Supplementary Material to Extracting information from S-curves of language change

Fakhteh Ghanbarnejad,[1,*] Martin Gerlach,[1,*] José M. Miotto,[1] and Eduardo G. Altmann[1,†]
[1] *Max Planck Institute for the Physics of Complex Systems, Dresden,Germany*

I. DATA

The data for the timeseries (available in Ref. [1]) is obtained from the latest version of the google-ngram database [2], which is an extension of the original data [3] enriched with more books and syntactic annotations. Given two or three linguistic variants, denoted by $q = 1, 2$ or 3, we count the total number of occurrences of each variant by $n_q(t)$ in each year $t \in [1800, 2008]$ irrespective of its capitalization. From this we estimate the fraction of adopters $\rho(t)$ by the relative usage of variant '1' for the given word (w):

$$\rho_w(t) = \frac{n_1^w(t)}{\sum_q n_q^w}. \tag{1}$$

Figures 1 to 4 show S-curves of individual words. We also calculate the average relative usage of one variant as an average over all tokens of an ensemble of words $w = 1, ..., W$.

$$\rho_{\text{avg}}(t) = \frac{\sum_w n_1^w}{\sum_w \sum_q n_q^w}. \tag{2}$$

The relative frequency of a word in the ensemble of W words, $f_w^{\text{rel}}(t)$, is measured as

$$f_w^{\text{rel}}(t) = \frac{\sum_q n_q^w(t)}{\sum_w \sum_q n_q^w(t)}. \tag{3}$$

While the frequency of a word, $f_w(t)$, is measured as the weight of all the variants combined divided by the size of the database (in word tokens):

$$f_w(t) = \frac{\sum_q n_q^w(t)}{\#\text{ tokens in the corpus at year t}} \tag{4}$$

We associate an error $\sigma_\rho(t)$ to each datapoint $(t, \rho(t))$, which we split into two parts, i.e.

$$\sigma_\rho^2 = \sigma_s^2 + \sigma_0^2 \tag{5}$$

in which σ_s is due to finite sampling of the data, and σ_0 subsumes additional uncertainties from exogenous perturbations. The introduction of the latter is necessary, because only considering the finite-sampling effect does not account for the observed fluctuations in the frequency of the most common words, which we assume to be stationary.

The effect of finite sampling, σ_s, is approximated by assuming that n_1 and n_2 ($q = 1, 2$) are the outcomes of a binomial process with $n = n_1 + n_2$ samples where variant '1' is drawn with probability $p = n_1/N$ and variant '2' is drawn with probability $1 - p$. From this we can calculate the error σ_s:

$$\sigma_s(t)^2 = \frac{n_1(t)n_2(t)}{(n_1(t) + n_2(t))^3}. \tag{6}$$

For the estimation of σ_0, which we treat as constant and independent of the sample size $n(t) = n_1(t) + n_2(t)$ in each year, we look at the timeseries of the relative frequency of the most frequent word, "the", in the English language. Assuming that this timeseries is stationary, we estimate σ_0, such that 95% of the points of the timeseries lie within the 95% confidence-interval assuming Gaussian errors according to Eq. (5). This gives $\sigma_0^2 = 0.002$, which we use in all cases.

---

*Both authors contributed equally to this work.
†Electronic address: fakhteh,gerlach,jmiotto,edugalt@pks.mpg.de

A. German orthographic reforms

In this section we focus exclusively on the competition between the letters 'ß' (s-sharp, $q = 2$) and 'ss' ($q = 1$) encoding the sound for voiceless s in the German orthography. The official set of rules concerning the usage of each variant changed twice in the orthographic reforms of 1901 and 1996 [4]. We investigate the usage of each variant over time for $W = 2960$ words as being affected by the orthographic reform of 1996 [5]. We consider the timeseries of four representative cases: (i) $\rho_{\text{avg}}(t)$; and three individual words as the most frequent (ii) word 'dass', $\rho_{\text{dass}}(t)$; (iii) verb 'muss', $\rho_{\text{muss}}(t)$; and (iv) noun 'einfluss', $\rho_{\text{einfluss}}(t)$.

B. Russian Names

In this section we focus on two Russian name-suffixes: 'ов' and 'ев'. The letter 'в' has been written in Roman script languages like English (en) and German (de) by 'v', 'w' or 'ff'. Here we consider the competition between the letter 'v' ($q = 1$) and two others together 'w'+'ff' ($q = 2$ and $q = 3$). We investigate the usage of each variant over time for $W = 50$ common Russian names which are listed below. We present the timeseries of six representative cases: (i) the average over all words, $\rho_{avg}(t)$; (ii) the five most used words $\rho_w(t)$.

a. German: Charkov, Saratov, Romanov, Stroganov, Tambov, Pirogov, Godunov, Katkov, Aksakov, Demidov, Semenov, Lermontov, Saltykov, Kornilov, Stepanov, Lobanov, Bulgakov, Krylov, Melnikov, Annenkov, Turgenev, Kostomarov, Filatov, Grekov, Putilov, Titov, Vinogradov, Danilov, Sobolev, Nikiforov, Kamenev, Novikov, Kondakov, Martynov, Rykov, Melikov, Platonov, Karpov, Lazarev, Balabanov, Krasnov, Nabokov, Dolgorukov, Kirov, Leonov, Maklakov, Naumov, Frolov, Mitrofanov, Fedotov

b. English: Saratov, Demidov, Pirogov, Tambov, Charkov, Katkov, Kornilov, Lazarev, Novikov, Melikov, Lermontov, Aksakov, Godunov, Turgenev, Menshikov, Stepanov, Vinogradov, Semenov, Kutuzov, Lebedev, Suvorov, Lomonosov, Mendeleev, Lavrov, Melnikov, Lobanov, Annenkov, Volkhov, Balakirev, Lvov, Bazarov, Shuvalov, Grigoriev, Titov, Yakov, Nekrasov, Mikhailov, Gorchakov, Morozov, Zubov, Chekhov, Sakharov, Dragomirov, Andreyev, Danilov, Chirikov, Yermolov, Bulgakov, Vasiliev, Saltykov

To make the lists, the primary list of common Russian names ending 'ов' and 'ев' was created according to the English Wikipedia pages including: List of surnames in Russia, List of Russian-language writers, scientists, composers, leaders of the Soviet Union and Marshals of the Soviet Union; Also list of cities and towns in Russia was counted in this list. The words which have been used at least 10 times for more than 100 years (75 years for German data) in this period were included in the initial list. Furthermore, in order to guarantee that these words are right competitors, we removed words satisfiying one of the following conditions:

- First letter written mostly by small letters instead of capital letters ($\frac{\sum_{t=1800}^{2008} f_w^{small}(t)}{\sum_{t=1800}^{2008} f_w^{captal}(t)} \geq 0.01$)
- Sudden peak at the late 20 century and before that were rarely used ($\frac{\sum_{t=1950}^{2000} f_{word}(t)}{0.99*\sum_{t=1850}^{2000} f_{word}(t)} \geq 1$), e.g. *Gorbachev*.
- Entries in Wikipedia not corresponding to the Russian origin e.g. *Rostow* which refers to Americans and *Romanow* which refers to Polish places

C. Regularization verbs in English

In this section we focus on the regularization of English verbs [6]. In addition to the regular past form of a verb, which is generated by adding -ed (laugh → laughed), there exists a small number of verbs which are conjugated irregularly, e.g. burn → burnt. However, all irregular forms coexist with a corresponding regular variant as listed in Ref. [3]. We investigate the competition between the regular ($q = 1$) and the irregular ($q = 2$) form for 281 verbs with a recently attested irregular form [3]. As an example, for the verb 'write', $n_1(t) = n(\text{writed}, t)$, and $n_2(t) = n(\text{writ}, t) + n(\text{written}, t) + n(\text{wrote}, t)$, since we have to combine the usage of past participle and preterit to capture all irregular past forms.

The following filterings were employed. We discarded any verb, where the irregular past form is the same as the infinitive since it would not be possible to distinguish between a verb that is used as a past form or a present form, e.g. for the verb 'beat' the irregular preterit is 'beat'. We condition the counts on those forms that are identified as verbs by the associated part-of-speech tag (already provided in the data). We then selected the 10 verbs that exhibit the largest relative change $\mid \rho_1 - \rho_0 \mid$, where ρ_0 and ρ_1 is the average over the 20 datapoints in the beginning ($t \in [1800 - 1819]$) and the end ($t \in [1989 - 2008]$) of the timeseries respectively. These verbs

are: abide (abided/abode), burn (burned/burnt), chide (chided/chid,chidden), cleave (cleaved/clove,cloven), light (lighted/lit), smell (smelled/smelt), spell (spelled/spilt), spill (spilled/spilt), thrive (thrived/throve,thriven), wake (waked/woke,waken).

II. SURROGATE DATA

In the following we describe how to apply our notion of endogenous and exogenous influence to several paradigmatic models of innovation spreading on complex networks.

A. Approximate Master Equation

We formulate the dynamics in the framework of approximate master equations (AME) [7], which describe the stochastic binary dynamics in an uncorrelated network with a given degree distribution P_k. Nodes can either be potential adopters (susceptible) or have already adopted the innovation (infected) and are grouped into classes $\{k, m\}$, where k denotes the degree and m the number of infected neighbors. The dynamics is specified by the function $F_{k,m}$ ($R_{k,m}$), the rate at which susceptible (infected) $\{k, m\}$-nodes become infected (susceptible). In this work we only consider monotone dynamics, i.e. $R_{k,m} = 0$, which leads to the following system of ordinary differential equations for the fraction of susceptible $\{k, m\}$-nodes, $s_{k,m}$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} s_{k,m} =& - F_{k,m} s_{k,m} - C\,(k-m)\, s_{k,m} \\ &+ C\,(k-m+1)\, s_{k,m-1},\end{aligned} \tag{7}$$

with $C = \langle (k-m)\, F_{k,m} s_{k,m} \rangle / \langle (k-m)\, s_{k,m} \rangle$ where $\langle \cdot \rangle = \sum_k P_k \sum_{m=0}^{k}$. From $s_{k,m}(t)$ we can calculate the timeseries for the total fraction of infected nodes, $\rho(t)$, according to:

$$\rho(t) = 1 - \sum_k P_k \sum_{m=0}^{k} s_{k,m}. \tag{8}$$

Assuming that at time t_0 a randomly chosen fraction of nodes, $\rho(t_0) = \rho_0$, is infected, we get as initial conditions for $s_{k,m}$ [7]:

$$s_{k,m}(t_0) = \binom{k}{m} \rho_0^m \left(1-\rho_0\right)^{k-m} \left(1-\rho_0\right). \tag{9}$$

We investigate two special cases of functions $F_{k,m}$ with two parameters a and b:

Bass-Model: In the Bass-model the probability of becoming infected is proportional to the number of neighbors that are already infected:

$$F_{k,m} = a + b\frac{m}{k}, \tag{10}$$

Threshold-model: In a threshold-model a node becomes infected with probability 1 if the fraction of infected neighbors exceeds a certain threshold:

$$\text{Threshold: } F_{k,m} = \begin{cases} a, & m/k < 1-b \\ 1, & m/k \geq 1-b \end{cases}. \tag{11}$$

B. Exogenous and Endogenous Influence

The formulation of the spreading dynamics in the framework of AME allows us to calculate exactly the 'ground truth' of the exogenous and endogenous contributions for any given $F_{k,m}$. Following the approach in Sec. II, main

text, we can now calculate exactly the individual contributions:

$$\begin{aligned} G^j &= \frac{1}{N}\sum_{i=1}^{N}\frac{g^j\left(t_i^*\right)}{g\left(t_i^*\right)}, \\ &= \frac{1}{N}\sum_k\sum_{i\in\{k\}}\frac{g^j\left(t_i^*,m_i^*,k\right)}{g\left(t_i^*,m_i^*,k\right)}, \\ &= \sum_k P_k\sum_{m=0}^{k}\int_0^\infty \frac{g^j\left(t,m,k\right)}{g\left(t,m,k\right)}\Delta_{k,m}\left(t\right)\mathrm{d}t, \end{aligned} \tag{12}$$

where $\Delta_{k,m}(t) = F_{k,m}s_{k,m}$ is the actual fraction of $\{k,m\}$-nodes that changed from susceptible to infected at time t. Noting that the total rate of change is given by $g(k,m) = F_{k,m}$, it follows that

$$G^j = \sum_k P_k\sum_{m=0}^{k}\int_0^\infty g^j\left(t,m,k\right)s_{k,m}\left(t\right)\mathrm{d}t, \tag{13}$$

Assuming that the exogenous contribution is given by transitions that occur when no neighbor is infected, i.e. $g^{\mathrm{exo}}\left(k,m\right) = F_{k,0}$, the exogenous and endogenous contribution yields:

$$\begin{aligned} G^{\mathrm{exo}} &= \sum_k P_k\sum_{m=0}^{k}\int_0^\infty F_{k,0}s_{k,m}\mathrm{d}t, \\ G^{\mathrm{endo}} &= \sum_k P_k\sum_{m=0}^{k}\int_0^\infty \left(F_{k,m}-F_{k,0}\right)s_{k,m}\mathrm{d}t. \end{aligned} \tag{14}$$

C. Numerical Implementation

Given a degree-sequence $k \in [k_{\min}, k_{\max}]$, a degree-distribution P_k, and one of the $F_{k,m}$ from Eqs. (10,11), we can solve the set of differential equations for $s_{k,m}$ numerically according to Eq. (7). We use scipy's [8] odeint-implementation to get the timeseries $\rho(t)$ from Eq. (9) and the true exogenous and endogenous influence from Eq. (14) for a particular trajectory. We set as parameters $\rho_0 = \rho(t_0 = 0) = 10^{-3}$ and sample the trajectory $\rho(t)$ at discrete points $t \in \{t_0 + i\cdot dt\}$ for $i = 1,\ldots,N$ with $dt = 0.01$ and $\rho(t = Ndt) \geq 1-\rho_0$.

III. TIME SERIES ESTIMATORS

In this section we explore three different methods of how to quantitatively assess the endogenous and exogenous factor in the spreading dynamics of a linguistic variant within the population. We want to restrict ourselves to the analysis of the timeseries of the total fraction of adopters of the innovation, $\rho(t)$, with $\rho \in [0,1]$ given a set of N observations $D = \{t_i, \rho_i\,\sigma_i\}$ with $i = 1,\ldots,N$, where t_i is the time, ρ_i the relative usage of one variant over the other, and σ_i the error associated to ρ_i, for details in the real data, see Sec. I

Our starting point for all three methods is the assumption that the dynamics of the total fraction of adopters, $\rho(t)$, can be effectively described by a generalized population-dynamics model:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = \left[1-\tilde{\rho}\left(t\right)\right]g\left(\tilde{\rho}\left(t\right)\right), \tag{15}$$

which means that the rate of change of $\tilde{\rho}$ is determined by an arbitrary function $g(\tilde{\rho})$ only affecting the fraction of susceptibles, $1-\tilde{\rho}$. Further, we want to account for the fact that the fraction of adopters is bounded by the two asymptotic values y_0 and y_1 such that $\rho(t\to-\infty) = y_0$ and $\rho(t\to\infty) = y_1$, which gives for the dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = \begin{cases}\left[y_1-\tilde{\rho}\right]g\left(\tilde{\rho}\right), & \tilde{\rho}\in[y_0,y_1]\\ 0, & \text{else}\end{cases}. \tag{16}$$

This rescaling of the asymtoyics is necessary to compare the models (in which $y_0 = 0$ and $y_1 = 1$) to data (in which typically $y_0 > 0$ and $y_1 < 1$). Additional parameter t_0 sets the characteristic timescale, such that $\tilde{\rho}(t_0) = \frac{1}{2}(y_0+y_1)$,

which is equivalent to specifying the initial condition. Assuming a parametrization of $g(\tilde{\rho} \mid \theta)$ by the set of parameters θ we calculate the Least-Squared-Error, $\Delta(t_0, y_0, y_1, \theta)$ between data $D = \{t_i, \rho_i, \sigma_i\}$ and the resulting curve $\tilde{\rho}(t \mid t_0, y_0, y_1, \theta)$ from our model

$$\Delta(t_0, y_0, y_1, \theta) = \sum_{i=1}^{N} \left(\frac{\rho_i - \tilde{\rho}(t_i \mid t_0, y_0, y_1, \theta)}{\sigma_i} \right)^2 . \tag{17}$$

From this we can infer the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta})$:

$$(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta}) = \underset{(t_0, y_0, y_1, \theta)}{\operatorname{argmin}} \Delta(t_0, y_0, y_1, \theta) \tag{18}$$

A. Method 1: S- vs Exponential Curve

The simplest version of a purely exogenously (endogenously) driven population dynamics spreading process assumes $g(\rho) = a$, $(g(\rho) = b(\rho - y_0))$. We want to determine which curve provide a better description of the data. For this, we calculate the relative likelihood of each model, which is used in an information-theoretic approach to model selection [9].

Real Data: With these choices of $g(\rho)$ we can solve Eq. (16) analytically which yields a four-parameter curve for each case:

$$\rho_{\text{exo}}(t \mid t_0, y_0, y_1, a) = \begin{cases} y_1 - \frac{1}{2}(y_1 - y_0)\, e^{-a(t-t_0)}, & t \geq t^* \\ y_0, & t < t^* \end{cases} , \tag{19}$$

$$\rho_{\text{endo}}(t \mid t_0, y_0, y_1, b) = y_0 + \frac{y_1 - y_0}{1 + e^{-b(y_1 - y_0)(t - t_0)}}, \tag{20}$$

with

$$t^* = t_0 - \frac{\ln 2}{a}. \tag{21}$$

Given our observational data D we can then find the best choice of parameters for each case and calculate the Least-Square-Error $\Delta_{\text{exo}}(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a})$ and $\Delta_{\text{endo}}(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{b})$ according to Eqs. (17,18).

In order to decide which of the two models (endogenous or exogenous) fits the data better, we employ the Bayesian information criterion (BIC) [10] used in model selection [9, 11] which is given by

$$BIC_{\text{exo}} = \Delta(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a}) + \log(N) K_{\text{exo}} \tag{22}$$

$$BIC_{\text{endo}} = \Delta(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{b}) + \log(N) K_{\text{endo}} \tag{23}$$

where $K_{\text{exo}} = K_{\text{endo}}$ is the number of fitted parameters in each model, and N the number of data points. From this we can calculate the relative likelihood, L_{exo} (L_{endo}), quantifying the evidence of the exogenous (endogenous) model among the selection of the two models (exogenous and endogenous) for the given data [9]:

$$L_{\text{exo}} = \frac{e^{-1/2 BIC_{\text{exo}}}}{e^{-1/2 BIC_{\text{exo}}} + e^{-1/2 BIC_{\text{endo}}}} \tag{24}$$

$$L_{\text{endo}} = \frac{e^{-1/2 BIC_{\text{endo}}}}{e^{-1/2 BIC_{\text{exo}}} + e^{-1/2 BIC_{\text{endo}}}} \tag{25}$$

with $L_{\text{exo}} + L_{\text{endo}} = 1$.

Surrogate Data: For the surrogate data, see Sec. II, we know that $y_0 = 0$ and $y_1 = 1$ by construction. We further specify the initial condition for the spreading process, $\rho(t = t_0) = \rho_0$, which reduces the above curves to one-parameter models:

$$\rho_{\text{exo}}(t \mid a) = \begin{cases} 1 - (1 - \rho_0) e^{-a(t-t_0)}, & t \geq t^* \\ 0, & t < t^* \end{cases} , \tag{26}$$

$$\rho_{\text{endo}}(t \mid b) = \frac{1}{1 + e^{-b(t - t_0)}} \tag{27}$$

with $t_* = t_0 + \ln(1 - \rho_0)$.

B. Method 2: Mixed Curve

In this section we want to extend our previous model by assuming that, both, exogenous and endogenous driving is present in the spreading dynamics simultaneously, i.e. $g(\rho) = a + b\,(\rho - y_0)$.

Real Data: Solving Eq. (16) yields a 5-parameter curve for ρ:

$$\rho_{\text{mixed}}(t \mid t_0, y_0, y_1, a, b) = \begin{cases} \frac{-(a-by_0)(y_1-y_0)+y_1(2a+b(y_1-y_0))e^{[a+b(y_1-y_0)](t-t_0)}}{b(y_1-y_0)+(2a+b(y_1-y_0))e^{[a+b(y_1-y_0)](t-t_0)}}, & t \geq t^* \\ y_0, & t < t^* \end{cases}, \tag{28}$$

with

$$t^* = t_0 - \frac{\ln\left(2 + \frac{b}{a}\,(y_1 - y_0)\right)}{a + b\,(y_1 - y_0)}. \tag{29}$$

We note that the special case $a = 0$ yields $t^* \to -\infty$, which means that for all finite t: $\rho(t) > y_0$ and only in the limit $\rho(t \to -\infty) = y_0$. Given the data D we estimate the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a}, \hat{b})$ using Eqs. (17,18).

For the given choice of $g(\rho) = a + (b - y_0)\rho$ we define the exogenous and the endogenous influence as

$$g^{\text{exo}}(\rho) = g(\rho = \hat{y}_0) = \hat{a}, \tag{30}$$

$$g^{\text{exo}}(\rho) = g(\rho) - g^{\text{exo}}(\rho) = \hat{b}(\rho - \hat{y}_0). \tag{31}$$

From this we can calculate the total exogenous and endogenous influence in the spreading process as the fraction of the population that switches at time t, $\dot{\rho}(t)$, weighted by the relative exogenous influence, $g^{\text{exo}}(\rho)/g(\rho)$, and relative endogenous influence, $g^{\text{endo}}(\rho)/g(\rho)$, respectively, integrated along the complete trajectory $\rho(t)$

$$G^{\text{exo}} = \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{-\infty}^{\infty} \mathrm{d}t \dot{\rho}(t) \frac{g_{\text{exo}}(\rho(t))}{g(\rho(t))} \tag{32}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{g_{\text{exo}}(\rho)}{g(\rho)} \tag{33}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{\hat{a}}{\hat{a} + \hat{b}\,(\rho - \hat{y}_0)} \tag{34}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \frac{\hat{a}}{\hat{b}} \ln\left[\frac{\hat{a} + \hat{b}\,(\hat{y}_1 - \hat{y}_0)}{\hat{a}}\right] \tag{35}$$

and

$$G^{\text{endo}} = \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{-\infty}^{\infty} \mathrm{d}t \dot{\rho}(t) \frac{g_{\text{endo}}(\rho(t))}{g(\rho(t))} \tag{36}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{g_{\text{endo}}(\rho)}{g(\rho)} \tag{37}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{\hat{b}\,(\rho - \hat{y}_0)}{\hat{a} + \hat{b}\,(\rho - \hat{y}_0)} \tag{38}$$

$$= 1 - \frac{1}{\hat{y}_1 - \hat{y}_0} \frac{\hat{a}}{\hat{b}} \ln\left[\frac{\hat{a} + \hat{b}\,(\hat{y}_1 - \hat{y}_0)}{\hat{a}}\right], \tag{39}$$

in which we choose the prefactor $1/(\hat{y}_1 - \hat{y}_0)$ such that we have the normalization $G^{\text{exo}} + G^{\text{endo}} = 1$.

Surrogate Data: For the surrogate data, see Sec. II, we know that $y_0 = 0$ and $y_1 = 1$ by construction. We further specify the initial condition for the spreading process, $\rho(t = t_0) = \rho_0$, which reduces the above curve to a two-parameter model:

$$\rho_{\text{mixed}}(t \mid a, b) = \begin{cases} \frac{-a(1-\rho_0)+(a+b\rho_0)e^{(a+b)(t-t_0)}}{b(1-\rho_0)+(a+b\rho_0)e^{(a+b)(t-t_0)}}, & t \geq t^* \\ 0, & t < t^* \end{cases}, \tag{40}$$

with

$$t^* = t_0 + \frac{1}{a+b} \ln \frac{a\,(1-\rho_0)}{a + b\rho_0}. \tag{41}$$

C. Method 3: Nonparametric Curve

In this section we want to infer the exogenous and the endogenous influence non-parametrically not assuming any specific functional form of $g(\rho)$. The idea is to infer $g(\rho)$ from the timeseries directly according to Eq. (16)

$$\hat{g}(\rho) := \frac{\dot{\rho}(t)}{y_1 - \rho(t)}. \tag{42}$$

From this we can infer the exogenous and the endogenous influence along the trajectory ρ:

$$g^{\text{exo}} = \hat{g}(\rho = y_0) \tag{43}$$

$$g^{\text{endo}} = \hat{g}(\rho) - \hat{g}(\rho = y_0) \tag{44}$$

which gives for the total exogenous and endogenous contribution

$$G^{\text{exo}} = \frac{1}{y_1 - y_0} \int \mathrm{d}\rho \frac{g^{\text{exo}}}{\hat{g}(\rho)} \tag{45}$$

$$G^{\text{endo}} = \frac{1}{y_1 - y_0} \int \mathrm{d}\rho \frac{g^{\text{endo}}}{\hat{g}(\rho)} \tag{46}$$

Surrogate Data: For the surrogate data we can infer $g(\rho)$ *directly* with the timeseries $\rho(t)$ being sampled at a given resolution in discrete time, $t = (t_i)$ with $i = 1..N$, such that we can approximate the time-derivate of $\rho(t)$ by finite differences, e.g.

$$\dot{\rho}(t_i) \approx \frac{\rho(t_{i+1}) - \rho(t_i)}{t_{i+1} - t_i} \tag{47}$$

for $i = 1...N-1$. Assuming that $\rho(t)$ is a monotone function in t, i.e. $t = t(\rho)$, we can express the time derivative as

$$\dot{\rho}(t) \overset{t=t(\rho)}{\longrightarrow} \dot{\rho}(\rho) \tag{48}$$

such that we can evaluate $\hat{g}(\rho)$, see Eq. (42), from the timeseries $\rho(t)$ and its derivative $\dot{\rho}$ via:

$$\hat{g}(\rho) := \frac{\dot{\rho}\left[t(\rho)\right]}{1 - \rho} \tag{49}$$

Real Data: Real data is only available with a given resolution in t and is subject to fluctuations, therefore the direct calculation of $\dot{\rho}$ in Eq. (42) does not lead to meaningful results. Instead, we want to infer $g(\rho)$ *indirectly*, i.e. find a particular choice of $g(\rho)$ that yields the best description of the data by solving Eq. (16) for $\rho(t)$ and then applying Eqs. (17,18). Our approach is to parametrize $g(\rho)$ by means of a natural cubic spline $s(\rho)$ [11]. Therefore, we divide the support of $g(\rho)$, $\rho \in [y_0, y_1]$, into n intervals of equal length $h = \frac{y_1 - y_0}{n}$, $\{[y_0 + (i-1)h, y_0 + ih\}$ for $i = 1...n$. In each interval i we define a cubic polynomial, such that the resulting curve $s_n(\rho)$ is piecewise-polynomial of order 4 and has continuous derivatives up to order 2. Furthermore, we restrict ourselves to natural cubic splines which implies that $s_n''(\rho = y_0) = s_n''(\rho = y_1) = 0$. The resulting spline $s_n(\rho)$ contains $n+1$ parameters $\theta = (\theta_i)$ with $i = 1...n+1$ and two additional parameters (y_0, y_1) specifying the asymptotic values for $\rho(t \to \pm\infty)$ which we denote by $s_n(\rho \mid (y_0, y_1, \theta))$.

For any given n we can infer $\hat{g}_n(\rho) = s_n(\rho \mid \hat{y}_0, \hat{y}_1, \hat{\theta})$ by Eqs. (16,17,18), which requires an extra parameter t_0 setting a characteristic time-scale of the change of $\rho(t)$ in time. In total, for a parametrization of $g(\rho)$ by a natural cubic spline on n intervals, we have $K = n + 4$ parameters. Finally, the exogenous and the endogenous influence in the spreading are calculated via Eqs. (43-46).

The crucial step then is to decide which value to choose for n as we have to find a trade-off between the most accurate description of the data and the problem of overfitting known as model selection [11]. We infer the best model by means of the Bayesian information criterion (BIC), which penalizes models with additional parameters according to:

$$BIC = \Delta + \log(N)K, \tag{50}$$

where Δ is the Least-Square error of the best fit of a given model according to Eqs. (17,18), K is the number of parameters estimated, and N is the number of datapoints. Due to computational constraints we restrict ourselves to the cases $n = 1, ...10$.

D. Numerical Implementation

The above mentioned methods require the minimization of the least-square error, see Eq. (18), in the space of parameters (t_0, y_0, y_1, θ). We find the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta})$ numerically using the 'L-BFGS-B'-algorithm [12] from scipy's optimization package [8]. The algorithm allows to impose additional constraints on a parameter x, such that we ensure that $x_{\min} \leq \hat{x} \leq x_{\max}$. In our case we choose the following constraints:

1. t_0 is unconstrained,
2. $0 \leq y_0, y_1 \leq 1$ since these parameters describe the asymptotic values of the fraction of adopters, i.e. $\rho(t \to \pm\infty)$,
3. $0 \leq a, b$ for method 1 and 2 considering positive exogenous and endogenous contributions,
4. $0 \leq \theta_i$ for $i = 1, \ldots, n+1$ for method 3 in order to guarantee that $\hat{g}(\rho) \geq 0$.

Addressing the issue of local minima, for each timeseries we perform the minimization task 100 times with different randomly chosen initial conditions in parameter space and select the global minimum.

We calculate the confidence intervals from standard bootstrapping [11], i.e. performing the same analysis for a number of B surrogate datasets obtained from random sampling with replacement of the original data (here $B = 200$).

In Fig. 1-4 we show the individual timeseries from the data described in Sec. I, the best fit of the mixed curve, method 2 from Sec. III B, and the results for assessing the L and the exogenous factor, G^{exo}, from all three methods, Sec. III A-III C. Timeseries and fitting script are available in Ref. [1].

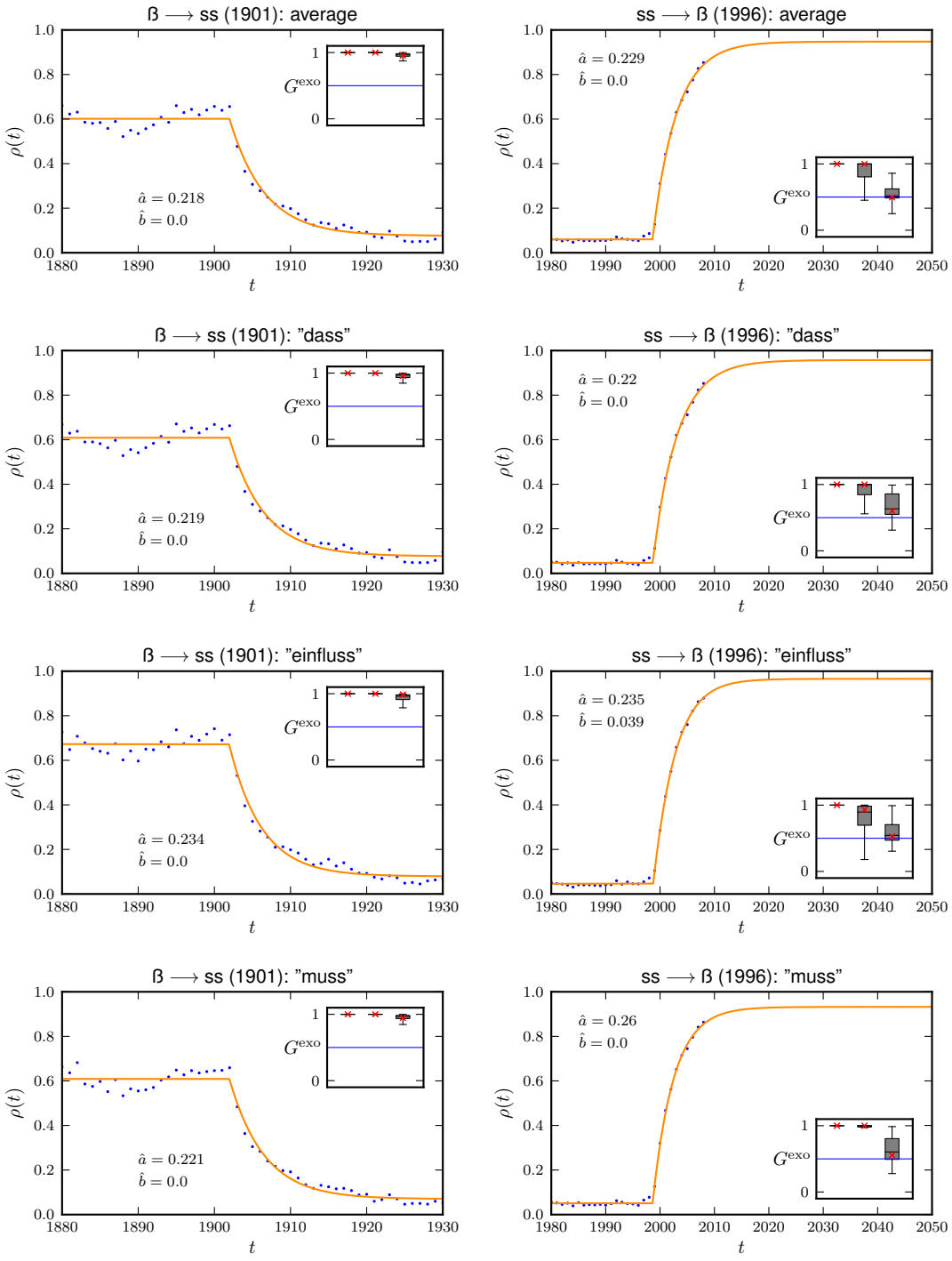

Figure 1: Orthographic Reform of 1901 and 1996. The timeseries show the data (dots), the best fit of method 2 (line) with the values of its two parameters $\hat{a}$ and $\hat{b}$, and the boxplot for the estimation of G^{exo} (inset) for all three methods: method 1 (left), method 2 (middle), and method 3 (right) with the result for the full data (red cross) and the 97.5%-, 75%-, 50%-,25%-, and 2.5%- percentiles from bootstrapping (black lines). Timeseries and fitting script are available in Ref. [1].

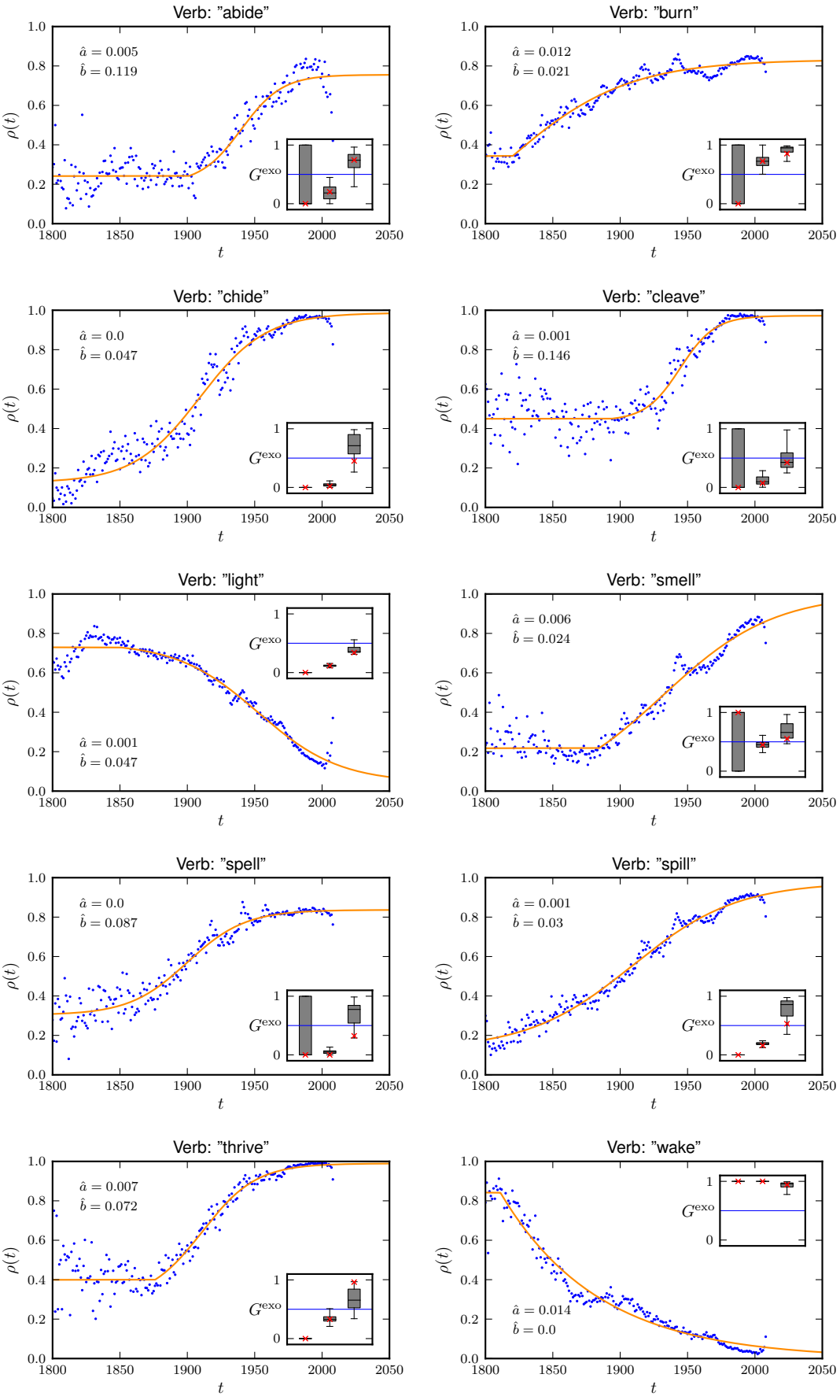

Figure 2: Regularization of English Verbs. Description see Fig. 1.

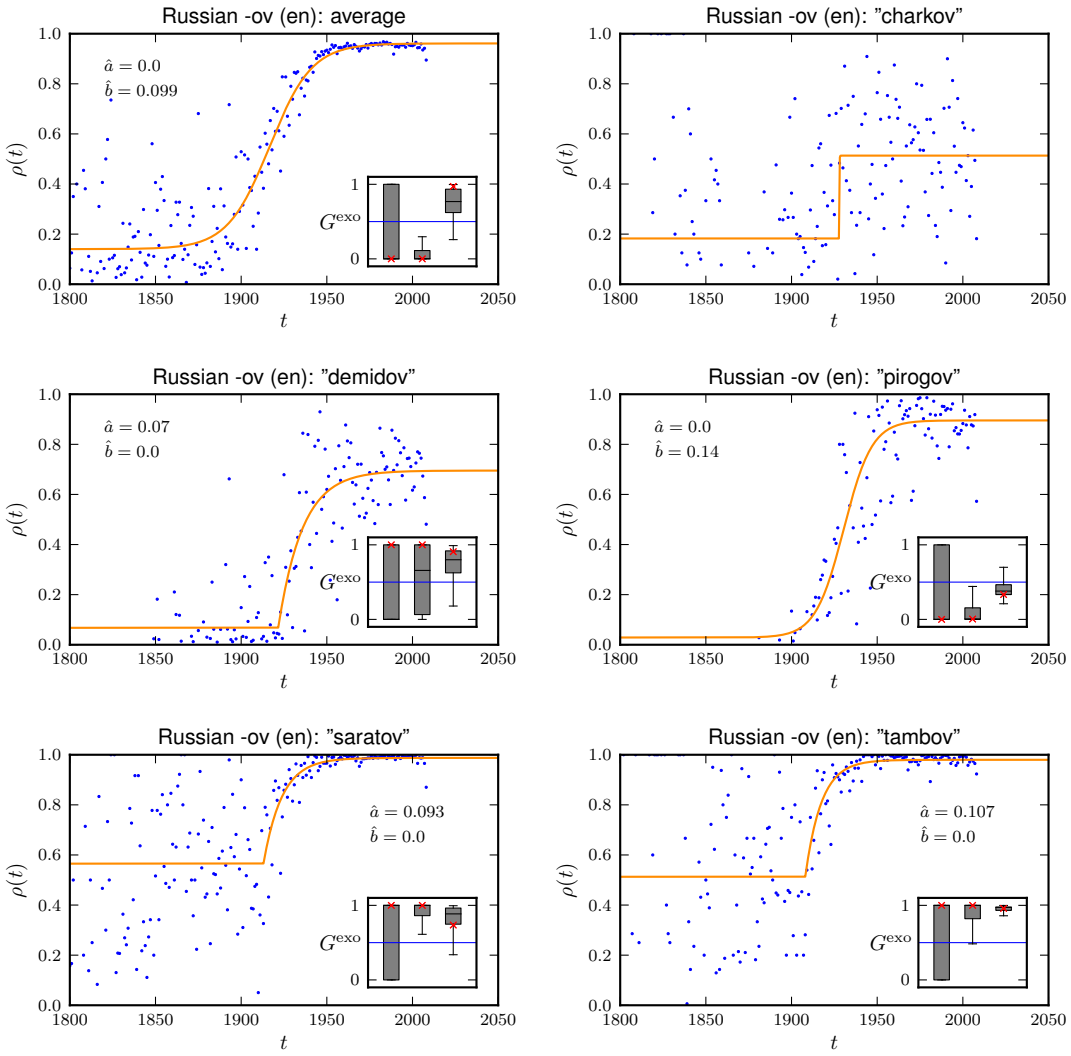

Figure 3: Transliteration of Russian -ов (en). Description see Fig. 1. For the word "charkov" the best fit is a step function meaning that there is no local minimum in the Least-Squared-Error, Eq. 17, for finite values of the parameters a, b; therefore, the numerical estimation of G^{exo} is omitted.

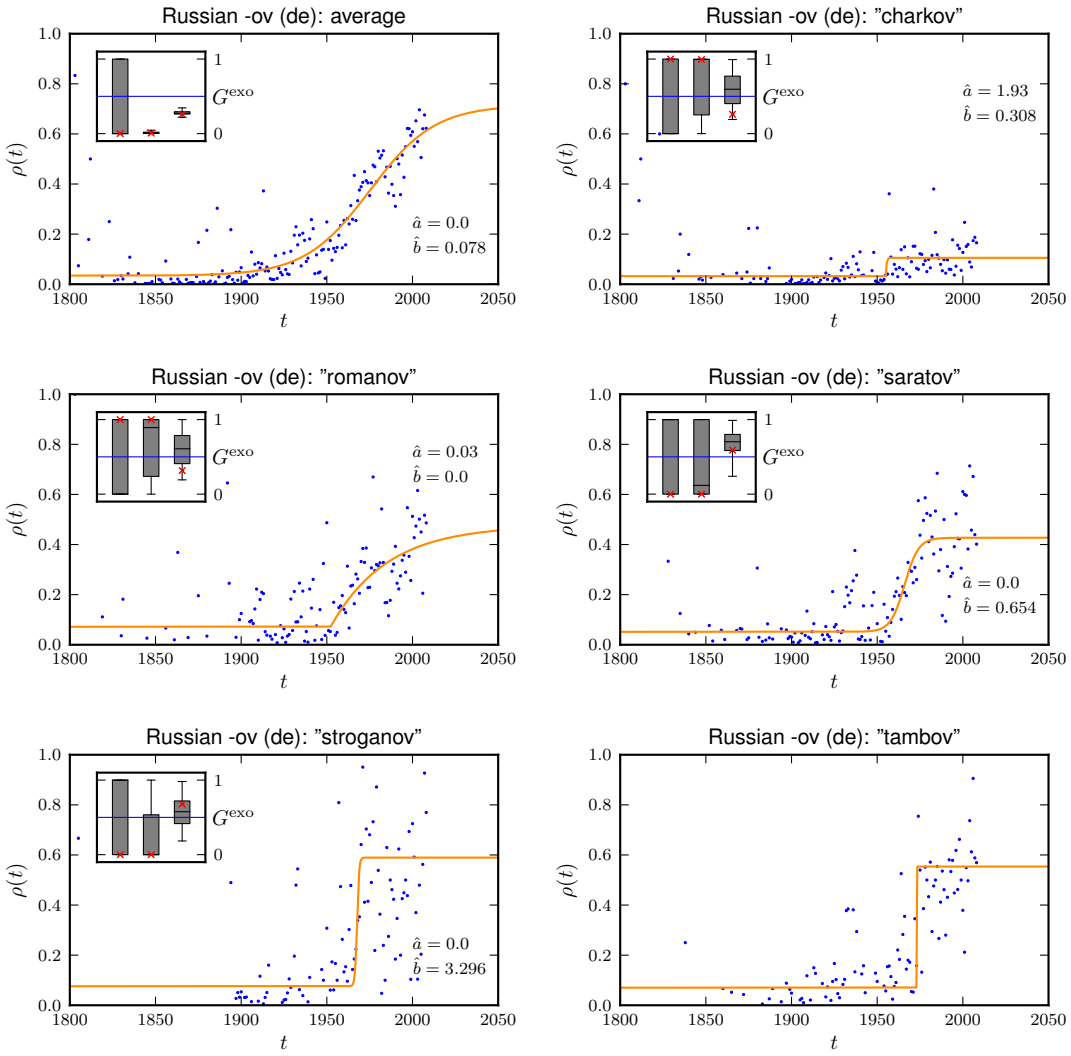

Figure 4: Transliteration of Russian -ов (de). Description see Fig. 1. For the word "tambov" the best fit is a step function meaning that there is no local minimum in the Least-Squared-Error, Eq. 17, for finite values of the parameters a, b; therefore, the numerical estimation of G^{exo} is omitted.

[1] F. Ghanbarnejad, M. Gerlach, J. M. Miotto, and E. Altmann (2014), URL http://dx.doi.org/10.6084/m9.figshare.1172265.

[2] Y. Lin, J.-B. Michel, E. Lieberman Aiden, J. Orwant, W. Brockman, and S. Petrov, in *Proceedings of the ACL 2012 System Demonstrations* (Association for Computational Linguistics, 2012), pp. 169–174, URL http://www.aclweb.org/anthology/P/P12/P12-3029.pdf.

[3] J.-B. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al., Science **331**, 176 (2011), URL http://www.ncbi.nlm.nih.gov/pubmed/21163965.

[4] S. Johnson, *Spelling trouble: Language, ideology and the reform of German orthography* (Multilingual Matters, Clevedon, UK, 2005).

[5] Canoonet, *German dictionaries and grammar [http://www.canoo.net]; accessed 03/04/2013.*

[6] S. Pinker, *Words and Rules: The Ingredients of Language* (Basic Books, New York, NY, US, 1999), URL http://psycnet.apa.org/psycinfo/1999-04277-000.

[7] J. Gleeson, Physical Review X **3**, 021004 (2013), URL http://link.aps.org/doi/10.1103/PhysRevX.3.021004.

[8] E. Jones et al., *SciPy: Open source scientific tools for Python* (2001–), URL http://www.scipy.org/.

[9] K. Burnham and D. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2002), 2nd ed.

[10] G. Schwarz, The Annals of Statistics **6**, 461 (1978), URL http://projecteuclid.org/euclid.aos/1176344136.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, 2009), 2nd ed.

[12] R. Byrd, P. Lu, and J. Nocedal, SIAM Journal on Scientific and Statistical Computing **16**, 1190 (1995).