Quantifying the driving factors for language shift in a bilingual region

Katharina Prochazka*1 and Gero Vogl*

*Department of Condensed Systems, Faculty of Physics, University of Vienna, 1090 Vienna, Austria

Edited by Barbara H. Partee, University of Massachusetts at Amherst, Amherst, MA, and approved February 13, 2017 (received for review November 2, 2016)

Many of the world’s around 6,000 languages are in danger of disappearing as people give up use of a minority language in favor of the majority language in a process called language shift. Language shift can be monitored on a large scale through the use of mathematical models by way of differential equations, for example, reaction–diffusion equations. Here, we use a different approach: we propose a model for language dynamics based on the principles of cellular automata/agent-based modeling and combine it with very detailed empirical data. Our model makes it possible to follow language dynamics over space and time, whereas existing models based on differential equations average over space and consequently provide no information on local changes in language use. Additionally, cellular automata models can be used even in cases where models based on differential equations are not applicable, for example, in situations where one language has become dispersed and retreated to language islands. Using data from a bilingual region in Austria, we show that the most important factor in determining the spread and retreat of a language is the interaction with speakers of the same language. External factors like bilingual schools or parish language have only a minor influence.

Significance

Languages are an important part of our culturally diverse world, yet many of today’s languages are in danger of dying out. To save endangered languages, one must first understand the dynamics behind language shift: what are the driving factors of people giving up one language for another? Here, we model language dynamics in time and space starting from empirical data. We show that it is the interaction with speakers of the same language that fundamentally determines spread and retreat of a language. This means that a minimum-sized neighborhood of speakers interacting with each other is essential to preserve the language.

Author contributions: K.F. and G.V. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The digitized census data on language use for southern Carinthia, 1880–1910, from the Austrian/Austro-Hungarian census reported in this paper have been deposited in figshare (https://figshare.com/articles/Language_use_in_Carinthia/435399).

*To whom correspondence should be addressed. Email: katharina.prochazka@univie.ac.at.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617232114/-/DCSupplemental.
reaction–diffusion equation following Fisher’s equation for advantageous genes (21):

$$\frac{\partial u_G}{\partial t} = D_G \cdot \nabla^2 u_G + k \cdot u_G \cdot \left(1 - u_G\right).$$  \[1\]

Here, $D_G$ is the diffusivity of German language and $k$ is the conversion rate from Slovenian to German. The fraction of Slovenian speakers is given by $u_S = 1 - u_G$.

In Carinthia, the language front between German and Slovenian essentially advances only southward motivating a one-dimensional treatment. Eq. 1 then results in a traveling front of the higher status language (in this case German) with velocity $v$:

$$v = 2 \sqrt{D_G \cdot k}. \quad \[2\]$$

We defined the front as the line bordering all cells with more than 50% Slovenian speakers each without including outlier cells detached from the contiguous language area (Fig. 2A).

From the data, the velocity of the language front can be derived (Table S1) and the product of diffusivity and conversion rate $D_G \cdot k$ can be determined (Supporting Information). If data on the conversion rate are available, the diffusivity of the majority language can be estimated.

In the period 1971–2001, no contiguous language area exists due to the large decrease in the number of Slovenian speakers (Fig. 2B). The cells with significant fractions of minority language have become dispersed and the minority language has retreated to language islands. Hence there is no continuous language front that clearly separates the two language areas of Slovenian and German. From this, it is evident that the concept of a language front fails when the minority language no longer covers a contiguous area, and reaction–diffusion equations and the resulting language front are not applicable in this case.

Noting the limits of treatment by reaction–diffusion equations, we take a different approach. We simulate the microdevelopment on the basis of the smallest registered population units, hence providing a spatially much more detailed description of language spread and retreat than the macroscale description by reaction–diffusion equations. As a result, we can follow not only global development over time and space but also local processes such as the deviating dynamics in urban areas.

Materials and Methods

We start by taking the data of the censuses in the former Austro-Hungarian Empire that were held in 10-y intervals beginning from 1880 until 1910. Such censuses were continued in the Federal Republic of Austria from 1971 until 2001. In between, the two world wars and the after-war turbulences prevented consistent censuses and data with the same level of detail is not available. In the censuses together with several other items, the vernacular language of each person was asked for and registered. In the censuses until 1910, only one language could be recorded in the questionnaire, whereas in the censuses after 1971 also bilingualism could be indicated. For the sake of simplicity and consistency and data with the same level of detail is not available, we included bilingual Slovenian–German speakers in the minority group, that is, Slovenian speakers. In this paper, we do not consider bilingualism as a separate speaker state. We neither consider the different Slovenian and German dialects that are not encoded in the census data.

From the census results, we read the number of German and Slovenian speakers in each of the ~1,500 population units (hamlets, villages, and towns) in southern Carinthia.

Mapping the Data. For our study of language dynamics in Carinthia, the area investigated is subdivided into a quadratic grid with cells sized 1 x 1 km. Carinthia spans ~2.6° of longitude or 184 km (east border to west border) and 0.8° of latitude or 84 km (north border to south border). We thus cover the total area of Carinthia with a regular grid of 84° of longitude or 184 km cells of 1 x 1 km each. Of these 15,456 cells, 9,549 comprise the actual area of Carinthia. A total of 1,170 of the cells inside the Carinthian borders is populated by data for speakers in southern Carinthia. Data for the number of speakers was extracted from census records for the period from 1880 to 1910 (22–25) and the raw dataset for 1971–2001 (special evaluation commissioned from Statistics Austria). Digitized data can be accessed on figshare (https://figshare.com/articles/Language_use_in_Carinthia/4535399). Speaker data are assigned to cells based on the geographic coordinates of the population unit. Geographic coordinates (WGS 84) were obtained from KAGIS Kärnten Atlas (https://gis.ktn.gv.at/). The 10-y periods between the censuses are divided into 1-y time steps for the simulation.

Language Shift Model. As a first basic model, we assume that in each cell after 1 y’s time the probability $\alpha_{\beta}(r, t + 1)$ to speak a language $\alpha$ will be proportional to the local number of speakers of that language in the preceding period.

**Table 1.** Fit values for the diffusivities $D_{\alpha}$ for both time periods

| Period | Diffusivity 1880–1910 | Diffusivity 1971–2001 |
|--------|-----------------------|-----------------------|
| $D_G$  | 0.1356                | 0.1078                |
| $D_S$  | 0.0939                | 0.0926                |

Diffusivities $D_\alpha$ according to simulations for both periods. Changing both diffusivities in the same direction and by the same amount changes little in the quality of the fit, whereas a small change to the ratio of diffusivities has a big impact on fit quality. A 10% change of both diffusivities in the same direction leads to a 1% change in the sum of absolute errors. On the other hand, if only one diffusivity is changed by 10% (i.e., the ratio between diffusivities changes as well), this leads to a 10% change in the sum of absolute errors.

Fig. 1. Percentage of Slovenian speakers in southern Carinthia according to census results. (A) Geographic location of southern Carinthia. The study area is indicated by the orange rectangle in the map at Bottom Left. (B) Census results for 1880. (C) Census results for 2001. Cells without any Slovenian speakers (0%) are shown in white. Each square marks a 1 x 1-km cell. Contour lines are in brown, and rivers and lakes are in gray. The two biggest Carinthian towns, Villach and the capital Klagenfurt, are encircled according to present administration.
and 4, assuming a linear population development between censuses. To evaluate the goodness of fit between simulated data and census data, we use ordinary least squares (OLS) to minimize the squared sum of errors:

\[ OLS = \sum_{t=0}^{m-1} \sum_{i=1}^{n} (O_i - E_i)^2, \]

where \( O_i \) is an observed data point (census data) and \( E_i \) is an estimated data point (simulated data). \( t \) is the number of times the observed data can be compared with the estimated data. \( t \) runs from 1 to \( m-1 \), where \( m \) is the number of censuses within the period. The data from the first census in each period are excluded as they are equivalent to the initial state of the system; hence there is no error for the initial state and we sum only over the remaining censuses. Optimization was done using the Nelder–Mead method (26). Additionally, we used least absolute errors (LAE) as follows:

\[ LAE = \sum_{t=0}^{m-1} \sum_{i=1}^{n} |O_i - E_i|, \]

that is, minimizing the sum of the absolute errors to check the reliability of the fit. General model performance is evaluated by comparison with a baseline. Comparison values can be found in Table S2.

Results

Language Shift in the Period 1880–1910. The widths of both Gaussians, and hence the diffusivities for German and Slovenian, are fitted to the number of speakers in each cell as given by the census data. Fits to the census data were performed for the period from 1880 to 1910. The best solution was achieved with the values given in Table 1.

Fig. 3 shows the increase (red) and decrease (blue) of the number of Slovenian speakers in southern Carinthia for census data and simulated data. We obtain satisfactory agreement between the empirical data and the predicted data on a microscopic scale. In detail, this can be seen in Fig. S1 where the model’s errors are shown for cells with different numbers of Slovenian speakers. The total number of Slovenian speakers as predicted by the simulation also agrees with the census data (Fig. S2). Fig. S3 shows the residuals (census data minus simulated data). Thus, our model is able to follow how either language has spread and retreated in the time period 1880–1910.

Extension of the Model Through Habitat Parameters. In a second step, for the period from 1880 to 1910, the influence of habitat conditions, such as the influence of urban areas, that is, major towns, the language of schools, and language in parishes, was investigated. To this end, we introduced a habitat parameter \( h_i \) into Eq. 3, which modifies the effect of local speakers by an exponential function with the argument \((\pm H h_i)\). The multiplicative factor \( H \) indicates the presence \((H = 1)\) or absence \((H = 0)\) of a local habitat condition, that is, \( H = 1 \) for the two largest towns Klagenfurt and Villach or if a bilingual school or Slovenian parish existed. Otherwise, \( H \) is set to zero. The exponential function was chosen as a modifier because it is a simple function, which for small \( h_i \) adds \( nh_i \) to the speaker effect if \( H = 1 \), while recovering the basic model (Eq. 3) if \( H = 0 \). We obtain the following equation for the probability \( p_o(r, t + 1) \) of speaking a language \( o \) at position \( r \) and time \( t + 1 \):

\[ p_o(r, t + 1) = \frac{n_o(r, t) \cdot \exp((\pm H h_i) + F_o(r, t))}{n_S(r, t) \cdot \exp((-H h_i) + F_S(r, t)) + n_G(r, t) \cdot \exp((H h_i) + F_G(r, t))}, \]

where \( D_o \) are the diffusivities of each language, that is, measures for their spread. The diffusivities can also be seen as a measure for the region of influence of a language. We set \( \Delta t = 1 \) year because \( c_o \) is calculated individually for each year from the result of the preceding year.

The Gaussian function is a simple choice to model the interaction with neighboring cells and provides a good fit with the census data. In an extension of our model, this interaction could be modeled by other functions such as leptokuritic (long-tail) distributions or combinations of functions to describe more complex interaction patterns, for example, both long-range and short-range interaction.

Evaluation Procedure. Simulations were performed using GNU Octave 4.0.0. The data from the first census in each period (1880 and 1971) were set as the initial state from which the number of speakers in each cell changes according to Eqs. 3
After simulating the language dynamics during the Austro-Hungarian Empire, we now turn to language development in the second period after the two world wars. The development from 1971 to 2001 was first pursued using the same basic model (Eqs. 3 and 4). Table 1 shows the numerical results and Fig. 4 A and B shows the increase (red) and decrease (blue) of the number of Slovenian speakers in three regions in the districts of Völkermarkt, Klagenfurt, and Villach to search for local differences in the diffusion behavior: is there a difference between rural and urban regions? In all three regions in the districts of Völkermarkt, Klagenfurt, and Villach, the diffusivity of the Slovenian language $D_S$ is between 25 and 50% lower than the diffusivity of the German language $D_G$, with the largest difference (factor of 2) for the urban region of Klagenfurt. We suppose that the discrepancies between the different regions are due to local differences in language spread and retreat because of differences in geography and population distribution: faster diffusion in urban areas, German diffusing particularly faster than Slovenian in urban region of Klagenfurt.

Conclusion and Discussion

Macroscopic vs. Microscopic Models. In the past, language dynamics have been commonly described on a macroscopic scale by reaction–diffusion equations that model the fraction of speakers of a language in the population. However, this treatment breaks down when the spread of one language and the retreat of the other one no longer follows a traveling front because one language has become dispersed and has retreated to language islands (Fig. 2B). Additionally, reaction–diffusion equations are not applicable at all in areas without any language front movement such as behind mountain chains (green arrows in Fig. 2A). In contrast, the development can still be followed and predicted in both cases with our microscopic model. The microscopic model also takes into account the interaction with all neighboring cells, whereas in the case of a macroscopic language front the interaction only comes from one direction. Thus, microscopic models yield a more detailed and complete description of language spread and retreat than macroscopic treatment by reaction–diffusion equations.

A challenge for microscopic models on a realistic basis is obviously the need for empirical detailed data (as were at hand for this work) from which to determine the diffusivities. For this reason, language censuses have to be conducted in regular intervals and with fine-grained spatial resolution.

What Drives Language Shift? We have shown that the data predicted by our basic model (Eqs. 3 and 4) show satisfactory agreement with the historical data for the period between 1880 and 1910. Even in different socioeconomic conditions (the second period between 1971 and 2001), the predicted data still match the empirical data. This means that the basic model can reliably reproduce language dynamics of the studied language competition between Slovenian and German.

The model is also able to reveal similarities between physical phenomena like atomic diffusion and social phenomena like...
language shift: by modeling linguistic interaction as a Gaussian function as in models of physical diffusion, we obtain good agreement between the predicted and the empirical data. Thus, we have illustrated that it is possible to use physical models to simulate social dynamics on a large scale over time and space.

The basic model uses only two parameters to calculate the probability of speaking a language: the number of speakers in the preceding year and interaction between speakers. Both of these can be directly calculated from census data, ensuring our model is applicable even in situations where data on other factors influencing language use (e.g., perceived status of a language) is not available or even possible to obtain. Without interaction (i.e., using only the number of speakers), the probability of speaking a language (Eq. 3) remains constant. Consequently, interaction with other speakers is an essential drive for the linguistic change in each cell. This point has been argued by linguists (27) and is validated by our simulation. The number of speakers of a language in the population units (hamlets, villages, towns) neighboring the given cell is therefore an important influence on language dynamics. This means that a minimum-sized neighborhood of speakers of the minority language interacting with each other is necessary to preserve the language.

In addition, the simulation shows that other habitat conditions (the language of schools, and in parishes) are of minor influence. There is, however, a noticeable effect of urban areas, which have their own dynamics: between 1880 and 1910, Slovenian decays slightly faster in the larger towns than predicted by the basic model; between 1971 and 2001, the development is reversed. That is, the number of Slovenian speakers increases at a higher rate in large towns than predicted by the basic model.

ACKNOWLEDGMENTS. We thank A. Gehart and W. Zöllner as well as A. Bauer (Statistics Austria) and P. Bouinig (Department of Statistics, Government of the State of Carinthia) for their critical comments on the manuscript. The geographical data (shapefiles) are gratefully acknowledged. We thank C. Dellago for critical comments on the manuscript. The geographical data (shapefiles) used for the figure backgrounds and contour lines are provided by Land Kärnten (https://www.data.gv.at/aufliste/organisation-land-kaernten) under a CC-BY-3.0 license. Diverging color scale is based on www.ColorBrewer.org. K.P. is supported by a unidoc fellowship from the University of Vienna.

1. UNESCO Ad Hoc Expert Group on Endangered Languages (2003) Language Vitality and Endangerment. Available at unesco.unesco.org/images/0018/001836/183699e.pdf. Accessed January 16, 2017.
2. Weinreich U (1953) Languages in Contact (Linguistic Circle of New York, New York).
3. Trunova T (2005) Language Endangerment and Language Revitalization. An Introduction (Mouton, Berlin).
4. Amano T, et al. (2014) Global distribution and drivers of language extinction risk. Proc Biol Sci 281(1793):20141574.
5. Nettle D (1998) Explaining global patterns of language diversity. J Anthropol Archaeol 17: 354-374.
6. Schulze C, Stauffer D, Wichmann S (2008) Birth, survival and death of languages by Monte Carlo simulation. Commun Comput Phys 3:271–294.
7. Kandler A (2009) Demography and language competition. Hum Biol 81(2-3):181–210.
8. Zhang M, Gong T (2013) Principles of parametric estimation in modeling language competition. Proc Natl Acad Sci USA 110(24):9698–9703.
9. Hegselmann R (1996) Understanding social dynamics: The cellular automata approach. Social Science Microsimulation, eds Troitzsch KG, Mueller U, Gilbert N, Doran JE (Springer, New York), pp 282–306.
10. Gilbert N (2008) Agent-Based Models (Sage, Los Angeles).
11. Busch B (2001) Slovenian in Carinthia — a sociolinguistic survey. The Other Languages of Europe: Demographic, Sociolinguistic and Educational Perspectives, eds Extra G, Gorter D (Multilingual Matters, Clevedon, UK), pp 119–137.
12. Abrams DM, Strogatz SH (2003) Linguistics: Modelling the dynamics of language use. Nature 424(6951):900.
13. Patriarca M, Leppänen T (2004) Modeling language competition. Physica A 338: 295–299.
14. Patriarca M, Heinlau E (2009) Influence of geography on language competition. Physica A 388:174–186.
15. Kandler A, Steele J (2008) Ecological models of language competition. Biol Theory 3:164–173.
16. Kandler A, Unger R, Steele J (2010) Language shift, bilingualism and the future of Britain’s Celtic languages. Philos Trans R Soc Lond B Biol Sci 365(1559):3855–3864.
17. Walters CE (2014) A reaction-diffusion model for competing languages. Meccanica 49:2189–2206.
18. Fort J, Pérez-Losada J (2012) Front speed of language replacement. J Theor Biol 318:275–294.
19. Murray JD (1996) Mathematical Biology: An Introduction (Springer, New York).
20. Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugen 7:353–369.