HOW HUMANS TRANSMIT LANGUAGE: SUPPLEMENTARY MATERIAL

A MODEL FOR LANGUAGE INTERNALISATION AND TRANSMISSION.

J. BRYDEN, S. WRIGHT, AND V. A. A. JANSEN

1. The bag of words model

We will present a basic model for language transmission. The basic premise is that language production is based, in part, on an internalised distribution of words. This internalised distribution will be formed dependent on the frequency of words in the language that an individual is exposed to, and this internalised distribution will in turn be used for the production of language. We will assume that this part of the internalised language functions like a bag (also known as a multiset): a set of which can have multiple instances of its elements can appear more than once. In a bag of words the same word can appear multiple times. We assume that the bag has got a fixed size.

Language is produced by sampling from this bag. As a consequence, for sufficiently long text the word frequency in the text will tend to the frequency in the bag. Language is internalised through sampling from the words that an individual receives. Our model assumes that words are sampled from the incoming text, and placed in the bag. To keep the size of the bag constant, a random element is then removed from the bag.

Individuals are in contact, and communicate with other individuals. These contacts span up a network on which communication takes place and messages are exchanged. The words produced are sampled from the bag of words of the individual that sends the message. A sample of these words will be internalised by the receiver of the message.

We consider a population of n individuals. Individual i sends words to individual j with rate r_{ij}. Let the bag of words of individual i be given by the set $\{x_{i1}, \ldots, x_{iw}\}$ where the x_{ik} is number of copies of the kth word that individual i has in its bag of words. The numbering of the words is the same for all individuals and the w is the total number of words in the population. The size of the bag, s, is constant and the same for all individuals, and hence $\sum_{k=1}^{w} x_{ik} = s$ for all $1 \leq i \leq n$.

Through the internalisation of words received, the internalised word numbers $x_{k,i}$ change over time. We will describe this change through a stochastic process similar to a Moran process (Blythe 2012, BLythe and McKane 2007). The Moran process, in its simplest form, describes the change in a population of genes through selection and random drift caused by replacement through birth in a finite population. Here, we will apply a similar logic to describe the changes in the bag of words. Words are internalised with a rate proportional to the words received. Individual i receives messages from individual j at rate r_{ij}, and we will assume that individuals do not act on messages received from self and hence take all $r_{ii} = 0$. The rate with word m is internalised is $\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s}$, where α_m is the rate constant for word m, or the probability per word of type m received that it will lead to an internalisation event. In every internalisation event a random word is removed from the receiver's bag. The rate with which a copy of word k is removed is therefore $\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} \frac{x_{ik}}{s}$.

TABLE S1. Parameters used in the model

Parameter	Description
n	Number of individuals
w	Number of words in population
s	Size of the bag
r_{ij}	Rate individual i sends words to individual j
x_{ik}	Number of copies of the kth word that individual i has in its bag of word
α_m	Probability that a word of type m is internalised

Thus the event:

$$x_{ik} \to x_{ik} - 1; x_{im} \to x_{im} + 1$$

takes place with rate

$$\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} \frac{x_{ik}}{s}$$

1.1. **Simulating change.** The model above is a generic description of the transmission and internalisation of word frequencies. It can be used to simulate networks of communicating individuals to study their changes in word frequencies. This can be done as follows: first, set up a network through specifying the rates of communication r_{ij}. Initialise the network by assigning an initial bag of words to all nodes on the network, which specifies the network at $t = 0$. Internalisation events take place with total rate

$$\sum_{i=1}^{n} \sum_{m=1}^{w} \sum_{k=1}^{w} \alpha_m \frac{x_{ik}}{s} \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} = \sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s}.$$

Using the Gillespie algorithm, the waiting time until the next event is exponentially distributed with mean

$$\left(\sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right)^{-1}.$$

We can draw the waiting time until the next event and update time from this distribution. The chance that the next event is in individual i is

$$\left(\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right) \left(\sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right)^{-1},$$

and use this to identify i. Word k is removed with probability $\frac{x_{ik}}{s}$, use this to lower the number of x_{ik} by one. Refill the gap with word m. Word m is now chosen with probability

$$\left(\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s}\right)\left(\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s}\right)^{-1}.$$

Note that these expressions simplify considerably if the internalisation rates are the same for all words: if $\alpha_m = \alpha$ then $\sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} = \alpha$

1.2. **Ensemble means of word frequencies.** We will next derive the behaviour of the system if it would be averaged over many simulations, of the same network, all starting from the same initial conditions. We will refer to one such simulation as a single realisation of the process. To find the average over many realisations we will write down the master equation for this process.

To do so we will consider the probability of the system to be in a certain state as the process develops. The system consist of n individuals, who each have bags of words with s spaces, and we need to keep track of all possible permutations of words possible. To help us do the book keeping we need some extra notation. All communication rates can be put together in a $n \times n$ matrix $\mathbf{R} = (r_{ij})$. We will represent the state of the system by a $n \times w$ matrix $\mathbf{X} = (x_{im})$, which has as elements the word counts of all individuals. Let $P(\mathbf{X})(t)$ be the probability for the system to be in state $\mathbf{X}$ at time t. This probability changes over time according to the master equation, given by:

$$\frac{\mathrm{d}P(\mathbf{X})}{\mathrm{d}t} = \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s}\left(\sum_{k=1}^{w} \frac{x_{ik}+1}{s} P(\mathbf{X}-\mathbf{E}_{i,m}+\mathbf{E}_{i,k}) - \sum_{k=1}^{w} \frac{x_{ik}}{s} P(\mathbf{X})\right)$$

where $\mathbf{E}_{i,m}$ is a $n \times w$ matrix, in which all elements are zero, expect from element i, m, which is one. We will use the master equation to calculate how the ensemble mean of the word frequencies change over time. The ensemble mean of all the word counts of all individuals $\mathbf{X}$ is given by $\sum_{\mathbf{X}\in\Omega} \mathbf{X} P(\mathbf{X})$, where Ω is the set that contains all the different states that $\mathbf{X}$ can be in. For example, if we would have only two places in the bag ($s = 2$) and only two words ($w = 2$), and 2 individuals then Ω would be:

$$\left\{\begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 2 & 2 \end{bmatrix}\right\}.$$

We will collect the ensemble means of the word frequencies of individuals (i.e. the values of $\frac{x_{im}}{s}$) in a $n \times w$ matrix $\mathbf{F}$, with elements $f_{im} = \frac{1}{s}\sum_{\mathbf{X}\in\Omega} x_{im} P(\mathbf{X})$.

The ensemble means change over time as:

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} &= \frac{1}{s}\sum_{\mathbf{X}\in\Omega} \mathbf{X}\frac{P(\mathbf{X})}{\mathrm{d}t} \\
&= \sum_{\mathbf{X}\in\Omega}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}\left(\sum_{k=1}^{w}\sum_{m=1}^{w} \alpha_m \frac{x_{ik}+1}{s}\frac{x_{jm}}{s}\mathbf{X}P(\mathbf{X}-\mathbf{E}_{i,m}+\mathbf{E}_{i,k}) - \sum_{k=1}^{w}\sum_{m=1}^{w} \alpha_m \frac{x_{ik}}{s}\frac{x_{jm}}{s}\mathbf{X}P(\mathbf{X})\right)
\end{aligned}$$

We will next substitute $\mathbf{Y} = \mathbf{X} - \mathbf{E}_{i,m} + \mathbf{E}_{i,k}$ in the first sum, and denote with Ω' the set of all states the $\mathbf{Y}$ can be in:

$$
\begin{aligned}
\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} &= \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \left(\sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} (\mathbf{Y} + \mathbf{E}_{i,m} - \mathbf{E}_{i,k}) P(\mathbf{Y}) - \sum_{\mathbf{X}\in\Omega} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{X} P(\mathbf{X}) \right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} (\mathbf{E}_{i,m} - \mathbf{E}_{i,k}) P(\mathbf{Y}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,m} P(\mathbf{Y}) - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,k} P(\mathbf{Y}) \\
&= \sum_{i=1}^{n} \sum_{\mathbf{Y}\in\Omega'} \sum_{j=1}^{n} \sum_{m=1}^{w} r_{ij} \alpha_m \frac{y_{jm}}{s} \mathbf{E}_{i,m} P(\mathbf{Y}) \sum_{k=1}^{w} \frac{y_{ik}}{s} - \sum_{\mathbf{Y}\in\Omega'} \sum_{i=1}^{n} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,k} P(\mathbf{Y}) \sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \\
&= \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{R} \cdot \mathbf{Y} \cdot \mathbf{diag}(\boldsymbol{\alpha})) P(\mathbf{Y}) - \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{diag}(\mathbf{R} \cdot \mathbf{Y} \cdot \boldsymbol{\alpha}) \cdot \mathbf{Y}) P(\mathbf{Y}) \\
&= \mathbf{R} \cdot \mathbf{F} \cdot \mathbf{diag}(\boldsymbol{\alpha}) - \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{diag}(\mathbf{R} \cdot \mathbf{Y} \cdot \boldsymbol{\alpha}) \cdot \mathbf{Y}) P(\mathbf{Y})
\end{aligned}
$$

where $\boldsymbol{\alpha}$ is a vector with elements α_i.

The second term contains second order elements, and we can see that if there is selection of certain words the ensemble means do not form a closed system. If, however, all words are internalised with the same rate, as the results in the main paper suggest, and there is no selection between words we have $\alpha_m = \alpha$ for all m, then the ensemble means obey:

$$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} = \alpha \left(\mathbf{R} \cdot \mathbf{F} - \mathbf{diag}(\hat{\mathbf{r}}) \cdot \mathbf{F} \ \right),$$

where $\hat{\mathbf{r}}$ is a vector with elements $\hat{r}_i = \sum_{j=1}^{n} r_{ij}$. The ensemble means now are a closed system and change over time as a linear system of ordinary differential equations. As the results in the main paper suggest that internalisation rate is independent of word frequency, we will from hereon assume that all these rates are the same.

The ensemble mean of an individual's word frequency changes as:

$$\frac{\mathrm{d}f_{im}}{\mathrm{d}t} = \alpha \hat{r}_i \left(\sum_{j=1}^{n} \frac{r_{ij}}{\hat{r}_i} f_{jm} - f_{im} \right)$$

The word frequencies of a user change in response to other users in the network connected to this user, but are independent of the frequencies of other words. It is straightforward to solve this system of ODEs. However, to gain insight we will, rather than providing a general solution, solve a number of cases of special interest.

2. An individual exposed to constant word frequencies

As a first example we will study how an individual adjust its word frequencies to that of its environment. We will therefore assume that an individual communicates with a large number of individuals, and that the word frequencies in the communication received are constant. Let the word frequency of word m that the focal individual i is exposed to be $h_{im} = \sum_{j=1}^{n} \frac{r_{ij}}{\hat{r}_i} f_{jm}$, which is assumed constant. We thus study the following w differential equations for the ensemble means of the word frequencies of the focal individual i:

$$\frac{\mathrm{d}f_{im}}{\mathrm{d}t} = \alpha \hat{r}_i (h_{im} - f_{im})$$

These systems have as equilibrium $f_{im} = h_{im}$, meaning that all word frequencies converge to the word frequencies that they are exposed to. The solution of the differential equations is

$$f_{im}(t) = h_{im} - (h_{im} - f_{im}(0))\, e^{-\alpha \hat{r}_i t}.$$

The word frequency of the words an individual is exposed to depends on the way that the individual is embedded in the social network. What is the effect of community structure on the word frequency if word frequencies transmit between communicators? Assume that individuals communicate predominantly with a subset of individuals which we call a community. By organising the network such that individuals' communication maximises the communication within communities, we can describe the network as a collection of communities. Such communities, it has been shown, tend to have distinct word frequencies (Bryden et al. 2013). Let the set c_k contain all the members of the k^{th} community, and that there are g such communities. The rate with individual i receives messages from members of c_k is $\hat{r}_i^k = \sum_{j \in c_k} r_{ij}$. In the messages that i receives from community c_k the word frequency of word m in the communication received is $h_{im}^k = \sum_{h \in c_k} \frac{r_{hi}}{\hat{r}_i^k} f_{im}$. The total word frequency received can now be partitioned as follows:

$$h_{im} = \sum_{k=1}^{g} \frac{\hat{r}_i^k}{\hat{r}_i} h_{im}^k.$$

The word frequency that a speaker will converge to in constant world is the weighted word frequency of all the communities that (s)he converses with.

If there is a dominant group in an individual's environment, and typically this would be the group that the focal individual is a member of, the word frequency of the dominant community will have most influence on the individual. This is compatible with the observations in Bryden et al. (2013), where it was found that 91% of Twitter conversation was within a community, and that the word frequencies of atypical words is shared within communities. It also is commensurate with the results of Tamburrini et al. (2015) where it is shown that outgoing messages from a community are more similar to the receivers, in terms of word frequencies than internal ones: if some members within a community communicate more with the outside world than others, those individuals will adjust their language more to the outside world then others. These individuals will be overrepresented in the outgoing communication. Outgoing messages from a group should therefore have word frequencies that are more like those of external groups than messages exchanged within the community.

There is one further aspect that this model can show. The word frequencies of an individual will change if the individual changes its environment, for instance, if it changes community. If that is the case there will be a change in the word frequencies received. Say that the new frequency of the word m that i receives is h_{im}. Let the word frequency of word m at the beginning of this process be equal to

$f_{im}(0)$ which is likely to be different from h_{im} because the individual was exposed to a different environment before. We will now describe how far the word frequency is removed from its final value.

The difference between the current word frequency, and the one that will be finally attained is $\Delta f_{im}(t) = h_{im} - f_{im}(t) = \Delta f_{im}(0)e^{-\alpha \hat{r}_i t}$. This leads to the observation that the relative convergence towards the frequency of the received messages is given by

$$\frac{\Delta f_{im}(t)}{\Delta f_{im}(0)} = e^{-\alpha \hat{r}_i t}.$$

What is remarkable is that the right hand side for this equation is independent of m: the relative convergence is the same for all words.

3. Two individuals converge over time, the more they communicate with one another

What if two individuals communicate? They will both change, yet also be subject to the environment that they are embedded in. If the word frequency coming from the environment is assumed to be constant, we have a system of two equations for the word frequencies of word m. At this point we will make our life easy and assume that $\hat{r}_1 = \hat{r}_2 = \hat{r}$, and $r_{21} = r_{21} = r$, for no other reason than that it simplifies the maths and makes the derivation easier to follow. It is not hard to relax this assumption and derive equivalent results. The frequencies of word m change over time as:

$$\begin{aligned}\frac{\mathrm{d}f_{1m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)h'_{1m} - \hat{r}f_{1m} + rf_{2m}\right)\\ \frac{\mathrm{d}f_{2m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)h'_{2m} - \hat{r}f_{2m} + rf_{1m}\right),\end{aligned}$$

where $h'_{im} = \sum_{h=3}^{n} \frac{r_{ih}}{\hat{r}_i - r_{i1} - r_{i2}} f_{hm}$, that is the word frequency of word m received by i from all individuals but individuals 1 and 2. It is further helpful to note that $\sum_{m=1}^{w} h'_{im} = 1$. This follows from the fact that $\sum_{m=1}^{w} f_{im} = 1$, which, in turn, follows from the fact that $\sum_{m=1}^{w} x'_{im} = s$.

To analyse this, we define $w_{1m} = (f_{1m} - f_{2m})/2$ and $w_{2m} = (f_{1m} + f_{2m})/2$. The equation for w_{2m} expresses how the overall frequency in the group community consisting if individuals 1 and 2 behaves. The w_{1m} equation shows how the two individuals converge towards each other. In these new variables the system changes as

$$\begin{aligned}\frac{\mathrm{d}w_{1m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)\frac{h'_{1m} - h'_{2m}}{2} - (\hat{r}+r)w_{1m}\right)\\ \frac{\mathrm{d}w_{2m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)\frac{h'_{1m} + h'_{2m}}{2} - (\hat{r}-r)w_{2m}\right)\end{aligned}$$

Now w_{2m} has an equilibrium at $(h'_{1m} + h'_{2m})/2$, which is simply the average of the frequencies received, and w_{1m} has an equilibrium at $\frac{(\hat{r}-r)}{\hat{r}+r}(h'_{1m} - h'_{2m})/2$. The solutions to these differential equations are:

$$\begin{aligned}\frac{\hat{r}-r}{\hat{r}+r}\frac{h'_{1m} - h'_{2m}}{2} - w_{1m}(t) &= \left(\frac{\hat{r}-r}{\hat{r}+r}\frac{h'_{1m} - h'_{2m}}{2} - w_{1m}(0)\right)e^{-\alpha(\hat{r}+r)t}\\ \frac{h'_{1m} + h'_{2m}}{2} - w_{2m}(t) &= \left(\frac{h'_{1m} + h'_{2m}}{2} - w_{2m}(0)\right)e^{-\alpha(\hat{r}-r)t}\end{aligned}$$

We can now reconstruct the solutions by using $f_{1m} = w_{2m} + w_{1m}$ en $f_{2m} = w_{2m} - w_{1m}$. The equilibria are for $f_{1m} : (\frac{\hat{r}}{\hat{r}+r} h'_{1m} + \frac{r}{\hat{r}+r} h'_{2m})$ and for $f_{2m} : (\frac{r}{\hat{r}+r} h'_{1m} + \frac{\hat{r}}{\hat{r}+r} h'_{2m})$.

The absolute convergence of the speakers is given by:

$$w_{1m}(t) - w_{1m}(0) = \left(\frac{(\hat{r} - r)}{\hat{r} + r} \frac{h'_{1m} - h'_{2m}}{2} - w_{1m}(0) \right) \left(1 - e^{-\alpha(\hat{r}+r)t} \right)$$

The right hand side depends r: the bigger the r the faster you converge, but also the further you have to travel. So the result combines two components: the fact that the convergence is faster, and the fact that the end points are much closer together. Note that the right hand side also depends on m: this measure is different for different words.

One way of quantify how much two users differ in their language is through the use of appropriate measures. A widely used measure to assess similarity is the Bray-Curtis similarity measure. The Bray-Curtis similarity, for two identical sized bag of words is the sum over the lowest frequency frequency found in the bags. We can calculate this as

$$\sum_{m=1}^{w} \min(f_{1m}, f_{2m}) = \sum_{m=1}^{n} \frac{f_{1m} + f_{2m} - |f_{1m} - f_{2m}|}{2} = \sum_{m=1}^{n} w_{2m} - |w_{1m}| = 1 - \sum_{m=1}^{n} |w_{1m}|.$$

The last step used $\sum_{m=1}^{n} w_{2m} = \sum_{m=1}^{n} \frac{f_{1m}+f_{2m}}{2} = 1$. The increase in the similarity measure over a period t is given by $\sum_{m=1}^{w} -(|w_{1m}(t)| - |w_{1m}(0)|)$, which is

$$\sum_{m=1}^{w} |w_{1m}(0)| - |w_{1m}(t)| =$$
$$\sum_{m=1}^{w} |w_{1m}(0)| - \sum_{m=1}^{w} \left| w_{1m}(0) e^{-\alpha(\hat{r}+r)t} + \frac{(\hat{r} - r)}{\hat{r} + r} \frac{h'_{1m} - h'_{2m}}{2} \left(1 - e^{-\alpha(\hat{r}+r)t} \right) \right|.$$

To proceed we next assume that the sum of all words received, $\hat{r}$, exceeds the words received from the conversation partner. If $\hat{r} \gg r$ then the above expression simplifies to

$$\sum_{m=1}^{w} |w_{1m}(0)| - \sum_{m=1}^{w} \left| w_{1m}(0) e^{-\alpha(\hat{r}+r)t} + \frac{h'_{1m} - h'_{2m}}{2} \left(1 - e^{-\alpha(\hat{r}+r)t} \right) \right|.$$

If we now define c_2 as the derivative of the above expression with respect to $e^{-\alpha r t}$:

$$c_2 = -e^{-\alpha \hat{r} t} \sum_{m=1}^{w} \left(w_{1m}(0) - \frac{h'_{1m} - h'_{2m}}{2} \right) \operatorname{sign} \left(w_{1m}(0) e^{-\alpha(\hat{r}+r)t} + \frac{h'_{1m} - h'_{2m}}{2} \left(1 - e^{-\alpha(\hat{r}+r)t} \right) \right).$$

If $w_{1,j}(0)$ and $m'_{1j} - m'_{2j}$ have the same sign this is a constant that does not depend on r. If the difference within most pairs is not too far from the point to which they will be eventually converge, then we can assume c_2 to be approximately constant. By defining

$$c_1 = \sum_{m=1}^{w} |w_{1m}(0)| - |\frac{h'_{1m} - h'_{2m}}{2}|$$

we now approximate the change in the Bray-Curtis similarity by

$$\sum_{m=1}^{w} |w_{1m}(0)| - |w_{1m}(t)| \approx c_1 + c_2 e^{-\alpha r t}. \tag{S1}$$

4. Stochastic process for measuring incorporation rate

In our formalisation, we model each person as having an internal collection of words which are updated through communication with other people. The model is a Moran process where, with probability α, an encountered word (from another person) replaces a random word already in the internal collection. On Twitter we have data over a period of time of which words are received by a focal user and which words they broadcast. Consequently, we should be able to infer the incoporation rate α by modelling this internal collection as a hidden variable.

4.1. **Method.** The user's internal collection is defined as a collection of size s words. Many words are in the collection more than once with many repetitions of the more common words. For a word, we maintain a probability distribution $p(i)$ $(0 < p < 1)$ that an individual has i representations of the word within their internal collection. Over time, we update the internal distribution according to words messaged to the *Focal User* from a second *Incoming User* and a proposed value of α. Following the logic of the Numerically Integrated State Space method (NISS - de Valpine and Hastings 2002), we use a procedure that generates a likelihood of the internal distribution given the outgoing messages of a user (see Fig. S1). Consequently, this method is able to calculate the extent to which the word usage of the Incoming User has influenced the word usage of the Focal User.

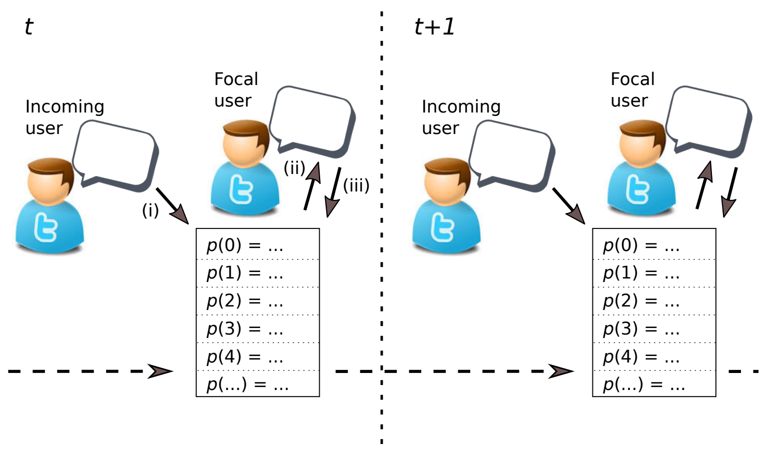

Figure S1. Diagram outlining our method for measuring the incorporation rate. A probability distribution is maintained for a Focal User's usage frequency of a word. (i) The distribution is updated according to input from an Incoming User's usage. (ii) A likelihood is calculated using the probability distribution and the output of the user. (iii) The probability distribution is updated according to the likelihood as specified by NISS.

The internal model is updated using incoming language through counting the number of mentions of the word (k) in the total number of N words in tweets directed at the user from the Incoming User. Because these words are batched together, and a sampled of the input to the incoming user, we assume the focal individual internalises the words at rate αk. Consquently, the probability of n copies of the target word being internalised by the user is $Poiss(n, \alpha k)$, where $Poiss$ is the Poisson probability density function[1]. For each word internalised, we update the internal distribution by setting $p(i+1) = p(i)(1-i/s)$ for $0 < i < s-1$. We subtract i/s because this is the probability that we have replaced the same word. The frequency of the first word internalised $\left(\sum_{n\geq 1} Poiss(n, \alpha k)\right)$, and, in general, of the mth word internalised $\left(\sum_{n\geq m} Poiss(n, \alpha k)\right)$. Put together, for each word internalised, we update the internal distribution as follows,

$$\begin{aligned}
&\forall i \in \mathbb{Z}, 0 < i < s \\
&\forall m \in \mathbb{Z}, 1 < m < s \\
p(i) =& p(i) + \sum_{n\geq m} Poiss(n, \alpha k) p(i-1)(1-(i-1)/s) \\
&- \sum_{n\geq m} Poiss(n, \alpha k) p(i)(1-i/s)\ ,
\end{aligned}$$

taking $p(-1) = 0$ to resolve the boundary. It is computationally too expensive to go through every different possibility of the order in which the N words could be incorporated, so we incorporate the target words first and then the $N-k$ non-target words. If anything, this would penalise α. We use a similar procedure to update the internal model for these non-target words. This time the probability of hitting a non-target word is $(1-i/s)$ so we set $p(i) = p(i+1)((i+1)/s)$ for $0 \leq i < s$.

$$\begin{aligned}
&\forall i \in \mathbb{Z}, 0 \leq i < s \\
&\forall m \in \mathbb{Z}, m < i < s \\
p(i) =& p(i) + \sum_{n\geq m} Poiss(m, \alpha k) p(i+1)(i+1)/s \\
&- \sum_{n\geq m} Poiss(m, \alpha k) p(i) i/s\ .
\end{aligned} \tag{S2}$$

Given the internal model of a user's propensity to use a word, we are able to generate the probability of the language they generate at a point in time t. For a user who has used k mentions of a specific word in N words, we calculate the probability of the internal model for the word as,

$$Pr_t(k, N) = \sum_i p_t(i) B(k; N, i/s), \tag{S3}$$

where B is the binomial probability density function. According to the NISS mathod, the internal model can now be updated to reflect this new information for the next time point.

$$\forall i, p_{t+1}(i) = \frac{p_t(i) B(k; N, i/s)}{\sum_i p_t(i) B(k; N, i/s)}\ . \tag{S4}$$

[1]An alternative to this is that the words are internalised with a Binomial distribution $Binom(n; N, \alpha k)$. Tests with this distribution found it gave similar results.

We can now calculate the likelihood of the model and parameters (H) given the data (D) using Equation (S3) as,

$$\mathcal{L}(\mathcal{H}|\mathcal{D}) = \prod_t Pr_t(k, N) \ . \tag{S5}$$

4.2. **Tests.** The method was tested by randomly generating pairs of two users, each member of the pair assigned with different internal frequencies of a specific word. At each time segment, both users output a random number of words with the frequency of the specific word picked from a binomial distribution according to their internal frequency of the word. One of the users is the focal user, who receives languague from the other, and updates their internal model according to a predefined value of α, dubbed α_{in}. We then run the method against the language generated by the two users to find the value of α (dubbed α_{out}) with a maximum likelihood (Eq. S5) and compare that against the predefined value - see Fig. S2.

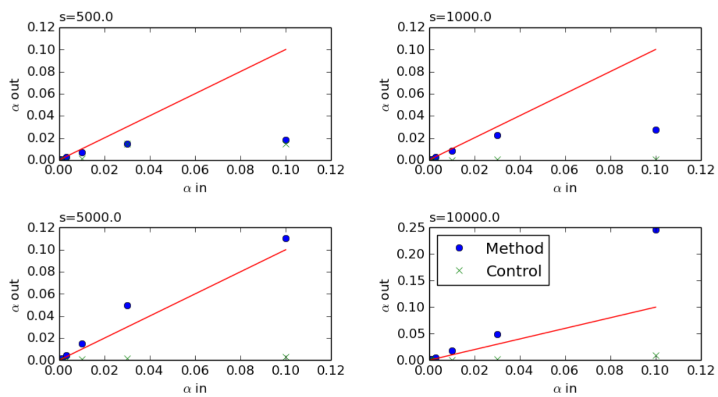

FIGURE S2. Four different values of s were tested. At low s, and higher levels of α, the focal user quickly adopted the other's usage. At the highest level of $s = 10,000$, the value of α was over estimated.

4.3. **Implementation.** We looked at pairs of users, a *target* user for whom we maintain an internal collection, and an *incoming* user from whom we monitor their incoming tweets. To some extent, two users will mirror one another in conversation, and so we would expect to see some transitory usage of the same words. We can account for this by ignoring tweets where a focal user sent messages directed back to the incoming user. We then looked at time segments where both the outgoing and the incoming user posted. We used a time segment of one month.

The initial condition of the internal collection was set at a binomial distribution, the mean of which is calculated by the relative frequency for the specific word over the first half of the time segments (multiplyed by s). There was then a burn in period for half the time segments where α was set to 0.0. At this point the internal frequency should reflect that of the usage of the target user. We then introduce update of the internal model from incoming language (Eq. S2). A control was generated where the outgoing tweets of the focal user were randomly shuffled so that the time signal was lost.

We focussed on 1,000 words randomly sampled from all the word instances (so including copies of words) we had in our complete sample of Twitter. We looked at 10,000 pairs of users, where both the users had tweeted at least one tweet to each other and at least 500 conversational tweets to all users. For a given value of α we calculated the log-likelihood by summing the log-likelihoods for each pair of users. For each word, we then used Nelder-Mead optimisation (Nelder and Mead 1965) to find the value of α with the maximum likelihood.

We found, when running optimisations, that increasing the level of s in the process increased the maximum-likelihood level of α. The reason for this is due to the difference in ways in which we update the probability distribution according to whether we update it from output from the focal user, or input from the incoming user. When we update the internal model according to output from the focal user (see Eq. S4), the amount of change is independent of the value of s. However, when we update the internal model due to incoming language (Eq. S2), larger values of s will mean the internal frequencies of words are updated more slowly. There is an increase in the value of α needed to compensate for this effect. To compensate for this, we generated a control by randomly shuffling the time signal of the incoming user. This shuffling will remove the influence the incoming user's language might have over the focal user. Our final value of α was calculated by subtracting the value of α generated for the shuffled control from the value of α generated with the unshuffled time signal. From our test runs, and due to computational time constraints, we settled on a value of $s = 2,000$ for the optimisation runs presented in the main manuscript.

5. References

- Blythe, R. A. (2012). Neutral evolution: a null model for language dynamics. Advances in Complex Systems, 15(03n04), 1150015.
- Blythe, R. A., & McKane, A. J. (2007). Stochastic models of evolution in genetics, ecology and linguistics. Journal of Statistical Mechanics: Theory and Experiment, 2007(07), P07018.
- Bryden, J., Funk, S., & Jansen, V.A.A. (2013). Word usage mirrors community structure in the online social network Twitter. EPJ Data Science, 2:3, 1-9.doi:10.1140/epjds15.
- de Valpine P. & Hastings A. (2002). Fitting population models incorporating process noise and observation error. Ecological Monographs, 72:5776.
- Nelder J.A & Mead R. (1965). Simplex Method for Function Minimization. The Computer Journal, 7:30813.
- Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. Social Networks, 40:84-89.