BRILL

LANGUAGE DYNAMICS AND CHANGE 8 (2018) 219–252

brill.com/ldc

Testing an agent-based model of language choice on sociolinguistic survey data

Andres Karjus
University of Tartu, University of Edinburgh
andres.karjus@ut.ee

Martin Ehala
University of Tartu
martin.ehala@ut.ee

Abstract

The paper outlines an agent-based model for language choice in multilingual communities and tests its performance on samples of data drawn from a large-scale sociolinguistic survey carried out in Estonia. While previous research in the field of language competition has focused on diachronic applications, utilizing rather abstract models of uniform speakers, we aim to model synchronic language competition among more realistic, data-based agents. We hypothesized that a reasonably parametrized simulation of interactions between agents endowed with interaction principles grounded in sociolinguistic research would give rise to a network structure resembling real-world social networks, and that the distribution of languages used in the model would resemble their actual usage distribution. The simulation was reasonably successful in replicating the real-world scenarios, while further analysis revealed that the model parameters differ in importance between samples. We conclude that such variation should be considered in parametrizing future language choice and competition models.

Keywords

language competition – language choice – agent-based modelling – sociolinguistics – social networks

 | DOI:10.1163/22105832-00802004

1 Introduction

The issues of language competition and extinction have long been in the focus of sociolinguists and language sociologists. Language extinction occurs mainly through language shift—a process in which users of one language gradually adopt another language for communication so that the old one falls out of use. Several complex theories have been developed to explicate the causal factors in this process (cf. Giles et al., 1977). While there is consensus about the major factors that affect language shift, "no instrument powerful enough to assess language shift adequately on a large scale has yet been devised" (Clyne, 2003). Since the seminal paper by Abrams and Strogatz (2003), computational modelling of language competition has become increasingly widespread, bringing a new perspective into the discussion and potentially providing such an instrument.

The mathematical model proposed by Abrams and Strogatz (2003) was quite simple, assuming populations of uniformly connected monolingual speakers of two languages, and explaining language extinction as an outcome of the interaction of two abstract variables (status/prestige and volatility, i.e., the speed of the shift), validated against census data of the proportions of speakers in several communities. Since then, numerous additions, variations and revisions to the model have been proposed, with methods drawn from the fields of differential equations, computational agent-based modelling, and game theory (for previous research, cf. Patriarca and Leppänen, 2004; Isern and Fort, 2004; Kandler et al., 2010; Minett and Wang, 2008; Patriarca et al., 2012; Castelló et al., 2013; Zhang and Gong, 2013; also, cf. Beltran et al., 2009 for a seemingly independent, yet very similar study). Agent-based modelling (cf. Sterling and Taveter, 2009 for an overview) is particularly suitable for modelling language shift, since it enables modelling emergent, individual-level phenomena that are at the core of most social dynamics. This trend of computational modelling of language dynamics is also mirrored in research concerned with the competition of linguistic elements within languages (which may itself lead to creolization or signal impending language shift) (cf. Baxter et al., 2009; Jansson et al., 2015).

Ultimately the extinction of a language is a cumulative consequence of speakers choosing one language over the other for communication in a particular setting. The choice of a language depends on the language competencies and attitudes that are in turn influenced by a myriad of environmental factors. Mathematical or computational models of language competition are thus only rough approximations of the actual social forces that are in operation in the cases of language shift. In the aforementioned previous research, populations

(or agents) are commonly assumed to be homogenous across their respective communities, sharing the same language proficiency levels and opinions on the prestige of the languages available to them—a situation unlikely to be true in the real world. Another complicating matter is that language shift in time may occur over several generations of speakers who are likely to have variable language competencies, usage preferences and views on the prestige of the competing languages. However, because of the relatively long duration of the process of language shift, there is a shortage of reliable datasets that could be used to assess the accuracy of competing models. As a consequence, the validation of previous models of language shift has commonly relied on census data, which only reflects the coarse proportions of speakers (or as a proxy, ethnicities) in a given region over time.

The construction of a more precise model of language shift should start with the construction and realistic parametrization of a model of grassroots language choice, i.e. a synchronic model that can replicate adequately the use of competing languages in an existing multilingual society or community. Synchronic data are also much easier to obtain than diachronic data on language shift. If a synchronic model is found to adequately replicate language choice in different settings, it would be reasonable to allow for the expansion of the model to include a mechanism of change in language competencies that would affect language choice in favor of one or the other competing language, and lead to large-scale diachronic changes.

The model proposed in this article aims at this first goal—to build an agent-based model for simulating language choice by a set of speakers that have different levels of language competencies, social attitudes and linguistic preferences. We made use of data from a large-scale sociolinguistic survey conducted in Estonia to inform and validate the model (Ehala et al., 2015), as well as to construct agents that reflect the attributes of the survey participants. To test the model, we sampled speakers from three qualitatively and quantitatively different multilingual communities and assessed the accuracy of the model to predict the linguistic choices of the speakers. We assume that a realistic model of language ecology should incorporate a social network, as presumably speakers in the real world do not sample their communication partners from the whole population at random. To that end, we implemented a mechanism that allows networks to emerge among the simulated speakers.

In summary, as a departure from previous modelling research on language competition and shift, we propose a model of synchronic language choice, and validate it against synchronic (self-reported) language usage data. Also in contrast with previous research, we model the agents of the simulation using the reported attributes of the participants of the survey, instead of assuming

homogenous groups of speakers. We measure and report both the fit of the model in terms of the language use distribution and the level of correspondence of the emergent networks to metrics argued in current research to reflect the properties of real-world social networks.

The proposed model is the first step in building an agent-based model of Estonian language ecology that could eventually be, with a certain level of confidence, used to predict the long-term consequences of possible language-political or sociological interventions. For this purpose, another, similar survey should be conducted in the future to allow for a point of comparison and validation of the diachronic expansion of the model.

The paper is structured as follows. The first section gives a short overview of the data we used. The second section explicates the theoretical sociolinguistic model underlying our simulation and, in the least technical way possible, the technical aspects of the simulation model (which are further described in detail in the Appendix). The third section presents results, followed by discussion and conclusions.

2 The survey

This section serves to provide a brief but by no means exhaustive overview of the sociolinguistic situation of Estonia and the data we will be using to inform and validate our model of language choice. The official language of Estonia is Estonian (which about 69% of the 1.3 million population speak as their mother tongue, according to the 2011 census), but there is a considerable Russian-speaking minority (30% of the population). Our model is based on a survey conducted in 2015 that included 1006 participants across the country, aged between 15 and 74 (Linguistic attitudes in Estonia 2015 dataset, Ehala et al., 2015; the data and further details are available at the repository link listed in the references). We model the choice between three languages—Estonian and Russian as the native tongues, with English as a possible lingua franca option. In terms of second languages spoken, English is a widely known second language in Estonia, particularly among the younger generation. Estonian is a common second language among native speakers of Russian (which in turn is primarily a second language among the older generation of native Estonian speakers). Other common second languages are Finnish and German. Most people in Estonia speak at least one other language in addition to their native tongue, at least to some extent.

The data from the following questions in the survey questionnaire are used in the current study. Note that all the questionnaire data is self-reported in

nature. There were two versions of the questionnaire—one in Estonian and the other in Russian. The only difference between them was the language in which the questions were presented and, crucially, what was referred to as the second language in questions regarding language, as well as the 'other' ethnicity in the respective questions (these are represented in square brackets below).

The respondents were asked to report their language proficiencies in the following form: *Please assess your level of fluency in the Estonian | Russian | English language on a scale of 1–7* (*where 1 = don't know at all and 7 = full fluency*). In our model, homophily (see the model section below) is expressed by ethnic preference, which is derived from the mean value of the answers to the following three questions: *On a scale of 1 to 7, how easy is it, in your opinion, to:* (1) *... become friends with* [*Estonians | Russian speakers*]*?* (2) *... communicate with* [*Estonians | Russian speakers*] *in the workplace/at school?* (3) *... find basic common ground with* [*Estonians | Russian speakers*]*?* In our model, agents have different probabilities to accept the language choice of their interlocutor (refer to the model section for details), based on the mean value of answers to the following questions: *On a scale of 1–7, rate the extent to which you agree with the following statements:* (1) *I will always choose the language which the other participants of the conversation understand best.* (2) *If my conversation partner's proficiency in* [*Estonian | Russian*] *is low, then I would prefer to communicate in a language that they know better*.

For each respondent in the survey, the proportion of their language use (used for the validation of the models) was calculated by taking the mean of the questions asking about their language choice in communication with friends, hobby and sports companions, and service personnel. The questions measuring language choice had the following structure, with the language in question depending again on the language of the questionnaire: *In which language do you communicate with your friends/acquaintances? 1—only in* [*Estonian|Russian*]*, 2—mostly in* [*Estonian|Russian*]*, 3—more in* [*Estonian|Russian*] *than in another language, 4—equally in* [*Estonian|Russian*] *and in another language, 5—more in another language than in* [*Estonian|Russian*]*, 6—mostly in another language, 7—only in another language*. If the respondent chose options 2–7, they were further asked *What other language do you have in mind?* For the purposes of the analysis, the values were scaled to a range of 0 to 1 and reversed, so that 1 stands for using only the native language, and 0 for using only another language (or languages).

3 The model

This section lays out the basic premise of the model. The purpose of the agent-based simulation tool developed for this study is to test a possible (albeit simplified) model of language competition in the society, stemming from individual decisions at the speaker level and leading to the distribution of the language used, observable at the community level. We assume certain properties of human communication (as described below) and set up the simulation in a way that allows social networks to emerge among the agents representing the speakers. Although the simulation relies on a necessarily simplified model of human communication, we hope to capture the most important variables affecting language choice, and propose a method to validate the model and the relative importance of its parameters. In broad terms, the simulation cycles through a sufficiently large number of iterations, and on each iteration, two agents from the population attempt to find a common language to communicate in, by proposing the languages they know to their partner. Agents that succeed in finding a common language form a social tie or network link between them, which makes them more likely to be paired up for another communication attempt in the forthcoming iterations. Given favorable global parameters governing the weights of the input variables (see below for details), this may lead to the formation of a relatively stable social network among the agents. We validate the model by counting the languages used by agent pairs in the network and compare this distribution to the distribution of language use reported in the sociolinguistic survey.

3.1 *Properties of the agents*

Based on previous sociological work on social networks and intergroup communication, we hypothesized that the formation of a social tie between two individuals is affected by the following five factors, which will be explained in more detail below: 1) the availability of shared languages between them (language competencies); 2) ethnolinguistic identity; 3) the number of friends they already share, an effect often referred to as 'triadic closure'; 4) the extent of preference for interaction with people of similar ethnic background, or 'homophily'; and 5) the willingness to accept the communication partner's choice of language, or 'linguistic accommodation.' Each agent in the model is based on a participant in the survey.

Language competencies characterize the ability of individuals to use languages for communication. We assume that each adult individual has competency in at least one language, but may know further languages at different levels of fluency. In our model, agents can know up to three languages, based

on the most commonly spoken languages in Estonia: Estonian, Russian and English (cf. the survey section for details).

Ethnolinguistic identity signifies individuals' belonging to groups. Our model assumes two possible ethnolinguistic identities, based on the two main ethnolinguistic groups in Estonia: Estonians and Russians. The ethnolinguistic identity was, as a necessary simplification, determined by the choice made by the respondents between the two versions of questionnaire (Estonian or Russian). Note that, while ethnolinguistic identity is a categorical variable with two values, language competencies are continuous and provide fine-grained information about different patterns of multilingualism.

Triadic closure is a social regularity according to which people are more likely to interact and make friends with people with whom they share common friends (Bianconi et al., 2014; Stark, 2015). This factor was not measured by the sociological survey. However, the agents in the simulation were programmed to take triadic closure into account when choosing communication partners, i.e., the probability of interaction between agents that have a link to a shared third agent was set higher than between agents without triadic closure. Triadic closure enhances the emergence of clustering in the simulated network, which is among the assumed properties of real-world networks.

Homophily expresses a social tendency that people prefer to interact with people that share the same racial, cultural or ethnic background (McPherson et al., 2001; Stark, 2015; cf. Leetmaa et al., 2015 for an Estonian account). In our model, homophily is expressed by ethnic preference, based on the relevant questions in the survey (cf. Section 2 above for details). As the survey indicated variation among respondents on the scale of ethnic preference, the agents in our model are also set to prefer interactions with their co-ethnics, according to the values from the data.

Linguistic accommodation is a phenomenon whereby a communicator aligns to the language of the interlocutor (Giles, 2008). Research has shown that people have different propensities to accommodate, depending on the situation, the communication partner or intergroup attitudes. In previous language competition research, this has been commonly reflected by the more abstract notion of language status or prestige. In our model, agents have different probabilities to accept the language choice of their interlocutor, derived from the answers of the survey participants (cf. Section 2 for details).

3.2 *Networks in the model*

In addition to attributes of the agents directly derived from the survey as described above, we also implemented social networks into the model. This is based on the assumption that real-world networks exhibit community struc-

ture (Luthi et al., 2008). They are neither fully connected—such that everybody has an equal chance to talk to anyone, as is assumed in earlier models (following Abrams and Strogatz, 2003)—nor regular lattices, such that everyone only has a fixed number of neighbors to talk to (Castelló et al., 2013). Community structure in a social network is also analogous to ecological niches in that it reduces competition and allows for the segregated co-existence of languages (cf. Patriarca et al., 2012 for an overview of various network models).

However, our data consists of the answers of randomly selected respondents in the population of Estonia and does not include information about social networks, besides their self-reported attitudes and accounts of social activity. Various theoretical models for creating networks with realistic structure have been proposed (Toivonen et al., 2006), and while the pre-generation of networks for the simulation could be an avenue of future research, we opt for allowing for the emergence of networks in the model by incrementally increasing the strength of the connections between agents upon successful interactions (and decreasing upon failures) (cf. Gong et al., 2004 for a similar approach in the field of language-society coevolution).

3.3 *Overview of the simulation*

The simulation is built to iterate a chosen number of times. Each iteration (after the initialization phase) consists of pairing up two agents and an attempt of communication between them. Note that all the equations pertaining to the simulation model, along with a more precise definition of the algorithm, can be found in Appendix A. In broad terms, an iteration of the simulation proceeds as follows:

- An agent is chosen randomly from the set of all agents to be the 'starter' (the initiator of the act of interaction).
- Another agent is chosen randomly, but with weighted probabilities (see below), from the set of all agents (minus the starter), to be the 'partner' for the starter in the act of interaction.
- The two agents interact for up to three rounds, taking turns proposing a language to be used for communication; this has two possible outcomes:
  - A common language for communication is successfully agreed upon, the interaction ends with success
  - No common language can be agreed upon, the interaction ends with failure
- Depending on the outcome of the interaction, the following values are updated:
  - Link strength between the two agents (success strengthens the link, failure weakens it)

- Their memory of a common shared language (upon the next interaction, this will be the first proposal by whichever of the two will have the role of the starter agent)
- Their knowledge of each other's ethnicity (they are now aware of each other's true ethnic identity; upon first contact, they made a guess; cf. Appendix B for details).

In the model initialization stage, the self-reported values (for language proficiencies and behavioral preferences) of each respondent in the chosen sample are used to build one (or potentially more) agent(s). The samples in our experiment were chosen to be approximately the same size (about 70–80 agents each), based on the subsample sizes in the survey. Dealing with large and variable population sizes is of course an avenue for future research. Each agent has the following properties derived from the questionnaire data: 1) self-reported proficiency values for the three languages considered in the model; 2) ethnic self-categorization; 3) extent of ethnic homophily; 4) level of accommodation. The simulation furthermore includes global parameters that determine the weight of the impact that these properties have on language and partner choice (more below). The rest of this section describes the subroutines that make up the model—for a more detailed, algorithmic description (along with values for constants and further explanation of the modelling choices), please refer to the Appendix.

On each iteration of the simulation, two agents are partnered up using the Partner Selection subroutine. The two agents engage in an interaction (using the Interaction subroutine), which can have two outcomes, success or failure, as outlined above. Upon success, the language they used to successfully interact will be recorded as the common language of this pair of agents, and the link between them increases (but not over the maximum value of 1)—this is the low-level process that gives rise to networks in the model. Upon failure, any previous common language record is removed and the link between them weakens. Regardless of the outcome, two things occur in the end of each iteration: each agent in the chosen pair becomes aware of the other's ethnic identity, and all links between all agents weaken by a small decrement (but not below 0). The latter is based on the assumption that, if people do not interact for a while, their relationship suffers over time (but note that we handle 'time' only in the sense of iterations, not in a real-world sense, and all conclusions are drawn after a large number of such iterations, by which language choices have become more or less stable). If any link reduces to 0 again, the common language between these agents is erased and their knowledge (or rather, now a guess) of each other's ethnic identity is set back to the initial probabilistic value derived from the ethnic composition of the setting.

The Partner Selection subroutine works as follows: the starter agent is chosen randomly, and it is assigned a partner semi-randomly, the chance of becoming a partner being weighted. The weighting depends on the strength of the existing link between the two, the existence of mutual "friends" (triadic closure), and the ethnic preferences (level of homophily) of the starter agent, with the probabilities taken from a logistic (note, not logarithmic) transformation of the value (the value is always > 0, i.e., all agents are in principle capable of choosing anyone else in the sample). The global parameters controlling the weight of ethnic preference and the shape of the curve are discussed in the next section (cf. Figs 1, 2).

The Interaction subroutine consists of up to three rounds (since there are currently three languages in the model), with the agents taking turns proposing languages to use. In the first round, the starter agent proposes the language. If a common language has been recorded from an earlier interaction, this is automatically proposed to the partner first—the rationale being that once people get used to using a language among themselves, it would be reasonable to expect they will use it again the next time. Otherwise, the starter proposes the language it knows best (or randomly one of the languages it knows). The partner either accepts it (Interaction ends successfully) or not, whereby the Interaction continues into round two, with the partner agent now proposing a language, and so on. If the responding agent accepts a proposal at any point, the Interaction ends and success is reported. An agent always accepts its native language (maximal proficiency value) and always rejects a language it does not know (proficiency value 0). The linguistic accommodation value (and respectively the global parameter controlling its importance) plays a role when an agent has an intermediate proficiency value. In this case, the acceptance of a proposal is decided probabilistically: if the acceptance value (calculated from the language proficiency and accommodation level of the respondent, weighted by the aforementioned global parameters) is larger than a random number between (but excluding) 0 and 1, drawn from a uniform distribution, then the language is accepted. If no language has been accepted after three rounds, the Interaction ends in failure.

For the purposes of this experiment, all models ran for 500,000 iterations, which we found to be enough for the observed language choice values and network properties to become stable. Each model (parameter combination) was rerun 10 times to assure the generalizability of the results. All features and statistics discussed in this paper are based on measuring the outcomes of the last iteration.

TABLE 1 The parameters and their assigned values in the models. The effects of these values on communicative partner selection and interactions between agents are illustrated in Figs 1 and 2 (see Appendix A for more).

Parameter	Abbr.	Affects	Low extreme	Hypothesized reasonable	High extreme
importance of accommodation level in Interaction	*c*		0.1	2	1000
curve parameter in Interaction	*b*	*c*	0.001	10	10,000
weight of ethnic preference in Partner Choice	*e*		0	3	100
curve parameter in Partner Choice	*d*	*e*	0	3	100
passive link decay constant	*pd*		0.0001	10	

3.4 *Selecting the simulation parameters*

There are a number of global parameters that control the weight of the sociolinguistic factors described above. However, with the parameters being both continuous and unbounded in their values, the possible space of combinations is effectively infinite. Therefore, we configured the combinations to represent extremes of the functions, contrasted with what we perceived as reasonable values for the parameters (the middle value in each triplet), based on experimentation and sociolinguistic intuition.

Table 1 shows the combinations of global parameters and their respective values that were tested in the simulations (see Appendix A for details and equations).

In the *c* parameter (importance of accommodation), higher values give more weight to the accommodation level of the speaker (as reported in the survey) in the formula determining whether or not to accept a proposed second language, while values close to zero nullify the accommodation effect, leaving the decision dependent on language proficiency only. The accompanying curve parameter *b* controls how language proficiency (in combination with accommodation) affects accepting a second-language proposal—high values produce a step function with a sharp cutoff in the middle (low proficiency: always refuse, high: always accept), while lower values produce a smoother curve. Values close to zero nullify the effect of language proficiency on the decision to accept or refuse a proposal.

Weight of the ethnic preference (*e*) parameter controls how much the homophily of the agent affects its choice of communication partners. A value of zero nullifies the effect of homophily, making the agents regard all other agents, regardless of ethnicity, as equal communication partners. A high value of *e* amplifies homophily, making even slightly biased agents much more likely to choose agents with the same ethnicity as communication partners. The curve parameter *d* controls the shape of the function similarly to *b*.

The passive link decay constant, or *pd*, controls how fast already formed links decay in the network as the simulation proceeds. This is used to control network formation—a value considerably above zero allows links to survive for a while without being used, while a very small value removes every link as soon as it is formed, effectively disabling any network formation in the model (so in models with *p*=0.0001, *d* has no effect; *e* is then kept constant at 0). A very high value would yield a fully connected network (which would be, again, unrealistic).

Leaving out combinations that would produce identical or near-identical results (signified by empty plots below), we reach a parameter space with 58 combinations. We hypothesize that the combinations illustrated by the central positions on the two 3×3 plot grids in Figs 1 and 2 would yield the best model in terms of correspondence with the real-world language choice data (i.e., *c*=2, *b*=10; *e*=3, *d*=3; *pd*=10). Figures 1 and 2 illustrate possible outcomes of the different values for the parameters controlling the interaction (language choice) and partner choice.

The simplest baseline model is the one illustrated by the functions on the top left subplots in Figs 1 and 2—where there is no preference for partner selection and language acceptance chance is always at 50–50 (with the exceptions of native or maximal proficiency language and no proficiency)—with the additional attribute of disabling network formation by setting the passive decay parameter to a very low value. Depending on the parameters of the model, the simulations yielded various kinds of networks with different topologies, some of which are exemplified in Fig. 3. The presence of a (strong) link between any two agents reflects the fact that they have managed to communicate successfully in the recent past (iterations) and are likely to do so in the future (should the simulation continue).

3.5 *Validation procedure*

We hypothesize 1) that the simulation of interactions between initially unconnected agents (based on the questionnaire data), endowed by the language competencies and principles of interaction described in the previous section, would give rise to a social network structure resembling actual social networks,

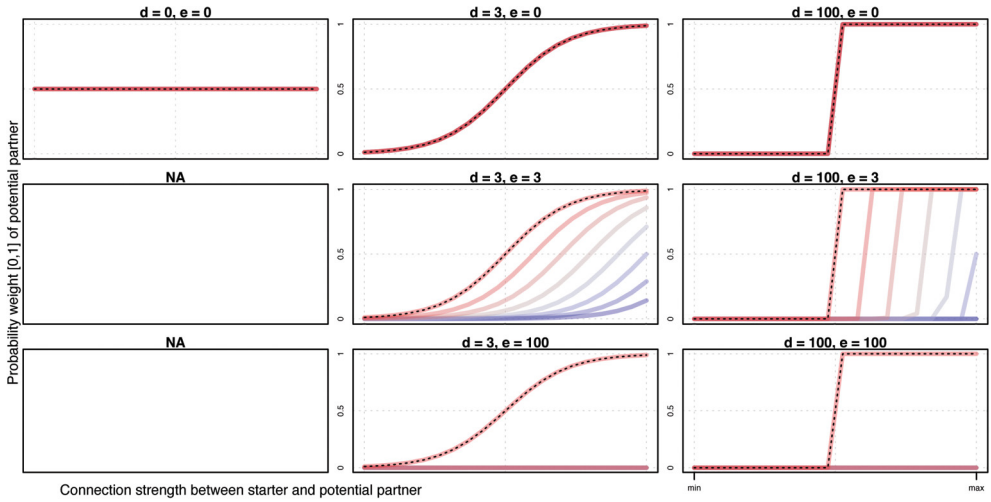

FIGURE 1 The Partner Choice function with different combinations of parameters. The strength of the connection between the agents (their link and their "common friend" links) is shown on the horizontal axis, with the potential weights for partners (for the semi-random choice) on the vertical axis. The colored lines exemplify the effect of the ethnic preference value of an agent, with red standing for less and blue for more preference to interact with members of their own ethnic group only. The red and black dashed line marks no preference (0; this is also in effect in the case of perceived same identity), and the blue stands for the strongest ethnic preference value across the samples (~0.7), with continuous gradience (illustrated by lines) in between. Setting the ethnic preference importance value (*e*) to 0 removes ethnic preferences from the model, setting it at a high value makes all even slightly ethnically biased agents try to avoid other ethnicities and only select for ethnic in-group partners.

and 2) that the extent of the languages (Estonian, Russian, and English) used in this social network would quantitatively resemble the extent to which these languages are used in corresponding communities. This calls for a way to quantify the usage of languages in the model, an operationalization of the self-reported language usage data from the questionnaires, and a metric to compare the resulting distributions.

For each respondent in the survey, the proportion of their language use was calculated by taking the mean of the questions asking about their language choice in communication with friends, hobby and sports companions, and service personnel (see Section 2 for details). In the simulation model, the distribution of languages was calculated based on the mean of the Common Language values between each agent and their linked partners. In other words, each agent contributes a value between 0 and 1 to the distribution, where 1 stands for having used only their native language as the language of communication with

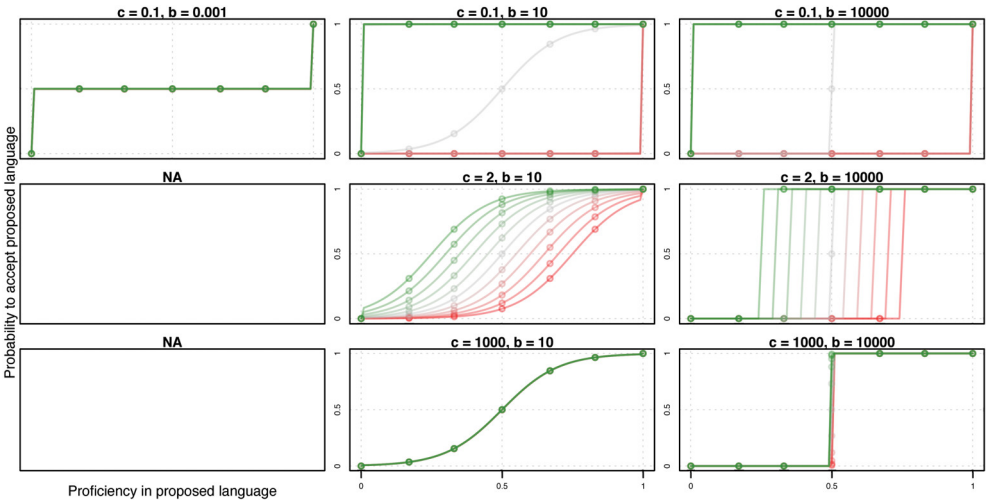

FIGURE 2 The Interaction function with different combinations of the parameters. The horizontal axis corresponds to the language proficiency of the agent in the proposed language (self-reported using a 7-point scale in the questionnaire; no proficiency on the left, native proficiency on the extreme right). The colored dots (connected by lines for visual aid) represent the interaction between language proficiency and various levels of accommodation: green is the maximum, or eagerness to speak the language of the other; red is the minimum, or aversion to speaking other languages; the grey in the middle stands for 0.5 or no strong preference—neither eagerness to speak the language of the other nor avoidance of such interactions.

all their currently linked partners and 0 stands for the opposite, having always agreed to use one of the other languages with all their partners.

The language usage distribution was compared to the behavior exhibited by the agent population. The respective empirical cumulative distribution functions were compared by utilizing the widely-used two-sample Kolmogorov-Smirnov distance statistic, yielding a simple numeric value for each simulation run, representing how closely the simulation approximated the real world (by the last iteration). Simply put, we compare the distribution of language choices (how much each agent communicates in their first language and how much in other languages) to the distribution of language choices in the survey samples. The closer the simulated distribution to the survey distribution, the more realistic the simulation. This allowed for the comparison of the combinations of parameters and provided a way to evaluate how well each parameter combination performed. For consistency, each individual parameter combination was repeatedly simulated ten times.

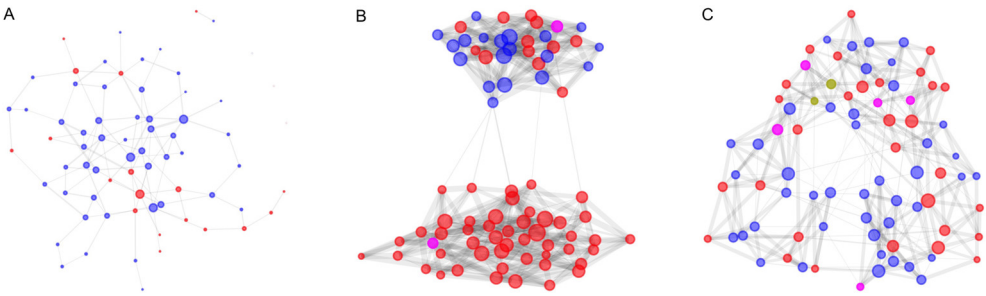

FIGURE 3 Various networks produced by the simulations. The nodes represent agents, and the color their native language. Red stands for Russian, blue for Estonian (and pink for self-professed bilinguals, yellow for other). The size of a node reflects the number of its edges (links) and the width of an edge corresponds to its strength. Network A: a sparsely connected network with mostly weak links (Tartu sample [see below for sample descriptions], with mostly Estonian speakers; c=0.1, b=0.001, e=0, d=0). B: a segregated network with two well-connected subgroups (Narva sample, c=2, b=10, e=3, d=3). C: a well-connected "small world" type network of a mixed language community (Tallinn sample, c=2, b=10, e=0, d=3).

4 Analysis of the results

We tested the model on three samples derived from three different linguistic environments: the South-Estonian town of Tartu with an Estonian-speaking majority and a small Russian-speaking community (70 agents), the Russian-language-dominated North-Eastern Estonia (represented by the towns of Narva and Kohtla-Järve; 81 agents) and a subset of Tallinn, the capital, with only a slight Estonian language majority (the districts of Haabersti and Mustamäe; 80 agents). The results are illustrated in Fig. 4. It is immediately visible that models without networks (dark blue on the right side of each panel in Fig. 4) tend to perform the worst, as expected, while the best models tend to have more densely connected networks (agents having 7–9 links on average). The hypothesized reasonable parameter combination (the red one marked with the red arrow in Fig. 4) is among the top performers—although it does not necessarily perform the best, there is not much difference among the top models either.

There seems to be considerable variation among the three samples regarding the performance of the model across parameter combinations—in the Narva sample (middle subplot in Fig. 4), the different parametrizations do not seem to make significant difference compared to the other two samples, while there is a clearer divide between networked and non-networked models (marked with dark blue). It should be kept in mind, though, that Narva is very much a one-language (Russian) dominated sample, hence it is "easier" for a model

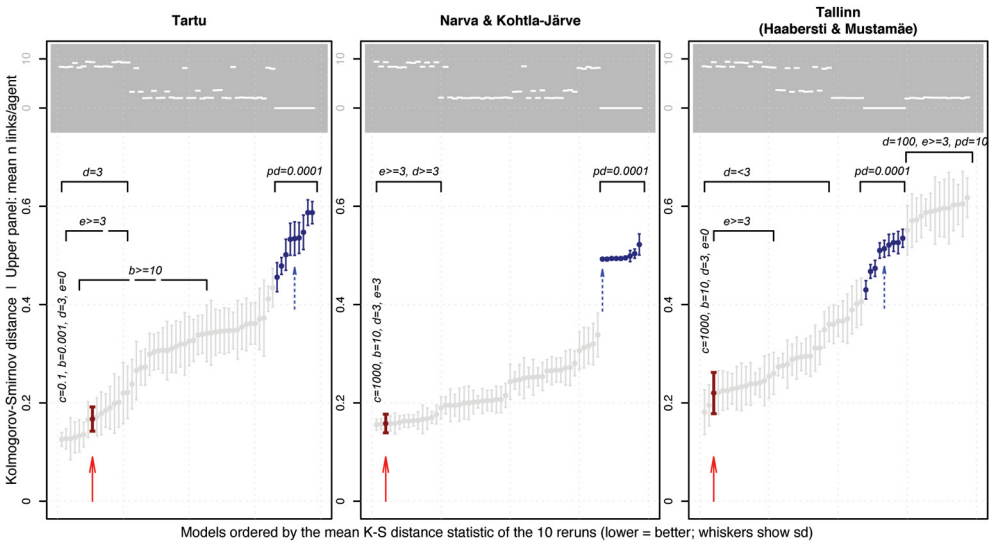

FIGURE 4 Results of the 58 model tests on the three sample populations, ordered by the mean of the Kolmogorov-Smirnov statistic. The vertical axis corresponds to the value of the Kolmogorov-Smirnov statistic, where 0 would indicate an exact match to the real-world usage distribution as attested by the survey. The gray dots with error bars (whiskers) on the white background visualize the results for each parameter combination. The dot in the middle of the bar is the mean for that parameter combination (across the 10 test runs of each model), the error bars stand for the mean ± its standard deviation. The dark red bold bar, emphasized with the solid red arrow, marks the hypothesized reasonable parameter combination. The dark blue ones correspond to the models without network formation. The small dashed blue arrows on the right mark the hypothesized simplest baseline models for each sample (no effect of ethnic preference, no effect of accommodation, no network formation). The top performing combination is labeled in italics with its parameters on the far left. Additionally, some interesting groups of models that happen to form a contiguous ordering and share the same parameter pattern have been labelled with black horizontal segments, with values in italics. The darker panel on top shows the mean number of links per agent in the models for each parameter combination.

to yield reasonably good results—the probability that an agent is paired with an agent with the same native language is much higher (and native language proposals are always accepted).

Finally, there is a subset of models with networks in the Tallinn sample that actually perform worse than non-networked models. While network formation seems to be a prerequisite for good results in the other samples, bad parametrization apparently causes the subset of networked models to perform worse in this particular case. Closer inspection reveals that, for all of them, the

value of the weight of ethnic preference (*e*) parameter is above 0 (3 or 100, i.e., agents with some homophily rather prefer partners from their own ethnic group) and the corresponding curve parameter *d* uniformly at 100, i.e., agents strongly prefer partners with whom they have above-average links and strongly avoid agents with below-average links (leading to sparsely connected networks; cf. the step function illustrated in Fig. 2, on the two bottom rightmost subplots). While having the step function in Partner Choice perform badly is not a surprise, it seems it has a greater effect in this sample than in the other two. These results call for a closer look at the contributions of the five global parameters to the outcome.

Figure 5 further visualizes the difference between three distinct parametrizations of the model in terms of network properties and correspondence with the real-world data. It appears that, given reasonable parameters, a relatively stable and connected network (high transitivity, more than a few links per agent) forms rather quickly. The Kolmogorov-Smirnov distance (K-S) to the real-world language usage distribution, while sensitive to the stochastic processes in the model, remains in a certain range. On the other hand, lower network connectivity (subplot B in Fig. 5.; or lack of it, cf. subplot C) appears to lead to less realistic language usage (indicated by a higher K-S distance).[1]

4.1 *Assessing the parameters*

In order to measure the importance of the parameters in terms of the final outcome of the model, we used a relatively straightforward machine learning tool, conditional random forests (Hothorn et al., 2006; Strobl et al., 2008). Random forests consist of a number of conditional inference tree classifiers (similar to decision trees in broad terms). An advantage of conditional inference trees is that they consider all possible interactions between the predictors while controlling for possible correlations between predictors, and they have been shown to perform well with categorical variables. Random forests repeat the classification procedure by growing a large number of such trees, allowing for the measurement of the average relative importance of predictors in terms of classifying the response. We grew 10,000 trees for each of the three city samples

1 Note that the apparent stability (of the K-S statistic value) of the model without network (bottom plot) is a technical artifact: since there is no network, there is no memory, hence there are no common languages. To average language use over iterations in these models, we simply record all successful language choices as a separate structure akin to the common language matrix, and calculate the fit with real-world data based on that. In a network-forming model, links can disappear over iterations, so the matrix of common languages is constantly changing—hence the variation.

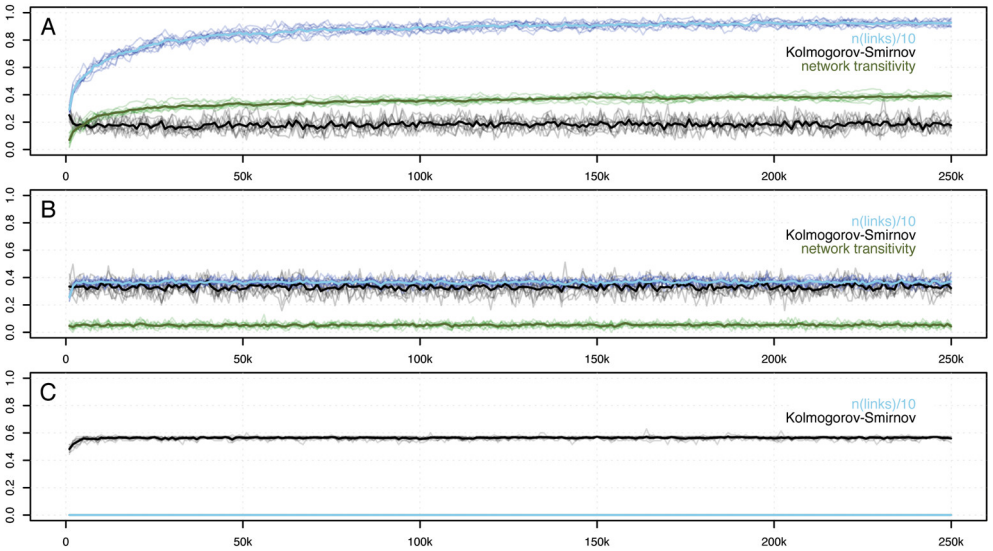

FIGURE 5 The first 250,000 iterations (horizontal axis) of three models from the Tartu sample. The top one (plot A) corresponds to the hypothesized reasonable parameter combination, the middle one (B) to a hypothetically inferior combination but with network formation enabled (c=0.1, b=10,000, d=0, e=0, pd=10) and the bottom one (C) to a model without network formation. The lighter lines correspond to the values of each of the 10 runs of the model, with the darker bold line showing the average across the 10. The statistics are measured once in every 1000 iterations, starting from the 1000th iteration. Light blue is for the average number of links per agent (divided by 10 for visualization purposes), black is for the K-S statistic (lower = closer to real data) and dark green for the transitivity of the network (the probability that the adjacent vertices of a vertex are connected, also known as graph clustering coefficient—this corresponds to the notion of triadic closure).

to ensure stability of the results, and used the training settings for unbiased forests as suggested by Strobl et al. (2007).

The Kolmogorov-Smirnov statistic is set as the (continuous) response variable and the parameter values as (categorical) predictors (while they are all numeric, treating them as continuous values would be counterproductive). The passive link decay parameter is excluded, along with the models without network formation, since they function differently and do not make use of all the parameters, as discussed above, making direct comparison difficult—moreover, they were already shown to perform noticeably worse in the previous section. In short, we are interested in the relative contribution of the rest of the four parameters in the models with network formation (n=490 per regional sample: 10 runs of each parameter combination, minus the models without networks).

FIGURE 6 Normalized relative importance of the simulation parameters in predicting the outcome of the simulation, i.e., the similarity to real-world data as measured with the Kolmogorov-Smirnov statistic. Only models with network formation are considered. Parameters on the right side of the bold gray line (at 0) can be considered significant predictors. As before, *c* is the importance of accommodation level in Interaction; *b* is the curve parameter in Interaction; *e* is the weight of ethnic preference in Partner Choice, and *d* stands for the curve parameter in Partner Choice.

The results from the random forests for each of the three samples (Fig. 6) show the relative contribution of each of the aforementioned simulation parameters across the three samples. The figure shows the importance values normalized to the range between 0 and the value of the most important variable, since the actual raw importance values of random forests are not directly comparable across models.

Note, of course, that these results should be viewed in the context of the parameter space of the models (which consisted of extreme values, contrasted with what we assumed to be more reasonable values, as explained above). The importance scores reflect the association of these values with the model outcome (Kolmogorov-Smirnov) value.

The parameter *c* controlling the importance of linguistic accommodation in Interaction, i.e. deciding whether to accept a proposed language or not, turns out to be of low importance in the case of Tartu, and next to none in Tallinn and Narva. In other words, linguistic accommodation (given the current parameter space, at least) does not make much of a difference in terms of model fit. We initially hypothesized that the attitude towards speaking other languages should influence language choice, yet it seems it only does so to some small extent, at least in this implementation and these samples. All the questionnaire data are self-reported and the results generally indicated high levels of accommodation. It may be that the questions did not quite capture the essence of linguistic accommodation, or the population in Estonia is actually very accommodating. This factor may still turn out to be significant in some other settings in which higher levels of non-accommodation have been reported (cf. Clément et al., 2003).

As for the curve parameter *b*, which appears to be of low importance (and next to no importance in Tallinn), it is important to bear in mind that the agents

always accept their native language and reject languages in which they have no proficiency. Apparently, what happens with in-between proficiencies does not affect the model very much.

The two simulation parameters that are consistently important appear to be the weight of ethnic preference *e* in choosing an interaction partner, and the curve parameter *d*, which controls the likelihood of choosing amongst potential partners. In other words, having good parameters for forming realistic linguistic networks (and enough simulation time for agents to seek out suitable communication partners by trial and error) seems to be almost enough to produce a realistic model of a language community (with the assumption that speakers will accept their native tongues and reject unknown ones).

The random forest results are also consistent with the observations in the previous section (cf. Fig. 4) concerning the Tallinn sample, where different values of the *d* parameter cause a noticeable difference. In the sample drawn from the Russian-language-dominated North-Eastern Estonia (Narva and Kohtla-Järve), the most important parameter is that of importance of ethnic preference, which takes second place in the two other models.[2] Observing, in this case, the individual parameter combinations and their outcomes reveals that turning off ethnic preference in the model ($e=0$) is associated with worse performance (higher Kolmogorov-Smirnov distance value).

4.2 *Analyzing network structure effects*

It has been observed that realistic social networks follow certain statistical tendencies. Figure 5 above already illustrated that better models may be associated with a higher network transitivity value. Having observed that the parametrization of the mechanism dictating partner choice (and hence network formation) plays a crucial role in our model of language choice, it would be valuable to see if the performance of our models in terms of language choice distribution correlates with values deemed to characterize realistic social networks in previous research. The following naturally only applies to models with network formation. Furthermore, we restrict the analysis to networks with a large number of links, since the metrics discussed below make little sense in networks with a large number of isolates or very little connectivity. As the models are clustered quite clearly in all three samples between those having a high mean number of links per agent and those in which this number is low (cf. Fig. 4),

2 However, it should be noted that, unlike the other samples, there is a small but still significant correlation between the ethnic preference and accommodation values among the participants of the Narva sample ($R^2=0.12$, $p=0.002$). The random forest method gives more importance to the better predictor in case of multicollinearity.

below we simply consider only those models with more links (than the grand mean; incidentally, that is uniformly 210 models—including re-runs—per sample).

Real-world social networks have been observed to exhibit a number of statistical properties. For instance, they tend to have low values of average path distance between nodes (the "small world" effect, or low degrees of separation) and high values of transitivity (or clustering, correspondent with the notion of triadic closure) (cf. Borgatti et al., 2009; Toivonen et al., 2006; Dekker, 2007; Tsourakakis, 2008). Furthermore, high values of modularity have been observed (modular communities within communities; cf. Newman and Girvan, 2004, Clauset et al., 2004).[3] Since low values of the Kolmogorov-Smirnov statistic correspond to realistic language use in our models, if we were to hypothesize that realistic usage is associated with realistic networks, then the expected correlations of the abovementioned three metrics with K-S would be, respectively, positive (average distance), negative (transitivity) and negative (modularity). We construct a linear regression model for each of the three samples, with the three metrics as predictors and K-S as the response variable.

In the Tartu sample (unlike in the others), transitivity and modularity are collinear; testing them separately shows that both are significantly negatively correlated with K-S, while average distance is expectedly positively correlated in both (adjusted R^2=0.29 for transitivity+distance, 0.1 for modularity+distance, $p<0.001$ for both models). In the Narva sample, transitivity and modularity correlate significantly with K-S (adjusted model R^2=0.33, $p<0.001$), but transitivity does so positively, unexpectedly. In the Tallinn sample, only modularity is significant, with an expected negative sign (adjusted model R^2= 0.1823, $p<0.001$). These results seem to be in alignment with the results of the random forest analysis and the distributions of the models (cf. Fig. 4). While network formation affects the results differently in Narva (with networks with triadic closure apparently yielding somewhat worse simulation scores), it is important for model success in Tartu; all the while something different seems to be going on in the Tallinn sample, where modular networks fare better but the other metrics are not associated with performance.

3 For isolated or single-link agents, we assign a transitivity value of 0; the average path length calculation only considers connected agents when traversing the graph; and we use 4-step random walks in the graph to determine communities in an unsupervised manner (cf. Pons and Latapy, 2005) as a basis for modularity calculation (Clauset et al., 2004; making use of the implementations by Csárdi and Nepusz, 2006 for all the metrics).

5 Discussion

We assumed that a realistic social network is necessary for a realistic model of language choice to yield reasonable outcomes. This turns out to be the case to some extent—models with no network formation process (or parameters strongly inhibiting network growth) performed noticeably worse compared to models that allowed for networks to form. In the Tallinn sample, it was observed that incorrect parametrization can make networked models even worse than models without networks in terms of performance. Further analysis of the models (with various parameter combinations) indicated that the simulation also performs rather differently given different samples of data, as the contribution of the parameters varies across samples. While we noticed that models initialized with the parameter combination hypothesized to yield reasonably good results indeed did do so, it appears that, depending on the sample, other parameter combinations may perform as well or better.

We observed positive correlations between more realistic network structure and model performance, but the correlations tended to be on the weak side. Then again, what we implemented was a rather simplistic model of human communication, and notably the formation of links in our model is entirely dependent on finding a common language, while in the real world, there are other factors that presumably affect social network formation.

A possible complication of our model is that it attempts to kill two, albeit possibly related birds with one stone: to generate a realistic network between the community members, but also to allow the communities to make (maximally realistic) language choices. We observed that often these goals converge, while there was also considerable variation among the samples, so that the parameters controlling the language choice and partner selection (network formation) mechanisms had differing importance in the three samples. Furthermore, we only tested the model for relatively small samples of agents, and also kept some underlying parameters constant (see Appendix A for details). Future research should address those issues and generalize the model to work on agent populations of various sizes (we noted that there is considerable difference in convergence time given agent populations of different sizes, hence we opted to test only similarly sized samples to obtain comparable results).

Another avenue for future research involves the notion of linguistic accommodation, or willingness to speak the language of "the other," which we hypothesized should have an effect on the language choice process. Our simulation results demonstrate that the effect of accommodation is rather negligible in terms of the final outcome of the model. There are multiple possible explana-

tions in this case. It may be that the questionnaire data did not quite capture the actual linguistic attitudes of the respondents or the population is indeed very accommodating. It is also possible that the formula underlying the simulation did not capture the complexity of the process (or, on the contrary, should be simplified instead, possibly collapsing the accommodation and homophily values), or that the assumed shape of the function or the tested values of the corresponding weight parameter are simply not close enough to the ideal. This observation calls for further experimentation and possible amelioration of the interaction part of the model.

6 Conclusions

We constructed an agent-based simulation model for language choice in multilingual communities and tested its performance on three samples of data drawn from a questionnaire carried out among native speakers of Estonian and Russian in Estonia. In contrast with previous work on diachronic language competition models, our purpose was to create and test an individual-driven model of synchronic language choice (reflective of language competition on the community level), grounding our model in recently collected sociolinguistic survey data.

The simulation incorporates a simplified model of human interaction where pairs of agents negotiate the language to be used in a hypothetical conversation. As the simulation progresses, a semi-stable network may or may not be formed among the agents, who may then proceed (depending on the global parameters of the model) to communicate more likely with agents they already have communicated with before. We tested a large number of global parameter combinations on three samples of agents, based on three sociolinguistically different areas of Estonia. We found that models with different parameters do behave differently, and that what we hypothesized to be a reasonable combination did perform rather well, presumably indicating that the model (given reasonable parametrization) does reflect the corresponding real-world language situation to a considerable degree. However, the parameters were found to be unequal in terms of differentiating the outcomes of the model, both within and between the three samples tested. We noted that the presence (and more so, realistic attributes) of social networks may contribute to a more accurate model of language competition, but again, the results differed across the three samples.

All in all, these findings point to the necessity of considering the variability between communities in building sociolinguistic language choice and compe-

tition models (particularly if they underlie diachronic language survival models), the need to reliably model grassroots-level interactions, as well as the importance of social networks in models of linguistic communities.

Acknowledgments

The research leading to these results received funding from the Estonian Research Council [grant number IUT20–3]. Andres Karjus is furthermore supported by the national scholarship program Kristjan Jaak, funded and managed by the Archimedes Foundation in collaboration with the Ministry of Education and Research of Estonia. The authors would like to thank Richard Blythe, Kenny Smith and Kuldar Taveter for useful comments on the earlier drafts, as well as the three anonymous reviewers of Language Dynamics and Change for comments and questions that led to numerous improvements in the paper.

References

Abrams, Daniel M. and Steven H. Strogatz. 2003. Modelling the dynamics of language death. *Nature* 424: 900.

Baxter, Gareth J., Richard A. Blythe, William Croft, and Alan J. McKane. 2009. Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change* 21: 257–296.

Beltran, Francesc S., Salvador Herrando, Doris Ferreres, Marc-Antoni Adell, Violant Estreder, and Marcos Ruiz-Soler. 2009. Forecasting a language shift based on cellular automata. *Journal of Artificial Societies and Social Simulation* 12: 5.

Bianconi, Ginestra, Richard K. Darst, Jacopo Iacovacci, and Santo Fortunato. 2014. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E* 90: 42806. doi: 10.1103/PhysRevE.90.042806.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. Network analysis in the social sciences. *Science* 323: 892–895.

Castelló, Xavier, Lucia Loureiro-Porto, and Maxi San Miguel. 2013. Agent-based models of language competition. *International Journal of the Sociology of Language* 2013: 21–51.

Clauset, Aaron, M.E.J. Newman, and Christopher Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70: 66111.

Clément, Richard, Susan C. Baker, and Peter D. MacIntyre. 2003. Willingness to communicate in a second language: The effects of context, norms, and vitality. *Journal of Language and Social Psychology* 22: 190–209.

Clyne, Michael G. 2003. *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge: Cambridge University Press.

Csárdi, Gábor and Tamás Nepusz. 2006. The igraph software package for complex network research. *InterJournal Complex Systems*: 1695.

Dekker, Anthony. 2007. Realistic social networks for simulation using network rewiring. In Lex Oxley and Don Kulasiri (eds.), *MODSIM 2007. International Congress on Modelling and Simulation*, 677–683. Canberra: Modelling and Simulation Society of Australia and New Zealand.

Ehala, Martin, Kadri Koreinik, Anastassia Zabrodskaja, Andres Karjus, Birute Klaas-Lang, Kristiina Praakli, Maarja Siiner, and Tõnu Tender. 2015. Linguistic attitudes in Estonia 2015. Online data repository. http://doi.org/10.15155/1-00-0000-0000-0000-00162L.

Giles, Howard. 2008. *Communication Accommodation Theory*. Thousand Oaks, CA: Sage.

Giles, Howard, Richard Bourhis, and Donald M. Taylor. 1977. Towards a theory of language in ethnic group relations. In Howard Giles (ed.), *Language, Ethnicity, and Intergroup Relations*, 307–348. New York: Academic Press.

Gong, Tao, Jinyun Ke, James W. Minett, and William S.-Y. Wang. 2004. A computational framework to simulate the co-evolution of language and social structure. In Jordan Pollack, Mark Bedau, Phil Husbands, Takashi Ikegami, and Richard A. Watson (eds.), *Artificial Life IX. Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, 214–219. Cambridge, MA: MIT Press.

Hothorn, Torsten, Peter Bühlmann, Rine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7: 355–373.

Isern, Neus and Joaquim Fort. 2014. Language extinction and linguistic fronts. *Journal of the Royal Society Interface* 11. doi: 10.1098/rsif.2014.0028.

Jansson, Fredrik, Mikael Parkvall, and Pontus Strimling. 2015. Modeling the evolution of creoles. *Language Dynamics and Change* 5: 1–51.

Kandler, Anne, Roman Unger, and James Steele. 2010. Language shift, bilingualism and the future of Britain's Celtic languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 3855–3864.

Leetmaa, Kadri, Tiit Tammaru, and Daniel B. Hess. 2015. Preferences toward neighbor ethnicity and affluence: Evidence from an inherited dual ethnic context in post-Soviet Tartu, Estonia. *Annals of the Association of American Geographers* 105: 162–182.

Luthi, Leslie, Enea Pestelacci, and Marco Tomassini. 2008. Cooperation and community structure in social networks. *Physica A: Statistical Mechanics and its Applications* 387: 955–966.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27: 415–444.

Minett, James W. and William S.-Y. Wang. 2008. Modelling endangered languages: The effects of bilingualism and social structure. *Lingua* 118: 19–45.

Newman, Mark E. and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69: 26113.

Patriarca, Marco, Xavier Castelló, José Ramón Uriarte, Victor M. Eguíluz, and Maxi San Miguel. 2012. Modeling two-language competition dynamics. *Advances in Complex Systems* 15: 1250048.

Patriarca, Marco and Teemu Leppänen. 2004. Modeling language competition. *Physica A: Statistical Mechanics and its Applications* 338: 296–299.

Pons, Pascal and Matthieu Latapy. 2005. Computing communities in large networks using random walks. In Pinar Yolum, Tunga Güngör, Fikret Gürgen, and Can Özturan (eds.), *International Symposium on Computer and Information Sciences—ISCIS 2005*, 284–293. Berlin: Springer.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Scott, David W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Stark, Tobias H. 2015. Understanding the selection bias: Social network processes and the effect of prejudice on the avoidance of outgroup friends. *Social Psychology Quarterly* 78: 127–150.

Sterling, Leon and Kuldar Taveter. 2009. *The Art of Agent-oriented Modeling*. Cambridge, MA: MIT Press.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8: 25.

Toivonen, Riitta, Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, and Kimmo Kaski. 2006. A model for social networks. *Physica A: Statistical Mechanics and its Applications* 371: 851–860.

Tsourakakis, Charalampos E. 2008. Fast counting of triangles in large real networks without counting: Algorithms and laws. In Fosca Giannotti, Dimitrios Gunopulos, Franco Turini, Carlo Zaniolo, Naren Ramakrishnan, and Xindong Wu (eds.), *ICDM 2008. Eighth IEEE International Conference on Data Mining*, 608–617. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Zhang, Menghan and Tao Gong. 2013. Principles of parametric estimation in modeling language competition. *Proceedings of the National Academy of Sciences of the U.S.A.* 110: 9698–9703.

Appendix

A *A more formal summary of the model*

The simulation model is characterized by:

- K, the set of languages in the model (currently {Estonian, Russian, English})
- L, the set of links between every possible pair of agents in the model
- z, the number of agents in the set A
- n, the number of iterations to iterate
- r, the number of rounds to interact (propose languages) in Interaction: currently this is 3, corresponding to the number of two presumed native languages (Estonian and/or Russian) and one additional secondary language; the introduction of more native languages into the model would increase r

There are a number of global parameters that influence a simulation:

- Importance of accommodation level a in Interaction, c
- S-curve parameter in the language choice process in Interaction, b
- Weight of ethnic preference in Partner Choice, e
- Curve parameter in Partner Choice, d
- Passive link decay constant, pd
- The increment/decrement in link strength change I (currently constant at 0.1)

Each agent $i \in A$ is characterized by the following values:

- A proficiency value $o_{i_k} \in [0, 1]$ for all languages $k \in K$; 0 = no proficiency, 1 = native or full proficiency
- Accommodation level $a \in [0, 1]$, reflecting willingness to speak in other languages; 0 = strong aversion, 1 = strong eagerness; only plays a role in cases where $0 < o_{i_k} < 1$, where k is a language offered to i in the course of the Interaction process
- A link strength value between it and every other agent $l_{ij} \in L \in [0, 1]$ (initially all 0; effectively plays no role if pd is too small to allow network formation)
- Its ethnic identity eid
- A value of "ethnic closedness" $ec \in [0, 1]$, a measure of (not) being open to interaction with the "other" ethnicity; 0 = views "own" and "other" ethnicities equally in terms of potential communication partners, 1 = maximally adverse to communicating with the "other"
- A common language value between i and every other agent $com_{ij} \in K$ (initially all NA); only plays a role if pd is sufficiently large to allow network links (and corresponding memory) to be preserved
- An "ethnic memory" set E_i consisting of the identity values for all other agents that i may (come to) know (initially all unknown, NA)

– A parallel "pseudo-memory" set F_i consisting of pseudo-identity values for all other agents that i might interact with (generated semi-randomly upon simulation initialization, but approximately reflecting the ethnic proportions of the sample); this structure determines what i assumes to be the ethnicity of agent j, if $l_{ij} = 0$; E and F effectively only play a role if $e > 0$ & $ec_i > 0$, and E only plays a role if pd is sufficiently large to allow network links (and corresponding memory) to be preserved.

The simulation consists of the following components.

Main loop

iterate for n iterations:

choose a random agent ("starter," s) from the set of all agents A

choose a "partner" (p_i) → cf. Partner Choice

interact for up to r rounds → cf. Interaction

if the Interaction is successful:

link l_{sp_i} increases $+I$ (up to 1)

the selected language is recorded as a Common Language com_{sp_i}

else:

set com_{sp_i} back to NA (next Interaction will again start with probabilistic language choice)

link l_{sp_i} decreases $-I$ (but never below 0)

regardless of outcome:

s and p_i are now aware of the true ethnicities of one another (update E_s and E_{p_i})

each and every link l_{ij} decreases slightly (currently $0.1/(an \times pd)$)

if any link disappears ($l_{ij} = 0$) by the end of the iteration, then:

knowledge of common language is removed: com_{ij} set to NA

knowledge of the ethnicity of j is removed from i ($E_{i_j} = NA$), and vice versa

output values from the last iteration of the simulation

Partner choice

the starter s is randomly assigned a partner p_i from the set of potential partners P ($P = A - s$); some members of P may have a considerably higher chance, but all members of P have some chance to be chosen (although it may be very low); already having a link and having "common friends" (triadic closure; "a friend of yours is a friend of mine") increases the chance of being chosen:

if there is at least one such common link ($l_{sp_i} > 0$ & $l_{sp_x} > 0$ & $l_{p_ip_x} > 0$), then:

if s and p_i have more than one common friend $p_x, \ldots, p_z$, then:

$t_{sp_i} = \max(l_{sp_x}, \ldots, l_{sp_z})$
else:
$t_{sp_i} = l_{sp_x}$
else: $t_{sp_i} = 0$
if the identity p_i is unknown to s ($E_{s_{p_i}} = NA$), then:
take eid_{p_i} from F_s (the pseudo-identity memory)
else:
take eid_{p_i} from E_s (the actual ethnic identity memory)
based on identities, determine the binary value $ex := [eid_{p_i} = eid_s]$
each member of P is assigned a weight w: $w_{p_i} = (1 + l_{sp_i}) \times (1 + t_{sp_i}) - (ex \times (ec_s \times e))$
partner choice is done by random sampling, the probabilities C for each potential partner $C_{p_i} = \frac{1}{1+e^{-d\times(w_{p_i}-2.5)}} \in (0, 1)$

Interaction
the starter s and the chosen partner p_i interact for up to r rounds, taking turns proposing a language of communication, with s proposing first; the common language com_{sp_i} may be instantiated (or removed) as a result of the Interaction, depending on the outcome, under the following conditions:
all languages $k \in K$ in the model may be considered; o_{k_x} stands for proficiency in language k_x; an agent may have maximum (i.e., native or near-native) proficiency in one or more k_x
every time a k_x is offered, the value of accepting it depends on the o_{k_x} and a of the responding agent, and is determined as:
$v_{k_x} = ((\frac{1}{1+e^{-b(o_{k_x}-(1-f(a)))}}) \times [o_{k_x} > 0])^{1-[o_{k_x}=1]}$
where $f(a) = \frac{a-((1-c)/2)}{c}$
any time a k_x is offered, a random number $rnd \in (0, 1)$ is generated; if $v \geq rnd$, then the interaction is considered successful and stops
the current model implements 3 rounds (we assume 2 native languages and 1 secondary language in the model); in the final round, there is a chance that s may or may not, depending on its a, offer any language not yet offered during this Interaction: whether or not s will make the offer is determined by the same language acceptance formula as described above, where, if $v_{k_x} > rnd$, then an offer for k_x is made (which the responding p_i may or may not accept)
practically, in the currently implementation, only one language option is left at this point, since native language offers are always accepted
if multiple languages are available, s will offer the language where it has higher proficiency (if equal, choose randomly); s cannot offer a language it does not know

success in the final round (the third language being stored as common language) allows for scenarios where a speaker with low proficiency keeps trying to communicate in a language they want to learn/use/etc.

an Interaction proceeds as follows:
first round of interaction; s offers a language k_x as follows:
if a previous value for com_{sp_i} exists, then:
  $k_1 = com_{sp_i}$
else:
  $M = \max(o_{s_k})$; if $|M| > 1$ then:
    k_1 is randomly chosen among M
  else:
    $k_1 = M$
if p_i accepts k_1, then:
  stop, report success
else:
  second round; p_i offers a language k_2 as follows:
  $M = \max(o_{p_{i_k}})$; if $|M| > 1$ then:
    k_2 is randomly chosen among M
  else:
    $k_2 = M$
  if s accepts k_2, then:
    stop, report success
  else:
    third round; s may make a final offer of k_3:
    $N = K - \{k_1, k_2\}$; if $|N| > 1$ then:
      $M = \max(o_{s_{N_1}}, \ldots, o_{s_{N_z}})$; if $|M| > 1$ then:
        k_3 is randomly chosen among M
      else:
        $k_3 = M$
    else:
      $k_3 = N$
    if s accepts k_3 (see above), then:
      if p_i accepts k_3, then:
        stop, report success
      else:
        stop, report failure
    else:
      stop, report failure

B *Additional notes on the model and its properties*

We implemented a number of simplifications and kept certain parameters constant to maintain the parameter space within manageable limits and allow for the analysis of the initial results. This section describes some further technical details that were left out of the main text to maintain readability.

Constants in the model

The passive link decay *pd* is currently implemented as a function of the number of agents, with a constant controlling the resulting decay rate; we used a value that was observed to yield more or less stable networks in a reasonable amount of time. Too large values prevent the formation of networks, too small values allow for links to remain between agents who almost never interact. Notably, the passive decay is linear (which could be argued against), but Partner Choice operates on an S-curve, meaning that minor decay in "stranger" (low strength) and "good friend" (high strength) links do not matter much in terms of probability of being partnered up again; yet the effect begins to accumulate once the link is left unused for a while. The increment *I* that is added or deducted from the link strength value between any two agents upon successful or unsuccessful interaction, respectively, is kept constant (0.1) as well throughout the simulation runs. Implementation of the effect of communication failure and the question whether it is inversely symmetric to the effect of a successful interaction in terms of the relationship between the participants is left as a venue for future research.

Perception of ethnicity

Knowledge of the ethnic identity of others is initially set semi-randomly. We assign an additional pseudo-identity to each agent—where the probability of being assigned to each group corresponds to the actual ethnic makeup of the given sample—and generate knowledge for each agent pair from that (e.g., in a sample based on a community of 60% Estonian and 40% Russian speakers, the probability of being assigned an initial Russian pseudo-identity would be 0.4). As the simulation proceeds and agents interact, they take the pseudo-identity at face value upon first contact, but keep the correct identity of the partner in mind for future interactions. The rationale is the following: upon meeting a person, but before interacting, people make assumptions about the identity of the other. This guess may or may not be correct, but it would be reasonable to assume that if two people meet in a place where one ethnicity is known to be the majority, then their guesses will be influenced by that knowledge.

Randomness

There are multiple sources of stochasticity in the model, meaning that every simulation run yields slightly different results, which prompts running each parameter combination for several times and then observing their average outcome (and variability). Partner Choice is stochastic, since the starting agent is chosen at random, and its partner is chosen at semi-random (potential partners may have very different chances of being chosen). Similarly, whether an agent accepts a language during the Interaction phase or not is semi-random, but the chance of acceptance varies accordingly with their proficiency and accommodation values (dependent on model parameters). As described above, the pseudo-identity mechanism introduces another source of randomness.

Validation data

For each respondent, the proportion of their language use was calculated by taking the mean of the questionnaire questions asking about language choice in direct communication (i.e., excluding online communication and media consumption such as television) with friends, hobby and sports companions, and service personnel. It should be noted that the questionnaire also contained questions about communication with other groups. The question concerning language use in the family circle was excluded in this study, as the questionnaire only consists of respondents over the age of 15. Concurrently, there are no "children" in the agent population, and therefore family communication cannot be accurately simulated at this stage. The agent population can be thought of as consisting of a sample of the adults in a given city. Questions regarding school and workplace communication were excluded, as the simulation does not feature organized clusters of agents that would emulate such settings. The simulation does feature interactions that could be equated with real-life interactions with "strangers"—agents that meet only once, communicate, but do not meet again soon enough to form a lasting link. However, the questionnaire question concerning communication with strangers on the street was not included, on the assumption that such communications are less frequent than those with the other groups and would bias the validation data. The questionnaire did not include questions about the proportion of their time the respondents would spend communicating with friends, family, and other groups. With all that in mind, we only used the Common Language values of links with a strength > 0.1 in the validation process (the increment constant being 0.1, agents that would have been in contact recently more than once would mostly have links > 0.1, i.e., are not quite strangers anymore).

We used the Kolmogorov-Smirnov test statistic as the measure of distance between distributions of language use from the models and the real-world

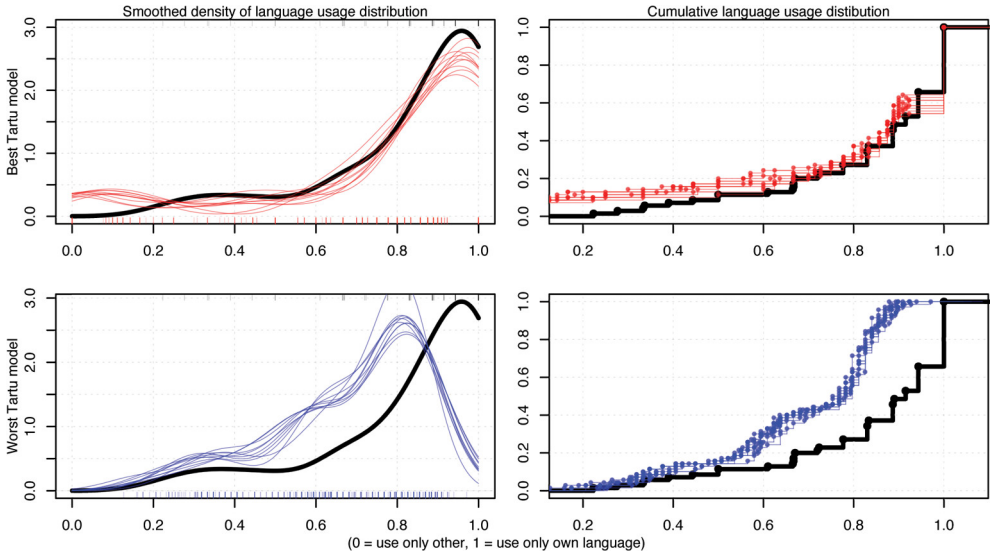

FIGURE 7 Example language distributions from the Tartu models. The best one, in red (cf. Fig. 4 for comparison), is on the top and the worst one on the bottom, in blue. The actual real-world distributions are shown as thick black lines, the model results (10 runs for each combination) as the colored thin lines. On the left are the smoothed density plots, and on the right, the empirical cumulative distributions, which are used to calculate the Kolmogorov-Smirnov statistic used in model validation. Note that care should be taken with interpreting kernel-smoothed density functions, since they can easily be misleading, depending on the chosen kernel (Gaussian here) and bandwidth (here both: 0.08, following the rule of thumb method from Scott, 1992).

distributions from the survey, being suitable for our data (cf. Fig. 7) as a non-parametric, distribution-free statistic, calculated from the empirical cumulative distribution of the data. The actual distribution consists of a fairly small number of unique values, and any sort of smoothing (which would allow for the direct comparison of distributions) carries the danger of misinterpretations stemming from over-generous or over-conservative smoothing values.

Finally, it could be argued that sharing a common language does not immediately mean that two people would be able to communicate or get along in the real world. This is a necessary idealization of the model: partner selection is only affected by the partner selection process (as outlined above), and we do not aim to model all the possible (but ultimately intractable) personal preferences, moods or desires that the agents could potentially exhibit, being representatives of real-world human questionnaire respondents.

Technical acknowledgements

The model and all the additional statistics were implemented in the R programming language (version 3.3.1; R Core Team, 2016), also making use of the packages party (Hothorn et al., 2006) and igraph (Csárdi and Nepusz, 2006).