How humans transmit language: Horizontal transmission matches word frequencies amongst peers on Twitter

By John Bryden[1,*], Shaun P. Wright[1], Vincent A. A. Jansen[1,*]

[1] *School of Biological Sciences, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom*

1. Abstract

Language transmission, the passing on of language features such as words between people, is the process of inheritance that underlies linguistic evolution. To understand how language transmission works we need a mechanistic understanding based on empirical evidence of lasting change of language usage. Here, we analysed 200 million online conversations to investigate transmission between individuals. We find that frequency of word usage is inherited over conversations, not just the binary presence or absence of a word in a person's lexicon. We propose a mechanism for transmission whereby for each word encountered there is a chance that it will be used more often. Using this mechanism, we measure that one word in around every hundred someone encounters will be used more often. Since more commonly used words are encountered more often, this means that it is the frequencies of words which are copied. Beyond this, our measurements indicate that this per-encounter mechanism is neutral and applies without any further distinction as to whether

* Corresponding authors John Bryden (john.bryden@rhul.ac.uk) and Vincent Jansen (vincent.jansen@rhul.ac.uk).

a word encountered in a conversation is commonly used or not. An important consequence of this is that frequencies of many words can be used in concert to observe and measure language transmission, and our results confirm this. These results indicate that our mechanism for transmission can be used to study language patterns and evolution within populations.

2. Keywords

Language transmission; Linguistic evolution; Evolution of language; Mathematical model; Moran process; Horizontal transmission; Iterated transmission; Word-heritability.

3. Introduction

Language use is constantly in flux and language evolution can happen at many spatial and temporal scales. Historical evidence shows how population groups experience wholesale changes in word usage and language syntax across many generations (Bloomfield 1933; Dunn *et al.* 2005; Lieberman *et al.* 2007; Gray *et al.* 2009; Pagel 2009). A broad theoretical background has been developed which explains how these large-scale and dynamic language patterns can be generated by language change at the individual level (Dunn *et al.* 2005; Lieberman *et al.* 2007; Gray *et al.* 2009; Pagel 2009; Nowak *et al.* 2001, 2002; Steels & Kaplan 2002; Castellano *et al.* 2009; Chater & Christiansen 2010; Kirby *et al.* 2014; Eisenstein *et al.* 2014). These studies assume that language elements are repeatedly transmitted between individuals in a population, and then use mathematical models or computer simulations to show that a macroscopic language pattern is generated from iterations of this individual behaviour. This makes it plausible that macroscopic changes follow from an accumulation of individual transmission events. However, these are 'plausibility arguments' (Castellano *et al.* 2009) and most theoretical efforts to explain language evolution suffer from not having been confronted with data, and are often unverifiable

(Hauser *et al.* 2014). The origins and mechanism of the evolution of language – arguably the most distinctive form of human behavior – remain a mystery.

Darwin already noted the similarity between biological and linguistic evolution (Darwin 1883). This similarity inspired Labov (Labov 2001, 2010) in explaining linguistic change. Whereas the similarity in homology of descent between the two processes is similar, in biological evolution the mechanism of descent is the transmission of genetic material. The mechanism of linguistic change is much harder to pinpoint. Of course children acquire their first language from parents or caretakers, but in a later phase children's language use diverges from that of their original, and adults change their language use, indicating transmission of language elements between speakers (Labov 2001, 2010). It has been posited that words transmit like alleles (Reali & Griffiths 2010), but evidence for this hypothesis has so far been scarce.

At an individual level, we adopt elements of our language throughout our lives. As children we acquire the majority of our language from our parents, but as we grow older we increasingly pick up language from our peers (Bloomfield 1933; Labov 2001, 2010). This form of cultural transmission between peers is called horizontal transmission (Cavalli-Sforza & Feldman 1981). While language acquisition early in life (known as vertical transmission) can be easily observed, the effect of horizontal transmission later on is more subtle and more difficult to detect. It has been known for several decades that word-usage patterns, as well as other linguistic variables, are imitated between interlocutors (Brennan 1996; Pickering & Garrod 2004; Gallois *et al.* 2005; Danescu-Niculescu-Mizil *et al.* 2011; Tamburrini *et al.* 2015). This imitation can be transient or reflective. This is due to people mirroring language while conversing or talking about similar conversation topics. To look for lasting changes we need to look for iterated transmission where people adopt words and use them in other conversations, which has been observed under laboratory conditions (Kirby *et al.* 2014). How language elements transmit in a lasting way between peers in natural situations is hard to measure, in part because there is a weak effect per conversation.

A possible clue to the mechanism of language element transmission lies in the observation that speakers often demonstrate probability matching: if different variants of a word or phoneme exist in a population, learners tend to match the frequency of these variants in their language use (Labov 2010). This indicates that the process of transmission does not just involve the adding of words to a lexicon, but the frequency with which these words are used is somehow stored and internalized.

Here, we will provide evidence of horizontal language element transmission. Our method detects lasting changes in language due to conversations between online individuals. However, to eliminate transient effects that can happen within conversations, we detect transmission by looking for changes in language sent to third parties which were not involved in the original conversations. To detect this weak signal, we need to use a large corpus of online conversations. The transmission of language elements is often assumed to be analogous to the spread of genetic traits (Pagel 2009). We therefore use techniques from the toolbox developed within evolutionary biology, on the interface between population genetics and linguistics (Cavalli-Sforza & Feldman 1981; Wang 1976). We study horizontal language transmission by investigating the change in the use of words following exposure to the language of other people. This assumes that, beyond simply having a lexicon, we have some internal language representation which influences which words we choose and how often we use them (Wang 1976). We cannot directly observe this representation, but we can infer it from word usage frequencies in a person's outgoing communication (Pagel *et al.* 2007; Labov 2010; Michel *et al.* 2011; Newberry *et al.* 2017). We will show here how it is possible to identify a change in the representation over time and then show, using advanced statistical methods, that this change happens due to conversations with another individual.

We will use a simple model for the internal representation of language which incorporates transmission of language between individuals. Because our aim is to study how word frequencies change, this highly simplified internal representation does not place any specific importance on grammar, syntax or word order. We simply treat communication as a multiset or a 'bag of words' (Salton & McGill 1983): how often a person uses a word is reflected by the number of copies of the word in their bag. Word instances received from conversation partners can occasionally replace other words in the bag, changing the internal representation and allowing the frequency of stored words to change in response to conversation (see figure 1). This model forms a Moran process and can be analysed using well understood techniques (Blythe 2012). Our analysis of the model (see electronic supplementary material) shows how the word frequencies used will equilibrate over time towards the frequencies received from conversation partners in a way that is very similar to osmosis (figure 1). The model predicts that an individual's word-usage patterns change through conversations with others and that this change will manifest itself in the word frequencies that the individual then uses to other people. Although in this model language changes in response to all language received, the effect of a conversation with a particular conversation partner will leave its mark, even if this conversation is only a relatively small part of all their conversations.

4. Results

We first show that word frequencies used by an individual change in response to the language used by a conversation partner, as predicted by our model. We studied a data set of conversations formed from a sample of 200 million messages sent publicly between users of the Twitter web site (Bryden *et al.* 2013) (see Methods). To eliminate any transient imitation that others have found in online communication (Danescu-Niculescu-Mizil *et al.* 2011; Tamburrini *et al.* 2015), we excluded any mutually directed messages between a pair being studied in our analysis. Motivated by the result from our model that the difference between users is important, we looked at the influence that the difference between a focal user and their partner's early usage of a word has on any later change of the focal user's usage of the word.

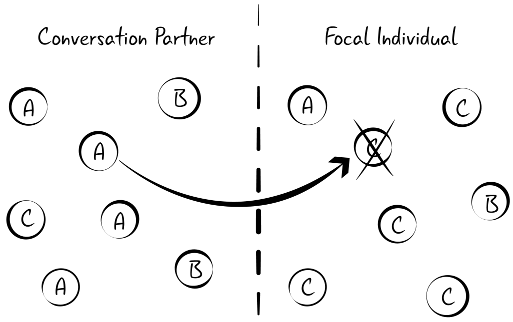

Figure 1. **An osmosis-like process for horizontal language transmission used in our model.** The two halves of the diagram show the internal language representations of two individuals as bags of words. The figure shows how an individual in our framework copies and stores a word from their conversation partner; an instance of word A is incorporated, replacing an instance of word C. The number of instances of a particular word defines how likely someone is to use the word in a given situation. In our model of this process, each bag contains s words; user i sends a word to user j at a rate r_{ij} and the recipient replaces a randomly chosen word in their bag with a received word with incorporation rate α. Since the likelihood of a word being replaced depends on its frequency in the bag, word frequencies change similarly to osmosis in that over time the frequencies of words in both halves will tend to equilibrate.

Since this is mathematically related to the heritability of genetic traits (Falconer & Mackay 1995) we dub this *word-heritability*. Over the 1,000 words tested (see Methods), we found that mean word-heritability was significantly greater for pairs of users that had sent each other messages than for control-pairs that had not (see figure 2). This indicates that an individual changes their word-usage toward that used by their conversation partner.

Within our model, when a focal individual encounters word instances used by another individual, a proportion of these incoming word instances will be incorporated replacing word instances within the focal individual's internal representation. We dub the proportion of word instances incorporated as the *incorporation rate* (α), and have developed a method to measure this rate. To do this, we implemented the model as a stochastic process. Focussing on an individual's usage of a word, we maintain a probability distribution of the word's frequency in the bag of words. We update this distribution with input received by the user according to the incorporation rate α, and then maximise the likelihood of produced frequencies of the word with respect to α and the input received (see electronic supplementary material for precise details).

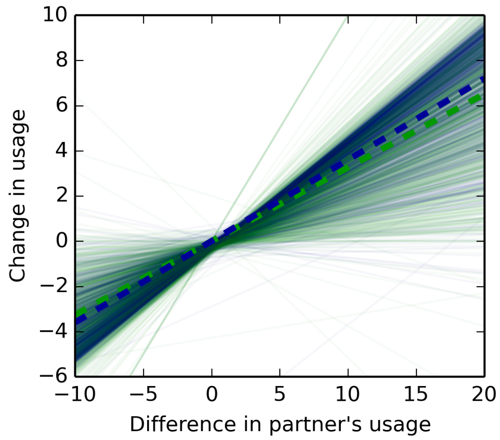

Figure 2. **Word heritability between conversing partners is greater than that for non-conversing partners.** For each test word, we plot regressions (see Methods) for data from conversing partners (blue solid lines) and non-conversing partners (green solid lines). The regression lines were superimposed by translucently plotting lines for each regression, interleaving between the two data sets. We found relatively high levels of word-heritability in non-conversing partners due to word-usage changing at population levels. A Mann-Witney U test indicated that the slopes for conversing partners tend to be steeper than those of non-conversing partners ($p_{\mathrm{MW}} < 9.5 \times 10^{-10}$). The two dashed lines (same colours) are slopes regressed over data collected for all of the words, the difference between these values was $W = 0.0340$ which is a measurement of word-heritability due to Twitter conversations. We tested that $W > 0$ using a bootstrap ($p_{\mathrm{B}} < 0.001$, see Methods).

We tested 1,000 different words (see Methods) and found the most likely value of α for each word.

It is important to find out if the incorporation rate of a word is dependent in any way on the frequency of usage of a word (Church 2000). If the relationship is neutral, then studies of language change can make measurements over many words in concert. Given the heavy tailed distributions of word usage characterised by Zipf's Law, one might expect that instances of more commonly used words are more likely to be incorporated than those less commonly used. Interestingly, we found that the rate of a word instance being taken up in our model is independent of word frequency across a wide range of word frequencies (see figure 3). This indicates that we are as likely to adopt an instance of a frequent word as much as we are to adopt an instance of an infrequent (and therefore conversation specific) word. This suggests that we have found a perspective whereby word transmission is a neutral process; a view consistent with some models that generate the heavy tailed distributions of word frequencies predicted by Zipf's Law (Blythe 2012).

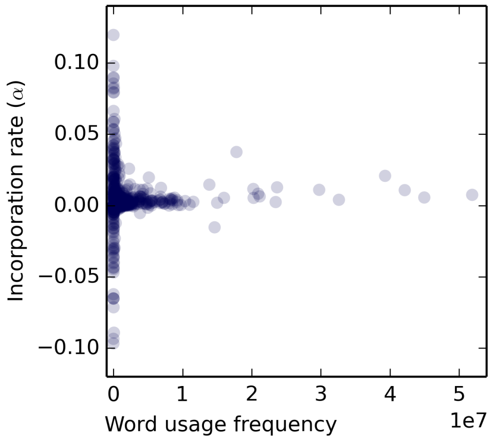

Figure 3. **The rates with which words are incorporated is independent of usage frequency.** Each circle is a word's incorporation rate (circles have translucency of 30%). Linear regression finds no correlation between the word's usage frequency (over the whole sample) and the incorporation rate (two-tailed Pearsons': $r^2 = 0.00040$, $p = 0.54$). The mean value of the word incorporation rate α is 0.0043, which we found significantly greater than zero ($p = 0.0083$, bootstrapping with 10,000 resamples of 100 values, and calculating the proportion of resamples with mean greater than zero). The high variance for very low frequencies is due to sampling effects.

Our finding that the incorporation rate of a word is not dependent on the word's usage frequency means that we can study transmission of many words in concert. We can therefore investigate the prediction, by our model (Equation 1 in Section 3 of the electronic supplementary material) and others which use a Moran process (Blythe & McKane 2007), that the frequencies of usage of two communicating individuals will converge exponentially over time. We did this by investigating if the Bray-Curtis similarity (Bray & Curtis 1957) of pairs of users increases over time according to the number of messages sent between the two users. We found a highly significant, positive correlation between the change in the proportion of word instances shared between two users and the number of messages sent between them; as well as a close quantitative fit with our model (see Experimental Procedures) and the data (figure 4). We tested our transmission model against a null model ($\alpha = 0$) using the Akaike

Information Criteria finding essentially no support for the null model compared with the transmission model (Burnham & Anderson 2002, see Methods). The value of the word incorporation rate, α, found was 0.01, a similar order of magnitude to the mean incorporation rate found in figure 3. These measurements indicate that we subconsciously incorporate approximately one in every 100-200 words that we experience.

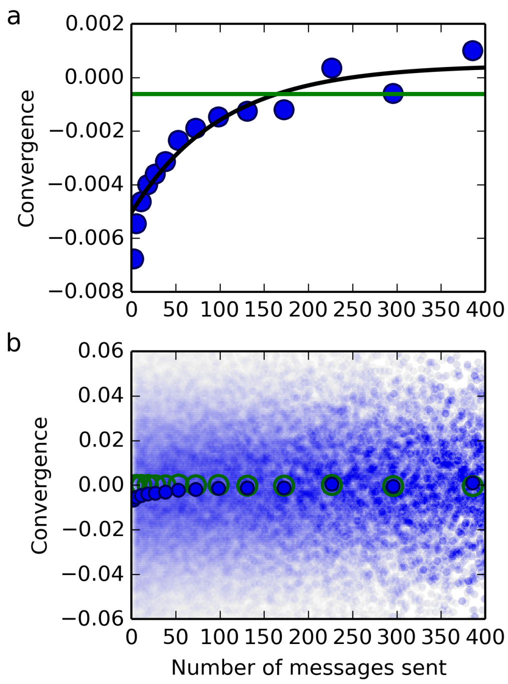

Figure 4. **The more messages were sent between two users, the more their language converged.** (a) Plot of the means of bins of conversation pairs (binned along the x-axis showing x, y means of each bin) and fitted models (black line is tranmission model, green dashed line is null model, see Methods). The fitted line of our model crosses zero at approximately 310 messages sent. (b) Illustratrion of the large variance in the data (unbordered translucent circles which are superimposed). The convergence of 500 conversation pairs (sampled with replacement) are plotted per bin on the x-axis (bordered blue circles). Control values are also shown (bordered green circles).

5. Discussion

Our results demonstrate that humans adopt lasting changes in their language usage upon conversation. These changes are consistent with the existence of an internal representation of word frequencies, where words are incorporated in a Moran process. We found that the per-encounter rate at which words are incorporated is independent of how commonly the word is used. We also found that this per-encouter rate is greater than zero, rejecting the null-model where the per-encounter rate is equal to zero. This means that we have developed a method whereby transmission can be detected and measured on changes of individual word frequencies, or many words in concert. Put together, this means that the more two individuals converse, the more they will use similar language outside their conversations. A corollary of this is that the word usage of two isolated, or weakly connected groups, will drift apart on this time scale.

The use of large quantities of data, gleaned from online conversations, allows us to detect evidence for an underlying process of language transmission. Through identifying this process, we fill a gap in our understanding of how language is shaped and evolves (Croft 2000; Pagel *et al.* 2007; Pagel 2009; Chater & Christiansen 2010; Hauser *et al.* 2014). We demonstrate a process which has subtle effects at the individual level (see figure 4b). However, when this process is iterated many times within a population, large scale social patterns can develop. For instance, it follows from our results that groups which interact more with one another will share similar and distinctive language patterns; which is borne out by evidence from online conversations (Bryden *et al.* 2013). The relatively high level of word-heritability amongst non-conversing partners (see figure 2) indicates that iterated transmission happens at a large scale in populations, which may explain increased regularisation of language found amongst larger populations (Kam & Newport 2005; Lupyan & Dale 2010; Dale & Lupyan 2012) while smaller populations are the most susceptible to language change (Trudgill 2005, 2011). Furthermore, our model and methods can be applied to date changing language usage of groups. This can make inferences as to dynamical changes in population structure and where possible linking these changes to genetic changes, especially regarding whether groups have become more integrated or more isolated, and make future predictions (Barbujani *et al.* 1994; Hunley & Long 2005; Hunley *et al.* 2007; Lieberman *et al.* 2007; Hunley *et al.* 2008; Kutanan *et al.* 2014; Longobardi *et al.* 2015; Srithawong *et al.* 2015; Creanza *et al.* 2015; Karafet *et al.* 2016).

The process of transmission demonstrated here, being peer-to-peer in nature, forms a basis for horizontal transmission (Cavalli-Sforza & Feldman 1981). Indeed, our results reject a model that human language use can solely be explained by vertical transmission as we have shown that horizontal transmission does take place. Furthermore, the mechanism of lasting transmission we have identified can go beyond horizontal transmission and may underlie vertical transmission whereby children acquire vocabularies from their parents, and oblique transmission whereby children acquire vocabularies from older generations. From this perspective, we propose that vertical transmission can work in much the same way as horizontal transmission but with an inequality between parents and children whereby parents are much less likely to pick up words from their children than *vice versa*. With an understanding of both forms of transmission, the model and evidence that we have presented can be applied to understand how word frequencies can change across several generations of a population.

Language transmission is a cognitive process with an underlying neurological mechanism. Our evidence that word frequencies are transmitted from person-to-person points to insights which can inform neuroscience about the sorts of brain structures, mechanisms and memory that are necessary for language uptake and storage, and may be awaiting discovery. For example, an internal, mutable representation of word frequency suggests a reinforcement process and directs neuroscientists towards plasticity theories; a conclusion supported by various studies showing a role for plasticity and/or Hebbian learning in language therapy (Sarasso *et al.* 2014), acquisition (Kim *et al.* 1997) and processing (Chee *et al.* 2002; Wennekers *et al.* 2006).

There are no genes for words, or other specific language features, yet languages change in a way that is very reminiscent of biological evolution. This suggests that there is something which is inherited and which is passed on like a gene, even if we do not know what this something is. Here we show how word frequencies can be stored and passed on. This forms a quantifiable basis for studying descent with modification of language: a requirement for language evolution.

6. Methods

(*a*) *Data acquisition*

We used conversations between users recorded on the social networking site Twitter. Online conversations on social networks allow the observation of natural, everyday language within its social context in a way that more formal, written media does not. The informal style of this language, and its short, back-and-forth nature, makes it much closer in form and appearance to spoken language than most other forms of written language. Communication on Twitter replicates the heterogeneity in usage that is found in spoken language (Eisenstein *et al.* 2014; Tamburrini *et al.* 2015; Bryden *et al.* 2013). The ubiquity of the use of online social media for human interaction allows the gathering of this data at a large scale and in quantities that are not normally achieved for spoken language. While there are likely to be differences, Twitter conversations are more like regular conversations than other, written forms of communication.

The data were recorded from the Twitter website during December 2009. A snowball sampling process was used to gather users as follows: for each user sampled, all their tweets that mentioned other users (using the '@' symbol) were collected directly from their profiles, meaning that we hopefully had a full history of their tweets. Any newly referenced users were added to a list of users from which the next user to be sampled was picked. Starting from a random user, conversational tweets (time-stamped between January 2007 and November 2009) were sampled, yielding over 200 million messages from over 189,000 users. We ignored messages that were copies of other messages (so called retweets, which are identified by a case-insensitive search for the text 'RT').

(*b*) *Test words*

The following tests were done using a list of 1,000 different test words. These words were selected randomly from the complete collection of all text in the sample.

(*c*) *Word heritability analysis*

Messages were temporally split into 'early' and 'late' halves around the median time. An 'early sample' was created by randomly sampling 1,000 words from the amalgamated early tweets. This was repeated with the amalgamated late tweets to create a 'late sample'.

Word heritability was measured by regressing over a series of points: each calculated on the basis of a single given word, and a randomly shuffled pair of users. For the first axis of the regression, we recorded the difference of the first user's usage of the word compared with that of the other user during the two early halves. For the second axis, we recorded the amount which the first user changed their usage of that word over time between their early and late halves. Two regressions were plotted for each word: one for conversing partners and one for non-conversing partners.

To test for significance, we did a bootstrap by generating two resamples of 500K points from the conversing and non-conversing data sets and regressed a line through each sample. We then measured the difference between the two slopes and recorded the proportion (reported in the main text as p) of the 1,000 bootstrap resamples for which the slope for non-conversing individuals exceeded the slope for conversing individuals. To test that we had used enough resample points, we confirmed that similar results could be achieved with smaller resample sizes.

In all we recorded approximately 500 million data points between conversing partners. To generate controls, we randomly generated pairs of users and checked that they had never sent one-another messages in our data set. We used 9 million pairs for our control which was sufficient to capture its distribution for our bootstrap and for the Mann-Whitney U test.

(*d*) *Convergence analysis*

The convergence analysis required a method that calculates how similar the language is of a pair of users. The Bray-Curtis similarity measure (Bray & Curtis 1957) was used because it takes frequency into account rather than simply binary presence/absence. Words are converted to lower case and stripped of punctuation (see Wright 2017, for more information). We divided each of the two users' language into early and late time periods and sampled 1,000 words (with replacement) from each time period. To measure convergence data points, we calculated the Bray-Curtis similarity between the samples from the late time periods and subtracted the Bray-Curtis similarity between the samples from the early time periods. For the control data points we took the early and late samples from the complete time period without division.

The transmission model fitted to the convergence data points was Eq. (1) from the Supplemental Material:

$$y = c_1 + c_2 e^{-\alpha x}$$

The null model was with $\alpha = 0$ which was simply

$$y = c_3 \ .$$

Fitting was done against the points sampled for display in figure 4 using a least squares method. The values found were: $c_1 = 0.000478$, $c_2 = -0.00552$, $\alpha = 0.00982$

and $c_3 = -0.000617$. The Akaike Information Criteria was calculated as:

$$AIC = 2k - 2\sum_i \ln\left[pdf\_norm(y_i; \mu_i, \sigma^2)\right]$$

where k is the number of parameters in the model, y_i are the model predictions and μ_i are the corresponding data points, σ^2 is the variance of the data points and pdf_norm is the probability distribution function of the normal distribution. We found $AIC_{transmission} = 1535774$ and $AIC_{null} = 1536263$ which means there is essentially no support for the null model in light of the transmission model.

7. Author Contributions

All of the authors contributed equally to the work.

References

Barbujani, G., Whitehead, G. N., Bertorelle, G. & Nasidze, I. S. 1994 Testing Hypotheses on Processes of Genetic and Linguistic Change in the Caucasus. *Hum. Biol.*, **66**(5), 843–864.

Bloomfield, L. 1933 *Language.* University of Chicago Press.

Blythe, R. A. 2012 Neutral evolution: A null model for language dynamics. *Advances in Complex Systems*, **15**(03n04), 1150 015.

Blythe, R. A. & McKane, A. J. 2007 Stochastic Models of Evolution in Genetics, Ecology and Linguistics. *J. Stat. Mech.*, **2007**(07), P07 018–P07 018.

Bray, J. R. & Curtis, J. T. 1957 An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, **27**(4), 325–349.

Brennan, S. E. 1996 Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, **96**, 41–44.

Bryden, J., Funk, S. & Jansen, V. A. A. 2013 Word usage mirrors community structure in the online social network Twitter. *E.P.J. Data Science*, **2**(3), 3.

Burnham, K. P. & Anderson, D. R. 2002 *Model selection and multimodel inference: a practical information-theoretic approach.* Springer.

Castellano, C., Fortunato, S. & Loreto, V. 2009 Statistical physics of social dynamics. *Rev. Mod. Phys.*, **81**(2), 591–646.

Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural transmission and evolution: a quantitative approach.* 16. Princeton University Press.

Chater, N. & Christiansen, M. H. 2010 Language Acquisition Meets Language Evolution. *Cogn. Sci.*, **34**(7), 1131–1157.

Chee, M. W., Hon, N. H., Caplan, D., Lee, H. L. & Goh, J. 2002 Frequency of concrete words modulates prefrontal activation during semantic judgments. *Neuroimage*, **16**(1), 259–268.

Church, K. W. 2000 Empirical Estimates Of Adaptation: The Chance Of Two Noriegas Is Closer To P/2 Than P2. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 180–186. Association for Computational Linguistics.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W. & Ramachandran, S. 2015 A comparison of worldwide phonemic and genetic variation in human populations. *Proc. Natl. Acad. Sci. U.S.A.*, **112**(5), 1265–1272.

Croft, W. 2000 *Explaining Language Change : an Evolutionary Approach.* Pearson Education.

Dale, R. & Lupyan, G. 2012 Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Adv. Complex Syst.*, **15**(03n04), 1150 017.

Danescu-Niculescu-Mizil, C., Gamon, M. & Dumais, S. 2011 Mark My Words!: Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 745–754. New York, NY, USA: ACM.

Darwin, C. 1883 *The descent of man and seletion in relation to sex*. John Murray.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science*, **309**(5743), 2072–2075.

Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. 2014 Diffusion of Lexical Change in Social Media. *PLoS ONE*, **9**(11), e113 114.

Falconer, D. S. & Mackay, T. F. C. 1995 *Introduction to Quantitative Genetics*. Longman, 4th edn.

Gallois, C., Ogay, T. & Giles, H. 2005 Communication accommodation theory: A look back and a look ahead. In *Theorizing About Intercultural Communication*. (ed. W. B. Gudykunst), pp. 121–148. Thousand Oaks, CA: Sage.

Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, **323**(5913), 479–483.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N. & Lewontin, R. C. 2014 The mystery of language evolution. *Front. Psychol.*, **5**.

Hunley, K., Cabana, G., Merriwether, D. & Long, J. 2007 A formal test of linguistic and genetic coevolution in native Central and South America. *Am. J. Phys. Anthropol.*, **132**(4), 622–631.

Hunley, K., Dunn, M., Lindström, E., Reesink, G., Terrill, A., Healy, M. E., Koki, G., Friedlaender, F. R. & Friedlaender, J. S. 2008 Genetic and Linguistic Coevolution in Northern Island Melanesia. *PLOS Genet.*, **4**(10), e1000 239.

Hunley, K. & Long, J. C. 2005 Gene flow across linguistic boundaries in Native North American populations. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(5), 1312–1317.

Kam, C. L. H. & Newport, E. L. 2005 Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Lang. Learn. Dev.*, **1**(2), 151–195.

Karafet, T. M., Bulayeva, K. B., Nichols, J., Bulayev, O. A., Gurgenova, F., Omarova, J., Yepiskoposyan, L., Savina, O. V., Rodrigue, B. H. *et al.* 2016 Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan. *J. Hum. Genet.*, **61**(3), 181.

Kim, K. H. S., Relkin, N. R., Lee, K.-M. & Hirsch, J. 1997 Distinct cortical areas associated with native and second languages. *Nature*, **388**(6638), 171–174.

Kirby, S., Griffiths, T. & Smith, K. 2014 Iterated learning and the evolution of language. *Curr. Opin. Neurobiol.*, **28**, 108–114.

Kutanan, W., Ghirotto, S., Bertorelle, G., Srithawong, S., Srithongdaeng, K., Pontham, N. & Kangwanpong, D. 2014 Geography has more influence than language on maternal genetic structure of various northeastern Thai ethnicities. *J. Hum. Genet.*, **59**(9), 512.

Labov, W. 2001 *Principles of linguistic change volume 2: Social factors.* John Wiley & Sons.

Labov, W. 2010 *Principles of linguistic change volume 3: Cognitive and cultural factors.* John Wiley & Sons.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. 2007 Quantifying the evolutionary dynamics of language. *Nature*, **449**(7163), 713–716.

Longobardi, G., Ghirotto, S., Guardiano, C., Tassi, F., Benazzo, A., Ceolin, A. & Barbujani, G. 2015 Across language families: Genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.*, **157**(4), 630–640.

Lupyan, G. & Dale, R. 2010 Language Structure Is Partly Determined by Social Structure. *PLOS ONE*, **5**(1), e8559.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D. *et al.* 2011 Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, **331**(6014), 176–182.

Newberry, M. G., Ahern, C. A., Clark, R. & Plotkin, J. B. 2017 Detecting evolutionary forces in language change. *Nature*, **551**(7679), 223.

Nowak, M. A., Komarova, N. L. & Niyogi, P. 2001 Evolution of universal grammar. *Science*, **291**(5501), 114–118.

Nowak, M. A., Komarova, N. L. & Niyogi, P. 2002 Computational and evolutionary aspects of language. *Nature*, **417**(6889), 611–617.

Pagel, M. 2009 Human language as a culturally transmitted replicator. *Nat. Rev. Genet.*, **10**(6), 405–415.

Pagel, M., Atkinson, Q. D. & Meade, A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, **449**(7163), 717–720.

Pickering, M. J. & Garrod, S. 2004 Toward a mechanistic psychology of dialogue. *Behav. Brain. Sci.*, **27**(2), 169–190.

Reali, F. & Griffiths, T. L. 2010 Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**(1680), 429–436.

Salton, G. & McGill, M. J. 1983 *Introduction to modern information retrieval.* McGraw-Hill.

Sarasso, S., Määttä, S., Ferrarelli, F., Poryazova, R., Tononi, G. & Small, S. L. 2014 Plastic Changes Following Imitation-Based Speech and Language Therapy for Aphasia A High-Density Sleep EEG Study. *Neurorehabil. Neural Repair*, **28**(2), 129–138.

Srithawong, S., Srikummool, M., Pittayaporn, P., Ghirotto, S., Chantawannakul, P., Sun, J., Eisenberg, A., Chakraborty, R. & Kutanan, W. 2015 Genetic and linguistic correlation of the Kra-Dai-speaking groups in Thailand. *J. Hum. Genet.*, **60**(7), 371–380.

Steels, L. & Kaplan, F. 2002 Aibo's first words: The social learning of language and meaning. *Evolution of communications*, **4**(1), 3–32.

Tamburrini, N., Cinnirella, M., Jansen, V. A. A. & Bryden, J. 2015 Twitter users change word usage according to conversation-partner social identity. *Soc. Networks*, **40**, 84–89.

Trudgill, P. 2005 Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguist. Typol.*, **8**(3), 305–320.

Trudgill, P. 2011 Social structure and phoneme inventories. *Linguist. Typol.*, **15**(2), 155–160.

Wang, W. S.-Y. 1976 Language Change. *Ann. N. Y. Acad. Sci.*, **280**(1), 61–72.

Wennekers, T., Garagnani, M. & Pulvermueller, F. 2006 Language models based on Hebbian cell assemblies. *J. Physiol. Paris*, **100**(1-3), 16–30.

Wright, S. 2017 Tuning in to terrorist signals. Ph.D. thesis, Royal Holloway, University of London.