Linguistic stability and change under small-scale egalitarian language contact: a mixture model approach

Chundra Aroor Cathcart (chundra.cathcart@uzh.ch)
Department of Comparative Language Science/Center for the Interdisciplinary Study of Language Evolution
University of Zurich

Joanne Yager (joanne.yager@ling.lu.se)
Centre for Languages and Literature
Lund University

Abstract

This paper investigates the outcomes of small-scale egalitarian language contact in an attempt to address whether different linguistic domains exhibit different degrees of stability and resistance to convergence among cohabitant speakers of Jahai and Jedek, two closely related Aslian (Austroasiatic) language varieties spoken in northern Peninsular Malaysia. Using non-parametric Bayesian mixture models, we find that basic vocabulary items show a signal that strongly matches the linguistic identity of individuals, while data from other domains do not. This result is in agreement with other findings from the study of language contact: basic vocabulary is said to be a domain where distinctions in linguistic identity are often emphasized and maintained, while other parts of the vocabulary may be less salient for the purposes of indexing speaker identity, and are thus more prone to the effects of convergence. We demonstrate that this finding is an artifact of neither data coverage nor model choice; at the same time, we are able to identify variation in basic vocabulary items across linguistic groups which is suppressed by the model we use, and outline alternative methods for analyzing data of this sort.

Keywords: Linguistics; Language contact; Language change; Bayesian modeling

Introduction

The forces which drive the growth of linguistic diversity are poorly understood. In this paper, we investigate factors which promote and constrain linguistic diversification and convergence in a multilingual community of foragers in small-scale egalitarian contact. Our chief goal is to investigate whether data from different linguistic domains are more prone to convergence in the face of contact between speakers of highly similar speech varieties. Additionally, we investigate the variation in stability displayed by individual features within linguistic domains, in this case basic vocabulary. These issues bear heavily on the study of cross-linguistic diversity, as loci of stability and change within individual speakers' linguistic profiles speak to predictions regarding the evolution of languages over time.

To address this question, we investigate patterns in the language production of speakers of two closely related Aslian (Austroasiatic) speech varieties in the village of Rual, a resettlement site that is home to bands of Jedek- and Jahai-speaking foragers since the 1970s. We use non-parametric Bayesian mixture models to group individuals into clusters according to the patterns they display in language production data from four linguistic domains, including basic vocabulary. We find that cluster labels inferred from basic vocabulary agree strongly with individuals' linguistic identity, but this is not the case for other types of linguistic data used to describe spatial relations, caused motion, and reciprocal events. This finding is in line with insights from the literature on language contact; speakers tend to enhance and maintain distinctions that index group membership in basic vocabulary, while other domains are more prone to convergence across linguistic groups in the context of multilingualism. This finding holds when we control for differing sample sizes and degrees of coverage across data sets; additionally, posterior predictive checks demonstrate that our results are not an artifact of differing degrees of goodness of fit between the model we choose and the different data sets we employ. At the same time, the method we have chosen may not account for all patterns displayed by the data we analyze, as evidenced by the presence of variation in the data that the model does not capture; we explore this variation and outline alternative ways of addressing the questions asked here.

Background

An understanding of the outcomes of small-scale language contact is highly important for the purpose of refining our knowledge of language change on a global scale, as language contact in small-scale contact scenarios has been an important driving force for language evolution and change throughout human history (Evans, 2010). Egalitarian contact informs our understanding of linguistic diversification, as social stratification can often lead to linguistic disparity: for instance, Vanuatu and Samoa, two geographically proximate island groups with similar chronologies of settlement, differ according to the presence of hierarchical hereditary chiefdoms and in terms of linguistic diversity. Samoa, with its hierarchical political structure, has very little linguistic diversity, in contrast to the widespread and striking linguistic diversity of Vanuatu (Pawley, 1981). Understanding the role played by individuals is key, as individual speakers are responsible for the diffusion of innovations through speech communities as well as the formation of new speech communities.

An open question concerns whether certain linguistic domains and features are resistant to contact-induced change (Curnow, 2011; Matras, 2007). In the context of egalitarian contact, large-scale structural convergence between languages is often found, accompanied by extensive divergence

in lexical forms. This structural convergence is thought to be due to the widespread multilingualism often found in egalitarian contact scenarios, while the lexical differentiation between speech varieties is thought to be due to an emblematic pressure to mark identity (François, 2011). At the same time, not all lexical items behave the same way; there is ample evidence that frequent words are more resistant to replacement by borrowing and other types of contact-induced change (Wieling, Nerbonne, & Baayen, 2011; Monaghan & Roberts, 2019). With this in mind, we test the view that basic vocabulary serves as a strong signal of linguistic identity. We also seek to determine which parts of the basic vocabulary may be more stable than others in an egalitarian multilingual environment.

Data

The data used in this study consist of lexical speech production data collected from Jedek and Jahai speakers at Rual (Yager & Burenhult, 2017; Yager, 2020). The data consist of descriptions of the Topological Relations Picture Series ([T]OPOLOGICAL [R]ELATIONS, Bowerman & Pederson, 1992), the Reciprocal Constructions and Situation Type film clips ([R]ECIPROCAL [E]VENTS, Evans, Levinson, Enfield, Gaby, & Majid, 2004), the PUT project film clips ([C]AUSED [M]OTION EVENTS, Bowerman, Gullberg, Majid, & Narasimhan, 2004), and [B]ASIC [V]OCABULARY collected using a list of meanings based on work by Swadesh (1952). The four data sets contain different but partially overlapping samples of individuals and have the following numbers of speakers (TR $= 38$, RE $= 46$, CM $= 49$, BV $= 10$). For convenience, we refer to each data set as containing data from J speakers consisting of D features, each attesting a number of possible variants. We remove features that are invariant across individuals.

Model

We wish to assess the extent to which individuals' language production corresponds to the linguistic identity with which they associate, and whether this correspondence differs across the four data sets. To address this question, we fit separate Dirichlet process Mixture Models (Gelman et al., 2013, DPMM) to the data for each data set. The DPMM, a non-parametric Bayesian clustering model, assigns a non-hierarchical mixture component or CLUSTER LABEL to each individual in a given data set, but does not assume the number of clusters *a priori*, inferring this number from the data instead. At a high level, individuals are assigned a cluster label based on their language production across the items of the data set in addition to the FEATURE DISTRIBUTIONS inferred for the cluster.

Because the features we analyze consist of categorical data (i.e., a given feature, such as a particular spatial relation, can elicit one of two or more outcomes for a given individual), we employ priors appropriate for categorical (and multinomial) distributions for these feature distributions. The Dirichlet distribution is a commonly used prior over categorical distributions. However, the Dirichlet distribution cannot explicitly capture similarities across categorically distributed outcomes. As an example, three forms for 'mouth' are attested in the data (**hãɲ**, **hɛ̃ɲ**, and **tnɨt**). Of these forms, the first two are cognate candidates, and their potential cognacy is meaningful from the perspective of lexical usage, but their similarity cannot be expressed by the Dirichlet. For this reason, we explore the use of the Logistic Normal (LN) distribution for representing cluster-level word distributions. The fact that the LN distribution is underlyingly multivariate normal means that covariance can be expressed within distributions; we model covariance between two forms in a distribution on the basis of the normalized edit distance between them (with lower distance corresponding to higher covariance). Crucially, this modeling choice encourages (but does not require) cluster-level probabilities of near-identical forms to be similar. There are a number of ways in which to measure phonological similarity, as well as risks associated with such measures. Employing models that are sensitive to phonological similarity as well as ones that are not allows us to assess the joint evidence for our hypothesis produced by both types of models; in the event that models produce different patterns, posterior predictive checks can be used to assess the validity of one model type over the other.

We fit a DPMM for each of the four data sets placing a Dirichlet prior and LN prior over the cluster-level feature distributions, yielding eight models in total. Additionally, to ensure that the cluster configurations that we infer are not simply artifacts of different sample sizes across data sets, we fit the models a second time, excluding speakers not found in the basic vocabulary data set, which contains considerably fewer speakers than the other data sets. Details of model specification and inference are found in the Appendix.[1]

Results

In general, the inference procedure finds small numbers of clusters among individuals, displayed for each data set and prior in Table 1. Models with different priors may infer different numbers of clusters on the basis of the same data set; for instance, the model fit to reciprocal event data which employed a LN prior inferred only one cluster a majority of the time, whereas the model fit to basic vocabulary data using a Dirichlet prior discovered three clusters; at times, the number of clusters inferred matches the the number of linguistic identities subscribed to by individuals (that is, two: Jedek and Jahai). We assess how well these clusters agree with the linguistic identity of each individual in the sample, and in addition how meaningfully these clusters capture variation across individuals in the data.

V-measure

We use samples from the inferred posterior distribution to determine how well the cluster labels inferred across indi-

[1]Data and code used in experiments are found at the following link: https://github.com/jo-yager/Models-of-linguistic-convergence-in-a-hunter-gatherer-community

	TR	RE	CM	BV
Dirichlet	2 (3)	2 (4)	2 (3)	3
LN	2 (2)	1 (2)	2 (2)	2

Table 1: Maximum a posteriori (MAP) number of clusters inferred from topological relations, reciprocal event, caused motion, and basic vocabulary data sets, for models using Dirichlet and Logistic normal priors. Numbers in parentheses represent the MAP number of clusters when speakers not in the basic vocabulary data set were excluded.

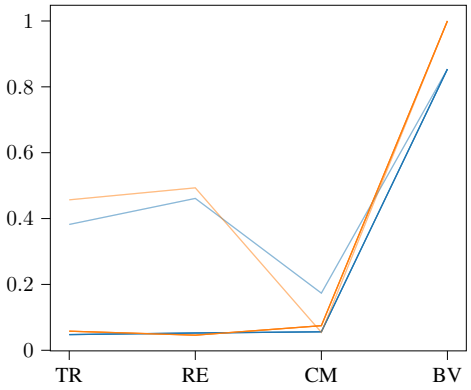

Figure 1: V-measures for cluster labels inferred from topological relations, reciprocal event, caused motion, and basic vocabulary data sets for models using Dirichlet (blue) and Logistic Normal (orange) priors over cluster-feature distributions, with respect to linguistic identity. Lighter lines represent results that exclude speakers not present in the basic vocabulary data set.

viduals reflect their linguistic identity (Jahai or Jedek). We employ the V-MEASURE (Rosenberg & Hirschberg, 2007), which quantifies the goodness of fit of an inferred configuration of cluster labels to a ground truth set of labels. A V-measure of 1 indicates that two clustering configurations partition data points identically, regardless of the labels assigned to clusters.

Figure 1 gives V-measures for each of the models described above averaged across posterior samples, clearly showing that clusters inferred on the basis of basic vocabulary data give a near-perfect match with the linguistic identity of speakers in the data set. This finding is in line with predictions from the literature on linguistic diversification: basic vocabulary is thought to be a linguistic domain where language users tend to index their social distinctiveness, but this is not necessarily the case for the kinds of items contained in the remaining three data sets. Our analysis does not allow us to explicitly represent whether this result stems from attempts on the part of speakers to actively maximize or simply maintain distinctions (a comparison with speakers of Jahai varieties from localities other than Rual suggests the latter phenomenon); nevertheless, patterns of basic vocabulary usage display a distinctive profile that aligns with divisions based on the linguistic identity of speakers.

Cluster labels inferred on the basis of the remaining three data sets do not show agreement with speakers' linguistic identities. Interestingly, this does not appear to be due to a wholesale convergence of speakers at Rual on a unified linguistic profile; in most cases, more than one cluster is inferred, but these clusters do not agree with with the Jahai/Jedek division according to speakers' language identities. This may be an artifact of model choice, or due to the incipient formation of cohesive groups between interacting speakers with different linguistic identities. Alternatively, the language identities of individuals at Rual may index something other than the features of their language production.

The lighter colored lines in Figure 1 represent V-measures based on models which only included speakers found in the basic vocabulary data set. While it is clear that the difference in V-measures is not as pronounced when we control for differences in the samples, the description provided above still holds, with data sets other than basic vocabulary yielding considerably lower V-measures.

Posterior predictive checks

If we are correct in our assumption that each individual's linguistic behavior is associated with a single cluster or mixture component for a given data set, then we may conclude that speakers employ basic vocabulary more than other types of data for the purpose of indexing their linguistic identity. But how appropriate is this assumption? Within a given data set, there may be conflicting cluster label configurations across individuals. Even within the domain of basic vocabulary, some concepts may be more stable than others (Pagel, Atkinson, & Meade, 2007); e.g., less frequent items may be vulnerable to contact-induced change (Monaghan & Roberts, 2019).

We carry out POSTERIOR PREDICTIVE CHECKS to assess whether the model used is appropriate for all data types, or whether our results may stem from the fact that the DPMM is a more appropriate choice for basic vocabulary data than other data types. Posterior predictive checks seek to identify areas of model misspecification by assessing how well data simulated from a fitted model can capture various properties of the observed data; mismatches indicate that our models have failed to capture certain parts of the structure of the data. Our concern is that the overall pattern learned by the DPMM, which assigns one label per individual, may detect inter-speaker connections for certain linguistic features, but may ignore conflicting inter-speaker connections based on other linguistic features. If this idea is confirmed, then an alternative model may be more appropriate than a DPMM, such as an admixture model (Pritchard, Stephens, & Donnelly, 2000; Blei, Ng, & Jordan, 2003; Chang & Michael, 2014).

We choose to compare PAIRWISE MATCHING COEFFICIENTS or the proportion of features in a data set shared by a pair of individuals of the observed data to matching coefficients found in simulated data. Pairs of individuals displaying similar behavior have matching coefficients closer to 1;

dissimilar individuals have matching coefficients closer to 0. If simulated data have a higher average matching component than the observed data, it means that the assumptions of the DPMM have oversimplified the structure of the data, picking up on some connections between pairs of individuals but ignoring others that conflict with the stronger clustering. If the discrepancy is negative, then it indicates that cluster-level feature distributions are not sparse enough to be informative, which may also imply that finer-grained structure among individuals is not being learned.

In general, the average matching coefficient for models using the LN prior tends to be below the observed discrepancy, while values for models using a Dirichlet prior tend to be higher (Figure 2, top row; $p < 1e-10$ for two-sided Z-tests between all simulated distributions and observed discrepancies). This indicates that the Dirichlet models generate data with lower variation than the observed data, while the LN models generate data with higher variation in patterns of linguistic usage among individuals; this can potentially be addressed by reparameterizing the LN prior to return sparser distributions and the Dirichlet prior to return smoother distributions.

Additionally, we evaluate the performance of our models by calculating the mean log-likelihood of held-out data conditioned on posterior parameters inferred on the basis of non-held-out data; we carry out K-fold cross validation, randomly splitting each full data set into $K = 4$ equal partitions. Mean log-likelihood values computed from posterior samples are found in Figure 2, bottom row. The LN models consistently outperform the Dirichlet models according to this metric.

In general, the Dirichlet models generate matching coefficients that are slightly (but not significantly) closer to the average matching coefficients for the observed data than those generated by models with LN priors. At this time, we have no clear explanation for why the Dirichlet model seems to show slightly better performance in terms of one PPC while a second PPC shows greater support for the LN model; it is likely that this discrepancy is linked to the priors we have chosen and requires further investigation. What is important is that patterns shown by our PPCs seem to hold across most of the data sets employed, indicating that the DPMM is no better or worse for one data set than it is for another, and that we do not find greater agreement between linguistic identities and cluster labels inferred on the basis of basic vocabulary data simply because the DPMM is a more appropriate choice for that particular data type; according to the matching coefficient PPC, the DPMM is a less-than-ideal choice for all data types (further exploration of PPCs and different model parameterizations is needed to determine exactly how poor a choice the DPMM is).

In the following section, we investigate the degree to which the DPMM suppresses interesting variation in the data in the process of assigning labels to individuals on the basis of their overall linguistic behavior. We investigate potential variation of this type, restricting our analysis to the basic vocabulary data, as the concepts contained in the data set are more interpretable than the visual elicitation stimuli of the remaining data sets.

Basic Vocabulary Patterns

To better understand the distribution of feature variants across individuals in the data, we visualize concepts in the basic vocabulary data set according to two variables, based on the results from our models. First, we wish to know whether for a given concept, the cluster assignment of the speaker is predictable on the basis of the form used. Additionally, we wish to know the predictability of the form used for a given concept, given a cluster label. Predictability/unpredictability along these dimensions has the potential to shed light on incipient patterns of convergence between individuals in the domain of basic vocabulary.[2]

We operationalize these types of unpredictability using conditional entropy, with the uncertainty of a cluster assignment given a form quantified by the measure $H(\text{cluster assignment}|\text{form})$ and the uncertainty of a form given a cluster label quantified as $H(\text{form}|\text{cluster assignment})$. We compute these measures for each posterior sample in each model, and average conditional entropy measures for each concept.

Mean entropies for each concept representing the unpredictability of a form given a cluster label and the unpredictability of a cluster label given a form are found in Figure 3. Concepts fall into roughly four or five clusters, though divisions are somewhat blurry. Group-specific items are found in the lower left quadrant, comprising concepts for which the cluster label is highly predictable given the form used and the form used is highly predictable given the cluster label. In the upper left quadrant are concepts that display variation in forms that is nested within cluster labels. The lower right quadrant contains common vocabulary shared across cluster labels; this invariance may be due to convergence between groups, may reflect more archaic or stable terminology that never diversified, or may reflect shared loans from Malay, the majority language of the region. The remainder of concepts in the upper right quadrant show variation which cuts across clusters; individuals in the different clusters may use the same term, and individuals in the same cluster may use different terms, potentially indicating contact-induced inter-group convergence and intra-group divergence.

If the pattern we have detected points to increasing convergence and has progressed to the extent that a mixture model is no longer appropriate for representing variation in the data, new approaches to characterizing patterns in the data will be needed, including admixture models such as the Hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2005), though this approach makes a number of problematic assumptions

[2]These questions can also be addressed on the basis of reported linguistic identities rather than cluster labels, given the high agreement between linguistic identities and cluster labels reported above. We use the results of our models in order to incorporate an estimation of uncertainty learned by our Bayesian analyses.

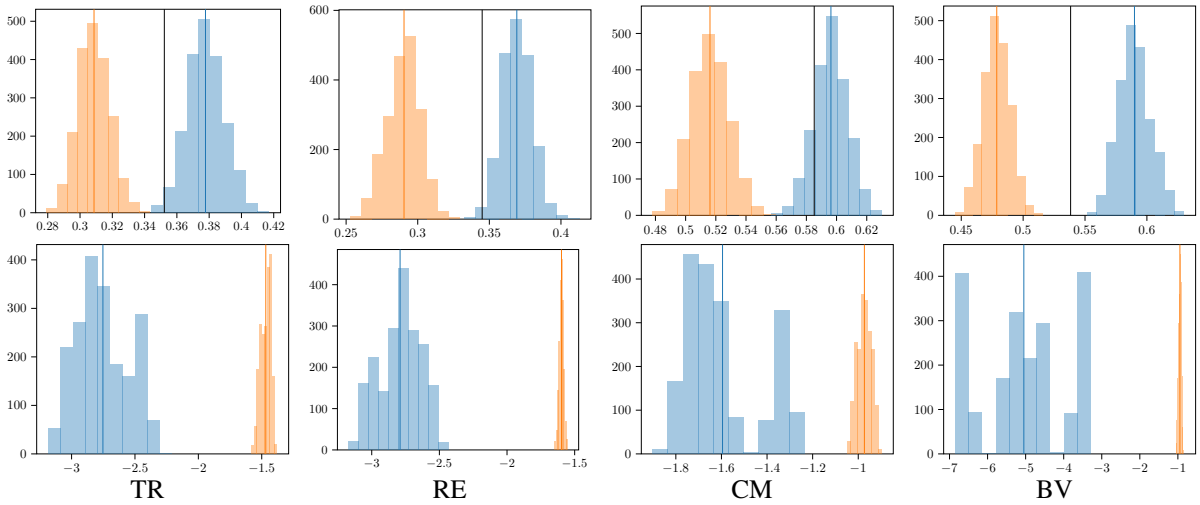

Figure 2: Above: observed and simulated matching coefficients for topological relations, reciprocal event, caused motion, and basic vocabulary data sets conditioned on posterior distributions from Dirichlet (blue) and Logistic Normal (orange) models. Black vertical lines represent the observed value. Below: average log-likelihood for held-out data points conditioned on posterior distributions from Dirichlet (blue) and Logistic Normal (orange) models. Blue vertical lines represent mean values for the Dirichlet model; orange vertical lines represent mean values for the LN model.

(Williamson, Wang, Heller, & Blei, 2010). Additionally, some of the independence assumptions made by the models used in this paper may do well to be relaxed.

Outlook

In this paper we investigated patterns of stability and instability under small-scale egalitarian contact across four data sets. We found that clustering configurations of individuals inferred on the basis of basic vocabulary speech production data agrees more strongly with their linguistic identity than for other data types, and demonstrated that this effect is not an artifact of differences in sample size and coverage across the data types used, or of model choice. At the same time, our result is relevant for only a single synchronic time slice; it is likely that the state of affairs will change over time, and since in this context we are dealing with contact between closely related language varieties, it remains difficult to tease apart the effects of convergence and shared genealogy. Longitudinal work along these lines is of extreme importance, and will shed light on the outcome of the incipient patterns we have identified — the variation seen may remain stable for a prolonged period of time, or single variants may become conventionalized standard forms. Additionally, tracking the forms which currently show nested variation within groups, a pattern that is not in conflict with the overarching assumptions of a mixture model, may help us understand the formation of new linguistic identities out of existing ones. Further research and data collection will be needed to link these findings to studies of contact-induced change based chiefly on languages of Western Europe which predict that less frequent words are more resistant to borrowing, as smaller languages generally lack resources containing information about word frequency that Western European languages such as English and Dutch possess. Continued data collection and the development of models appropriate for capturing a finer-grained picture of variation over time in settings such as the one described here will serve as a much-needed contribution to the study of the dynamics of language change.

Appendix: Model Specification and Formulae

Each data set consists of data points generated by J individuals. There are D features in each data set, each of which has two or more variants.

Priors

The models used in this paper employ either a Dirichlet or Logistic Normal prior over $\phi_{t,d}$, the distribution over feature outcomes for feature d in cluster t. All Dirichlet priors are symmetric with a concentration parameter of .1. Logistic Normal priors are generated in the following fashion: first, a multivariate sample $\psi_{t,d} \sim \text{Normal}(\mathbf{0}, \Sigma)$, is drawn. The matrix Σ contains covariances between each pair of variants i, j in each feature d, which we model as $\exp(-\delta_{ij})$, where δ_{ij} is the normalized edit distance between the two variants (i.e., the minimum number of insertions, deletions, and mutations needed to convert one string into the other, divided by the length of the longer string). The transformation $\phi_{t,d} = \text{softmax}(\psi_{t,d})$ results in a simplex of categorical probabilities summing to 1.

DP Mixture Model

Key parameters in a DPMM are θ, a global categorical distribution over the presence of clusters across individuals, and $\boldsymbol{\phi}$, the cluster-level feature distributions discussed above. We generate θ via truncated stick-breaking view of the Dirichlet

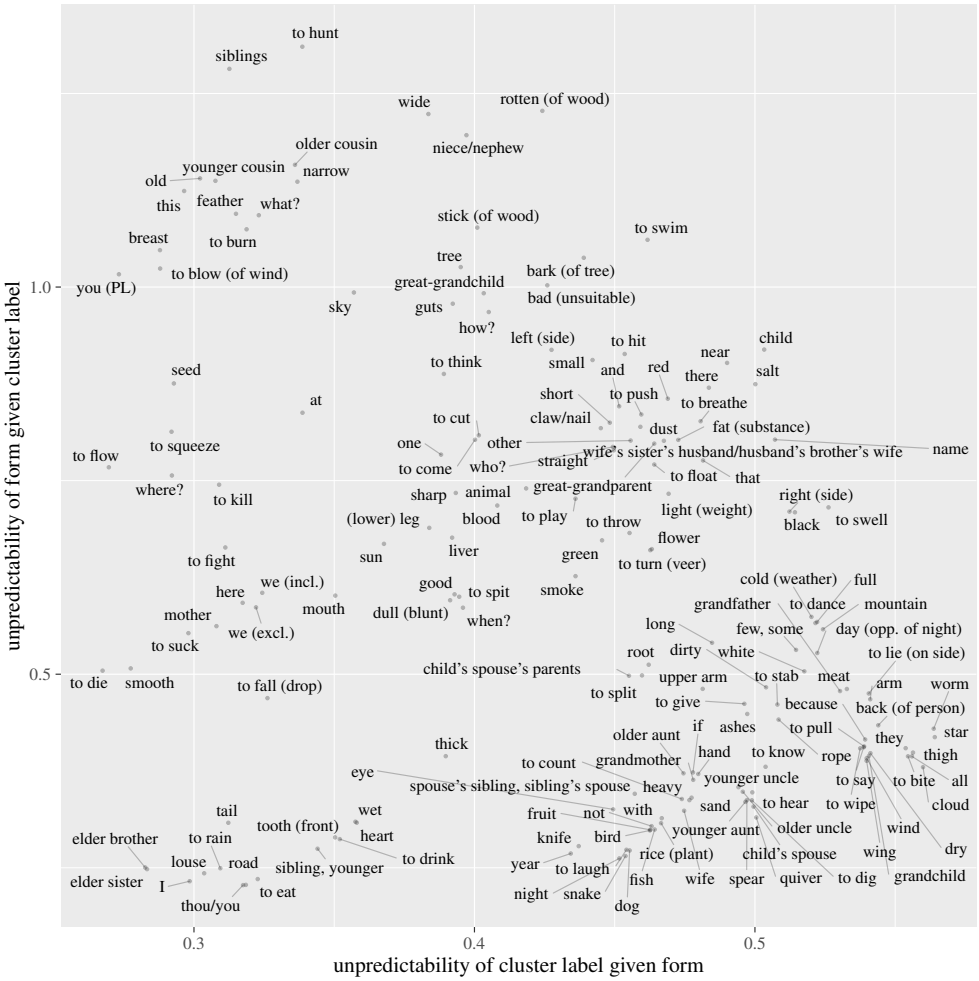

Figure 3: Unpredictability of speaker linguistic identity given form plotted against unpredictability of form given speaker identity, averaged by concept and transformed to rank.

Process (Ishwaran & James, 2001), which effectively allows a number of clusters to be learned from the data, setting the truncation level T to 5. The generative process of the DPMM is as follows:

$\gamma, \delta \sim \text{Gamma}(1,1)$; $\theta \sim \text{DP}(\gamma, \delta)$
For each individual $j \in \{1, ..., J\}$, sample a cluster label $z_j \sim \text{Cat}(\theta)$
  For each feature $d \in \{1, ..., D\}$, sample the observed feature $y_{j,d} | z_j = t \sim \text{Cat}(\phi_{t,d})$

This process yields the following likelihood, with the discrete parameter $\boldsymbol{z}$ marginalized out:

$$P(\boldsymbol{y}|\theta, \boldsymbol{\phi}) = \prod_{j=1}^{J} \sum_{t=1}^{T} \theta_t \prod_{d=1}^{D} \phi_{t,d,y_{j,d}}$$

For each of the four data sets, we carry out inference for the DPMM using Automatic Differentiation Variation Inference (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2017) as implemented in PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) across four separate initializations, monitoring convergence via the evidence lower bound (ELBO) and the potential scale reduction factor (Gelman & Rubin, 1992) for parameters that are not prone to label switching. We draw 500 samples from the fitted variational posterior distributions to approximate a posterior sample.

Following inference, the discrete cluster label for a given individual can be sampled according to the following probability:

$$P(z_j = t | y_j, \theta, \boldsymbol{\phi}) \propto \theta_t \prod_{d=1}^{D} \phi_{t,d,y_{j,d}}$$

For each data set, ϕ_t comprises the following numbers of distributions and parameters, respectively (TR: 58/184, RE: 63/488, CM: 59/305, BV: 171/483).

Posterior Predictive Check

We compute matching coefficients for each pair of individuals $i, j, i \in \{1, ..., J\}, j \in \{i+1, ..., J\}$ by counting the number of

agreeing features across each pair and dividing by the number of features for which data are attested for both individuals.

Conditional entropy measures

Given a posterior sample of $\boldsymbol{\theta}, \boldsymbol{\phi}$, we compute the conditional entropy of a cluster label given an observed word form (for a given concept, notation concerning which we excluded above, for brevity), $H(\text{cluster assignment}|\text{form})$, as well as $H(\text{form}|\text{cluster assignment})$ via the joint probability of cluster label $z = t$ and form variant v for a concept d $P(z = t, v|d) = P(z = t)P(v|z = t, d) = \theta_t \phi_{t,d,v}$. From this joint probability distribution, it is straightforward to compute $P(z = t|d) = \sum_{v \in V_d} \theta_t \phi_{t,d,v}$ and $P(v|d) = \sum_{t=1}^{T} \theta_t \phi_{t,d,v}$. This makes it straightforward to compute $H(z = t|v, d) = H(v, z = t|d) - H(v|d)$ and $H(v|z = t, d) = H(v, z = t|d) - H(z = t|d)$ for each concept.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning research*, *3*, 993–1022.

Bowerman, M., Gullberg, M., Majid, A., & Narasimhan, B. (2004). Put project: The crosslinguistic encoding of placement events. In A. Majid (Ed.), *Field manual, vol. 9* (p. 10-24). Nijmegen: Max Planck Institute for Psycholinguistics.

Bowerman, M., & Pederson, E. (1992). Crosslinguistic perspectives on topological spatial relationships. In *87th annual meeting of the American Anthropological Association, San Francisco, CA*.

Chang, W., & Michael, L. (2014). A relaxed admixture model of language contact. *Language Dynamics and Change*, *4*(1), 1–26.

Curnow, T. (2011). What language features can be 'borrowed'. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics* (p. 412-436). Oxford: Oxford University Press.

Evans, N. (2010). *Dying words: Endangered languages and what they have to tell us*. Chichester: Wiley-Blackwell.

Evans, N., Levinson, S. C., Enfield, N. J., Gaby, A., & Majid, A. (2004). Reciprocal constructions and situation type. In A. Majid (Ed.), *Field manual, vol. 9* (p. 25-30). Nijmegen: Max Planck Institute for Psycholinguistics.

François, A. (2011). Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence. *Journal of Historical Linguistics*, *1*(2), 175-246.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* 3rd edition. New York: Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.

Matras, Y. (2007). The borrowability of structural categories. In Y. Matras & J. Sakel (Eds.), *Grammatical borrowing in cross-linguistic perspective* (p. 31-73). Berlin & New York: Mouton de Gruyter.

Monaghan, P., & Roberts, S. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition*, *186*, 147-158.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717-20.

Pawley, A. (1981). Melanesian diversity and Polynesian homogeneity: a unified explanation for language. In J. Hollyman & A. Pawley (Eds.), *Studies in Pacific languages and cultures in honour of Bruce Biggs* (pp. 269–309). Auckland: Linguistic Society of New Zealand.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (p. 410-20). Prague: Association for Computational Linguistics.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, *96*, 452-463.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems* (pp. 1385–1392).

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), e23613.

Williamson, S., Wang, C., Heller, K. A., & Blei, D. M. (2010). The IBP compound dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on Machine Learning.* Haifa, Israel.

Yager, J. (2020). *Small-scale multilingualism and language contact in egalitarian foragers*. Unpublished doctoral dissertation, Lund University.

Yager, J., & Burenhult, N. (2017). Jedek: A newly discovered Aslian variety of Malaysia. *Linguistic Typology*, *21*, 493–545.