Information-theoretic causal inference of lexical flow

Johannes Dellert

Language Variation 4

language science press

Language Variation

Editors: John Nerbonne, Dirk Geeraerts

In this series:

1. Côté, Marie-Hélène, Remco Knooihuizen and John Nerbonne (eds.). The future of dialects.
2. Schäfer, Lea. Sprachliche Imitation: Jiddisch in der deutschsprachigen Literatur (18.–20. Jahrhundert). Press.
3. Juskan, Martin. Sound change, priming, salience: Producing and perceiving variation in Liverpool English.
4. Dellert, Johannes. Information-theoretic causal inference of lexical flow.

ISSN: 2366-7818

Information-theoretic causal inference of lexical flow

Johannes Dellert

language
science
press

Dellert, Johannes. 2019. *Information-theoretic causal inference of lexical flow* (Language Variation 4). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/233

ISBN: 978-3-96110-143-6 (Digital)
978-3-96110-144-3 (Hardcover)

ISSN: 2366-7818
DOI:10.5281/zenodo.3247415
Source code available from www.github.com/langsci/233
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=233

Cover and concept of design: Ulrike Harbort
Typesetting: Johannes Dellert
Proofreading: Amir Ghorbanpour, Aniefon Daniel, Barend Beekhuizen, David Lukeš, Gereon Kaiping, Jeroen van de Weijer,
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

5 Simulating cognate histories

After some general remarks on the place of simulation studies in computational historical linguistics as well as the in-silico approach to evaluation, this chapter presents the simulation model which I am using in parallel to the NorthEuraLex data to evaluate lexical flow inference methods.

Unlike more detailed existing simulation models like Hochmuth et al. (2008), which have components for explicitly generating and modifying phonetic strings and modeling the geographical spread of languages, my model limits itself to modeling contact in the form of transmitting discrete units, i.e. it models loanwords on the level of cognacy, without generating actual phonetic forms. This creates data different from what historical linguistics would apply to decide whether lexemes were inherited or borrowed, but the shape of the data is exactly what will be needed to evaluate lexical flow inference algorithms.

5.1 Simulation and in-silico evaluation

5.1.1 Advantages and shortcomings of simulation

A simulation model is an algorithm which models the behavior of some real-world system, and uses randomness to generate output which is similar to the output of the real system. For instance, an adequate model of an economy should generate time series of measures such as interest rates, inflation, and unemployment which behave just as erratically as their real-world equivalents. An adequate model of tree growth should generate trunk shapes and branch structures which look just like the ones we can observe on real trees.

A very popular application of simulation models is as a way of testing assumptions about how the actual data can be explained. If our model generates data which are indistinguishable from our real data according to some relevant measure, we can take this as evidence that we have correctly understood and formalized an interesting aspect of the problem.

Within linguistics, this paradigm has previously mainly been applied to language competition. Schulze et al. (2008) give an overview of different attempts to let a distribution of language sizes (by number of speakers) emanate from minimalistic models, with the goal of mirroring the observable distribution as closely as possible. Many of the more successful models are agent-based, modeling individual speakers which can choose to take over the language of neighboring speakers based on a prestige value, or just the dominant language in the neighborhood.

Combining previous models for explaining the distibution of language sizes, de Oliveira et al. (2008) arrive at a model which captures the observable distribution of language family sizes. While the final model given in the appendix of the paper is rather simple, the authors report that in additional experiments, adding more complexity to the models (e.g. by including the effects of war and similar historical contingencies) did not have any influence on the overall good fit with observed family sizes. I will take this as an indication that attempting to include such effects into my model is just as unlikely to lead to different behavior, allowing me to keep my own simulation simple as well.

Establishing knowledge about real-world systems on the basis of simple simulations is frequently criticized as too reductionistic, and spectacularly wrong predictions resulting from simulation models may have contributed to wide-spread scepticism towards modern economic theories. The practical and far less controversial advantage of simulation models is that they allow us to generate arbitrary amounts of data to test algorithms on. This in-silico evaluation allows us to compensate for insufficient amounts of actual test data, or as additional cross-validation of models developed on (and optimized for subsets of) actual data.

This is the paradigm in which Embleton (1986) already worked. Her simulation model represents an early attempt to adequately model the influence of borrowing between neighboring languages on cognate-based phylogenetic inference. The model is similar in spirit to the one I am presenting in this chapter, in that it operates on the level of individual cognate replacement events. Unlike my model, it assumes that the process of language split can be modeled by recursive subdivision of a two-dimensional area, precluding the possibility of geographical spread. Also, borrowing events are modeled as independent, i.e. for every new borrowing event sampled according to a global borrowing rate, a language picks one of the neighboring languages as the donor language at random, whereas the model I present in this chapter samples an additional level of contact channels in order to mirror the strong tendency for loans to occur in 'packages' triggered by historical events.

Murawaki (2015) presents another approach which explicitly simulates the transmission of lexical items by borrowing among neighboring language varieties, but does not have a phylogenetic component. The structures produced are thus very similar to my concept of contact flow networks. Based on the cognacy overlaps resulting from simulation on different network typologies, Murawaki then performs phylogenetic analysis, with the somewhat surprising result that the phylogenetic signal tends to look tree-like on tree-like spatial structures, even if inheritance is not modeled. This could indicate that the usually very good fit of tree models does not necessarily have to result from tree-like evolution, but that wave-like change can just as easily lead to tree-like signals in certain geographical configurations, which has interesting implications for the debate between family tree and wave models of language change.

Unlike the other models, the rather complex simulation model by Hochmuth et al. (2008) generates phonetic data, to which the authors then apply modern standard tools for phylogenetic tree and network inference. While they find the amount of simulated lateral contact to have little impact on the performance of tree inference algorithms, the behavior of phylogenetic network algorithms is described as very erratic.

The main difficulty in using simulation models is that they are necessarily based on a set of assumptions about the nature of the data, which might not be true in reality. What if the way in which we generate data fails to capture an important case that occurs in real data, and is then not covered by the algorithm which we developed and tested on simulated data? To keep this problem under control, it is always best to evaluate a system both against simulated and real data. In the realm of causal inference, there has been a very strong tendency to develop the theory and algorithms either on very well-studied toy examples, or on massive amounts of simulated data. This makes it difficult to assess the performance of these methods on large real-world datasets, a problem that we are going to be faced with again when evaluating their potential for lexical flow inference.

5.1.2 Principles of in-silico evaluation

When assessing the performance of a heuristic algorithm (i.e. one without provable properties), the classical framework is to collect a set of gold-standard data, and to let the algorithm run on the data, comparing the output to the gold-standard using a useful definition of true and false positives and negatives, and then quantifying the performance in terms of precision and recall. Since gold-standard data are often difficult to acquire in large quantities (the last chapter provides a very good example of the efforts that may be required), the conclusions made from evaluating an algorithm on real data often rest on unstable grounds.

In the absence of a large enough amount of gold-standard data, one can use simulated data to get an impression of how the algorithm would perform on other data of the same shape. To get informative results, quite a bit of effort needs to be invested into developing a simulation model which is adequate for the purpose. The main requirements are that the model should not be overly complex in order to decrease the risk of overfitting the algorithm to certain (possibly hidden) properties of the gold-standard data. In a simulation model, it is always tempting to capture all aspects of the real data, but such an approach will often require many decisions to be made with inadequate backing in data or theory.

For instance, the actual linguistic history of a region is shaped by many historical events such as invasions, political ideas, technical innovations, and the shape of trade networks. A simulation model could try to emulate all of these phenomena in order to arrive at realistic simulated histories, and use these events to generate linguistic data. The problem is of course that such a model would require a very explicit (and formalized or at least quantitative) theory of political events, predictions about the conditions under which they will occur, and many other components which would quickly explode into separate research projects if we want to justify all of the myriad decisions which would be involved in designing such a model. Keeping the model complexity low, and the number of design decisions at a minimum, helps to avoid introducing too many unwarranted assumptions.

To configure the parameters even of a small model, it is good scientific practice to use structural features or at least statistics estimated from real data to increase (and quantify) the amount of realism. For instance, we might want to put data on historically observed unemployment rates into an economic model, to estimate how strong we expect oscillations in this measure to be in reality. For a simulation of language history, we will need to estimate (and inform the model) how often languages tend to split, and how intensively they can borrow from neighboring languages.

5.2 Generating phylogenies

A core component of any simulation model in computational historical linguistics (and also of some of the more advanced statistical methods) is a generative description of possible tree shapes. In statistical methods, these models are used to efficiently sample the space of possible trees in order to find good phylogenies. In in-silico evaluation, some part of the generated tree is removed from the input data for an algorithm which tries to reconstruct the missing information, and can then be evaluated against the truth.

Evolutionary models of species trees in biology as in linguistics are minimally based on two modules: the first one describes how languages or species split, and the second one models the process of languages or species becoming extinct. Moreover, if we explicitly model a genome, at least the possible mutations during inheritance need to be modeled. In all these respects, the simulation model presented here makes very simple assumptions in order to avoid dependence on too many parameters and choices. If even a simple model yields cognate histories which are interesting enough for evaluation, there is no reason to introduce additional complexity.

5.2.1 Models of lexical replacement

On the level of cognate sets, the central evolutionary process to take into account is the gradual replacement of existing words for many basic concepts with new lexical material. Semantic change is a phenomenon which appears to occur even in geographically isolated languages, and should therefore be modeled as a language-internal process. Internal replacement of words is also the main mechanism which makes the descendants of an ancestral language which has split dissimilar over time.

While some results suggest that semantic change happens more quickly for some concepts than for others (Pagel et al. 2007), and that these different rates of lexical replacement have cognitive correlates (Vejdemo & Hörberg 2016), for simplicity we will assume both that semantic change occurs at equal speed to all concepts, and that the rates are constant across languages. Any other design choice would lead to additional parameter settings which are difficult to motivate on the basis of available literature, and at the time depth of 5,000 years we will be simulating, the possible existence of ultraconserved words in real data is not much of an issue.

A constant replacement rate ρ is the only parameter which defines the behavior of lexical replacement. ρ defines the probability for a given word in a given language to be replaced by a word from a new cognate set during the current simulated year. We are therefore not simulating semantic change that would lead to loss of differentiation between concepts, and since we will only consider cognate sets for each concept separately, we also do not model the fact that the new cognate set for a concept might arrive there by extension from a different concept.

In the simulations I am running, the base replacement rate is set to $\rho := 0.00036$. This is the rate we arrive at if we assume a retention rate of 70% after 1,000 years. I am setting the retention a bit lower than the 81% derived for a 215-concept list by Swadesh (1955) to account for the unavoidable presence of lower-stability concepts in a list of 1,000 concepts. Most of the early assumptions about retention rates, especially their constancy across time, has been rejected in many individual cases by subsequent research. In a computational study covering three large language families, Atkinson et al. (2008) substantiate the suspicion that lexical change tends to occur in bursts, rather than gradually. In order to keep the number of parameters low, I still stick to Swadesh's original model while acknowledging that the assumption of a constant replacement rate across languages, concepts, and time is a very rough approximation to a much more complex reality. It is worth noting, however, that this does not mean I am assuming a constant effective replacement rate. As I am going to demonstrate, the combination of a Swadesh-style fixed base rate with a fully developed model of borrowing does yield quite realistic variation in the effective replacement rate. I therefore see no immediate need to further complicate the model by an additional parameter, even though the vast majority of contemporary models tend to treat the replacement rate as a further parameter that itself varies across time and across language varieties.

5.2.2 Simulating how languages split and die

With language-internal lexical replacement in the model, all that is needed for a basic model of linguistic evolution is a model of the process by which a language splits into several descendants, and the disappearance of languages. For a realistic model of splits, we would need an explicit geographical model where every language takes up a certain territory, where a larger territory would make it more likely for the language to develop dialects and then split. Moreover, the process should be modified by the possible presence of stabilizing factors such as states, and the ease at which people can migrate throughout their language community. For a good model of extinction, we would similarly have to model at least the effect of armed conflicts, competition between languages of different prestige and cultures at different technology levels, natural disasters, and assimilation processes within states.

Instead of trying to model all these details (a process which would again involve many decisions that are difficult to justify), we resort to a popular basic model of species evolution in biology. A branching process is a Markov process describing the development of a number of nodes each of which generates some number of children with a given probability at each discrete time step. If the possible values for the number of children are 0, 1, and 2, we are modeling a process where branches of the population can die out, stay at the same size, or multiply.

These three options are sufficient to generate any binary branching language tree. The simplest parametrization results when we set the probabilities for each number of children to $p(2) := \sigma$, $p(1) := 1 - \sigma$. In this formulation, we can interpret the parameter σ as the split rate, expressing how likely it is for a language to disintegrate into two separate languages.

To simulate how languages become extinct, we could simply assume an extinction rate δ, and delete during each simulated time unit each language from the tree with this probability. However, previous research summarized by Holman (2005) has shown that the distribution of language family sizes and tree shapes generated by such a branching process differs significantly from the patterns we observe in actual language trees.

To arrive at more realistic datasets, it seems necessary to adapt at least a very simple model of geography in order to simulate at least some of the effects of competition between languages, and the survival of remnants of older language families in isolated geographical positions, such as islands and mountain valleys. My central modeling assumption is that languages never disappear if left alone, but only because speakers of another language migrate and become dominant (e.g. English in North America), a state conquers a new territory and imposes its language on the newly acquired population (e.g. the Roman Empire in Gaul, or later colonial empires), or a population shifts to a more prestigious language for economic reasons (e.g. from Livonian to Latvian, and many other minority languages). The crucial point is that I will assume extinction to happen exclusively due to the spread of another language. Even if exceptions to this rule might exist, I consider this a much more sensible default assumption than to assume that some languages just happen to become extinct without being in contact with other languages.

My model subsumes all of these situations by having a language that splits expand into a neighboring territory, which might previously have been occupied by another language, which then becomes extinct during the process. That said, a splitting language will always prefer to spread into an unoccupied territory first, so that the map will tend to become filled with languages before competition and replacement sets in. To create geographical niches in which less frequently splitting families can survive longer (the Caucasus), and hub areas were languages tend to replace each other much more frequently (the Steppes), only a randomly shaped island or continent of about half the size of a square grid of cells is treated as occupiable territory. The neighborhood relation is defined primarily by adjacency, but it also connects diagonals (i.e. a language can have up to eight neighbors). Many of the random continent shapes will feature drawn-out peninsulae with only one access point, or landbridges which serve as bottlenecks for expansion.

When creating a scenario, between two and ten initial language families (all unrelated, i.e. with cognate sets modelled as completely independent) are put in random positions on the landmass, and an overall split probability σ is selected uniformly from the range $[0.0004, 0.0001]$. The purpose of varying σ is to emulate the consequences of overall political instability or a geography prone to migrations in a single parameter that may vary between scenarios. For the simulation study, we will be operating on grids of 10 × 10 cells, with a random connected landmass occupying 50 of the 100 tiles. This means that only 50 languages can exist at any given time, a number which is in the tractable range for the algorithms I will develop. Depending on the split rate σ, many extinct languages and a very complex contact history can be hidden behind the final set of observable living languages.

5.3 Modeling lexical contact

5.3.1 Modeling the preconditions for contact

The simplest possible contact model would just establish contact between any pair of living languages with a small probability per simulated year, and would let contact break down again after a random number of years. Initial explorative analysis of such a model quickly showed that the possibilities for contact should be influenced by a model of geographical proximity, which is trivially given by our model which assigns a single cell to each language, in a grid which defines a neighbor relation between speakers or geographical positions which can be occupied by languages.

This type of geographical constraint also appears obvious on the basis of general considerations. If we imagine a historical contact situation were words were exchanged, the prototypical cases would be people from neighboring villages who meet and transfer the words for new concepts that the neighboring culture does not yet have. The more long-distance influences which happened through trade typically had much influence on the technological and sometimes cultural vocabulary, but did not tend to influence the basic vocabulary so much that we would necessarily have to model it.

5.3.2 A monodirectional channel model of language contact

Most contacts between languages which have severe consequences for one of the languages involved are monodirectional. While some lexical material might be mutually exchanged, e.g. to talk about trade goods derived from different modes of subsistence in different climate zones (as was the case for Nenets reindeer herding and Komi agricultural vocabulary), if the basic vocabulary is affected, this typically entails that one language is in a dominant position, and the other language is heavily influenced by the other. As Sankoff (2001) puts it, "language contacts have […] taken place in large part under conditions of social inequality resulting from wars, conquests, colonialism, slavery, and migrations — forced and otherwise". Typical examples of this within the NorthEuraLex sample are the contact of a technologically advanced civilization with a less advanced ethnic group (e.g. Chinese influence on Mongolian), the language of a conquering elite influencing the language of a population they control (e.g. Norman French and English), or both (colonial languages, like English influence on Hindi).

These general observations imply that monodirectionality is a reasonable default assumption for lexical flow affecting basic vocabulary. My simulation model emulates the window of time in which one language dominates and influences another by generating directed channels through which lexical items may flow at a certain rate, inheriting the channels through splits by handing them on to the daughter language which stays in place, and closing the channels again after some time. This does not make it impossible to model the historically rare case where two languages exchanged large amounts of lexical material on a relatively equal footing, as e.g. resulting from intensive trade contact of neighboring cities. The simulated histories will occasionally include such situations as well, since there is nothing to prevent that two monodirectional lexical transfer channels in reverse directions will be opened independently.

5.3.3 Opening and closing channels

The probability $\alpha_t(l_1, l_2)$ of a channel opening from language l_1 to l_2 can simply be modeled as dependent on the neighborhood relation. We could assign $\alpha_t(l_1, l_2) := \alpha(\|(x_t(l_1), y_t(l_1)) - (x_t(l_2), y_t(l_2))\|)$ for any function α assigning channel opening probabilities to any distance. In the current implementation, however, I am simply drawing a global α value for each scenario from a uniform distibution over the interval $[0.0001, 0.0003]$, and set $\alpha(l_1, l_2) := \alpha$ whenever l_1 and l_2 occupy neighboring cells, and to $\alpha(l_1, l_2) := 0$ for languages with a distance of more than one cell. To make this number easier to grasp, it means that if we have 49 living languages filling a square of 7 × 7 cells (the most compact configuration which can result from the simulation model), we expect between 0.0156 and 0.0468 new contacts channels to be opened during each simulated year, i.e. a new contact every 21 to 64 years. To justify these rates, it is necessary to compare them to the number of contacts arising in a similar cluster of real languages in a geographic area. One obvious option to do this is to reconsider the NorthEuraLex gold standard from the last chapter. Taking the gold standards for a low-contact area (Siberia) and a high-contact area (the Caucasus) together, we have the very convenient number of 44 languages in the set we consider. The gold standard contains 89 contacts which shaped the history of these languages during the past 3,000 years, with older contacts mostly being under the detectability threshold. On average, a new contact has therefore opened every 33.708 years, which is well within the range determined for a slightly larger language sample in the simulation model.

While a channel persists, lexical material will be transmitted from the donor language to the recipient language at a certain rate, randomly replacing cognate sets for different concepts. After some amount of simulated years, the channel might break down, and lexical influence might cease. The most obvious historical parallel to this is if the speakers of one language moved away from the speakers of a contact language, which is likely to decrease the intensity of contact between the speakers. A case in point would be the Turkic influence on Hungarian. Hungarian was subject to a lot of lexical influence from Bulgar or a related Turkic language on its way towards Middle Europe, and there was some additional (though far weaker) influence during the Turkish occupation of large parts of Hungary in the 17th century. For the past three hundred years, the Hungarians have not been neighbors to any Turkic nation, which has caused the lexical flow from Turkic to Hungarian to stop completely.

The straightforward idea resulting from these considerations is to model the closing of channels in a very similar framework to their opening. Again, we define a contact breakoff probability $\omega_t(d)$ which could be dependent on the geographic distance d at time t. For the simulated language histories generated at the end of this chapter, I am using a constant $\omega := 0.002$, i.e. each contact is expected to last for 500 years on average. Letting the duration of contacts vary is motivated by the different duration of real-word causes of language contacts, such as the frequency of migrations, the existence of states, or the stability of colonial rule. Given that during the time contact is established, its duration is not yet known, it makes sense to decide randomly on a year-to-year basis whether a contact persists. Simulating this with a constant rate that does not depend on any other factors is an answer to the basic requirement of keeping the number of parameters low. The choice for the value of ω is a little harder to justify than others, because it interacts heavily with subsequent choices for simulating channel behavior. Ultimately, the value mostly influences how many of the generated contacts will break down before they had the chance to leave noticeable traces. Since it would be much more economical to just open fewer contact channels instead of opening many which do not result in any borrowings, it makes sense to set ω low enough for most contacts to actually have visible consequences. On the other hand, if too many contacts persist for thousands of years, we risk creating the unrealistic scenario of a language's lexicon becoming almost completely replaced by that of a neighboring language, due to our not simulating differences in concept stability. Using these two constraints and experimenting with different values for ω, the chosen value yielded a good compromise where most contacts have noticeable consequences, while still leaving a large part of the recipient language's basic vocabulary intact.

5.3.4 Simulating channel behavior

To simulate the behavior of a channel, we simply transmit every word for each concept with a given probability. There is thus no notion of more or less stable concepts, and we also abstract away from the layered structure of loanwords (where e.g. month names are usually borrowed as a package). The only complex decision remaining is how to determine the strength $\tau_t(l_1, l_2)$ of the channel, which will influence the rate of transfer $\beta_t(l_1, l_2)$. The design decision in the model presented here is to generate a constant strength $\tau_t(l_1, l_2)$ for each channel when it is created, again dependent on the distance d of the languages at that time. In my implementation, $\tau_t(l_1, l_2)$ is only changed when a new channel is established, and is set to $\tau_t(l_1, l_2) := (1 - d) \cdot X$ for a random variable X that is uniformly distributed over $[0, 1]$. In the current implementation, the relation from channel strength to transfer rate is $\beta_t(l_1, l_2) := 0.01\tau_t(l_1, l_2)$. For languages with 1,000 basic concepts, this effectively sets the maximum possible transfer rate (for $X = 1$ and $d = 0$) to 10 loanwords per simulated year. This maximal rate makes it possible to generate very strong superstrate influences which occur within few generations, such as the introduction of the Norman French layer into English. At a more typical rate of 1 loanword per year, we expect that during an average contact lasting 500 years, about 40% of the recipient's lexicon will be replaced.

5.3.5 Overview of the simulation

To summarize, Algorithm 1 again specifies the entire simulation procedure in pseudocode. As is evident from the discussion above, very simple choices were made for the majority of the many parameters we introduced, although the simulation model could be made more complex in many places, leaving some potential for increasing the model's realism as additional quantitative results become available.

5.4 Analyzing the simulated scenarios

The final step towards establishing the quality of the simulated data as a test set is to inspect the results a posteriori, and see in how far they show the desired properties of being similar to the real data, while still displaying structural variability.

To make the inspection and tracing of simulated language histories easier, the following naming convention was adapted: identifiers of living languages start with a capital `L`, whereas dead languages have a `D` in that position. The second position in a language name is occupied by a numeric phylum ID, i.e. the independently generated ancestor language. Languages with an identical phylum ID are thus deeply related, whereas similarities between languages with different phylum IDs can only be explained by contact. The remainder of the language ID encodes the true phylogenetic tree by appending to the parent's name a pair of different random vowels or random consonants in order to produce the names for the two children resulting from a split event. As a result, we can tell at a glance that `D1fab` is a common ancestor of `L1fabu`, `L1fabewi`, and `L1fabexizo`, as well as a sister language of `D1faw`, and a descendant of `D1fa`.

To give an impression of the kind of histories arising from the simulation model, Figure 5.1 shows the trees of two language families in contact, where, just as in the gold standard visualizations, contact channels are represented by green arrows, and inheritance relationships are represented by black arrows. The thickness of arrows represents the size of the lexical contribution from each source. For instance, the language `L1ra` has diverged quite a lot from its sister language `D1re` due to being heavily influenced, among others, by `L0z` and `L1fevah`. In contrast, the layer of loans from `L1fabewu` in `L1fabexizo` is not very large.

To showcase the geographic model, the maps in Figure 5.2 shows the final positions of each language for the same scenario, as well as the state of the simulation after 1,800 and 3,400 simulated years (i.e. 3,200 and 1,600 years before the final

Algorithm 1 simulate_network($k, t_{max}, n, \rho, \delta, \sigma, \alpha(d), \omega(d), \beta(\tau)$)

```
$\mathscr{L} := \{(w_{i1}, \dots, w_{in}) \mid 1 \leq i \leq k\}$, ($k$ proto-languages of random words for $n$ concepts)

$t := 0$
while $t < t_{max}$ do
    for each $L \in \mathscr{L}$ do
        if $rnd() < \sigma$ then
            $L_1 := copy(L), L_2 := copy(L)$
            $\mathscr{L} := \mathscr{L} \cup \{L_1, L_2\}$
            $living(L) := false$
            $pos(L_1) := pos(L)$
            if $pos(L)$ has unoccupied neighbor $newpos$ then
                $pos(L_2) := newpos$
            else if $pos(L)$ has neighbor $newpos$ occupied by $L_3$ then
                $pos(L_2) := newpos$
                $living(L_3) := false$
            end if
        end if
    end for
    for each $L_i \in \mathscr{L}$ where $living(L_i)$ do
        for $1 \leq x \leq n$ do
            if $rnd() < \rho$ then
                $w_{ix} := w_*$ for a new cognate ID $w_*$
            end if
        end for
    end for
    for each $L_i, L_j \in \mathscr{L}$ where $living(L_j)$ do
        if $\tau(L_i, L_j) > 0$ and $rnd() < \omega(d(L_i, L_j))$ then
            $\tau(L_i, L_j) := 0$
        else if $\tau(L_i, L_j) = 0$ and $rnd() < \alpha(d(L_i, L_j))$ then
            $\tau(L_i, L_j) := rnd() \cdot (1 - d(L_i, L_j))$
        end if
    end for
    for each $L_i, L_j \in \mathscr{L}$ where $living(L_j)$ and $\tau(L_i, L_j) > 0$ do
        for $1 \leq x \leq n$ do
            if $rnd() < \beta(\tau(L_i, L_j))$ then
                $w_{jx} := w_{ix}$
            end if
        end for
    end for
    $t := t + 1$
end while
return $\mathscr{L}$
```

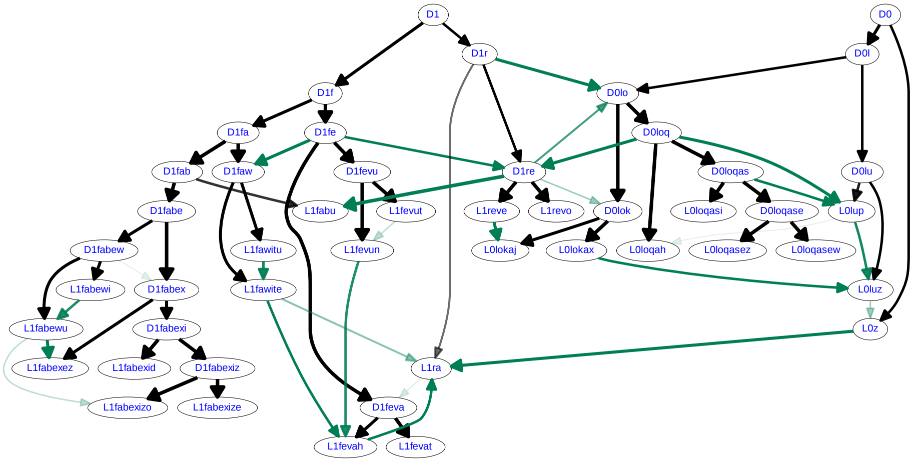
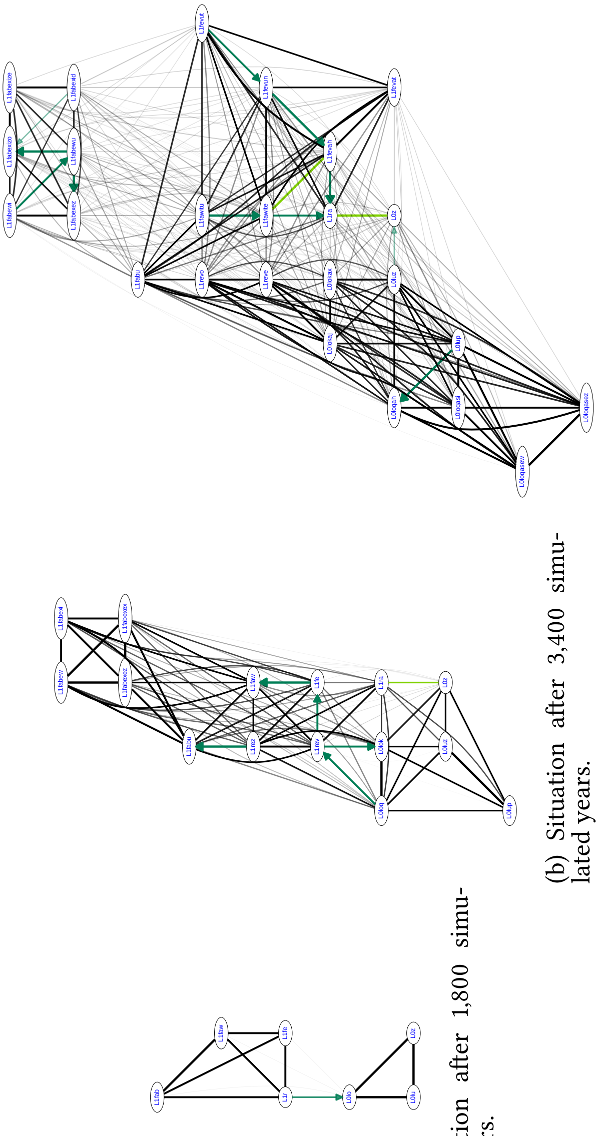


Figure 5.1: An example of a simulated scenario, with complex interactions

state). The three maps give an impression of how the initial population splits to fill the available space, and what type of contact is simulated by the model. Open contact channels are visualized in dark green (monodirectional contact) and light green (bidirectional, i.e. contact channels in both directions are open). The thickness of these lines represents the intensity of the contact, i.e. the rate at which lexical material is transmitted across each channel.

As in the cognate overlap maps used to visualize the shape of the inference problems in Chapter 4, the thickness of the black lines visualizes the strength of cognate overlaps between living languages at the respective point in time. In the comparison between the three stages, it becomes very obvious how some lines which were still strong thousands of years ago have faded into the background, reflecting the loss of similarity caused by lexical replacement.

To generate the evaluation set, a total of 50 scenarios were generated by simulation. Taken together, the simulated data contain 2,139 living languages distributed over 297 language families. In addition, a total of 7,128 intermediate (proto-)languages was modeled while producing the data for the living languages. In total, while generating the history of the languages, a total of 2,250,891 borrowing events were generated and logged, of which 380,571 events (16.9%) turn up in the stored etymologies of one of the 2,139,000 lexical items in living languages.



(a) Situation after 1,800 simulated years.

(b) Situation after 3,400 simulated years.

(c) Final situation after 5,000 simulated years.

Figure 5.2: Maps visualizing the situation of the example scenario at three points in time

In order to decide in which respects the simulation model does make sense (or not), the next section addresses the question whether the simulated data are similar enough to the NorthEuraLex data to be able to serve as additional test cases for validating my results. The section thereafter will answer the question whether the generated histories are non-trivial enough to provide some challenge to phylogenetic and lexical flow inference methods, and also varied enough to cover a wide array of situations we would expect to be faced with in actual linguistic histories.

5.4.1 Are the scenarios realistic?

The most trivial question to ask about the realism of the simulated histories is whether the distribution of cognate class sizes is similar enough to the one inferred from the NorthEuraLex data. The average size of cognate sets in the NorthEuraLex data is 2.253, whereas the simulation produces scenarios with average cognate sizes 2.164 ± 0.192, with the maximum being 2.656, and the minimum 1.801. This shows that the simulated cognate sets are similar in the size to the ones inferred by NorthEuraLex, indicating that the types of overlaps and therefore the information geometry will behave similarly. But since the average size of cognate sets heavily depends on the number of languages in the dataset, it might be more relevant to compare the average number of cognate sets per concept per language, and compare this measure across scenarios. In the simulated data, this number varies around 0.577 ± 0.095, with the minimum at 0.388 and the maximum at 0.800. The equivalent measure computed from the inferred correlate sets in NorthEuraLex is 0.497, again fitting very well into the distribution of simulated scenarios. Finally, Figure 5.3 shows the distribution of cognate class sizes in the simulated data next to the one for the classes automatically inferred from NorthEuraLex. It is clearly visible that both distributions are very similar, except for a much higher ratio of two-element cognate classes resulting from automated cognate detection. This fits well with the observations made when inspecting the inferred classes for FISH, and is very likely an artefact of UPGMA clustering, which is sensitive to spurious pairwise similarities between elements which should form singleton classes. This difference also becomes visible when fitting Pareto distributions to the observed counts. The maximum likelihood estimates of the alpha parameter are $\alpha = 0.284$ for the NorthEuraLex data, but only $\alpha = 0.189$ for the simulated data.

To get a first impression of how realistic the amount of contact in the simulated data is, we can compute the percentage of words in the final data which were borrowed at some point in their history. Across all scenarios, 23.92% of all attested

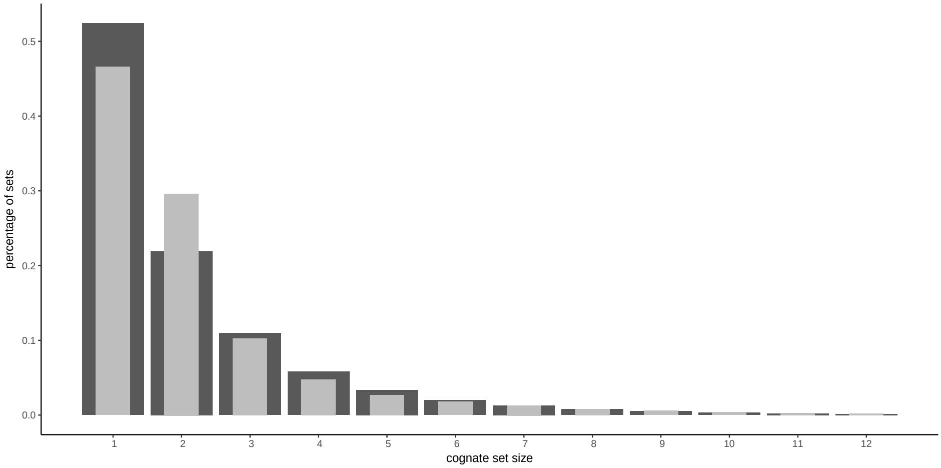


Figure 5.3: Distributions of cognate class sizes on NorthEuraLex (gray) and simulated data (black)

words have at least one borrowing event as part of their history. Between scenarios, the ratio varied between 1.27% and 38.53%, with the mean at 23.07 ± 8.71%. This is very much in line with the numbers derived from the WOLD database and summarized in Tadmor (2009), where the ratio of loans in content words varied between 1.3% (Mandarin Chinese) and 65.6% (Selice Romani), and the average ratio of loans across all languages was 24.2%. Therefore, both the overall frequency of loans and the variance of the ratio across simulated languages and scenarios are very realistic.

With respect to lexical replacement, we can compute the distribution of word ages from the logged histories. For each word in a living language, we can trace back the history to the point where the word came into existence as a word for the concept in question by a replacement event (a mutation, in biological terms), or back all the way to one of the initial languages. Across all scenarios, 16.93% of the words could be traced back to one of the initial languages. Unsurprisingly, the distribution almost exactly fits an exponential curve with our loss rate of about 0.036% per year, leading to e.g. 3.42% of words older than 4,500 years, 6.83% between 2,000 and 2,500 years old, and 16.29% younger than 500 years. The language-wide averages of word ages were distributed around 2357.43 ± 133.84. I have found it impossible to find even rough cross-linguisitic estimates for the word age distribution across entire lexica of modern languages. Still, we can use some etymological resources to get a first impression whether the numbers seem realistic. Sammallahti (1988a) provides the most up-to-date overview of the known lexicon of Proto-Uralic as well as some later proto-languages of branches such as Finno-Ugric. He counts 124 Uralic stems as being reconstructable for Proto-Uralic (perhaps 7,000 years ago), about 290 additional ones for Proto-Finno-Ugric (5,000 years ago), and 150 more for Proto-Finno-Permic (a contested subgrouping of perhaps 4,000 years). Janhunen (1977) reconstructs about 700 stems in total for Proto-Samoyedic, at about 2,500 years of age. A typical dictionary of a fully known language that is sorted by lexical roots, such as de Vaan (2008) for Latin or Lehtisalo (1956) for Nenets, covers the history of about 2,000 roots. We can thus assume that this is roughly the number of etyma which we can assume for an unwritten language. Fitting an exponential curve to these five data-points, our replacement rate of 0.036% per year, or about 30.24% per millennium, fits almost perfectly. Calculating the distribution of word ages across this curve (and counting every word older than PFU as 5,000 years old in the same way as I need to do it for the simulated data), we arrive at a mean age of 2317.5 years, which again fits very well into the distribution of values derived from the simulated data. The distorting effect of borrowings on reconstructability therefore seems to have a very negligible influence.

Another interesting question to ask about the distribution of word ages is whether there are conservative languages which tend to conserve more ancient words across the basic lexicon, whether due to limited lexical contact, or an inherent tendency to slower lexical replacement. This is equivalent to the question whether different languages can have vastly different average replacement rates across a timescale of thousands of years. While there is considerable evidence that replacement rates can vary a lot on the short term (Atkinson et al. 2008), it is unclear whether these differences are random fluctuations which might equal out with time, or whether they are inherent properties of language systems which remain in place for millennia. On the simulated data, while the underlying model of divergence operates at a global replacement rate for all languages, replacement by borrowing leads to very different retention rates. In the most conservative simulated language, the average word age was 3475.92, as opposed to 1892.89 in the language with the least stable vocabulary. This shows that a mixture of a variance in global replacement rate for each scenario already leads to an interesting and realistic range of measurable conservativity in the simulated languages.

The next relevant point of comparison is the shape of the trees. An essential observable property of binary trees is their balancedness, which can be defined in many different ways, each capturing a different type of asymmetric behavior. Holman (2005) analyzes linguistic trees in this framework, making the very interesting observation that language trees have different structure than would be generated by simple birth and death processes, even if we allow different diversification rates for each branch. Holman uses the weighted imbalance score by Purvis et al. (2002), to show that actual language trees are much more unbalanced than such a model would predict. The imbalance score for a binary node l with children l_1 and l_2 is computed as $I(l) := \frac{B - \lceil S/2 \rceil}{S - \lceil S/2 \rceil - 1}$, where S is the number of nodes on the subtree under l, and B is the maximum of the sizes of the two subtrees under l_1 and l_2. $I(l)$ will be 0 for a maximally balanced node (in the sense that a node with this number of descendants could not be more balanced), and close to 1.0 if one of the children is a leaf and many other nodes are descendants of the other child. For an entire tree, the weighted imbalance score is a weighted mean over these node-based scores, where the weight $w(l)$ is 1 if S is odd, $w = (S - 1)/S$ if S is even and I > 0, and $w = 2(S - 1)/S$ if S is even and $I = 0$ (such that completely balanced nodes count twice). This score is defined in such a way that the expected value for trees generated by a birth and death process is 0.5. As Holman shows, the weighted imbalance scores for actual language trees are significantly higher than that, clustering around 0.7. The simulated trees have weighted imbalance scores between 0.587 and 0.820, with the mean at 0.698 ± 0.053, which fits Holman's results surprisingly well. It seems that a simple death-by-replacement model on a constrained geography is all that is needed to explain the imbalance in empirically observed trees, without any need to allow for branch-specific diversification rates or similar devices.

There are some additional phenomena which are so commonly observed in historical linguistics that realistic datasets should contain some instances of them. One of these are isolates, which can here simply be defined as phyla with only one surviving descendant language. According to the Glottolog classification, roughly half of the world's families are isolates, and this is also roughly the ratio of isolates in NorthEuraLex (9 out of 21 families). While isolates are not interesting for evaluating phylogenetic methods, they can still be involved in some interesting contact scenarios, and should therefore be present at least in a few scenarios. So does the simulation model, where isolates can only occur if one of the initial languages never splits during 5,000 simulated years, or if all but one of the potentially many languages from a family are replaced by neighboring families branching and expanding into its territory, produce a significant number of isolates? Across all scenarios, only 16 of the 270 families generated by the simulation model are isolates, a number which is unrealistically low compared to the large numbers of isolates we observe in many regions of the world. While this is not a problem for evaluation purposes (as long as some isolates are present), it might still be worthwhile to speculate why so many more isolates occur in reality. One possible explanation is that the over-simplified geography (without remote mountain valleys) does not generate enough niches for smaller families to survive. Moreover, the constraint of having exactly one language per place will counteract the arising of true isolates. In reality, if two villages with closely related dialects of an isolate are surrounded by completely unrelated languages, the two villages might prefer close contacts among each other, counteracting the divergence into separate languages.

As a final point, a problem of realistic complexity should contain substrates. To recapitulate, a substrate relationship is one possible result of language shift, when the speakers of one language rapidly shift to another language, but the shift is incomplete in leaving traces of the ancestral language (a substrate) in the new language. In historical linguistics, the term *substrate language* is often used in the sense of an otherwise unattested language whose existence can only be reconstructed from a layer of words which have no etymology in the respective family. With access to the simulated history of each word in a large datastructure, it becomes possible to compute the ratio of words which were borrowed at some point from a substrate language, and the number of languages without living descendants which became sources of such borrowings. Analyzing the histories of all words in the simulated dataset, we find that 5.04% of all words have a substrate history by our definition. Averaging across scenarios, 65.16 ± 37.05 languages, i.e. slightly less than half of the 142.56 ± 62.66 extinct languages, played the role of a substrate language during the history of at least one word. Of 78.62 ± 37.41 contacts, 26.14±17.24 occur with a donor language which leaves no living descendant in an average scenario. It is difficult to assess how realistic these numbers are, because in the real world, words without an etymology are often difficult to attribute to one common substrate donor. Still, there tend to be many instances of unknown substrates even in the history of a very limited linguistic region such as Northern Europe. The most famous instance is an unknown substrate in Germanic consisting of mainly maritime vocabulary such as **strandō* 'beach' and **seglan* 'sail'. Originally assumed to comprise a third of the common Germanic lexicon, it has been shrinking in size as additional Indo-European etymologies for Germanic lexemes are being established. While some scholars like Hawkins (1990) continue to advocate it, it now seems to be on the way to becoming a minority position. Less contested instances of substrates in the North are a layer of pre-Uralic lexical material in the Saami languages (Aikio 2004), and a different pre-Uralic substrate which heavily influenced the Samoyedic languages (e.g. Helimski 1998). Given the ubiquity of such examples even in an area with a rather short post-glacial settlement history, it makes sense to have substrate relationships occur so frequently in the simulated data.

5.4.2 Are the scenarios interesting?

The second question about the adequacy of the simulation model is whether the simulated scenarios are difficult and varied enough to make the results of evaluation interesting. To answer the first question, this section analyzes how well the tree signal is recoverable from the overlap of cognate classes alone, and, answering the same question from a slightly different angle, how well the cognate class boundaries coincide with phylogenetic units. Then, the identifiability of contact events is analyzed in order to quantify the maximum performance that we could hope an ideal system to achieve on the lexical flow inference task. For an answer to the second question, the section goes through some phenomena that we might expect to occur in actual linguistic histories of geographical areas where several language families are neighbors for several millennia, and discusses to what extent these phenomena also occur in the simulated data.

A very direct way of assessing the difficulty of phylogeny inference is to measure the recoverability of the tree signal from the cognate data. If the cognate overlaps perfectly encode the tree structure, we should have $|c(A,B)| > |c(A,C)|$, $|c(B,C)|$ in all configurations where A and B are more closely related than either is with C, or more precisely, if the lowest common ancestor of A and B is a descendant of the lowest common ancestors of A and C as well as B and C. In the simulated scenarios, the cognate overlaps match the criterion for 89.32 ± 6.87% of such triples. On the automatically inferred cognates derived from NorthEuraLex and the reduced Glottolog tree, the value is 82.28%, i.e. comparable in complexity to the more difficult simulated scenarios. In the most difficult scenario, the inequality only holds for 64.91% of triples, and the easiest scenario (with only one contact in 5,000 years) has 100%. Given the presence of errors in automated cognate judgments, it is hardly surprising that the NorthEuraLex task is on the more difficult end of the scale.

A more strict measure of the difficulty of the inference task is the fit of cognate set boundaries to phylogenetic units. More precisely, we are interested in the percentage of cognate sets which exactly correspond to the descendants of a single phylogenetic node. Note that this correspondence is not only destroyed by borrowing, but also by lexical replacement in one language of the unit, though the latter situation will produce a new cognate set which is aligned to a phylogenetic unit of trivial size. On the simulated scenarios, the distribution of this percentage can be summarized as 17.33 ± 4.98%. The value of 10.06% on NorthEuraLex cognates and the Glottolog tree is close enough to this distribution to provide additional evidence that the simulated scenarios are quite realistic in difficulty. However, the NorthEuraLex data are by this measure more challenging even than the worst of 50 simulated scenarios at 10.28%. This is again easily explainable by the existence of erroneous automated cognacy judgments, as both false positives and false negatives will destroy perfect alignments of cognate classes with phylogenetic units. On the tree signal, the errors apparently almost cancel out, not detracting much from recoverability, whereas the matches of entire cognate sets are much more sensitive to uncertain cognacy judgments. This has problematic implications for algorithms building on cognate set overlaps, such as the lexical flow inference procedure I will be exploring.

An important question to ask about the data concerns the identifiability of contact events. How many of the contact channels still have visible consequences in the living languages, in the sense that they are part of the histories of enough words in living languages to go beyond the detection threshold of 20 loans? This measure gives us an upper bound on the performance we could hope to achieve with any algorithm which reconstructs previous reflexes of cognate sets and then tries to infer contact events. In the simulated data, 67.6% of all simulated contacts were still visible by this definition, resulting in an average of 49.68 ± 23.16 detectable contacts per scenario. This number implies that certainly enough interesting contact patterns occur in the simulated data, and the consequences of two thirds of these are still in principle detectable in the output data.

Finally, a few words can be said about the variability of patterns encountered in the simulated data. Using graph terminology, there are 27,164 unshielded triples in the gold standard flow graphs. 15.05% of these are colliders, where a single language is influenced by two different languages. 550 (13.45%) of these colliders connect languages which have not interacted in any way before. As we will see in the next chapter, this type of collider is easily detectable using v-structure tests, giving us very reliable directionality information. The problem is that out of 12,179 lateral connections whose directionality we will want to determine, only 870 take part in such strict colliders, and 4,088 lateral connections are part of any type of collider. This means that we only have completely reliable directionality information for 7.14% of these connections, and some direct evidence for 33.6%. The directionality of two thirds of our lateral links would thus have to be determined by some form of constraint propagation, which will cause some problems if our unshielded collider tests are not completely reliable. To sum up, the simulated instances are definitely challenging enough to provide useful information about the potential of causal inference for lexical flow inference.

Coming to phenomena of more immediate linguistic interest, we first consider chains of loans, where lexical material is transmitted from one language to another through a third language which serves as a bridge, whereas the first two languages are not directly in contact. Such chains would mirror the phenomenon of wanderwörter such as *sugar* or *wine*, which were not borrowed from one original language into every language which now uses them, but travelled from language to language together with the trade goods, reaching remote geographical areas via trading intermediates. In the simulated data, 16.4% of words with a borrowing history were borrowed twice, and 2.9% were borrowed three or more times. There is thus a fair amount of wanderwort-like word histories in the data, adding to the realism of the simulated scenarios.

Another phenomenon whose frequency is worthwhile to investigate is internal borrowing within the same language family, i.e. the situation in which a word can be replaced by a cognate word. In the simulated data, because related languages are more likely to be neighbors, internal borrowing happens quite often, so that in 26.280% of all generated borrowing events one cognate set was replaced by the same set, not changing anything about the data. Across scenarios, this percentage was pretty stable at 29.531 ± 5.80%, although there are some outliers with the maximum at 67.3%. Set-internal borrowings are obviously a problem for contact detection algorithms which work on the level of cognate sets, because such borrowings do not leave any detectable traces.

5.5 Potential further uses of simulated scenarios

In addition to their usage as test cases for evaluating my methods, other researchers might want to use simulated scenarios as generated by my model as well. For instance, because each event during the simulation is logged in a format which allows the complete history of each word to be tracked explicitly, it becomes possible to evaluate loanword detection algorithms on much larger datasets than the real datasets which are currently and will ever be available.

This also applies to the comparative evaluation of methods and algorithms for tasks wich are closely related to lexical flow inference, such as the inference of different types of phylogenetic networks. In this context, the simulation could be used to assess the impact of lateral transfer on the reliability of phylogenetic inference methods, e.g. for practical experiments reinforcing the empirical findings by Greenhill et al. (2009) or the mathematical results of Roch & Snir (2012), both indicating that phylogenetic inference is quite robust to realistic amounts of lateral transfer.

For all these purposes, the simulated scenarios are distributed together with this book in various standard formats (trees in Newick format, cognacy data in a Nexus format readable by the most common software tools in phylogenetic inference). Moreover, my Java programs for generating more scenarios of this type, as well as the parsers for the log files which are necessary to extract statistical information of the type I was covering in this chapter, will be packaged and released as standalone executables along with their source code in order to allow other researchers to adapt the simulation model to their requirements, or experiment with different parameter settings than the ones I have been operating on here.

References

Aikio, Ante. 2002. New and old Samoyed etymologies. *Finnisch-Ugrische Forschungen (FUF)* 57. 9–57.

Aikio, Ante. 2004. An essay on substrate studies and the origin of Saami. In Irma Hyvärinen, Petri Kallio & Jarmo Korhonen (eds.), *Etymologie, Entlehnungen und Entwicklungen: Festschrift für Jorma Koivulehto zum 70. Geburtstag* (Mémoires de la Société Néophilologique de Helsinki 63), 5–34. Helsinki: Uusfilologinen Yhdistys.

Aikio, Ante. 2006a. New and old Samoyed etymologies II. *Finnisch-Ugrische Forschungen (FUF)* 59. 5–34.

Aikio, Ante. 2006b. On Germanic-Saami contacts and Saami prehistory. *Journal de la Société Finno-Ougrienne* 91. 9–55.

Aikio, Ante. 2014. The Uralic-Yukaghir lexical correspondences: Genetic inheritance, language contact or chance resemblance? *Finnisch-Ugrische Forschungen (FUF)* 62. 7–76.

Anikin, A. E. & E. A. Helimskij. 2007. *Samodijsko-tunguso-man'čžurskie leksičeskie sv'azy*. Moskva: Jazyki slav'anskoj kul'tury.

Ánte, Luobbal Sámmol Sámmol. 2012. An essay on Saami ethnolinguistic prehistory. In Riho Grünthal & Petri Kallio (eds.), *A linguistic map of prehistoric Northern Europe* (Suomalais-Ugrilaisen Seuran Toimituksia 266), 63–117.

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill & Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319(5863). 588–588.

Baba, Kunihiro, Ritei Shibata & Masaaki Sibuya. 2004. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics* 46(4). 657–664.

Bailey, H. W. 1987. Armenia and Iran iv. Iranian influences in Armenian language. In Ehsan Yarshater (ed.), *Encyclopædia Iranica, vol. ii, fasc. 4-5*, 445–465. London: Encyclopædia Iranica Foundation.

Beckwith, Christopher I. 2005. The ethnolinguistic history of the early Korean peninsula region: Japanese-Koguryŏic and other languages in the Koguryŏ, Paekche, and Silla kingdoms. *Journal of Inner and East Asian Studies* 2(2). 34–64.

Bereczki, Gábor. 1988. Geschichte der wolgafinnischen Sprachen. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 314–350. Leiden: Brill.

Bergsland, Knut. 1959. The Eskimo-Uralic hypothesis. *Journal de la Société Finno-Ougrienne* 61. 1–29.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths & Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* 10.1073/pnas.1204678110.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. In Christian Chiarcos, Richard Eckart de Castilho & Manfred Stede (eds.), *Proceedings of the Biennial GSCL Conference*, vol. 156, 43–53. Tübingen, Germany: Gunter Narr Verlag.

Bowern, Claire. 2016. Chirila: Contemporary and historical resources for the indigenous languages of Australia. *Language Documentation and Conservation* 10. 1–44.

Bowern, Claire & Quentin D. Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845.

Bowern, Claire & Bethwyn Evans (eds.). 2015. *The Routledge handbook of historical linguistics.* London: Routledge.

Brown, Cecil H., Eric W. Holman & Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89(1). 4–29.

Buck, Carl D. 1949. *A dictionary of selected synonyms in the principal Indo-European languages.* Chicago, USA: University of Chicago Press.

Campbell, Lyle. 1999. *Historical linguistics: An introduction.* Cambridge, Massachusetts: The MIT Press.

Chaves, Rafael, Lukas Luft, Thiago O. Maciel, David Gross, Dominik Janzing & Bernhard Schölkopf. 2014. Inferring latent structures via information inequalities. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014).* 112–121.

Chickering, David Maxwell. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3(Nov). 507–554.

Claassen, Tom & Tom Heskes. 2012. A Bayesian approach to constraint based causal inference. In Freitas de Nando & Kevin P. Murphy (eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (UAI'12), 207–216. Catalina Island, CA: AUAI Press.

Collinder, Björn. 1940. *Jukagirisch und Uralisch*. Vol. 8 (Uppsala Universitets Årsskrift). Leipzig: Harrassowitz.

Colombo, Diego & Marloes H. Maathuis. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15(1). 3741–3782.

Colombo, Diego, Marloes H. Maathuis, Markus Kalisch & Thomas S. Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* 40(1). 294–321.

Corson, David. 1995. Norway's "Sámi Language Act": Emancipatory implications for the world's aboriginal peoples. *Language in Society* 24(4). 493–514.

Cover, Thomas M. & Joy A. Thomas. 2006. *Elements of information theory*. 2nd edn. Hoboken, New Jersey: John Wiley & Sons.

Dahl, Östen & Maria Koptjevskaja-Tamm (eds.). 2001. *Circum-Baltic languages – Volume 1: Past and present* (Studies in Language Companion Series 54). Amsterdam: John Benjamins.

de Oliveira, Paulo Murilo Castro, Dietrich Stauffer, Søren Wichmann & Suzana Moss de Oliveira. 2008. A computer simulation of language families. *Journal of Linguistics* 44. 659–675.

de Vaan, Michiel Arnoud Cor. 2008. *Etymological dictionary of Latin and the other Italic languages* (Leiden Indo-European etymological dictionary series 7). Leiden, The Netherlands: Brill.

Décsy, Gyula. 1988. Slawischer Einfluss auf die uralischen Sprachen. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 616–637. Leiden: Brill.

Dellert, Johannes. 2015. Compiling the Uralic dataset for NorthEuraLex, a lexicostatistical database of Northern Eurasia. In Tommi A. Pirinen, Francis M. Tyers & Trond Trosterud (eds.), *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2015)* (Septentrio Conference Series). Tromsø: UiT The Arctic University of Norway.

Dellert, Johannes. 2016a. Uralic and its neighbors as a test case for a lexical flow model of language contact. In Tommi A. Pirinen, Eszter Simon, Francis M. Tyers & Veronika Vincze (eds.), *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2016)*. Szeged: University of Szeged.

Dellert, Johannes. 2016b. Using causal inference to detect directional tendencies in semantic evolution. In Sean Roberts, Christine Cuskley, Luke McCrohon, Lluis Barceló-Coblijn, Olga Feher & Tessa Verhoef (eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).* New Orleans, LA: EvoLang Scientific Committee.

Dellert, Johannes & Armin Buch. 2015. Using computational criteria to extract large Swadesh lists for lexicostatistics. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics.* Tübingen: University of Tübingen.

Dol'gopol'skij, Aron B. 1964. Gipoteza drevnejšego rodstva jazykov Severnoj Evrazii. Problemy fonetičeskih sootvetstvij. In Sergej P. Tolstov (ed.), *VII meždunarodnyj kongress antropologičeskih i ètnografičeskih nauk*, 1–22. Moskva: Nauka.

Dunn, Michael. 2000. Planning for failure: The niche of standard Chukchi. *Current Issues in Language Planning* 1(3). 389–399.

Dunn, Michael. 2015. *Indo-European lexical cognacy database.* http://ielex.mpi.nl/ (Last accessed 2019-06-09.)

Dybo, Anna V. 2007. *Lingvističeskie kontakty rannih t'urkov: Leksičeskij fond prat'urkskij period.* Moskva: Vostočnaja literatura RAN.

Dyen, Isidore, Joseph B. Kruskal & Paul Black. 1992. An Indoeuropean classification. A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5). iii–132.

Ellison, T. Mark. 2007. Bayesian identification of cognates and correspondences. In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, 15–22. Prague, Czech Republic: Association for Computational Linguistics.

Embleton, Sheila M. 1986. *Statistics in historical linguistics* (Quantitative Linguistics 30). Bochum, Germany: Studienverlag Dr. N. Brockmeyer.

Feist, Timothy Richard. 2011. *A grammar of Skolt Saami.* Manchester, UK: The University of Manchester.

Felsenstein, Joseph. 2004. *Inferring phylogenies.* Sunderland, Massachusetts: Sinauer Associates.

Finkenstaedt, Thomas & Dieter Wolff. 1973. *Ordered profusion. Studies in dictionaries and the English lexicon.* Heidelberg: C. Winter.

Fisher, Ronald A. [1925] 1934. *Statistical methods for research workers.* 5th edn. (Biological Monographs and Manuals V). Edinburgh & London: Oliver & Boyd.

Fortescue, Michael D. 1998. *Language relations across Bering Strait: Reappraising the archaeological and linguistic evidence* (Open linguistics series). London & New York: Cassell.

Fortescue, Michael D. 2005. *Comparative Chukotko-Kamchatkan dictionary* (Trends in Linguistics. Documentation [TiLDOC]). Berlin: De Gruyter.

Fortescue, Michael D. 2011. The relationship of Nivkh to Chukotko-Kamchatkan revisited. *Lingua* 121. 1359–1376.

Fortescue, Michael D. 2016. How the accusative became the relative: A Samoyedic key to the Eskimo-Uralic relationship? *Journal of Historical Linguistics* 6(1). 72–92.

Fortescue, Michael D., Steven Jacobson & Lawrence Kaplan. 2010. *Comparative Eskimo dictionary: With Aleut cognates* (Alaska Native Language Center research papers). Fairbanks, Alaska: Alaska Native Language Center, University of Alaska Fairbanks.

François, Alexandre. 2014. Trees, waves and linkages. Models of language diversification. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 161–189. London: Routledge.

Geisler, Hans & Johann-Mattis List. 2010. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In Heinrich Hettrich (ed.), *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Wiesbaden: Reichert. (Unpublished manuscript.)

Goldberg, Yoav. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57(1). 345–420.

Grant, Anthony. 2009. English. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/13 (Last accessed 2019-06-09.)

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965). 435–439.

Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790). 1052–1055.

Greenhill, Simon J. 2015. TransNewGuinea.Org: An online database of New Guinea languages. *PLOS ONE* 10. e0141563.

Greenhill, Simon J., Robert Blust & Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271–283.

Greenhill, Simon J., Thomas E. Currie & Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B: Biological Sciences* 276(1665). 2299–2306.

Grünthal, Riho. 2007. The Mordvinic languages between bush and tree. In Jussi Ylikoski & Ante Aikio (eds.), *Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. Beaivve 2007* (Mémoires de la Société Finno-Ougrienne 253), 115–137. Helsinki: Finno-Ugrian Society.

Gruzdeva, Ekaterina. 1998. *Nivkh* (Languages of the World 111). Munich, Germany: Lincom Europa.

Guy, Jacques B. M. 1984. An algorithm for identifying cognates between related languages. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, 448–451. Stanford, California: Association for Computational Linguistics.

Häkkinen, Jaakko. 2006. Uralilaisen kantakielen tutkiminen. *Tieteessä tapahtuu* 1. 52–58.

Häkkinen, Jaakko. 2007. *Kantauralin murteutuminen vokaalivastaavuuksien valossa.* Helsinki: University of Helsinki, Faculty of Arts, Department of Finno-Ugrian Studies. (MA thesis).

Häkkinen, Jaakko. 2009. Kantauralin ajoitus ja paikannus: Perustelut puntarissa. *Journal de la Société Finno-Ougrienne* 92. 9–56.

Häkkinen, Jaakko. 2012. Early contacts between Uralic and Yukaghir. *Journal de la Société Finno-Ougrienne* 264. 91–101.

Halilov, Madžid Šaripovič. 1993. *Gruzinsko-dagestanskie jazykovye kontakty: (na materiale avarsko-cezskih i nekotoryh lezginskih jazykov).* Mahačkala: RAN. 51.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2015. *Glottolog 2.5.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org (Accessed 2015-06-13.)

Haspelmath, Martin. 2008. Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In Thomas Stolz, Dik Bakker & Rosa Salas Palomo (eds.), *Aspects of language contact*, 43–62. Berlin: Mouton de Gruyter.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *WOLD.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/ (Last accessed 2019-06-09.)

Hauer, Bradley & Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In Haifeng Wang & David Yarowsky (eds.), *Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011)*, 865–873. Chiang Mai, Thailand. November 8-13, 2011.

Hausenberg, Anu-Reet. 1998. Komi. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 305–326. London: Routledge.

Hawkins, John A. 1990. Germanic languages. In Bernard Comrie (ed.), *The major languages of Western Europe*, 58–66. London: Routledge.

Helimski, Eugene. 1998. Selkup. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 548–579. London: Routledge.

Hewitt, George. 2004. *Introduction to the study of the languages of the Caucasus* (LINCOM handbooks in linguistics 19). Munich: Lincom Europa.

Hewson, John. 1974. Comparative reconstruction on the computer. In John M. Anderson & Charles Jones (eds.), *Proceedings of the 1st International Conference on Historical Linguistics*, 191–197. Amsterdam.

Ho, Trang & Allan Simon. 2016. *Tatoeba: Collection of sentences and translations.* http://tatoeba.org/eng/ (Last accessed 2019-06-10.)

Hochmuth, Mirko, Anke Lüdeling & Ulf Leser. 2008. Simulating and reconstructing language change. (Unpublished manuscript.) https://edoc.hu-berlin.de/handle/18452/3133 (Last accessed 2019-06-10.)

Hock, Hans H. & Brian D. Joseph. 1996. *Language history, language change, and language relationship. An introduction to historical and comparative linguistics.* Berlin: Mouton de Gruyter.

Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London B: Biological Sciences* 269(1493). 793–799.

Holman, Eric W. 2005. Nodes in phylogenetic trees: The relation between imbalance and number of descendent species. *Systematic Biology* 54(6). 895–899.

Hruschka, Daniel J., Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel & Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1). 1–9.

Huelsenbeck, John P. & Jonathan P. Bollback. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology* 50(3). 351–366.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267.

Huson, Daniel H. & Celine Scornavacca. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61(6). 1061–1067.

Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2). 245–291.

Jäger, Gerhard & Johann-Mattis List. 2017. Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change* 8(1). 22–54.

Jäger, Gerhard & Pavel Sofroniev. 2016. Automatic cognate classification with a support vector Machine. Proceedings of the 13th Conference on Natural Language Processing (KONVENS).

Janhunen, Juha. 1977. *Samojedischer Wortschatz* (Castreanumin toimitteita 17). Helsinki: Helsingin Yliopisto.

Janhunen, Juha. 1996. *Manchuria: An ethnic history* (Suomalais-ugrilaisen seuran toimituksia 222). Helsinki: Finno-Ugrian Society.

Janhunen, Juha (ed.). 2003. *The Mongolic languages* (Routledge Language Family Series). London: Routledge.

Janhunen, Juha. 2005. Tungusic: An endangered language family in Northeast Asia. *International Journal of the Sociology of Language* 2005(173). 37–54.

Johanson, Lars & Éva Ágnes Csató. 1998. *The Turkic languages* (Routledge Language Family Series). London: Routledge.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H. Maathuis, Peter Bühlmann, et al. 2012. Causal inference using graphical models with the R package `pcalg`. *Journal of Statistical Software* 47(11). 1–26.

Kessler, Brett. 2001. *The significance of word lists. Statistical tests for investigating historical connections between languages.* Stanford, CA: CSLI Publications.

Key, Mary Ritchie & Bernard Comrie (eds.). 2015. *IDS*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ids.clld.org/ (Last accessed on 2019-06-10.)

Kobyliński, Zbigniew. 2005. The Slavs. In Paul Fouracre (ed.), *The New Cambridge Medieval History: Volume 1, c. 500 – c. 700*, 524–544. Cambridge: Cambridge University Press.

Koller, Daphne & Nir Friedman. 2009. *Probabilistic graphical models: Principles and techniques.* Cambridge, MA & London: MIT Press.

Kondrak, Grzegorz. 2002. Determining recurrent sound correspondences by inducing translation models. In Shu-Chuan Tseng, Tsuei-Er Chen & Liu Yi-Fen (eds.), *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, vol. 1, 1–7. Taipeh: Association for Computational Linguistics.

Kondrak, Grzegorz. 2005. N-gram similarity and distance. In *12th International Conference on String Processing and Information Retrieval (SPIRE 2005)* (Lecture Notes in Computer Science 3772), 115–126. Berlin & Heidelberg: Springer.

Kroonen, Guus. 2013. *Etymological dictionary of Proto-Germanic.* Leiden: Brill.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. Oxford: Blackwell.

Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg & Outi Vesakoski. 2014. Behind family trees – secondary connections in Uralic language networks. *Language Dynamics and Change* 4(2). 189–221.

Lehtisalo, Toivo. 1956. *Juraksamojedisches Wörterbuch* (Lexica Societatis Fenno-Ugricae 13). Helsinki: Suomalais-ugrilainen seura.

Lindén, Krister, Erik Axelson, Sam Hardwick, Tommi A. Pirinen & Miikka Silfverberg. 2011. HFST – framework for compiling and applying morphologies. In Cerstin Mahlow & Michael Piotrowski (eds.), *Second International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011)*, 67–85. Berlin & Heidelberg: Springer.

List, Johann-Mattis. 2012a. LexStat: Automatic detection of cognates in multilingual wordlists. In Miriam Butt, Jelena Prokić, Thomas Mayer & Michael Cysouw (eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125. Avignon: Association for Computational Linguistics.

List, Johann-Mattis. 2012b. SCA: Phonetic alignment based on sound classes. In Daniel Lassiter & Marija Slavkovik (eds.), *New directions in logic, language and computation* (Lecture Notes in Computer Science 7415), 32–51. Berlin & Heidelberg: Springer.

List, Johann-Mattis. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis, Simon J. Greenhill & Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). e0170046.

List, Johann-Mattis, Simon Greenhill, Tiago Tresoldi & Robert Forkel. 2018. *LingPy. A Python library for quantitative tasks in historical linguistics*. http://lingpy.org (Last accessed 2019-06-10.)

List, Johann-Mattis, Philippe Lopez & Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2, 599–605. Berlin: Association for Computational Linguistics.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler & William Martin. 2014. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays* 36(2). 141–150.

Lloyd, Stuart. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28(2). 129–137.

Martin, Samuel E. 1966. Lexical evidence relating Korean to Japanese. *Language* 42(2). 185–251.

Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir* (Mouton Grammar Library 27). Berlin: Walter de Gruyter.

Meek, Christopher. 1995. Causal inference and causal explanation with background knowledge. In Philippe Besnard & Steve Hanks (eds.), *Proceedings of the 11th conference on Uncertainty in Artificial Intelligence (UAI 1995)*, 403–410. San Mateo, CA: Morgan.

Menges, Karl Heinrich. 1995. *The Turkic languages and peoples: An introduction to Turkic studies.* Wiesbaden: Otto Harrassowitz Verlag.

Menovščikov, G. A. 1988. *Slovar' èskimossko-russkij i russko-èskimosskij.* 2nd edn. Leningrad: Prosveščenie.

Moravcsik, Edith A. 1975. Verb borrowing. *Wiener Linguistische Gazette* 8. 3–30.

Morrison, David A. 2011. *An introduction to phylogenetic networks.* Uppsala: RJR Productions.

Murawaki, Yugo. 2015. Spatial structure of evolutionary models of dialects in contact. *PLOS ONE* 10(7). 1–15.

Murawaki, Yugo & Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution* 3(1). 13–25.

Murayama, Shichirō. 1976. The Malayo-Polynesian component in the Japanese language. *Journal of Japanese Studies* 2(2). 413–436.

Myers-Scotton, Carol. 2002. *Language contact: Bilingual encounters and grammatical outcomes.* Oxford: Oxford University Press.

Needleman, Saul B. & Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3). 443–453.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler & Bui Quang Minh. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1). 268.

Nikolaeva, Irina. 2006. *A historical dictionary of Yukaghir* (Trends in Linguistics. Documentation [TiLDOC]). Berlin: De Gruyter.

Nikolayev, Sergei L. & Sergei A. Starostin. 1994. *A North Caucasian etymological dictionary.* Moscow: Asterisk Press.

Oakes, Michael P. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics* 7(3). 233–243.

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

Pakendorf, Brigitte & Innokentij Novgorodov. 2009. Sakha. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/19 (Last accessed 2019-06-09.)

Pearl, Judea. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.

Pereltsvaig, Asya & Martin W. Lewis. 2015. *The Indo-European controversy: Facts and fallacies in historical linguistics*. Cambridge: Cambridge University Press.

Piispanen, Peter S. 2013. The Uralic-Yukaghiric connection revisited: Sound correspondences of geminate clusters. *Journal de la Société Finno-Ougrienne* 94. 165–197.

Purvis, Andy, Aris Katzourakis & Paul-Michael Agapow. 2002. Evaluating phylogenetic tree shape: Two modifications to Fusco & Cronk's method. *Journal of Theoretical Biology* 214(1). 99–103.

Puura, Ulriikka, Heini Karjalainen, Nina Zajceva & Riho Grünthal. 2013. *The Veps language in Russia: ELDIA case-specific report* (Studies in European Language Diversity 25). Mainz: ELDIA (European Language Diversity for All).

Raghavan, Usha Nandini, Réka Albert & Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76. 036106.

Rama, Taraka. 2015. Automatic cognate identification with gap-weighted string subsequences. In Rada Mihalcea, Joyce Yue Chai & Anoop Sarkar (eds.), *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (HLT-NAACL 2015)*, 1227–1231. Denver, CO: Association for Computational Linguistics.

Rama, Taraka. 2016. Siamese convolutional networks based on phonetic features for cognate identification. *arXiv Computing Research Repository (CoRR)*. arXiv:abs/1605.05172.

Rama, Taraka, Johannes Wahle, Pavel Sofroniev & Gerhard Jäger. 2017. Fast and unsupervised methods for multilingual cognate clustering. *arXiv preprint*. arXiv:1702.04938 (Last accessed 2019-06-10.)

Ramsey, Joseph, Jiji Zhang & Peter L. Spirtes. 2006. Adjacency-faithfulness and conservative causal inference. In Rina Dechter & Thomas Richardson (eds.), *Proceedings of the 22nd annual conference on Uncertainty in Artificial Intelligence (UAI 2006)*, 401–408. Arlington, VA: AUAI Press.

Reichenbach, Hans. 1956. *The direction of time.* Berkeley: University of California Press.

Richardson, Thomas & Peter Spirtes. 2002. Ancestral graph Markov models. *The Annals of Statistics* 30(4). 962–1030.

Rießler, Michael. 2009. Kildin Saami. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/14 (Last accessed 2019-06-09.)

Roch, Sebastien & Sagi Snir. 2012. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis. In Benny Chor (ed.), *RECOMB 2012: Research in computational molecular biology* (Lecture Notes in Computer Science 7262), 224–238. Berlin & Heidelberg: Springer.

Róna-Tas, András. 1988. Turkic influence on the Uralic languages. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 742–780. Leiden: Brill.

Rosvall, Martin & Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4). 1118–1123.

Rot, Sándor. 1988. Germanic influences on the Uralic languages. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 682–705. Leiden: Brill.

Saitou, Naruya & Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4(4). 406–425.

Salminen, Tapani. 2002. Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. In *Lingvističeskij bespredel: sbornik statej k 70-letiju a. i. kuznecovoj.* 44–55. Moskva: Izdatel'stvo MGU.

Sammallahti, Pekka. 1988a. Historical phonology of the Uralic languages (with special reference to Permic, Ugric and Samoyedic). In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 478–554. Leiden: Brill.

Sammallahti, Pekka. 1988b. Saamic. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 43–95. London: Routledge.

Sankoff, David. 1972. Matching sequences under deletion/insertion constraints. *Proceedings of the National Academy of Sciences* 69(1). 4–6.

Sankoff, David. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* 28(1). 35–42.

Sankoff, Gillian. 2001. Linguistic outcomes of language contact. In Peter Trudgill, J. Chambers & N. Schilling-Estes (eds.), *Handbook of sociolinguistics*, 638–668. Oxford: Basil Blackwell.

Schmidt, Christopher K. 2009a. Japanese. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/21 (Last accessed 2019-06-09.)

Schmidt, Christopher K. 2009b. Loanwords in Japanese. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 545–574. Berlin: Mouton de Gruyter.

Schulte, Kim. 2009a. Loanwords in Romanian. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 230–259. Berlin: Mouton de Gruyter.

Schulte, Kim. 2009b. Romanian. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/8 (Last accessed 2019-06-09.)

Schulze, Christian, Dietrich Stauffer & Søren Wichmann. 2008. Birth, survival and death of languages by Monte Carlo simulation. *Communications in Computational Physics* 3(2). 271–294.

Senn, Alfred. 1944. Standard Lithuanian in the making. *Slavonic and East European Review. American Series* 3(2). 102–116.

Sergejeva, Jelena. 2000. The Eastern Sámi: A short account of their history and identity. *Acta Borealia* 17(2). 5–37.

Sicoli, Mark A. & Gary Holton. 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLOS ONE* 3(9). e91722.

Siegl, Florian. 2013. The sociolinguistic status quo on the Taimyr Peninsula. *Études finno-ougriennes* 45. 239–280.

Smolicz, Jerzy J. & Ryszard Radzik. 2004. Belarusian as an endangered language: Can the mother tongue of an independent state be made to die? *International Journal of Educational Development* 24(5). 511–528.

Sokal, Robert R. & Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38. 1409–1438.

Spirtes, Peter, Clark Glymour & Richard Scheines. 2000. *Causation, prediction, and search.* 2nd edn. Cambridge, MA & London: MIT Press.

Spirtes, Peter & Thomas Richardson. 1997. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In Padhraic Smyth & David Madigan (eds.), *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics. (AISTATS 1997).* Society for Artificial Intelligence & Statistics.

Steiner, Lydia, Peter Stadler & Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1(1). 89–127.

Steudel, Bastian, Dominik Janzing & Bernhard Schölkopf. 2010. Causal Markov condition for submodular information measures. In Adam Tauman Kalai & Mehryar Mohri (eds.), *Proceedings of the 23rd Annual Conference on Learning Theory*, 464–476. Madison, WI: OmniPress.

Suhonen, Seppo. 1973. *Die jungen lettischen Lehnwörter im Livischen* (Mémoires de la Société Finno-Ougrienne 154). Helsinki: Suomalais-ugrilainen seura.

Suhonen, Seppo. 1988. Die baltischen Lehnwörter der finnisch-ugrischen Sprachen. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences.* (Handbuch der Orientalistik 8), 596–615. Leiden: Brill.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American linguistics* 21(2). 121–137.

Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas Wahlberg. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica* 30(3). 323–352.

Taagepera, Rein. 2013. *The Finno-Ugric republics and the Russian state.* London: Routledge.

Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55–75. Berlin: Mouton de Gruyter.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics.* Berkeley & Los Angeles: University of California Press.

Thordarson, Fridrik. 2009. Ossetic language i. History and description. In Ehsan Yarshater (ed.), *Encyclopædia Iranica, online version.* http://www.iranicaonline.org/articles/ossetic (Last accessed 2019-06-10.)

Turchin, Peter, Ilja Peiros & Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3. 117–126.

Vajda, Edward J. 2009. Loanwords in Ket. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 471–495. Berlin: Mouton de Gruyter.

Vajda, Edward J. 2010. A Siberian link with Na-Dene languages. *Archeological Papers of the University of Alaska* 5(New Series). 33–99.

Vajda, Edward J. & Andrey Nefedov. 2009. Ket. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/18 (Last accessed 2019-06-09.)

van der Sijs, Nicoline. 2009. Dutch. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/12 (Last accessed 2019-06-09.)

van Hout, Roeland & Pieter Muysken. 1994. Modeling lexical borrowability. *Language Variation and Change* 6(1). 39–62.

Vejdemo, Susanne & Thomas Hörberg. 2016. Semantic factors predict the rate of lexical replacement of content words. *PLOS ONE* 11(1). 1–15.

Viires, Ants & Lauri Vahtre. 1993. *The red book of the peoples of the Russian empire*. Tallinn. http://www.eki.ee/books/redbook (Last accessed 2019-06-10.)

Viitso, Tiit-Rein. 1998. Fennic. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 96–114. London: Routledge.

Volodin, A. P. & K. N. Halojmova. 1989. *Slovar' itel'mensko-russkij i russko-itel'menskij*. Leningrad: Prosveščenie.

Volodin, A. P. & P. J. Skorik. 1997. Čukotskij jazyk. In A. P. Volodin, N. B. Vaxtin & A. A. Kibrik (eds.), *Jazyki mira: Paleoaziatskie jazyki*, 23–39. Moskva: Indrik.

Wells, John C. 1995. *Computer-coding the IPA: A proposed extension of SAMPA*. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm (Last accessed 2019-06-10.)

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. *The ASJP database (version 17)*. http://asjp.clld.org/ (Accessed 2017-05-22.)

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2018. *The ASJP database (version 18)*. http://asjp.clld.org/ (Accessed 2019-06-10.)

Wichmann, Søren & Jan Wohlgemuth. 2008. Loan verbs in a typological perspective. In Thomas Stolz, Dik Bakker & Rosa Salas Palomo (eds.), *Aspects of language contact*, 89–122. Berlin: Mouton de Gruyter.

Wiebusch, Thekla. 2009. Mandarin Chinese. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/22 (Last accessed 2019-06-09.)

Willems, Matthieu, Etienne Lord, Louise Laforest, Gilbert Labelle, François-Joseph Lapointe, Anna Maria Di Sciullo & Vladimir Makarenkov. 2016. Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16(1). 180.

Willems, Matthieu, Nadia Tahiri & Vladimir Makarenkov. 2014. A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology* 12(05). 1450024.

Yang, Ziheng, Sudhir Kumar & Masatoshi Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4). 1641–1650.

Yeung, Raymond W. 2008. *Information theory and network coding*. New York, NY: Springer Science & Business Media.

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft & Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* 113(7). 1766–1771.

Zachrisson, Inger. 2008. The Sámi and their interaction with the Nordic peoples. In Stefan Brink & Neil Price (eds.), *The Viking world*, 32–39. London: Routledge.

Zajceva, N. G. 2010. *Uz' vepsä-venäläine vajehnik = novyj vepssko-russkij slovar'*. Petrozavodsk: Periodika.

Zhang, Jiji. 2006. *Causal inference and reasoning in causally insufficient systems*. Pittsburgh, PA: Carnegie Mellon University. (Doctoral dissertation.)

Zhang, Jiji. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172(16). 1873–1896.

Name index

Language index

Subject index

Information-theoretic causal inference of lexical flow

This volume seeks to infer large phylogenetic networks from phonetically encoded lexical data and contribute in this way to the historical study of language varieties. The technical step that enables progress in this case is the use of causal inference algorithms. Sample sets of words from language varieties are preprocessed into automatically inferred cognate sets, and then modeled as information-theoretic variables based on an intuitive measure of cognate overlap. Causal inference is then applied to these variables in order to determine the existence and direction of influence among the varieties.

The directed arcs in the resulting graph structures can be interpreted as reflecting the existence and directionality of lexical flow, a unified model which subsumes inheritance and borrowing as the two main ways of transmission that shape the basic lexicon of languages. A flow-based separation criterion and domain-specific directionality detection criteria are developed to make existing causal inference algorithms more robust against imperfect cognacy data, giving rise to two new algorithms. The Phylogenetic Lexical Flow Inference (PLFI) algorithm requires lexical features of proto-languages to be reconstructed in advance, but yields fully general phylogenetic networks, whereas the more complex Contact Lexical Flow Inference (CLFI) algorithm treats proto-languages as hidden common causes, and only returns hypotheses of historical contact situations between attested languages.

The algorithms are evaluated both against a large lexical database of Northern Eurasia spanning many language families, and against simulated data generated by a new model of language contact that builds on the opening and closing of directional contact channels as primary evolutionary events. The algorithms are found to infer the existence of contacts very reliably, whereas the inference of directionality remains difficult. This currently limits the new algorithms to a role as exploratory tools for quickly detecting salient patterns in large lexical datasets, but it should soon be possible for the framework to be enhanced e.g. by confidence values for each directionality decision.

ISBN 978-3-96110-143-6

9 783961 101436