bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inferring linguistic transmission between generations at the scale of individuals

Valentin Thouzeau[†], Antonin Affholder[†], Philippe Mennecier[†], Paul Verdu[†], Frédéric Austerlitz[†]

[†] *CNRS, MNHN, Université Paris Diderot, UMR 7206 Eco-Anthropologie et Ethnobiologie, Paris 75016, France*

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Historical linguistics highly benefited from recent methodological advances inspired by phylogenetics. Nevertheless, no currently available method uses contemporaneous within-population linguistic diversity to reconstruct the history of human populations. Here, we develop an approach inspired from population genetics to perform historical linguistic inferences from linguistic data sampled at the individual scale, within a population. We built four demographic models of linguistic transmission at this scale, each model differing by the number of teachers involved during the language acquisition, and the relative roles of these teachers. We then compared the simulated data obtained with these models with real contemporaneous linguistic data sampled in Tajik speakers in Central Asia, an area known for its high within-population linguistic diversity, using approximate Bayesian computation methods. With these statistical methods, we were able to select the models that best explained the data, and inferred the best-fitting parameters under these selected models, demonstrating the feasibility of using contemporaneous within-population linguistic diversity to infer historical features of human cultural evolution.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

Several recent studies used linguistic data under a computational framework aiming at reconstructing various aspects of the cultural history of human populations (Atkinson, 2011; Bouckaert et al., 2012; Gray and Atkinson, 2002; Pagel et al., 2013). These data consist mainly of a set of presence or absence of items within a given set of contemporaneous languages, which can be found, for example, in databases such as the World Atlas of Language Structures WALS (Dryer and Haspelmath, 2013), or the Global Database of Cultural, Linguistic and Environmental Diversity D-PLACE (Kirby et al., 2016). Most studies consider languages at a macro-evolutionary scale, i.e. they deal only with differences among languages, neglecting the variability within each language. For instance, Gray and Atkinson (2002) used a set of Swadesh lists obtained for 87 languages to investigate the origin of the Indo-European linguistic family. Atkinson (2011) considered the number of phonemes used in 504 languages worldwide to test the hypothesis of a serial founder effect due to the Out-Of-Africa expansion. Reesink et al. (2009) used the linguistic diversity of the ancient Sahul continent (present day Australia, New Guinea, and surrounding islands) for 121 languages to infer the history of the structural characteristics of these languages.

These approaches rely implicitly on several assumptions. They require primarily a clear separation between several differentiated languages. Nevertheless, this notion of distinct languages is often irrelevant at a local scale, in particular in contexts of dialectal continuum or linguistic contacts (Heeringa and Nerbonne, 2001; Livingstone and Fyfe, 1999). Furthermore, most of these studies do not take into account the within-population linguistic diversity, since traditional linguistics often considers languages as unique and coherent systems (Pateman, 1983).

This assumption implies the loss of a large amount of information, knowing that the demographic phenomena at population level – different population sizes, bottlenecks, expansions –

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are expected to play a major role in language evolution (Vogt, 2009). Including contemporaneous within-population linguistic diversity in the reconstruction of the demographic history of human populations at a local scale should thus open a whole new dimension into the field of historical linguistic inferences.

In this context, Croft (1996) argued for a replacement of the 'essentialist' theory of language changes by a 'population' approach of language changes, and later proposed a detailed review of the "evolutionary linguistic" field and underlying paradigms (Croft, 2008). Nevertheless, very few studies deal with the contemporaneous within-population linguistic diversity in a historical reconstruction perspective. Some recent examples include the use of surnames in Austria as linguistic contemporaneous information (Rodriguez-Larralde and Barrai, 2000), the use of the family names in different contexts (Darlu et al., 2012), or the use of proportion of African words in free speech among Cape Verdean Kriolu speakers (Verdu et al., 2017).

In order to perform historical linguistic inferences from current linguistic data, we need to assume one or several possible model of linguistic transmission between generations, and a possible set of historical scenarios which produced these observed data. Nevertheless, there is no consensual theoretical framework allowing to handle within-population linguistic diversity data in order to infer the underlying historical scenarios and evolutionary mechanisms. It is possible to first assume a clear and delimited mechanism of linguistic evolution, and then to study the range of historical scenarios that could have produced the observed linguistic data. Nevertheless, the validity of the conclusions depends on the validity of the assumed mechanism. It is then crucial to determine the most relevant mechanism of linguistic evolution, in order to produce, ultimately, valid inferences.

We propose, in this article, to evaluate a series of models of linguistic evolution between generations at the individual scale. We did not study the history of higher-order objects such as "the languages", but the history of the linguistic diversity carried by individuals within a population among which communication events may occur over time. We aimed here at understanding how the

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evolution of linguistic diversity among generations is affected by demographic parameters such as population size (the number of individuals of a given speech community), and thus to assess whether it is possible to infer the best demographic scenario and its corresponding parameters from a set of linguistic data.

Approximate Bayesian Computation methods (ABC, Beaumont et al., 2002; Tavaré et al., 1997) provide a particularly well-adapted framework to tackle this problem. In this paper, we used the recently developed Approximate Bayesian Computation via Random Forest (ABCRF) algorithm to assess, among a set of possible competing scenarios, the scenario that best explains the observed data, and estimate the posterior parameters of this scenario (Breiman, 1999; Pudlo et al., 2016).

For this purpose, we implemented an individual-based simulation program, which simulates the evolution of linguistic items among generations, under different modes of linguistic transmission. These simulated data allowed us to perform the ABCRF procedure on a real dataset from Central Asia. This dataset consisted of 30 individuals interviewed for 185 words across 10 villages in Tajikistan. These villages are known to use the same language, but with some variability among individuals (Mennecier et al., 2016). We aimed at inferring the most probable models of linguistic transmission mechanisms between linguistic generations, under a demographic scenario of demographic expansion or contraction. We proposed four transmission models. The “Clonal model” assumes that each individual learns his/her linguistic items from only one teacher. The “Sexual model 1” assumes that each individual learns his/her linguistic items from two teachers (one male and one female), with specific items transmitted only by males and specific items transmitted only by females. The “Sexual model 2” assumes that each individual learns his/her linguistic items from two teachers (one male and one female), without specific items belonging to males or females. Finally, the “Social model” assumes that each individual learns his/her linguistic items from the whole population. We aimed then at inferring the best-fitting parameters under the chosen scenario: linguistic mutation rates, and populations sizes. Our aim was to demonstrate the feasibility of using

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contemporaneous within-population linguistic diversity to infer historical features in human cultural evolution.

2. Models

2.1. Production of utterances

We considered a linguistic population as a group of individuals that may potentially interact through linguistic communication. The mechanisms of linguistic communication and transmission may follow different modalities, which correspond to different models of linguistic evolution. Nevertheless, we considered that the unit of linguistic communication is the *utterance*, a production of linguistic items associated with a meaning.

Each linguistic item is a possible version from a class. There are several types of linguistic items, which can be related to various aspects of languages: vocabulary, grammar, structure…, etc. We developed here a general model of linguistic item transmission, which we applied in particular to the case of cognates, which correspond to words with different etymological origins that express the same meaning. For example, the Spanish word “Flor” and French word “Fleur” are two items of the class Flower of the same meaning and the same etymological origin, and are then cognates. The Spanish word “Multa” and French word “Papillon” are two items of the class Butterfly with the same meaning, but with different etymological origin, and are then not cognates. We considered here that cognates can vary among individuals within a population. This differs from the assumptions made in previous studies (Bouckaert et al., 2012; Gray et al., 2009; Thouzeau et al., 2017) where cognates are sampled at the language scale and for which individuals are considered as users rather than producers of this language.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2. Four models of acquisition of a new language

We developed a new simulation software *PopLingSim 2* (*PLS2*). This software implements an individual-based forward-in-time simulation model with discrete generations, in which we assumed that populations were composed of only two types of individuals: "learners" and "teachers". We assumed that the rules of utterance productions of a teacher depended only on the utterances that he/she heard when he/she was a learner. We assumed that each learner chose only one item from each class during the learning phase. Two learners could choose the same linguistic item. After the whole learning phase, each teacher was discarded and each learner became a teacher. Then, new learners appeared (exactly half male and half female in "*Sexual*" models, see blow).

We tested four models of linguistic acquisition during learning (Figure 1). These models differed by the number of teachers involved during the language acquisition, and the relative roles of these teachers.

In the first model, named the "*Clonal*" model, each learner had only one teacher, which was drawn at random in the teacher population. The learner copied "in a clonal way" every item that the teacher produced. In the second model, named the "*Sexual*" model, two different teachers (one "male" and one "female") were attributed at random to each learner. The learner then copied directly the first half of the items produced by teacher 1, and the second half of the items produced by teacher 2. Thus, a determined half of the items was always transmitted by one teacher, and the other half by the other teacher. In the third model, named the "*Sexual2*" model, two different teachers (one "male" and one "female") were attributed to each learner at random. For each item, the learner copied at random either the item from teacher 1 or teacher 2, with equal probabilities (½, ½). Thus, no particular item had a teacher-specific transmission, every item was transmitted from one teacher chosen at random. In the fourth model, named the "*Social*" model, for each class of meaning each learner copied an item drawn at random from all the items produced by all the teachers in the population.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For each model, we assumed that errors could occur during the transmission of each item, leading to the creation of a completely new item. We denoted such errors "linguistic mutations". The mean mutation rate $\bar{\mu}_L$ was drawn in a log-uniform prior distribution, between 10^{-6} and 10^{-1} mutations per lexical item per generation. For each item, its mutation rate was subsequently drawn in a beta distribution with a mean $\bar{\mu}_L$ and a shape $\beta = 2$, allowing us to simulate a set of linguistic items with a different rate of change.

2.3. Historical scenario

We focused here on a single linguistic population, defined as a language community, where the individuals have been sampled using a linguistic questionnaire. This linguistic population evolved first with a constant size N_0 until $t_0 = 5 \times N_0$, a time that, as we visually checked, was sufficient to reach an equilibrium between the production of linguistic diversity through mutation, and the reduction of this diversity through random sampling. This population then evolved with a new size N_1 during t_1 generations. The linguistic items were then sampled at the final generation. This model allowed simulating a range of histories, depending on the relative values of the parameters N_0 and N_1 and on the value of t_1. The population sizes N_0 and N_1 were drawn in a uniform distribution between 100 and 1000 individuals, this low upper bound being set to limit the large computational time requirement for completing these forward-in-time simulations. Time t_1 was drawn in a uniform distribution, between 0 and 1000 generations. The median, the minimum, the maximum, and the quantile 5% of the priors of the models are summarized in Table 1.

3. Materials

We sampled cognate variability for 30 individuals from 10 villages in Tajikistan (Figure 3) assuming that the individuals belonged to a single linguistic population. In contrast with our

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previous study, where we considered for each cognate only its most frequent variant in each locality (Thouzeau et al., 2017), we kept here the linguistic variant recorded for each individual. Thus, for each individual, we recorded the words used for 185 meanings from an adapted Swadesh list. We considered as "cognate" a group of words with the same etymological origin and the same meaning, such words being more likely to be related by a common ancestry. The classification of lexical data gathered on the field into cognates was performed by Philippe Mennecier following previous work (Mennecier et al., 2016; Thouzeau et al., 2017).

4. Analyses

4.1. Simulations

For each model, we performed 10 000 simulations using our newly-developed software *PopLingSim 2* (*PLS2*). We parallelized the simulations using 250 cores of the cluster station *Genotoul*, amounting to approximately 90 000 CPU hours. Most of this computation time was spent during the phase to reach equilibrium between mutation and drift at $t_0 = 5 \times N_0$ generations.

During the process of sampling linguistic items from our simulations, we simulated missing values by transforming cognates drawn at random into missing values. The total number of simulated missing values was set to the number of missing values in the real data set, to avoid the bias they may induce in the following ABC procedures.

4.2. Summary statistics

We constructed a new set of population linguistic summary statistics, some of which were inspired from classical population genetics statistics. After computing $p_{i,j}$, the proportion of

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individuals using the item *i* of the class *j*, we computed the linguistic diversity $D_j = 1 - \Sigma_i\ p_{i,j}^2$, analogous to genetic diversity (Nei, 1987).

Then, we computed across all items:

- The mean linguistic diversity, $\overline{D}$;
- The range of the linguistic diversity, R(D) ;
- The variance of the linguistic diversity, V(D) ;
- The number of strictly different lists of items, S ;
- The mean number of items in each class, $\overline{N}$;
- The variance of the number of items in each class, V(N) ;
- The frequency spectrum of the number of items per class, F.

4.3. Model selection

Before model selection, we performed a goodness-of-fit test to check if the simulations were able to produce data close to the real data using the function *gfit* from the *R* package *abc* (Csilléry et al., 2012) to verify that we simulated datasets close to the real dataset. We performed model selection using the *R* package *abcrf* with the RF algorithm and the function *abcrf (Pudlo et al., 2016)*. We graphically checked if a forest of 500 trees allowed a convergence of the error rate. We then performed a cross-validation analysis using an out-of-bag approach implemented in the package *abcrf*, evaluating if the algorithm was *a priori* able to distinguish between the four models.

4.4. Parameters estimation

We used the RF algorithm with the function *regAbcrf* of the package *abcrf* (Raynal et al., 2017) to estimate the expectation, the median, the variance and the quantiles 5% of the parameters N_1, N_0, t_1, μ_L and the composite-parameters $N_1 \times \mu_L$, $N_0 \times \mu_L$ and $t_1 \times \mu_L$. Note that the RF algorithm does not estimate the whole posterior distribution of the parameters directly, but estimates the quantiles of this distribution instead.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Results

5.1. Model selection

Using the goodness-of-fit test, we verified that there was no significant differences between the real and simulated datasets (p-value = 0.55, with 1000 replications). We performed the RF analysis using 500 trees, and we verified graphically that the error rate converged. The RF analysis rejected the *Clonal* and the *Sexual* models, and selected with equal probability the *Sexual2* and the *Social* models (Table 2).

The cross-validation analysis (Figure 4) indicated a good *a priori* differentiation between the *Clonal* model, the *Sexual* model and the group '*Sexual2* and *Social'* models. Nevertheless, the S*exual2* and the *Social* models could not be reliably distinguished. It was therefore impossibleto choose, based on our data, between the '*Sexual2'* and the '*Social'* models, but we may be confident in the rejection of the *Clonal* and the *Sexual* models.

5.2. Parameter estimation

For the two most likely models (*Sexual2* and *Social)*, we could not estimate separately the parameters N_0, N_1 and t_1*:* the estimated quantiles of their posterior distributions were similar to those of their priors (Tables 3 and 4). Nevertheless, the estimated quantiles of the parameter μ_L and of the composite parameters $N_1 \times \mu_L$, $N_0 \times \mu_L$ and $t_1 \times \mu_L$, were substantially narrower than those of their respective priors (Tables 3 and 4). Using the estimated posteriors for the *Sexual2* and *Social* models separately, we estimated that the linguistic mutation rate ranged between 1.98×10^{-4} and 1.44×10^{-3} mutations per cognate per linguistic generation.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Discussion

In this article, we built individual-based models simulating the linguistic evolution of a population, under a given demographic scenario, considering four possible kinds of linguistic transmission between generations. We used an ABC framework to compare the simulated data with a real dataset of 30 individuals in Tajikistan typed for 185 cognates, in order to estimate which models fitted best the data and estimate the parameters of these best-fitting models.

First, we showed that some of our models were able to produce simulated data close to the contemporaneously observed data. It meant that we were able to implement linguistic transmission models between generations at the individual scale, which were consistent with the linguistic diversity of the sampled populations.

We provided thus inferences of some features of linguistic history, selecting the most plausible mechanisms of linguistic transmission, and estimating the parameters of the selected models for our sample of Tajik-speaking individuals. The low posterior probabilities of the *Clonal* and S*exual* models compared to the S*exual2* and the S*ocial* models indicated that the mechanisms of linguistic acquisition followed, in this case, a process of linguistic recombination with several teachers, and not a process of transmission without recombination among utterances from different teachers.

In other words, we inferred that these individuals did not learn their basic vocabulary from only one individual, or from two individuals with “male”-specific and “female”-specific lexical items. They seemed to learn their vocabulary either from two individuals without “sex”-specific vocabulary, or from the whole population. This is consistent with the fact that Tajik populations are known to be cognatic (Krader, 1966), i.e. they inherit social status and material goods from their two parents. This symmetric role of parents may imply that they receive also linguistic items from both of them. It would be of great interest in future work to distinguish between a transmission following a S*exual2* model (with only two teachers), and a transmission following a *Social* model

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(with a whole community as a teacher). This is likely to require a substantially larger amount of linguistic data at the within-population scale.

Our estimates of the mean linguistic mutation rate of the lexical items of the Swadesh list ranged between 10^{-4} and 10^{-3} mutations per lexical item per generation. Our micro-evolutionary context (i.e. at the scale of the individuals within a language) may be compared with a macro-evolutionary context (i.e. at the scale of a whole language or a linguistic variety). The mutation rate estimated here fell in the same range than the mutation rate in macro-evolutionary studies (Pagel et al., 2007). Considering that languages at a global scale emerge from the interactions among individuals, our result led us to hypothesise that the mutation rate estimated globally emerges from the mutation rate at a local scale.

Our posterior estimations of population sizes did not differ from the priors of the simulations. It meant that our method could not directly evaluate the number of individuals in the current and ancestral populations, but only synthetic parameters such as $N_0\mu$. Such limitation has been also observed in population genetics, where it is also quite difficult to estimate directly effective population sizes (Wang, 2005). In this context, one of the more promising approach might be to use temporal samples, as it was shown in population genetics that it was one of the most efficient method for estimating recent population size, and/or to design specific statistics (like for instance sibship frequencies in population genetics, Wang, 2016).

In this study, unlike most other studies focusing on within-population linguistic diversity (Baxter et al., 2009; Danescu-Niculescu-Mizil et al., 2013; Kandler et al., 2010), we only used contemporaneous linguistic diversity. This method allowed us to perform historical inferences only based on sampling campaigns conducted in existing populations. The amount of information available depends only on the sampling effort, and not on the relatively limited historical records.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

There are nevertheless some theoretical obstacles remaining. First, the models of linguistic acquisition that we proposed here do not integrate the particular constraints of communication processes. In particular, we assumed a neutral production of variants without any constraints on linguistic communication. Some evolutionary linguists would argue for an integration of the particularity of languages as communication systems, associated with a strong set of constraints (Beckner et al., 2009). Indeed, individuals maximize the probability of being understood, as well as minimize the cost of communication, two features that will strongly affect linguistic evolutionary processes (Tamariz and Kirby, 2015). These constraints are particularly strong in the case of phonological, morphological, or syntactical systems, and we may wonder if lexical variants are subject to these constraints too. If so, theses particularities of linguistic systems may be at odds with inferences based on a model of neutral evolution, and should thus be taken into account for an accurate model of linguistic evolution at the individual scale, for historical inferences purposes.

Moreover, we assumed that linguistic transmission occurs between generations, ignoring the impact of iterated communication between individuals of the same generation. Moreover, we did not take into account global media as books, radio, internet, or television. We should thus consider in future investigations several alternative models of language evolution, where the acquisition of language results from a series of interactions between individuals rather than from a unique transmission event.

Finally, note that the formalism of our models are close to the formalism of population genetics. This should allow proposing joint inferences coupling genetic and linguistic data for the same set of populations and individuals, but some theoretical limits remain. We may wonder whether a speech community (a “linguistic population”) is identical to a reproductive group (a “genetic population”). It is far from obvious that human reproductive boundaries overlap language boundaries among human groups. A joint model between genetics and linguistics should then request clarifying and

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

articulating rigorously the concepts of population genetics with the concepts of population linguistics to propose robust joint inferences.

7. Acknowledgements

We thank the Genotoul bioinformatics platform (Toulouse, Midi-Pyrenees) for providing help, computing and storage resources. V.T. was financed by a PhD grant from the French 'Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche'. V.T. and F.A. received a travel grant from the NEFREX project funded by the European Union (People Marie Curie Actions, International Research Staff Exchange Scheme, call FP7-PEOPLE-2012-IRISES). This work was also partially funded by the Agence Nationale de la Recherche grant DemoChips (ANR-12-BSV7-0012).

8. Bibliography

Atkinson, Q.D. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. Science *332*, 346–349.

Baxter, G.J., Blythe, R.A., Croft, W., and McKane, A.J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. Language Variation and Change *21*, 257.

Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. Genetics *162*, 2025–2035.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. Language Learning *59*, 1–26.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A., and Atkinson, Q.D. (2012). Mapping the Origins and Expansion of the Indo-European Language Family. Science *337*, 957–960.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Breiman, L. (1999). Random forests. UC Berkeley TR567.

Croft, W. (1996). Linguistic Selection: An Utterance-based Evolutionary Theory of Language Change. Nordic Journal of Linguistics *19*, 99.

Croft, W. (2008). Evolutionary Linguistics. Annual Review of Anthropology *37*, 219–234.

Csilléry, K., François, O., and Blum, M.G.B. (2012). abc: an R package for approximate Bayesian computation (ABC): *R package: abc*. Methods in Ecology and Evolution *3*, 475–479.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In Proceedings of the 22nd International Conference on World Wide Web, (ACM), pp. 307–318.

Darlu, P., Bloothooft, G., Boattini, A., Brouwer, L., Brouwer, M., Brunet, G., Chareille, P., Cheshire, J., Coates, R., Dräger, K., et al. (2012). The Family Name as Socio-Cultural Feature and Genetic Metaphor: From Concepts to Methods. Human Biology *84*, 169–214.

Dryer, M.S., and Haspelmath, M. (2013). The World Atlas of Language Structures Online (Leipzig: Max Planck Institute for Evolutionary Anthropology).

Gray, R.D., and Atkinson, Q.D. (2002). Language-tree divergence times support the Anatolian theory of Indo-European origin. Geophysical Research Letters *29*.

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

Heeringa, W., and Nerbonne, J. (2001). Dialect areas and dialect continua. Language Variation and Change *13*, 375–400.

Kandler, A., Unger, R., and Steele, J. (2010). Language shift, bilingualism and the future of Britain's Celtic languages. Philosophical Transactions of the Royal Society B: Biological Sciences *365*, 3855–3864.

Kirby, K.R., Gray, R.D., Greenhill, S.J., Jordan, F.M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D.E., Botero, C.A., Bowern, C., Ember, C.R., et al. (2016). D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. PLOS ONE *11*, e0158391.

Krader, L. (1966). Peoples of central Asia (Indiana University [1966]).

Livingstone, D., and Fyfe, C. (1999). Modelling the evolution of linguistic diversity. Advances in Artificial Life 704–708.

Mennecier, P., Nerbonne, J., Heyer, E., and Manni, F. (2016). A Central Asian Language Survey. Language Dynamics and Change *6*, 57–98.

Nei, M. (1987). Molecular Evolutionary Genetics (Columbia University Press).

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

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pagel, M., Atkinson, Q.D., S. Calude, A., and Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. Proceedings of the National Academy of Sciences *110*, 8471–8476.

Pateman, T. (1983). What is a language? Language & Communication *3*, 101–127.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C.P. (2016). Reliable ABC model choice via random forests. Bioinformatics *32*, 859–866.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C.P., and Estoup, A. (2017). ABC random forests for Bayesian parameter inference. Peer Community in Evolutionary Biology 100036.

Reesink, G., Singer, R., and Dunn, M. (2009). Explaining the Linguistic Diversity of Sahul Using Population Models. PLOS Biology *7*, e1000241.

Rodriguez-Larralde, and Barrai (2000). Elements of the surname structure of Austria. Annals of Human Biology *27*, 607–622.

Tamariz, M., and Kirby, S. (2015). Culture: Copying, Compression, and Conventionality. Cognitive Science *39*, 171–183.

Tavaré, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. (1997). Inferring Coalescence Times from DNA Sequence Data. Genetics *145*, 505–518.

Thouzeau, V., Mennecier, P., Verdu, P., and Austerlitz, F. (2017). Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations. Proc. R. Soc. B *284*, 20170706.

Verdu, P., Jewett, E.M., Pemberton, T.J., Rosenberg, N.A., and Baptista, M. (2017). Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population. Current Biology *27*, 2529-2535.e3.

Vogt, P. (2009). Modeling interactions between language evolution and demography. Human Biology *81*, 237–258.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society of London B: Biological Sciences *360*, 1395–1409.

Wang, J. (2016). A comparison of single-sample estimators of effective population sizes from genetic marker data. Mol. Ecol. *25*, 4692–4711.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Median	**Min**	**Max**	**Quantile 2.5%**	**Quantile 97.5%**
N_0	550	100	1000	122	978
N_1	550	100	1000	122	978
t_1	500	0	1000	25	975
μ_L	3.165×10^{-4}	10^{-6}	10^{-1}	1.35×10^{-6}	7.73×10^{-2}
$N_0\times\mu_L$	0.150	10^{-4}	100	5.25×10^{-4}	44.5
$N_1\times\mu_L$	0.150	10^{-4}	100	5.25×10^{-4}	44.5
$t_1\times\mu_L$	0.116	0	100	2.80×10^{-4}	42.0

Table 1 – Summary of the prior distributions of the parameters for the four models.

Clonal	**Sexual**	**Sexual2**	**Social**
0.002	0.04	0.478	0.48

Table 2 – Proportion of votes for the four models of linguistic evolution.

	Expectation	**Median**	**Variance**	**Quantile 2.5%**	**Quantile 97.5%**
N_0	526	499	43331	126	968
N_1	645	714	65762	154	975
t_0	479	466	87448	21	937
μ_L	4.66×10^{-4}	3.23×10^{-4}	1.13×10^{-7}	2.18×10^{-4}	1.44×10^{-3}
$N_0\times\mu_L$	0.243	0.193	0.039	0.057	0.87
$N_1\times\mu_L$	0.255	0.244	4.10×10^{-3}	0.15	0.467
$t_1\times\mu_L$	0.239	0.177	0.064	8.092×10^{-3}	1.152

Table 3 – Summary of the posterior distributions of the parameters, assuming a *Sexual2* scenario.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Expectation	**Median**	**Variance**	**Quantile 2.5%**	**Quantile 97.5%**
N_0	544	542	60108	153	986
N_1	655	681	61907	148	966
t_0	353	290	109196	9	954
μ_L	4.26×10^{-4}	3.14×10^{-4}	1.03×10^{-7}	1.98×10^{-4}	1.28×10^{-3}
$N_0\times\mu_L$	0.203	0.175	0.028	0.074	0.553
$N_1\times\mu_L$	0.255	0.246	4.85×10^{-3}	0.122	0.432
$t_1\times\mu_L$	0.204	0.126	0.098	5.33×10^{-3}	1.09

Table 4 – Summary of the posterior distributions of the parameters, assuming a *Social* scenario.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

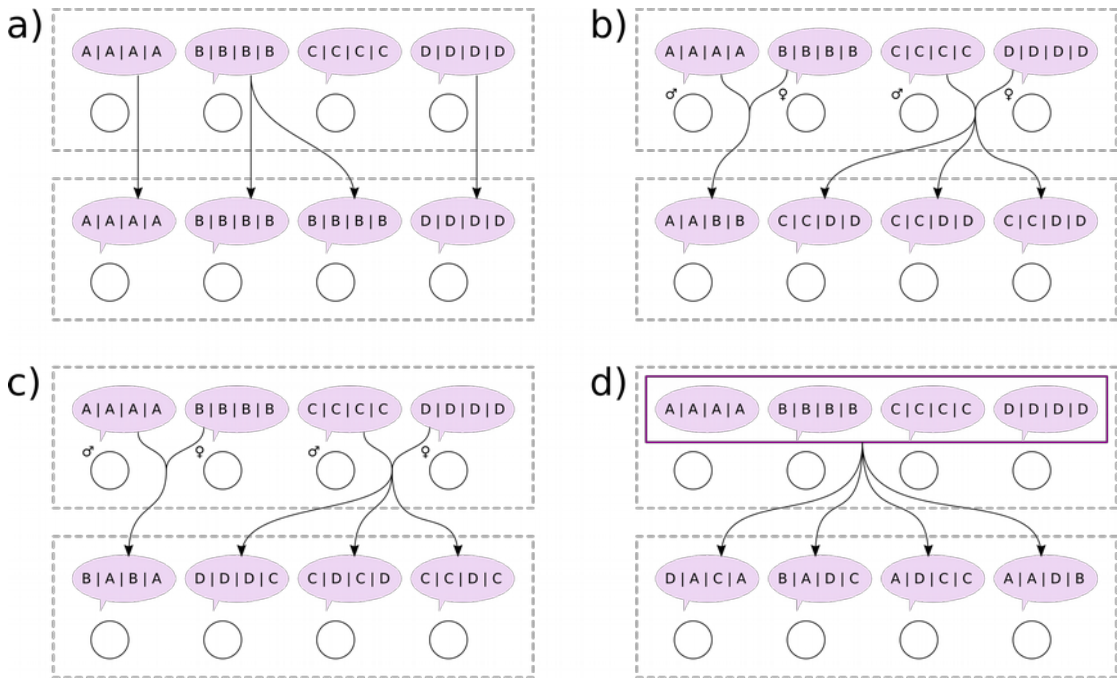

Figure 1 – Four models of linguistic transmission between generations. Each circle represents an individual. The utterances that individuals produce depend only on the utterances that their teachers produced at the previous generation, and on the mutations induced during the transmission. Four transmission modalities were considered: (a) a "C*lonal*" model with only one teacher per learner, (b) a "S*exual*" model with two teachers associated with a distinct set of vocabulary for each sex, (c) a "S*exual2*" model with two teachers without a distinct set of vocabulary for each sex, and (d) a "*Social*" model with the whole population as teacher for each learner.

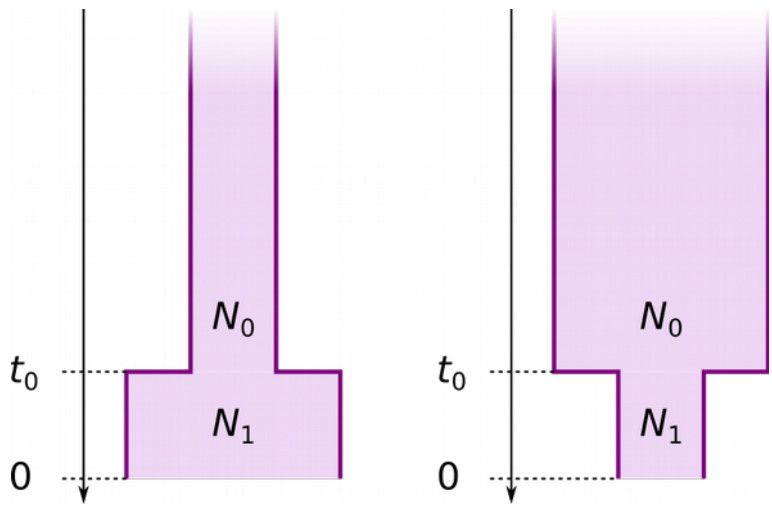

Figure 2 – Historical scenario. Its structure depends on the relative values of the parameters N_0 and N_1. If $N_0 = N_1$, we assumed a scenario of constant population size. If $N_0 < N_1$, we assume a scenario of expansion of the population. If $N_0 > N_1$, we assume a scenario of contraction of the population.

bioRxiv preprint doi: https://doi.org/10.1101/441246; this version posted October 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

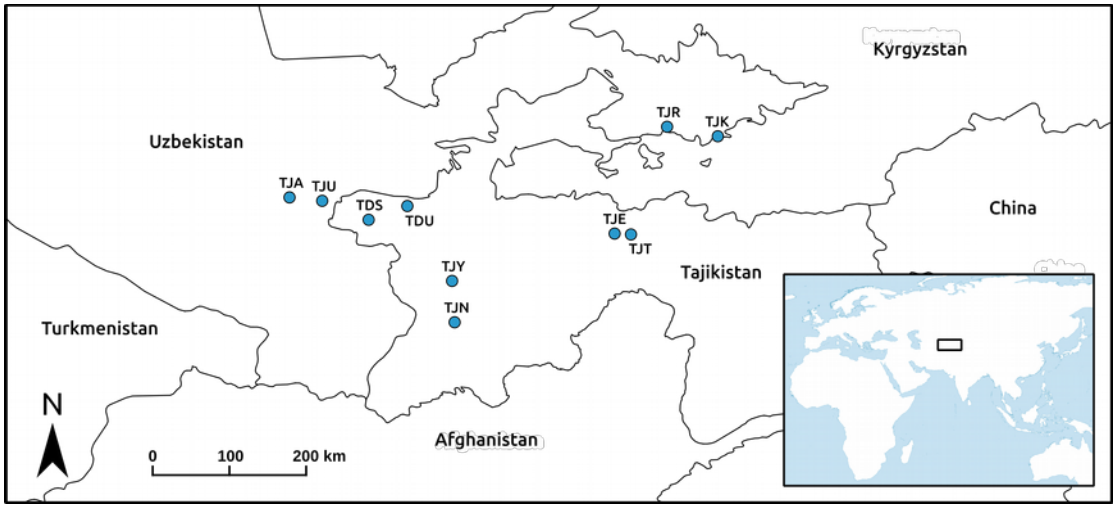

Figure 3 – Geographical distribution of the 10 sampled units under study.

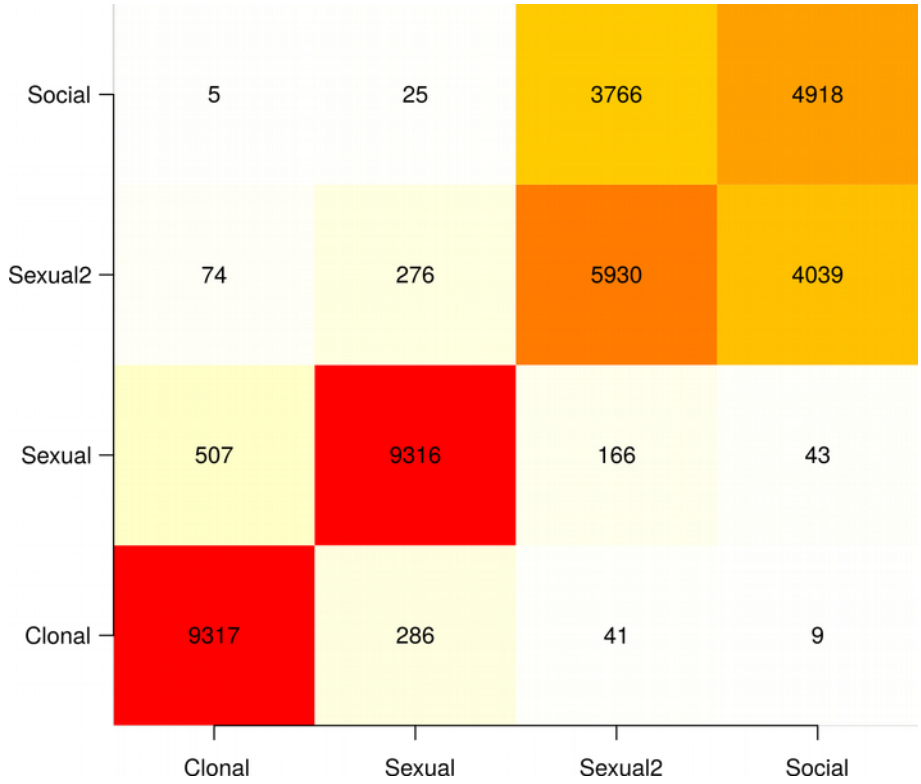

Figure 4 – Confusion matrices from the out-of-bag cross-validation analysis of the four models, using 10 000 pseudo-observed data.