\beta$ change in each language dataset. (A) The relation between $\alpha - \beta$ and fraction of each language in the Singapore whole country dataset. (B) The relation between $\beta$ and fraction of each language in the Singapore whole country dataset. (C) The relation between $\alpha - \beta$ and fraction of each language in Chinese community in Singapore. (D) The relation between $\beta$ and fraction of each language in Chinese community in Singapore. (E) The relation between $\alpha - \beta$ and fraction of each language in Indian community in Singapore. (F) The relation between $\beta$ and fraction of each language in Indian community in Singapore. (G) The relation between $\alpha - \beta$ and fraction of each language in Hong Kong. The relation between $\beta$ and the fractions of each language speakers in Hong Kong.
Figure 6: States of the language with the largest fraction of speakers as a function of increasing the majority preference and the minority aversion. "Dominate" denotes one language domination over others, otherwise, all languages shown in the figure coexist. (A) The language with the largest fraction of speakers in relation to $\alpha - \beta$ and $\beta$ in Chinese community in Singapore. (B) The language with the largest fraction of speakers in relation to $\alpha - \beta$ and $\beta$ in Indian Community in Singapore. (C) The language with the largest fraction of speakers in relation to $\alpha - \beta$ and $\beta$ in the Singapore whole country data set. (D) The language with the largest fraction speakers in relation to $\alpha - \beta$ and $\beta$ in Hong Kong.

2.6 PHASE DIAGRAM

As for the combined effect of majority preference and minority aversion, we still use the language initial fraction of speakers and parameters from the data and modeling section. Here, we only consider the language with the largest fraction of speakers. In Fig. 6, when majority preference $\beta \in [0, 0.98]$, three different kinds of patterns appear. In the first pattern, all three languages (English, Dialect, and Mandarin) coexist and Mandarin is the most popular. The next pattern has Mandarin dominant. The third pattern has Dialect dominant. As $\beta$ increases from 0 to 0.98, Dialect becomes more and more likely to become dominant because the range of $\alpha - \beta$ for Dialect to be in this role becomes larger and larger. When $\beta \in [0.98, 1.1]$, two different patterns arise. In the first, Mandarin is dominant while in the second, it is Dialect which is dominant. When $\beta$ exceeds 1.1, Dialect is dominant. Fig. 6B shows that for the majority preference $\beta \in [0, 1]$, again three different patterns arise. In the first one, three languages (English, Tamil, Malay) coexist and Tamil is the most popular. In the second pattern, English is dominant, while in the third it is Tamil that is dominant. Similar to Fig. 6A, the language with the largest utility (Mandarin in Fig. 6A and English in Fig. 6B) tends to be the most popular when the majority preference and minority aversion are small. With $\beta \in [1, 1.02]$, two different patterns are present. In the first, again three languages coexist, and English is the most popular. The second has English dominant. The third pattern arises when $\beta > 1.02$, Tamil, which is the language with the largest initial fraction of speakers, is dominant.

Fig. 6C shows that when minority aversion $\alpha - \beta \in [0, 0.04]$, four different patterns arise. In the first, again three languages (English, Dialect, Mandarin) coexist and Mandarin has the most speakers. The second has Mandarin dominant, while in the third pattern it is English that is dominant. Finally in the fourth case Dialect is dominant. In this case, even English can be dominant over a short range of parameters, because of its comparably high initial fraction and language utility; it has the second largest initial fraction and the second largest language utility. With
the first three patterns from the case of lowest β reappear. Again, Mandarin is the language with the largest utility and it is the most when minority aversion and majority preference are small. When α − β is greater than 0.56, Dialect is dominant. In Fig. 6D, with minority aversion \( \alpha − \beta \in [0.06, 0.56] \), four different patterns arise. In the first, four languages (English, Hakka, Hoklo, Sze Yap) coexist, and English is the most popular. In the second, English is dominant while in the third it is Hakka that is dominant. Finally, in the fourth, Sze Yap is dominant. When \( \alpha − \beta \in [0.68, 2] \), the previous patterns, the first, the second, and the fourth reappear. In this dataset, English has the largest language utility and owns the largest fraction of speakers when the majority preference and minority aversion are small.

When the majority preference and the minority aversion are relatively small, competing languages tend to coexist with each other and language with the largest utility tends to own the most speakers. Hence, when the majority preference and the minority aversion are relatively small, they affect language competition weakly. It is the language utility that plays an essential role in this competition. As the majority preference and the minority aversion increase, some languages become dominant. When the majority preference and the minority aversion further increase, the language with the largest initial fraction is dominant, indicating that when the majority preference and the minority aversion are large enough, they will favor the growth of the language with the largest initial fraction of speakers and make it dominant.

### 3 DISCUSSION

Here we provide a model and its validation using four real-world language competitions involving several languages that extends the Abrams-Strogatz model in the important direction. The model fits well with the real data, as shown in the first section, enabling us to further analyze factors affecting the language competition in detail. Our contributions can be summarized as follows.

1. We show that language utility affects the competitive evolution of communities using several languages. When the utility of a language is low enough, it might go extinct, but when this utility is high enough, it can be dominant and drive other languages to extinction. However, it is also possible that with the value of this utility in mid-range the system can be in a "coexistence state".

2. The relation between convergence time and state transition of languages shows that convergence time to steady state fractions of language users reaches a peak at the state transition tipping points. Such critical slowdown can be caused by similar competitiveness of the different competing languages. At the tipping points either one dominant language is replaced by another, or the system transitions from a "dominance state" to "coexistence state" or vice versa.

3. We demonstrate the influence that the majority preference and the minority aversion can separately exert on competing languages. When majority preference is small, a language with small initial fraction of speakers (including the language with the smallest initial fraction of speakers) can have the most speakers after the language competition, and even lead to the extinction of other languages. When the majority preference is large enough, the language with the largest initial fraction of speakers will win the language competition, usually driving all other languages to extinction. The simulations with varying the minority aversion yield similar results as described above.

4. We discuss also the influence that the majority preference and the minority aversion can together exert on the evolution of competing languages. When both of these biases are relatively low, a language with the small initial fraction of speakers can gain the most speakers in the steady state, and become dominant. When both biases are high enough, the language with the most speakers initially is most likely to be dominant. Moreover, there are variants of these conditions yielding results other than the two reported above, as shown in Fig. 6C where English is dominant as shown in Fig. 6D where Hakka is dominant.

Our simulations illustrate the results of the language competition under various conditions, providing examples of the impact of the language utility, the majority preference, and the minority aversion on the competition outcomes. Yet, analytical formulas defining quantitatively the competitive evolution of languages as a function of time and different parameters of the model are not known yet. Moreover, our simulations are completed without considering geographical [37], physiological [38], and other factors. Hence, constructing a conclusive formulation of the language competition in the real world requires futures future research.

### 4 DATA ACCESSIBILITY

All data needed to evaluate the conclusions in the paper are present in the paper itself or are available upon request from the authors.
5 AUTHOR’S CONTRIBUTIONS

Z.Z., B.K.S and J.G. conceptualized the study; Z.Z. simulated extensively; Z.Z., B.K.S and J.G. wrote the paper.

6 COMPETING INTERESTS

We declare we have no competing interests.

7 ACKNOWLEDGEMENTS

This study was supported in part by the Office of Naval Research (ONR) Grant No. N00014-15-1-2640, and by the Army Research Office (ARO) Grant No. W911NF-16-1-0524.

References

[1] Denice E Welch and Lawrence S Welch. The importance of language in international knowledge transfer. Management International Review, 48(3):339–360, 2008.

[2] Louise J Bracken and Elizabeth A Oughton. ‘what do you mean?’the importance of language in developing interdisciplinary research. Transactions of the Institute of British Geographers, 31(3):371–382, 2006.

[3] Denice Welch, Lawrence Welch, and Rebecca Piekkari. Speaking in tongues: The importance of language in international management processes. International Studies of Management & Organization, 35(1):10–27, 2005.

[4] Marco Patriarca and Teemu Leppäinen. Modeling language competition. Physica A: Statistical Mechanics and its Applications, 338(1-2):296–299, 2004.

[5] Elizabeth Bates, Brian MacWhinney, and B MacWhinney. Competition, variation, and language learning. Mechanisms of language acquisition, pages 157–193, 1987.

[6] Roberta Amato, Lucas Lacasa, Albert Díaz-Guilera, and Andrea Baronchelli. The dynamics of norm change in the cultural evolution of language. Proceedings of the National Academy of Sciences, 115(33):8260–8265, 2018.

[7] Marco Patriarca, Xavier Castelló, José Ramón Uriarte, Víctor M Eguíluz, and Maxi San Miguel. Modeling two-language competition dynamics. Advances in Complex Systems, 15(03n04):1250048, 2012.

[8] Bruce Bueno De Mesquita and Johanna DeStefano. When languages collide: Perspectives on language conflict, language competition, and language coexistence. Ohio State University Press, 2003.

[9] Daniel M Abrams and Steven H Strogatz. Linguistics: Modelling the dynamics of language death. Nature, 424(6951):900, 2003.

[10] Albert Costa, Àngels Colomé, Olga Gómez, and Nuria Sebastián-Gallés. Another look at cross-language competition in bilingual speech production: Lexical and phonological factors. Bilingualism: Language and Cognition, 6(3):167–179, 2003.

[11] Lari-Valteri Suhonen. Speak good english movement in singapore: Reactions in social and traditional media, 2011.

[12] Yang You. Language unification, labor and ideology. Technical report, Working Paper, 2018.

[13] Michael Tomasello. Constructing a language. Harvard university press, 2009.

[14] Kees De Bot, Wander Lowie, Marjolyn Verspoor, and Marjolijn Henriëtte Verspoor. Second language acquisition: An advanced resource book. Psychology Press, 2005.

[15] Ton Dijkstra, Walter JB Van Heuven, and Jonathan Grainger. Simulating cross-language competition with the bilingual interactive activation model. Psychologica Belgica, 1998.

[16] Juan Pablo Pinasco and Liliana Romanelli. Coexistence of languages is possible. Physica A: Statistical Mechanics and its Applications, 361(1):355–360, 2006.

[17] Ryosuke Nishi and Naoki Masuda. Collective opinion formation model under bayesian updating and confirmation bias. Physical Review E, 87(6):062123, 2013.

[18] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. science, 331(6014):176–182, 2011.
[19] Victor Ginsburgh and Shiomo Weber. How many languages do we need?: The economics of linguistic diversity. Princeton University Press, 2011.

[20] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. Reviews of modern physics, 81(2):591, 2009.

[21] Dietrich Stauffer, Xavier Castelló, Victor M Eguiluz, and Maxi San Miguel. Microscopic abrams–strogatz model of language competition. Physica A: Statistical Mechanics and its Applications, 374(2):835–842, 2007.

[22] Ricard V Solé, Bernat Corominas-Murtra, and Jordi Fortuny. Diversity, competition, extinction: the ecophysics of language change. Journal of The Royal Society Interface, 7(53):1647–1664, 2010.

[23] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. Quantifying the evolutionary dynamics of language. Nature, 449(7163):713, 2007.

[24] Menghan Zhang and Tao Gong. Principles of parametric estimation in modeling language competition. Proceedings of the National Academy of Sciences, 110(24):9698–9703, 2013.

[25] Vittorio Loreto, Andrea Baronchelli, Animesh Mukherjee, Andrea Puglisi, and Francesca Tria. Statistical physics of language dynamics. Journal of Statistical Mechanics: Theory and Experiment, 2011(04):P04006, 2011.

[26] Alexander M Petersen, Joel Tenenbaum, Shlomo Havlin, and H Eugene Stanley. Statistical laws governing fluctuations in word use from word birth to word death. Scientific reports, 2:313, 2012.

[27] Ryo Fujie, Kazuyuki Aihara, and Naoki Masuda. A model of competition among more than two languages. Journal of Statistical Physics, 151(1-2):289–303, 2013.

[28] Francesco Cavallaro and Bee Chin Ng. Language in singapore: From multilingualism to english plus. Challenging the monolingual mindset, 156:33, 2014.

[29] Chun Fat Lau and Daniel WC So. From multi-dialectal society to cantonese dominant: A survey of language shift in hong kong 1949–1971. The Journal of Chinese Sociolinguistics, pages 89–104, 2005.

[30] "use of language by hong kong population". https://www.bycensus2016.gov.hk/tc/Snapshot-08.html

[31] John Bacon-Shone and Kingsley Bolton. Charting multilingualism: Language censuses and language surveys in hong kong. Language in Hong Kong at century’s end, pages 43–90, 1998.

[32] NG Chin Leong. A study of attitudes towards the speak mandarin campaign in singapore. Intercultural Communication Studies, 23(3), 2014.

[33] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. Nature, 461(7260):53, 2009.

[34] Marten Scheffer, Steve Carpenter, Jonathan A Foley, Carl Folke, and Brian Walker. Catastrophic shifts in ecosystems. Nature, 413(6856):591, 2001.

[35] Cees Diks, Cars Hommes, and Juuxi Wang. Critical slowing down as an early warning signal for financial crises? Empirical Economics, pages 1–28, 2015.

[36] TM Lenton, VN Livina, V Dakos, EH Van Nes, and M Scheffer. Early warning of climate tipping points from critical slowing down: comparing methods to improve robustness. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370(1962):1185–1204, 2012.

[37] Marco Patriarca and Els Heinsalu. Influence of geography on language competition. Physica A: Statistical Mechanics and its Applications, 388(2-3):174–186, 2009.

[38] Patricia K Kuhl. Early language learning and literacy: neuroscience implications for education. Mind, Brain, and Education, 5(3):128–142, 2011.