

Information-theoretic causal inference of lexical flow

Johannes Dellert

Draft
of June 17, 2019, 18:08

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.

Information- theoretic causal inference of lexical flow

Johannes Dellert

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>

© 2019, Johannes Dellert

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

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

978-3-96110-144-3 (Hardcover)

ISSN: 2366-7818

DOI:[10.5281/zenodo.3247415](https://doi.org/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: Xe_{La}TeX

Language Science Press

Unter den Linden 6

10099 Berlin, Germany

langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität  Berlin

Contents

Preface	v
Acknowledgments	ix
1 Introduction	1
2 Foundations: historical linguistics	7
2.1 Language relationship and family trees	7
2.2 Language contact and lateral connections	11
2.3 Describing linguistic history	12
2.4 Classical methods	13
2.4.1 The comparative method	14
2.4.2 Theories of lexical contact	19
2.5 Automated methods	25
2.5.1 Lexical databases	26
2.5.2 Phylogenetic inference	30
2.5.3 Phylogeographic inference	34
2.5.4 Automating the comparative method	36
2.5.5 On the road towards network models	40
2.6 The lexical flow inference task	45
2.6.1 Phylogenetic lexical flow	45
2.6.2 Contact flow	45
2.7 The adequacy of models of language history	46
3 Foundations: causal inference	51
3.1 Philosophical and theoretical foundations	51
3.1.1 Correlation and causation	52
3.1.2 Causality without experiment	54
3.1.3 Conditional independence	56
3.1.4 Bayesian networks	61
3.1.5 Causal interpretation of Bayesian networks	63

3.2	Causal inference algorithms	64
3.2.1	Causal graphs	65
3.2.2	Determining conditional independence relations	71
3.2.3	The PC algorithm	76
3.2.4	The FCI algorithm	80
3.2.5	Alternative algorithms	87
4	Wordlists, cognate sets, and test data	89
4.1	NorthEuraLex	89
4.1.1	The case for a new deep-coverage lexical database	89
4.1.2	Selecting the language sample	90
4.1.3	Selecting and defining the concepts	91
4.1.4	The data collection process	94
4.1.5	Difficulties and future development	95
4.2	Transforming and encoding into IPA	97
4.2.1	Encoding cross-linguistic sound sequence data	97
4.2.2	Implementing orthography-to-IPA transducers	99
4.2.3	Tokenizing into reduced IPA	102
4.3	Information-Weighted Sequence Alignment (IWSA)	106
4.3.1	The case for information weighting	106
4.3.2	Gappy trigram models	107
4.3.3	Implementing IWSA	108
4.3.4	Inspecting the results of IWSA	110
4.4	Modelling sound correspondences	113
4.4.1	Perspectives on sound correspondences	114
4.4.2	Modeling sound correspondences as similarity scores	115
4.4.3	Inferring global correspondences from NorthEuraLex	116
4.4.4	Inferring pairwise correspondences for NorthEuraLex	119
4.4.5	Aligning NorthEuraLex and deriving form distances	123
4.5	Cognate clustering	124
4.5.1	The cognate detection problem	124
4.5.2	Approaches to cognate clustering	125
4.5.3	Deriving cognate sets from NorthEuraLex	128
4.5.4	Evaluation on IELex intra-family cognacy judgments	128
4.5.5	Evaluation on WOLD cross-family cognacy judgments	131
4.5.6	A look at the cognate sets	134
4.6	Deriving a gold standard for lexical flow	137
4.6.1	Defining the gold standard	138

4.6.2	Case study 1: the Baltic Sea area	139
4.6.3	Case study 2: Uralic and contact languages	144
4.6.4	Case study 3: the linguistic landscape of Siberia	149
4.6.5	Case study 4: a visit to the Caucasus	164
5	Simulating cognate histories	171
5.1	Simulation and in-silico evaluation	171
5.1.1	Advantages and shortcomings of simulation	171
5.1.2	Principles of in-silico evaluation	173
5.2	Generating phylogenies	174
5.2.1	Models of lexical replacement	175
5.2.2	Simulating how languages split and die	176
5.3	Modeling lexical contact	178
5.3.1	Modeling the preconditions for contact	178
5.3.2	A monodirectional channel model of language contact	179
5.3.3	Opening and closing channels	179
5.3.4	Simulating channel behavior	181
5.3.5	Overview of the simulation	182
5.4	Analyzing the simulated scenarios	182
5.4.1	Are the scenarios realistic?	186
5.4.2	Are the scenarios interesting?	191
5.5	Potential further uses of simulated scenarios	193
6	Phylogenetic lexical flow inference	195
6.1	Modeling languages as variables	196
6.1.1	Languages as phoneme sequence generators	196
6.1.2	Languages as cognate set selectors	197
6.2	A cognate-based information measure	198
6.3	Conditional mutual information between languages	201
6.4	Improving skeleton inference	202
6.4.1	Problem: stability on discrete information	202
6.4.2	Flow Separation (FS) independence	203
6.5	Improving directionality inference	204
6.5.1	Problem: monotonic faithfulness and v-structures	204
6.5.2	Unique Flow Ratio (UFR): flow-based v-structure testing	206
6.5.3	Triangle Score Sum (TSS): aggregating directionality hints	208
6.6	The phylogenetic guide tree	213
6.7	Deriving proto-language models	214
6.7.1	Ancestral state reconstruction algorithms	214

Contents

6.7.2	Evaluation of ASR algorithms on simulated data	220
6.8	Phylogenetic Lexical Flow Inference (PLFI)	223
6.9	Evaluation of PLFI	225
6.9.1	Evaluation metrics for phylogenetic flow	226
6.9.2	Overall quantitative results for NorthEuraLex data	228
6.9.3	Qualitative discussion of NorthEuraLex scenarios	230
6.9.4	Evaluation on simulated data	242
7	Contact lexical flow inference	251
7.1	The contact flow inference task	251
7.2	Advantages and disadvantages of contact flow	252
7.3	Difficulties in applying the RFCI algorithm	253
7.4	Significance testing for v-structures	255
7.5	Contact Lexical Flow Inference (CLFI)	257
7.6	Evaluation of CLFI	258
7.6.1	Evaluation metrics for contact flow	260
7.6.2	Overall quantitative results for NorthEuraLex data	261
7.6.3	Qualitative discussion of NorthEuraLex scenarios	263
7.6.4	Evaluation on simulated data	271
8	Conclusion and outlook	277
8.1	Summary	277
8.2	Future work	279
8.3	Final remarks	283
	References	289
	Index	305
	Name index	305
	Language index	309
	Subject index	315

6 Phylogenetic lexical flow inference

In many respects, this chapter represents the core of this book. It discusses both the idea of and the methodological difficulties in applying causal inference to data in the shape of overlapping cognate sets in detail, before defining PLFI (phylogenetic lexical flow inference), the first of two algorithms I describe in this book, and then evaluating it on both NorthEuraLex and the simulated scenarios.

This algorithm requires adding reconstructions of all known proto-languages to the model, which makes it causally sufficient, allowing me to work with the much simpler PC algorithm, and postponing the need to get into the complexities of the RFCI algorithm to the next chapter.

In §6.1, I discuss possible ways of modeling lexicostatistical data as statistical or information-theoretic variables, and motivate my decision to stick to coarse-grained cognate overlap judgments for the purposes of this book. §6.2 then introduces my cognate-based information measure, which allows me to derive a very natural measure of conditional mutual information between languages in §6.3.

Sections 6.4 and 6.5 are concerned with ways to specialize parts of the PC algorithm in order to balance out assumptions that are not met by conditional independence tests based on this measure. These include the development of an explicitly flow-based criterion for defining plausible separating sets, and of two heuristic criteria for aggregating evidence of directionality. These criteria become necessary because the separating set membership criterion of the PC algorithm is found to be too unstable on cognacy overlap data.

§6.6 describes how I arrived at the guide tree for my experiments on NorthEuraLex, which is then used in §6.7 to perform ancestral state reconstruction in order to create the data for the proto-languages. Before deciding on maximum-likelihood construction as the basis of the PLFI algorithm, I evaluate several alternatives on the simulated data from Chapter 5.

§6.8 then puts the results of the previous sections together to define the PLFI algorithm, which is finally evaluated in the last section of this chapter, both on the NorthEuraLex gold standard developed in Chapter 4, and the simulated data from Chapter 5.

6.1 Modeling languages as variables

Everything starts with the idea of detecting conditional independence relationships between sets of languages, making it possible to apply causal inference algorithms to the lexical flow inference task. To take up the idea foreshadowed by the language examples in Chapter 3, I will assume the lexicon of each language to quite literally be caused by the lexicon of its ancestor language and possibly other languages which influenced its development. Put differently, the model is intended to determine how the lexicon of a language came about as a mixture of lexical material from other languages, and summarize the results in a causal graph.

As the crucial step towards this goal, we need a formalism which makes it possible to treat languages as information-theoretic variables. Depending on a range of choices about how we model languages, there are many possibilities to define useful measures of information content and mutual information between languages. While a single rather simple measure on cognate sets will be used later, in this section the idea will be put into a wider context by discussing more generally the different ways in which languages might be treated as statistical or information-theoretic variables.

6.1.1 Languages as phoneme sequence generators

The reductionist premise of lexicostatistical databases is to view each language simply as a (possibly many-to-many) mapping from concepts to lexical realizations. This means we could treat languages as variables generating sequences of phoneme n -grams, and measure information-theoretically how much about the generated sequence in one language we know given the sequence generated by another language. For unrelated languages, we could expect the mutual information to be not significantly different from zero. A significance threshold for an independence test could be based on the amount of mutual information we would expect if the two word lists were randomly sampled from the two languages.

The question what a good sampling procedure would look like, is quite involved, and again depends on how we model the generated phoneme sequences. The most straightforward way to model the phoneme emission would be to use n -gram distributions, to treat concept realizations as events producing bags of n -grams, and measure the mutual information between the resulting distributions. This is a formal answer to the intuitive question: given the n -grams for a realization of some concept in language A, how much on average do we already know about the realization of the same concept in language B? To come back to one

of our examples from Chapter 3, the phonetic bigram representations of *slange* [slanɐ] and *Schlange* [ʃlanɐ], the Danish and German words for ‘snake’, share three out of four elements ($\{[la], [an], [ɐ]\}$), giving them a very high mutual information, whereas knowing either word does not help us at all to predict the Finnish equivalent *käärme* [kæærmɛ], which does not share a single bigram with the Germanic words.

no IPA
length
sign?

While this is an attractive idea, initial exploratory experiments quickly show that the information content of n -gram overlaps is not very high. For instance, the global bias towards CVCV-type syllable structures will lead to spurious mutual information, as e.g. CV-type bigrams such as [pa] or [ku] will always be more common than CC or VV-type n -grams.

The realization-based mutual information measure can be somewhat improved upon by building it only on aligned positions. Essentially, we optimally align all the realizations for each pair of languages, count how often each pair of n -grams (for reasons of data sparseness, only unigrams and bigrams are feasible) is aligned, and compare this to the distribution we would expect if the words were randomly chosen. Again, the threshold needs to be computed by resampling, because the two n -gram distributions computed in this way will not be independent due to the fact that the alignment algorithm will always find some vowels (and often some consonants) to align even in completely random and unrelated words. In exploratory experiments, the necessary threshold turned out to be so high that the common signal between languages from different branches of Indo-European could not be distinguished from noise. While it might be possible to arrive at a sufficiently sensitive independence test by refining this approach, the process is hampered by the fact that error causes are very difficult to track down and interpret in such a model.

6.1.2 Languages as cognate set selectors

A third possibility (and the one which I am going to build upon) starts with structures that are situated one step higher in the usual toolchain of computational historical linguistics. Assume we have a good cognate detection method in place. Then, we can use this module to group the realizations of each concept i into a set of cognate sets $Cog_i := \{cog_{i,1}, \dots, cog_{i,n_i}\}$.

In terms of a probabilistic model, this leaves us with a quite complex chain of random variables building on the basic view of languages as string generators. For each language j , we could start with a lexicon generator variable $Lex_j : \Omega \rightarrow \Sigma^*$ for some universal alphabet Σ of phonetic symbols. Possible observations of such a variable could be phonetic strings such as [æææ], [ktkæŋ], and infinitely

many other highly improbable strings. Alternatively (and especially to model re-sampling), Lex_j can be defined as selecting strings from a predefined set $L_j \in \Sigma^*$ containing all the strings of the lexicon, that is, in our case, a phonetic representation of all the lemmas in our database. On the NorthEuraLex data, the English variable would generate words such as [taʊn], [fi:və], [hɛvi], [aɪ], but the assignment of these forms to meanings would be assumed to be entirely random.

Any automated cognate detection procedure can now be conceptualized as a very complex function of all the lexicon generator variables which generates a set of cognate sets for each concept C_i :

$$cog(X_1, \dots, X_n) : \Omega \rightarrow \bigotimes_{i=1}^n \wp \left(\bigcup_{j=1}^m Lex_j \right)^* \quad (6.1)$$

Unrelated languages can now be seen as independently sampling one or several of these cognate sets for each given concept. For related languages, we should then be able to measure a dependence in the form of non-zero mutual information. Intuitively, the more closely two languages are related, the more knowing which cognate sets one language picked will help us to predict the sets picked by the other language. If we know for one Germanic language (such as German) that it picked its word for ‘honey’ from the cognate set of English *honey*, this will be much more helpful for predicting to which class the Icelandic word will belong, than knowing the cognate set of the equivalent in a more distantly related language like Greek.

Mathematically, we can now model each language L_j as a variable picking for each concept a random subset of the set of cognate sets, i.e.:

$$L_j : \Omega \rightarrow \bigotimes_{i=1}^n \wp(Cog_i) \quad (6.2)$$

From a probabilistic point of view, this is the shape of the variables I am going to operate on, although it would be very difficult to assign explicit joint probability distributions to sets of such variables. Instead, I will only use an information geometry on these variables.

6.2 A cognate-based information measure

We now turn to the question of how to estimate conditional mutual information between languages modeled as cognate set selectors. The basic idea is to define an

easy-to-compute and intuitive measure h on outcomes of cognate set selection, which mimics joint entropy in adhering to the basic properties of a submodular information measure. Based on this measure, I will then be able to define an equivalent of conditional mutual information between languages.

A simple information content measure i turns out to be easy to find: one can simply define $i(L_j)$ for a language variable L_j as the number of cognate sets selected by the language across concepts:

$$i(L_j) := \sum_{i=1}^n |L_{j,i}(\omega)| \quad (6.3)$$

I will treat this definition as the self-information of L_j , i.e. the outcome of a random variable measuring the entropy of language L_j . There is a very intuitive parallel between this measure and the view of entropy as a measure of descriptive complexity: given a set of concepts and a set of cognate sets for each concept, the minimum description length for the lexicon of L_j can be seen as the length of the minimal specification of the mapping from concepts to cognate sets, which will be linear in the number of cognate sets touched by the language.

The equivalent of the joint entropy $h(L_j, L_k)$ can now be defined as the number of cognate sets selected by one of the two languages, i.e. the union of the outcomes represented by both languages:

$$h(L_j, L_k) := \sum_{i=1}^n |L_{j,i}(\omega) \cup L_{k,i}(\omega)| \quad (6.4)$$

Analogously, we can define $h(\mathbf{Z})$ for all subsets $\mathbf{Z} = \{Z_1, \dots, Z_m\} \subseteq \mathcal{L}$ of our set of languages \mathcal{L} . In Appendix C, I show that this measure adheres to the elemental inequalities defining a submodular information measure. According to the theory introduced in §3.2.2.3, this establishes that the measure $h(\mathbf{Z})$ for a set of languages $\mathbf{Z} = \{Z_1, \dots, Z_m\}$, is similar enough to a measure of joint information to lead to a consistent definition of conditional mutual information.

We have seen previously how cognacy data can be pictured as representing each language as a long binary vector in which each dimension represents a cognate set, and the values 1 and 0 represent the presence or absence of each cognate set in that language. By the information measure just introduced, the information content $i(L_j)$ of a language L_j is then simply the number of ones in its vector representation, and the joint entropy is the number of ones in the disjunction of the language vectors, i.e. the vector which has a zero in the components

where every language vector has a zero, and a one in all components where at least one language vector has a one. This representation is of course very sparse, since there will sometimes be dozens of cognate classes for any given concept observed across a hundred languages. Measures based on such a representation can be expected to be very prone to sampling errors, e.g. if a cognate for a word exists in a language, but simply was not attested in our lexical resource, or it has shifted slightly in meaning, which is why we do not observe the full path.

Structurally, these problems are very similar to the reasons why the idea of learning more dense embeddings of such sparse vectors in a lower-dimensional space has gained a lot of traction in computational linguistics, leading to a variety of techniques for distributed representations which are summarized e.g. by [Goldberg \(2016\)](#). The general idea of applying embeddings here would be to find a good representation of the cognacy features for each concept as vectors in a lower-dimensional vector space, so that concepts which pattern similarly across languages will have similar representations. The representations of the cognate sets picked by a language would then be concatenated to form a more dense language vector. This procedure allows information sharing between cognate set presence or absence features, which would not only correct for the fact that the features are not independent (because generally, already having a word from cognate class A for some concept makes it less likely that it will also have a word from cognate class B for it), but also provide a way of smoothing out gaps in the data based on the knowledge provided by similar languages (e.g. predicting that the word for ‘honey’ in any Germanic language will belong to the same cognate class as English *honey*, even if for some reason there is a gap in the database).

On such embedded cognacy vectors, conditional independence tests could be defined e.g. through cosine similarities or Pearson correlations (after centering), but it would also very likely be possible to define a joint entropy measure $h(L_j, L_k)$ based on combinations of embedding-based language vectors. While given the experience in mainstream areas of natural language processing, such representations would very likely yield better results due to better information sharing, its results would no longer be interpretable in terms of discrete lexical transmission events, and the resulting notion of lexical flow would only be a very abstract information-theoretic measure instead of being directly mappable to the paths by which certain words traveled. In this book, I will therefore only explore the count-based joint entropy measure defined above, and leave the exploration of embedding-based measures to future work.

6.3 Conditional mutual information between languages

To arrive at the needed measure of conditional mutual information $i(L_i; L_j|\mathbf{Z})$, I can now simply use the standard definition on the information measure:

$$i(L_i; L_j|\mathbf{Z}) = h(L_i, \mathbf{Z}) + h(L_j, \mathbf{Z}) - h(L_i, L_j, \mathbf{Z}) - h(\mathbf{Z}) \quad (6.5)$$

Applying the definition of h I just developed, and writing $\text{cog}(L_1, \dots, L_k)$ for the set of cognate sets shared by the languages L_1, \dots, L_k , i becomes a count of conditional cognate overlap, which can be seen intuitively as the number of items in unblocked lexical flow. Note that the blocking of information flow was also the intuition behind the concept of d-separation which I introduced in §??, and the close correspondence between conditional independence constraints and d-separation in Bayesian networks is what will make it possible to infer networks via sequences of tests for vanishing conditional cognate overlap.

In order to be able to define a global threshold value for conditional independence tests, conditional mutual information needs to be normalized by the number of remaining cognates not touched by any of the conditioning languages, yielding the normalized conditional mutual information $I(L_i; L_j|\mathbf{Z})$:

$$I(L_i; L_j|\mathbf{Z}) := \frac{|\text{cog}(L_i, L_j) \setminus \{c \mid \exists \{Z_1, \dots, Z_k\} \subseteq \mathbf{Z} : c \in \text{cog}(Z_1, \dots, Z_k)\}|}{\max\{|\text{cog}(L_i, Z_1, \dots, Z_k)|, |\text{cog}(L_j, Z_1, \dots, Z_k)|\} - |\text{cog}(Z_1, \dots, Z_k)|} \quad (6.6)$$

Informally, $I(L_1; L_2|\mathbf{Z})$ thus quantifies the share of cognates between L_1 and L_2 which cannot be explained away by having been borrowed through a subset of the languages in \mathbf{Z} . To use this measure of dependence as a conditional independence test, we simply check whether $I(L_1; L_2|\mathbf{Z}) \leq \theta_{L_1, L_2}$ for a threshold θ_{L_1, L_2} , which could be derived from the number of false cognates between L_1 and L_2 which we expect due to automated cognate detection. In practice, I am setting $\theta_{L_1, L_2} := 0.025$ for all language pairs because the distribution of false cognates is difficult to estimate, and language-specific thresholds did not lead to better results in initial experiments on a smaller language set. On the NorthEuraLex data, this means that languages which share 25 cognates or less will be unconditionally independent, and every link the algorithm establishes will explain an overlap of at least 26 cognates.

Based on this conditional independence test, the first stage of the PC algorithm will derive a causal skeleton which represents a scenario of contacts between pairs of input languages that is only as complex as necessary to explain the lexical

overlaps. The model thus assumes that all similarities are primarily due to mutual influence, and never infers the existence of hidden common causes, such as proto-languages. As we shall see, to arrive at a phylogenetic network, we will need to introduce the proto-languages as additional variables.

6.4 Improving skeleton inference

When applying my newly developed conditional independence test to small test cases, it quickly becomes clear that they do not reflect the constraints defined by the gold standards very well. The main problem for skeleton inference is the way in which separating set candidates are selected in the PC and PC* algorithms. In this section, I develop an alternative to these two standard candidate selection techniques, and a comparison of the performance of all three variants will be part of the evaluation at the end of this chapter.

6.4.1 Problem: stability on discrete information

The PC algorithm as presented in §3.2.3 is only tractable because it tests separating set candidates by ascending order of cardinality, and builds on the assumption that any separating set must be a subset of immediate neighbours of L_i and L_j in the current skeleton. Conditioning on any set of neighbours is possible in the vanilla PC algorithm because the faithfulness assumption implies that true dependencies will always “shine through”, no matter how many intervening variables we condition on. For our model, this is a problematic assumption, because being allowed to select any neighbours makes it too easy to screen off languages from each other in our information geometry of limited granularity.

A small simplified example based on the NorthEuraLex data will show why this is a problem for skeleton inference. In the gold standard, there is a contact link between Norwegian (*nor*) and Southern Saami (*sma*). In the cognacy data, we have $|\text{cog}(\text{nor}, \text{sma})| = 114$, i.e. the two languages share 114 cognate sets, of which some are false positives, but most reflect North Germanic loans. At some stage of skeleton inference, *sma* is still connected to its neighbour Northern Saami (*sme*), and one of the remaining neighbours of *nor* is Swedish (*swe*). Now, most of the North Germanic material is of course also present in Swedish, leaving a quite high overlap $|\text{cog}(\text{sma}, \text{swe})| = 111$, a large part of which is shared between all three languages: $|\text{cog}(\text{nor}, \text{sma}, \text{swe})| = 96$. Using these overlaps, we first get a successful conditional independence test ($\text{nor} \perp\!\!\!\perp \text{sma} \mid \text{swe}$), because

$$\begin{aligned}
I(nor; sma|swe) &= \frac{|cog(nor) \cup cog(swe)| + |cog(sma) \cup cog(swe)|}{\max(|cog(nor) \cup cog(swe)|, |cog(sma) \cup cog(swe)|) - |cog(swe)|} \\
&\quad - \frac{|cog(nor) \cup cog(sma) \cup cog(swe)| + |cog(swe)|}{\max(|cog(nor) \cup cog(swe)|, |cog(sma) \cup cog(swe)|) - |cog(swe)|} \\
&= \frac{18}{1008} = 0.018
\end{aligned}$$

Now, after deleting the link *nor* — *sma* due to the successful test, the neighbour relation between Norwegian and Swedish persists, also giving us (*swe* \perp *sma* | *nor*), so that the resulting skeleton lacks any hint of the contact between North Germanic and Western Saami.

This example demonstrates that for skeleton inference to work on my coarse-grained data, the decision to remove a link between two languages should not be based only on the numbers of cognates shared with some set of immediate neighbours. The PC* algorithm already goes one step towards the solution by considering only neighbours on connecting paths, but this will not solve the problem in this case, either: There is a connecting path *nor* — *swe* \rightarrow *fin* \rightarrow *sme* — *sma*, which would also make *swe* a possible element for separating set candidates, causing exactly the same problem.

6.4.2 Flow Separation (FS) independence

We will now see how it is possible to at least partially correct for the lack of faithfulness by exploiting the fact that we have more than just a single conditional mutual information value to perform each conditional independence check. $I(L_i; L_j | \mathbf{Z})$ is computed from as many as 1,016 individual “concept stories” which provide us with a much richer picture of what is going on, and can help us to quantify the information flow much more precisely. To explain away a cognate that is shared between two languages L_i and L_j , it must have been possible for the lexeme in question to have travelled between the two languages on some other path. Therefore, any minimal separating set must form a union of acyclic paths between L_i and L_j . In effect, this constitutes an explicit model of the lexical flow helping us to decide more reliably which links can be deleted. In our example case, we now get $I(nor; sma | \mathcal{L} \setminus \{nor, sma\}) = 0.051$, i.e. the link will correctly not disappear whichever separating set candidate we condition on.

The adapted mutual information measure will be called *flow separation (FS)*. Adopting the convention of using two-letter shorthands for the different skeleton inference methods, PC will be used for the vanilla PC variant, and PS as a shorthand for PC* (“PC-Star”). My implementation of FS uses a depth-first search of the current graph to get all connecting paths which contain four nodes or less,

and generates all combinations of these paths which lead to separating set candidates of a given cardinality. Longer paths would need to be considered in theory, but did not lead to different results on my data, at a much higher computational cost. The cognate sets for each concept are tested separately against these paths in a highly optimized fashion, making the FS-based independence test not significantly slower than the PC and PS variants.

6.5 Improving directionality inference

Similar problems in applying the vanilla PC algorithm face us in directionality inference, the second stage of constraint-based causal inference algorithms. The two standard v-structure detection procedures which were already introduced in Chapter 3, will from now on be referred to as VPC (Vanilla PC) and SPC (Stable PC). Both of these variants build on the separating sets used to infer the skeleton, but the uncertain nature of independence checks on our data again forces us to explore alternative approaches. The first variant only replaces v-structure detection and then works with the propagation rules from Stage 3 of the PC algorithm, and the second variant I will introduce here tries to infer the directionality signal independently of the skeleton. In both cases, the basic idea behind directionality detection will still be informed by the theory of causal inference.

6.5.1 Problem: monotonic faithfulness and v-structures

To recapitulate, in the second stage of the PC algorithm and related algorithms, VPC and SPC directionality inference on the causal skeleton is performed by asking whether the central language B in each pattern of the form $A - B - C$ was part of the separating set that was used for explaining away the link $A - C$. The idea is that if B was not necessary to explain away any possible correlation between A and C (i.e. there is a separating set not containing B), this excludes all causal patterns except $A \rightarrow B \leftarrow C$. This is based on the assumption that in each of the three other possible causal scenarios, we would see some information flow between A and C if we do not condition on B .

This type of reasoning is again justified by the causal faithfulness assumption, which states that we can derive exactly the conditional independence relations implied by the true graph through d-separation. More specifically, the scenario $A \rightarrow B \leftarrow C$ would be characterized by the conditional independences $A \perp\!\!\!\perp C$ and $(A \perp\!\!\!\perp C \mid B)$, whereas $(A \perp\!\!\!\perp C \mid B)$ would hold in all other scenarios for the unshielded triple $A - B - C$.

Unfortunately, this version of faithfulness only holds in the probabilistic case, and does not apply to information-theoretic causal inference. If we have $A \perp\!\!\!\perp C$, it necessarily follows that $(A \perp\!\!\!\perp C \mid B)$, which means that we will never encounter the pattern characterizing $A \rightarrow B \leftarrow C$, because of a spurious independence $(A \perp\!\!\!\perp C \mid B)$ that is not induced by d-separation. In our application to languages, the problem can be made intuitive by stating that additional languages can only be used to “explain away” cognates for a given language pair, but we will never find additional cognates given the information from other languages.

Steudel et al. (2010) show that the independence relations derived from a sub-modular information measure still follow a weaker notion of faithfulness, which they call *monotone faithfulness*. Monotone faithfulness relaxes the enforced correspondence between d-separation and conditional independences by only requiring that $(A \perp\!\!\!\perp C \mid \mathbf{B})$ implies d-separation of A and C by a set \mathbf{B} if \mathbf{B} is minimal among all conditioning sets that render A and C independent. It turns out that the correctness proof for the PC algorithm can be adapted directly to show that it will return monotonically faithful representations if the input consists of monotonically faithful observations.

So what does this mean for the results of the PC algorithm on my cognate data? Even the weaker requirement of monotonic faithfulness still implies the very strong assumption that every scenario in which A has an influence on B and B on C , this would become visible as a dependence between A and C (because A and C would not be d-separated by the empty set). While this assumption may be unproblematic for continuous statistical variables, we cannot expect it to hold for my information-theoretic notion of independence. Again, the underlying problem is that I am modeling languages (and mutual information between them) as discrete sets of entities (or their overlap), which is far too coarse-grained to detect clean and consistent causal signals.

The assumptions behind the PC algorithm therefore imply the following statements which do not hold for languages:

- If two unrelated languages borrow from a third language ($A \leftarrow B \rightarrow C$), some words will always be borrowed into both languages ($I(A; C) \neq 0$).
- If one language influences another one which in turn influences a third one ($A \rightarrow B \rightarrow C$), this will always cause some lexical material to be transferred from the first one to the third ($I(A; C) \neq 0$).

While the first assumption might be defensible in my model (it is indeed very common that e.g. a language with state support leaves some common core of lexical items in all minority languages), the second one is extremely problematic,

since it is easily conceivable that if a language A borrows from a language B which in turn borrows from a language C , none of the lexical material from C will appear in A , especially if there is a temporal order to these contacts.

For our information-theoretic language variables, the inadequacy of the faithfulness assumption leads to many erroneous v-structures and a chaotic picture if we apply the second stage of the PC algorithm as is. In addition, since there will often be many separating sets of the same size, we are quickly faced with a well-known weakness of the PC algorithm: The results of its second stage are highly dependent on the order in which separating set candidates are tried out. In practice, this means that many possible orders have to be tested, often giving rise to conflicting evidence which needs to be reconciled. Crucially, this implies that we cannot directly rely on the separating sets to detect v-structures in a way that is robust enough for propagation as in the PC algorithm. Instead, we need a more robust way of detecting v-structures in cognate overlap patterns.

6.5.2 Unique Flow Ratio (UFR): flow-based v-structure testing

Above all, it is the lack of reliability in the independence tests which will lead to chaotic results if the separating set criterion is applied in the standard way. This lack of reliability is partially caused by erroneous cognacy judgments (which could be improved by hand-crafted annotations), but more importantly by a lack of statistical power in tests which involve distantly related languages, i.e. where the lexical overlaps only consist of a handful of items. But, even if the independence tests are correct (and they are more likely to be thanks to the FS criterion), there are typically many alternative separating sets, indicating that not much useful information can be extracted from the fact that a language was contained in some separating set.

To improve on the situation, and develop an alternative v-structure test which gets the maximum out of the little data I have available in my application, I need to get back to the motivation why the PC-algorithm and its variants detect v-structures by considering separating sets in unshielded triples. The essential idea justifying the inference of a v-structure $A \rightarrow B \leftarrow C$ was to decide whether B was necessary to separate A and C by inspecting the separating sets, and seeing whether B was in all (VPC) or the majority (SPC) of them.

Can this basic idea be applied more directly to the cognate set overlaps between languages? It turns out that a flow-based criterion can also help us here, as we can explicitly calculate how much of the overlap between A and C can only be explained via paths involving B on the causal skeleton. The essential idea is thus the same as the one behind the flow-separation independence test.

Let $c(L_1, \dots, L_k) := |\text{cog}(L_1, \dots, L_k)|$ be the number of cognate IDs shared between all the languages L_1, \dots, L_k , and define $c(A - B - C)$ (the *unique flow*) as the number of cognates which no path excluding B could have transported between A and C . Now, we can quantify the answer to the question how much B is needed to remove $A - C$ based on the answers of two simpler questions. The first question is whether there is as much unique flow as expected if B were needed for the separation. This can be captured by a single score measuring, with appropriate normalizations, the strength of unique flow in comparison to the flow we would expect if the true pattern were $A \rightarrow B \rightarrow C$ or $A \leftarrow B \rightarrow C$, i.e. if both arrows represented independent sampling of a certain number of items in the donor language, and the transmission of the sampled material in the recipient language. Due to the independence, the expected ratio of material shared between all three languages can be computed as the product of the two ratios on the links, leading to the following score quantifying how much of potential overlap is reflected by unique flow through B :

$$ufr_1 := \frac{\frac{c(A-B-C)}{\min(c(A), c(B), c(C))}}{\frac{c(A,B)}{\min(c(A), c(B))} \cdot \frac{c(B,C)}{\min(c(B), c(C))}} \quad (6.7)$$

A second question about the relevance of B for separation is how relevant the flow through B is for generating the actual overlap between the three languages, i.e. which ratio of the cognates shared can only be explained by transmission through B . This leads to a simple quotient as the second score:

$$ufr_2 := \frac{c(A - B - C)}{c(A, B, C)} \quad (6.8)$$

While $c(A - B - C)$ is quite costly to compute, it has the advantage of not requiring reuse of unreliable separating sets, and being determined only by the data and the skeleton. Based on these two quantitative answers to the partial questions, I can define a combined score by multiplication, as a good v-structure should score low on both measures:

$$ufr := ufr_1 \cdot ufr_2 \quad (6.9)$$

To observe the behavior of this *Unique Flow Ratio (UFR)* score, and to determine a good threshold for v-structure decisions, I produced five additional histories using the simulation model of Chapter 5, and extracted all triples of connected variables in the gold-standard graph. For these triples, it is thus known whether

they constitute colliders or not, providing me with a training set of 1,277 instances in order to determine the threshold value. To make the test cases more realistic, I emulated the noise level inherent in automatically clustered cognate data by adding a number of spurious cognates to each pairwise overlap. This noise for each link was sampled uniformly from between 0 and 20 additional cognates. The resulting distribution of *ufr* scores for both colliders and non-colliders is visualized in Figure 6.1. The distribution shows that as intended, almost all *ufr* values for colliders are very close to zero. In contrast, the values for non-colliders are distributed equally across the entire value range [0,1], with highest densities near 0.1 and 0.95, i.e. near the extreme values, but only a tiny fraction are as low as typical collider scores.

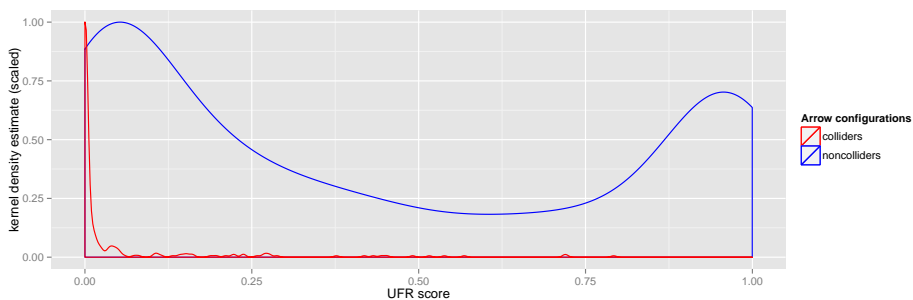


Figure 6.1: UFR scores for collider and non-collider test instances

Analysing the precision-recall tradeoff on the *ufr* values, the optimal threshold in terms of F-score was found to be as low as 0.004. This is the value I chose to adapt, so that the test $ufr < 0.004$ constitutes the UFR criterion for v-structure detection.

6.5.3 Triangle Score Sum (TSS): aggregating directionality hints

Another possibility to stabilize directionality inference is to move away from the framework of propagating binary v-structure decisions, instead embracing the fact that on noisy data, the different triples that each link in the skeleton takes part in may yield conflicting evidence of different strength. The obvious idea then is to quantify the directionality evidence present in each triple, and to combine these scores into an aggregate measure where conflicting evidence cancels out, and random spurious patterns in a single triple are overwritten by a larger number of more well-behaved triples. The basic quantification of directionality evidence in a triple can again be expressed in terms of the difference

between the three-way overlap we observe, and the overlap we would expect in a non-collider.

If we continue to only consider the unshielded triples that are still present in the skeleton, and try to aggregate an evidence score from these measures, we are frequently faced with the problem that there is an unshielded triple only for one direction, or that there is a severe imbalance in evidence strength for both sides. This means that small errors in the skeleton can still propagate into large errors in directionality inference.

This leads to the idea of not considering only unshielded triples, but all sets of three languages in the dataset for deciding each link, making the directionality inference step independent from skeleton inference. The resulting score infers for every pair of languages in the entire graph whether a connection between them would look directional, based on the triangle scores of that pair with every other language.

In the discussion that follows, shorthands will be used to compactly represent the relevant overlap quantities. For the shared material between each pair of variables, we use the Greek letter corresponding to the member of the triple that is not involved, i.e. $\alpha := \frac{c(B,C)}{\min(c(B),c(C))}$, $\beta := \frac{c(A,C)}{\min(c(A),c(C))}$, and $\gamma := \frac{c(A,B)}{\min(c(A),c(B))}$. In addition, I will use $\delta := \frac{c(A,B,C)}{\min(c(A),c(B),c(C))}$ for the amount of information shared between all three variables. In the notation, the hat diacritic will be used to denote expected overlaps, as opposed to observed values. I will be predicting the expected value $\hat{\delta}$ for the overlap based on the other observable overlap ratios, and then derive a quantification of the evidence against the assumed collider from the difference between observed δ and expected $\hat{\delta}$.

Let us now derive an approximate expression for $\hat{\delta}$ in the pattern $A \rightarrow B \leftarrow C$, which in the absence of latent variables is the only v-structure scenario. The only way in which a cognate set can come to be shared between all three languages in this scenario is if it was already shared between A and C , and was borrowed into B from one of the two languages. The percentage of material shared between A and C is given by β , and the percentage of material in B borrowed from A and C is simply γ and α , respectively. Assuming independent sampling, the percentage of items which end up in δ via the transfer $A \rightarrow B$ should be equal to $\gamma \cdot \beta$, and the percentage transmitted via $C \rightarrow B$ should be $\alpha \cdot \beta$. When we simply add up these percentages, we will count some of the expected transferred items twice. The probability for each element in $\text{cog}(A, B, C)$ to have been selected twice is simply β , because this is the probability that a random element picked for transfer from A or C is shared by both nodes. We therefore expect $\delta \cdot \beta$ items to have been

counted twice. Using this as a correction, we receive $\hat{\delta} = \gamma\beta + \alpha\beta - \hat{\delta}\beta$. Resolving this expression for $\hat{\delta}$, we arrive at the following equation for the expected three-way overlap in a collider:

$$\hat{\delta}(A \rightarrow B \leftarrow C) := \frac{\beta(\alpha + \gamma)}{1 + \beta} \quad (6.10)$$

Note that we considered the shielded case here. The situation of an unshielded triple is covered by $\beta := 0$, which causes the definition to collapse to $\hat{\delta} = 0$, capturing the intuition I already used in UFR, namely that a v-structure should result in zero three-way overlap that cannot be explained by other paths.

Turning the fitting of δ to $\hat{\delta}$ into a fit score could be done in a number of ways, but the easiest way turned out to be to form the quotient of the smaller by the larger of the two values, with special treatment for boundary cases to avoid division by zero:

$$ts(A \rightarrow B \leftarrow C) := \begin{cases} 1 - \frac{\min(\delta, \hat{\delta}(A \rightarrow B \leftarrow C))}{\max(\delta, \hat{\delta}(A \rightarrow B \leftarrow C))} & \text{if } \delta > 0 \text{ or } \hat{\delta}(A \rightarrow B \leftarrow C) > 0 \\ 0 & \text{if } \delta = 0 \text{ and } \hat{\delta}(A \rightarrow B \leftarrow C) = 0 \end{cases} \quad (6.11)$$

$ts(A \rightarrow B \leftarrow C)$ measures the strength of evidence which the cognacy overlaps between the three languages provide against the v-structure pattern. Because all triangle scores are on the same scale, but cover overlaps of different strengths, we cannot directly add up these triangle scores to aggregate the directionality information they encode into a global evidence score. To avoid strong influences from languages with little overlap to the pair in question, it is necessary to weight the contributions to the triangle score sum by their relevance to the link in question. One simple way to define these weights is by considering the strength of the connection of the third variable C to either of the two variables involved, and then normalizing these weights to keep the weighted sum in the $[0, 1]$ range. In practice, making the weight differences a little more pronounced helped to moderate the contribution of distant third languages, leading me to square the weights before normalization:

$$w'(A \rightarrow B; C) = \max\left(\frac{c(B, C)}{c(B)}, \frac{c(A, C)}{c(A)}\right)^2 \quad (6.12)$$

The normalization of weights then happens in the obvious way:

$$w(A \rightarrow B; C) = \frac{w'(A \rightarrow B; C)}{\sum_{D \notin \{A, B\}} w'(A \rightarrow B; D)} \quad (6.13)$$

Finally, the weighted sum of triangle scores over all third variables C gives us the definition of the *Triangle Score Sum* (TSS) after which the directionality inference method is named:

$$tss(A \rightarrow B) = \sum_{C \notin \{A, B\}} w(A \rightarrow B; C) \cdot ts(A \rightarrow B \leftarrow C) \quad (6.14)$$

The TSS can be calculated on each link for arrows in both directions, and the results have the same scale, yielding a natural decision criterion in terms of the evidence strength quotient:

$$sc(A \rightarrow B) = \frac{tss(A \rightarrow B)}{tss(B \rightarrow A)} \quad (6.15)$$

To understand better how TSS works, let us take a look at an example from the Baltic Sea scenario. For the link between Russian and Kildin Saami (*sjd*), the triangles with the heighest weight are given by Skolt Saami (*sms*), due to its high overlap with Kildin Saami, as well as by Polish (*pol*) and Belarusian (*bel*), both due to their high overlap with Russian. Together, these three triangles account for 62.3% of the total weight sum, meaning that if a strong tendency arises from these three triangles, it will not be inverted by the remaining low-overlap triangles. Let us start with $ts(sjd \rightarrow rus \leftarrow bel)$. If this were a true v-structure, we would expect a three-way overlap of 34.55 cognates. In reality the overlap is higher at 47, giving a moderate counterevidence score of $ts(sjd \rightarrow rus \leftarrow bel) = 0.265$. From the reverse perspective, we have a much better fit at $ts(rus \rightarrow sjd \leftarrow bel) = 0.068$, because the prediction in this case would be an overlap of 50.42 cognates. The first triangle thus delivers a score contribution that is almost four times higher for the arrow direction $rus \rightarrow sjd$. The scores for the other two triangles we consider point in the same direction: $ts(sjd \rightarrow rus \leftarrow pol) = 0.362$, but $ts(rus \rightarrow sjd \leftarrow pol) = 0.105$, and $ts(sjd \rightarrow rus \leftarrow sms) = 0.232$ is much higher than $ts(rus \rightarrow sjd \leftarrow sms) = 0.073$. From these three triangles, we can begin to approximate $tss(rus \rightarrow sjd) = 0.300 \cdot \frac{0.068}{0.265} + 0.165 \cdot \frac{0.105}{0.362} + 0.148 \cdot \frac{0.073}{0.232} + \dots \approx 0.2786$. The actual value with the full triangle sum is $tss(rus \rightarrow sjd) = 0.6658$, i.e. the missing triangles will equal the score out quite a bit, although the general tendency for more evidence against $sjd \rightarrow rus$ (i.e. a signal against the wrong directionality) remains.

The challenge of the TSS approach is to decide on a threshold for turning the evidence strength quotient $sc(A \rightarrow B)$ on each link into a directionality decision $A \rightarrow B$. For instance, is $sc(A \rightarrow B) = 0.667$, i.e. 50% more evidence in this direction than the other, enough to make the decision? Reusing the five scenarios generated to derive the best ufr threshold, I extracted the 211 links that were not part of the phylogenetic tree, of which 187 are monodirectional. For each link $A \rightarrow B$ I computed $sc(A \rightarrow B)$. If $sc(A \rightarrow B) < 1.0$, evidence pointed in the right direction, making this an instance where TSS worked correctly. Counterbalancing this are the cases where $sc(A \rightarrow B) > 1.0$, i.e. where the implied arrow was inverted. Figure 6.1 shows the distribution of $sc(A \rightarrow B)$ for the good instances, and the inverse for bad instances. The separation is disappointingly bad, indicating that the possible advantage of a pairwise criterion like TSS over a triple-based like UFR is much reduced by the more difficult classification task. Still, it is clearly visible that the correct arrows tend to cluster closer to 0, and the inverted arrows closer to 1. This gives me an empirical basis for deciding on a threshold value, because I want as many good instances as possible below the threshold, while keeping as many bad instances as possible above it, because I prefer not assuming an arrow at all to inferring wrong directionality of contact.

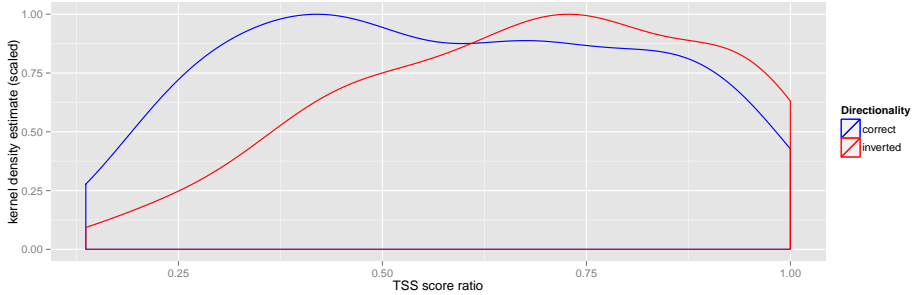


Figure 6.2: TSS scores for correct and inverted arrows

This is another instance of a precision-recall tradeoff, where in addition to the correct and inverted TSS scores, the bidirectional links filter in as additional false instances in either direction. Aiming for a precision of 70%, we can only get a recall of 21% with a threshold value of 0.424. On the other hand, if we want to find two thirds of all arrows (66% recall), we can only achieve that with a threshold of 0.982, i.e. we would have to add arrows very aggressively. A good compromise needs to be found somewhere between these values. Precision remains about constant at 64%, i.e. about two thirds of inferred arrows are correct, across a very

large range of threshold values between 0.555 and 0.720, which is when it starts to drop significantly. At the threshold value of 0.72, the maximum of this range, recall is at 47%, which seems to be a reasonable compromise given the overall low performance.

In these considerations, we have not considered how the performance of TSS and UFR varies with the number of shared cognates defining each link, i.e. with data sparseness. Unavoidably, the rather simple TSS method can easily be misled by noisy or sparse data. My general impression is that TSS works considerably better on the NorthEuraLex data than the test cases suggest, possibly because the relevant contacts shaping real datasets tend to include large numbers of cognates. The evaluation will shed more light on this question, showing that the intuition is largely, but not always, correct.

6.6 The phylogenetic guide tree

As explained at the start of this chapter, in order to infer fully general evolutionary networks via causal inference, I will need to infer data for the proto-languages. To establish the proto-languages that need to be reconstructed, and as a guiding datastructure for the reconstruction, I need some phylogenetic tree over the attested languages as a starting point. A fully integrated system that starts with word lists and returns a lexical flow network as a result, would therefore include at least a rudimentary component for phylogenetic tree inference. As we have seen in §2.5.2, phylogenetic tree inference is already a very well-established field, and any of the methods discussed by Felsenstein (2004) can in principle be used for this purpose.

The purpose of my investigation is to compare the performance of different network inference variants on the best possible guide tree, and not to provide a fully integrated software package. My software implementation therefore requires the user to specify cognate sets over the input data, and some phylogenetic tree over the same languages, with branch lengths. This leaves it to the user to plug in cognate detection or tree inference algorithms of their choice, making it possible for my system to profit from future advances in these two subfields, although basic methods for a one-pass processing from word lists to lexical flow network will be provided as part of the release version of my software.

For my experiments on the simulated data from Chapter 5, I will simply use the binary tree created by the simulated language split events, together with the branch lengths defined by the times at which these events occurred. This tree is then reduced to the languages which still lived at the end of the simulation, first

removing all the branches leading only to extinct languages, and then removing non-branching nodes in the resulting tree while maintaining consistent branch lengths.

For the NorthEuraLex data, I will use the expert tree defined by Glottolog, again reduced to only those leaves which are attested in the database. Adding branch lengths to this non-binary tree is one of only two places where I found it necessary to use existing phylogenetic inference tools. According to a suggestion by Gerhard Jäger, the inferred cognates were encoded as binary features and given to the IQ-Tree software (Nguyen et al. 2015) as input for inferring branch lengths on the unrooted Glottolog tree. For the output, the tree was re-rooted with Mandarin Chinese (*cmn*) as an outlier. None of the family-internal branch lengths inferred by the selected model GTR2+FO+ASC+G4 seemed unpalatable on inspection (see Appendix B.2 for a visualization), so that the resulting tree seems adequate as input for reconstruction methods.

6.7 Deriving proto-language models

If I want to be able to treat proto-languages as observed variables in the causal inference paradigm, I need to put some effort into deriving at least a good hypothesis about the presence or absence of each cognate class at each ancestral node in our phylogenetic tree. The idea essential to reconstruction then is to assume that the proto-language of some group of observable languages is most likely to have contained those cognate sets which are present in a large (and diverse) subset of its descendant languages.

This is very similar to the reconstruction of ancestral genomes in bioinformatics, from where we can take a variety of readily applicable algorithms. In addition to the mainstream ASR (Ancestral State Reconstruction) techniques in this tradition, I also present a naive threshold-based approach that will be used as a baseline. After evaluating the different methods on the simulated data, only the best two methods will be used to arrive at a usable reconstruction of ancestral cognacy in the NorthEuraLex dataset, and this reconstruction will be treated just like actual observations for phylogenetic network inference.

6.7.1 Ancestral state reconstruction algorithms

From the large number of ASR methods discussed in the literature, I will only sample a few very common and robust variants that are trivial to extend from nucleotides to cognacy data. The treatment of the methods cannot be complete,

and I neither have the space to give examples of each reconstruction algorithm here, nor to provide the full algorithmic details for each method. Still, my explanations should suffice to provide the reader with correct intuitions about each method, and the formal statements are precise enough to completely describe the core features of any implementation.

6.7.1.1 Naive threshold-based approaches

For the initial experiment, I opted for a simple recursive criterion. For each node in the expert tree, it includes those cognate sets that are present in a majority (more than 50%) of its immediate daughter languages. This implies we start at the observable languages and their cognate sets, and reconstruct upwards in the expert tree, arriving at the root language(s) in a single bottom-up pass through the entire tree.

If we introduce the notation $P(l = A)$ for the probability of cognate class A being reconstructed for some concept in the language l with children l_1, \dots, l_k , the majority-based reconstruction can be written as a recursive formula

$$P(l = A) := \begin{cases} 1 & \text{if } \frac{1}{k} \sum_{i=1}^k P(l_i = A) > 0.5 \\ 1 & \text{if } k = 0 \text{ and the word for the concept was assigned to class } A \\ 0 & \text{else} \end{cases} \quad (6.16)$$

This simple definition directly implements an important property of any useful reconstruction, namely that a word which was borrowed once at an intermediate stage and therefore now turns up in every language of a branch with many languages, will not end up in the proto-language if there is no other branch which also features a cognate. This usage of the tree to channel the information is superior to an even more naive criterion that would simply count the occurrences of each cognate class at the leaves under each ancestral node.

The main problem of this approach is of course that it tends to err very much on the safe side, as reconstruction stops as soon as there is a configuration of two subgroups for which different ancestral states are reconstructed, which occurs very often in real data. On the other hand, the few reconstructions this method arrives at tend to be extremely reliable. A threshold lower than 0.5 could be introduced for nodes which are more than binary-branching, but this would quickly lead to far too generous reconstruction.

In sum, threshold-based approaches can only be expected to show a very low performance, and do not offer many options to fine-tune them to the specific task. I will still use the majority-based approach as a baseline for comparison with more advanced alternatives.

6.7.1.2 Parsimony-based approaches

The most direct modern approach to ASR is based on maximizing parsimony, which can be seen as a formalization of Occam’s razor, i.e. the principle of selecting the simplest hypothesis which explains all the data. In the context of ASR, parsimony can simply be described by the number of state changes which the model needs to assume. If we reconstruct the ancestral states in such a way that the number of mutations the model needs to assume is minimized, we are maximizing the parsimony of our reconstruction.

The standard algorithm for maximum parsimony is the Sankoff algorithm, originally defined by [Sankoff \(1972\)](#), which uses dynamic programming to keep track of the minimal number of replacement operations which needs to be assumed for smaller subproblems, and fills a table for each node, storing the total number of replacement operations which each state at that node would imply. For the current optimal solution in each cell, backpointers are stored which make it possible to reconstruct the configuration of ancestral states which led to the minimum number of replacement events for the entire problem. In my application, there are two useful variants of this basic Sankoff algorithm, which differ in whether we consider the presence of each cognate set as an independent character, or treat the different cognate sets for one concept as the different values of a single multistate character.

In the first version, which I will call *multi-value MP reconstruction*, a separate run of this basic Sankoff algorithm is performed for each presence-absence character, meaning that the table for each node only has two cells (one for presence, one for absence). Formally, each cognate set A is reconstructed for the subset $L_A \subseteq L$ of all languages for which $\sum_{(l_i, l_j) \in E} 1_{l_i \in L_A, l_j \notin L_A}$ is minimal, under the condition that $l \in L_A$ holds for every language l where the cognate set A is attested, and $l \notin L_A$ for every language l where it is not. In this version, it is possible that absence is reconstructed for all cognate sets, leaving a node without reconstruction, or that presence is reconstructed for more than one cognate set at a given node, hence the name I am using.

The second version, which I will call *single-value MP reconstruction*, reconstructs exactly one cognate set from the set of ancestral nodes, out of a set of

candidates defined by the cognate sets occurring in the attested languages assigned to that phylogenetic unit. The Sankoff table has as many cells as there are candidate cognate sets. Formally, given a cost function $c(A, B)$ for the replacement of cognate sets (typically, $c(A, A) := 0$ and $c(A, B) := 1$ for $A \neq B$), this variant assigns a tuple of cognate sets (A_1, \dots, A_n) to all the nodes (l_1, \dots, l_n) such that that $\sum_{(l_i, l_j) \in E} c(A_i, A_j)$ is minimal, while keeping one set A_j of the cognate sets assigned to each attested language l_j fixed.

The main problem of parsimony-based ASR is that different branch lengths cannot be accounted for. This exploits existing knowledge only suboptimally, since if one of two languages forming some phylogenetic unit is known to be more conservative (which would be reflected by a shorter branch length or a lower replacement rate in phylogenetic tree models), this will make it more likely that the ancestral set survived in this language, making it more informative for the reconstruction. A related more general problem of MP is that there will often be a large number of maximally parsimonious reconstructions, i.e. parsimony alone does not give us a sufficient decision criterion for finding a single reconstruction.

6.7.1.3 Fully probabilistic approaches

More recent approaches to ASR work in the maximum-likelihood (ML) paradigm. These fully probabilistic methods treat the discrete states y at internal nodes as unknown parameters whose values need to be estimated given the data x we observe. An obvious choice is the maximum likelihood estimator, which maximizes $P(y|x)$, the probability of different parameter values given the observed data, with the help of Bayes' rule:

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_y P(y)P(x|y)} \quad (6.17)$$

While the denominator is hard to compute, it is independent of y and is therefore irrelevant for maximizing the expression. Maximum likelihood estimation assumes that no prior information about plausible values $P(y)$ is available, which reduces the task of maximizing $P(y|x)$ to maximizing the likelihood $P(x|y)$ that we see the data given parameter values y . A good ML estimator will typically converge to the most likely value of y , although it is possible that other values of y are almost as likely, and that the ML estimate \hat{y} is even an outlier in the space of plausible parameter values. Still, ML estimates of model parameters provably maximize the agreement of the model with the data in many types of in-

ference problems, given ML estimation a strong independent motivation outside the Bayesian paradigm.

In the application to ASR, optimization is based on an explicit parameterized evolutionary model $P_{ij}(\theta)$ which fully describes how each state i is likely to evolve along a given phylogenetic tree, and thereby assigns a probability $P(x|y, \theta)$ to the observed data for each set of parameter values. If the evolutionary model is Markovian (the probability of each state change only depends on the parent state, not on earlier states), dynamic programming can be used to efficiently derive the internal states y which maximize $P(x|y, \theta)$. For different applications, different evolutionary models are plugged into this basic paradigm.

Apart from the different evolutionary models, the main dividing line between approaches to ML-based ASR is in the method the optimal ancestral states at internal nodes are calculated. In the computationally simpler marginal reconstruction as introduced by [Yang et al. \(1995\)](#), reconstruction of states at an internal node A is done by re-rooting the tree such that A becomes the root, and then computing the likelihoods for the different states at each node in a bottom-up fashion, summing over all possible combinations of states in the children $ch(A) = \{B_1, \dots, B_n\}$ down to the leaf nodes:

$$\hat{y}_A := \arg \max_i L_A(i), \quad L_A(i) := \sum_{j_1, \dots, j_n} \prod_{k=1}^n P_{ij_k} L_{B_k}(j_k) \quad (6.18)$$

These values can again be computed by dynamic programming, meaning that the computation is only more time-consuming than the Sankoff algorithm by a factor linear in the number of ancestral nodes.

By contrast, in the more complex joint reconstruction, the likelihood is jointly maximized over reconstructed values at all nodes, which is computationally a lot more demanding. Moreover, according to [Yang et al. \(1995\)](#), this variant is less suitable for retrieving optimal reconstructions at each ancestral node (because suboptimal local solutions can be necessary for an optimal global reconstruction), which leads me to disregard it as an option for the current application.

Implementing marginal reconstruction from scratch is non-trivial, and the implementation details of existing systems are not fully specified in the literature. This makes ML reconstruction the second of the two places in my infrastructure where it seemed more prudent to rely on third-party software instead of engineering my own implementation to behave exactly like a reference implementation. For marginal ML reconstruction, I am using a somewhat brittle interface from my Java code into the R package *phangorn* via system calls to *Rscript*, and a

custom method on the R side for output in a Nexus format which can be read back into Java and mapped onto the original character information.

In analogy to multi-value MP reconstruction, *multi-value ML reconstruction* operates on binary characters which encode the presence or absence of each cognate set at each node. For non-attested nodes, the marginal reconstruction assigns probability values to each of the two values 0 (absence) and 1 (presence) of each character. If the probability of 1 at an ancestral node is above 50%, the corresponding cognate set is reconstructed for the respective proto-language. This variant shares its basic properties with multi-value MP reconstruction. Ancestral nodes can remain without any reconstructed cognate set if for each cognate set character, the probability of the value 1 was below 50%.

For *single-value ML reconstruction*, we only use one multi-state character for each concept, where each state encodes one of the possible cognate sets. The marginal reconstruction then produces a probability distribution over all the possible cognate sets at each ancestral node, and we only reconstruct the one with the highest probability. The behavior of the resulting method is again similar to single-value MP reconstruction in that exactly one cognate set will be reconstructed for each ancestral node. Due to limits in the phangorn implementation which are well-justified for its main field of application in biology, there is no support for ambiguity in the input data. This prevented me from modeling the synonyms in the NorthEuraLex data, so that the single-value ML reconstruction only builds on the first form in each list of concept realizations.

Maximum-likelihood methods for ASR work on a single phylogenetic tree with branch lengths, which is typically inferred by a different (often Bayesian) method. This can lead to problems because the likelihood of the individual tree hypothesis that the reconstruction is based on is usually quite low, and a reconstruction which takes more than a point estimate of plausible trees into account will be much more sound if computationally feasible.

To account for the uncertainty in the tree reconstruction, Bayesian methods which account for the uncertainty in both the ancestral characters and the tree structure have been developed. The hierarchical Bayes method by [Huelsenbeck & Bollback \(2001\)](#) is able to model uncertainty in the tree, branch lengths, and the substitution model at the same time, but is reported to lead to very high uncertainty in the results. The advantages of this more accurate quantification of uncertainty are unclear in an application where only a single optimal reconstruction can be used. This and the prohibitive computational complexity associated with fully Bayesian methods justify confining myself to the simpler maximum-likelihood paradigm.

6.7.2 Evaluation of ASR algorithms on simulated data

It is easy to evaluate the different ASR methods on the simulated data. Let us act as if we lost the true configurations of all dead languages in the simulated set, and are left with a reduced version of the true tree which only contains the living leaves, plus the internal nodes which are necessary to keep the branching structure over these leaves. We can then feed the reduced true tree and the data from the leaves into the different reconstruction algorithms.

To compare the results, we return to the true data we discarded, and compute the percentage of bits (representing presence or absence of each cognate class) at the reconstructed nodes that correspond to the true values, as well as precision and recall on the level of reconstructed classes. Finally, we analyse the difference in reconstruction quality for proto-languages of different age, to find out whether one of the algorithms is more robust at higher time depths.

The following five previously introduced ASR methods were compared in this way:

1. **Mjrty** (majority-based), i.e. using the naive criterion of reconstruction or presence in the majority of children
2. **MPsgl** (single-value MP), i.e. using the Sankoff algorithm to reconstruct exactly one correlate set for each concept at each node
3. **MPmlt** (multi-value MP), i.e. using the Sankoff algorithm to reconstruct a binary presence/absence value for each correlate set and concept
4. **MLsgl** (single-value ML), i.e. using a marginal estimator on multistate characters, and selecting the most likely cognate set for each ancestral node
5. **MLmlt** (multi-value ML), i.e. using a marginal estimator on binary presence/absence values, reconstructing the correlate set if presence is more likely than absence

Running the different reconstruction algorithms and comparing the results on our 50 simulated linguistic histories, we get the numbers in Table 6.1. The overall picture is clearly in favor of the MP and ML methods, but the differences between the two are not very pronounced. Only for the single-value variants, there is a clear advantage for ML over MP. For both MP and ML, the single-value variants shows higher recall than the multi-value variants, because they always produce some reconstruction even if evidence is not very strong, but this comes at a significant cost to precision. The very conservative majority method has the

highest precision, but achieves this at a much higher cost to recall than the MLmlt method almost equalling it in precision. Overall, MLsgl is clearly the best method, as it achieves the highest recall by a significant margin, without compromising too much on precision.

Table 6.1: Performance of different ASR algorithms on simulated data

#	Mjrty	MPsgl	MPmlt	MLsgl	MLmlt
Accuracy	0.9804	0.9774	0.9830	0.9803	0.9829
Precision	0.8463	0.5720	0.8199	0.6292	0.8255
Recall	0.3212	0.5874	0.4586	0.6289	0.4386
F-score	0.4656	0.5796	0.5882	0.6291	0.5813

In order to check the suspicion that the differences between the algorithms might become more pronounced at higher time depths, where the few really difficult reconstructions are, we can evaluate the performance separately on proto-languages of different ages. Figure 6.3 visualizes the performance of the five reconstruction methods across twenty different age ranges.

Whereas F-scores are satisfactory for reconstructed languages that go back only a few hundred simulated years, already at a time depth of 1,000 years substantial differences in the performance of the different algorithms start to appear. At higher time depths, recall becomes so low that the F-scores are already surprisingly close to zero. Still, there are some interesting developments at this highly problematic range. At a time depth of 5,000 years there is a clear split between three methods which essentially do not reconstruct anything useful any more (the majority method and the multi-state methods, all with recall under 10%), and the other three methods which still manage to reconstruct at least some cognate sets (recall of 20-30%), albeit with a high error rate (less than 40% of the reconstructed cognate sets are correct). We can conclude that the acceptable overall performance of all methods in the previous analysis was mainly due to the dominance of simple reconstruction tasks if we evaluate across entire trees with many more younger languages than old ones. The task of reconstruction at higher time depths is still very much an unsolved problems. At these higher time depths, the two multi-state methods perform better precisely because they are biased towards assuming some reconstruction even if not enough evidence is available. However, since the overall advantage of the maximum-likelihood methods remains stable for proto-languages of any age, we at least know that they are clearly our best methods for phylogenetic flow inference, and will therefore be the default reconstructions used for the further experiments. The superiority of

6 Phylogenetic lexical flow inference

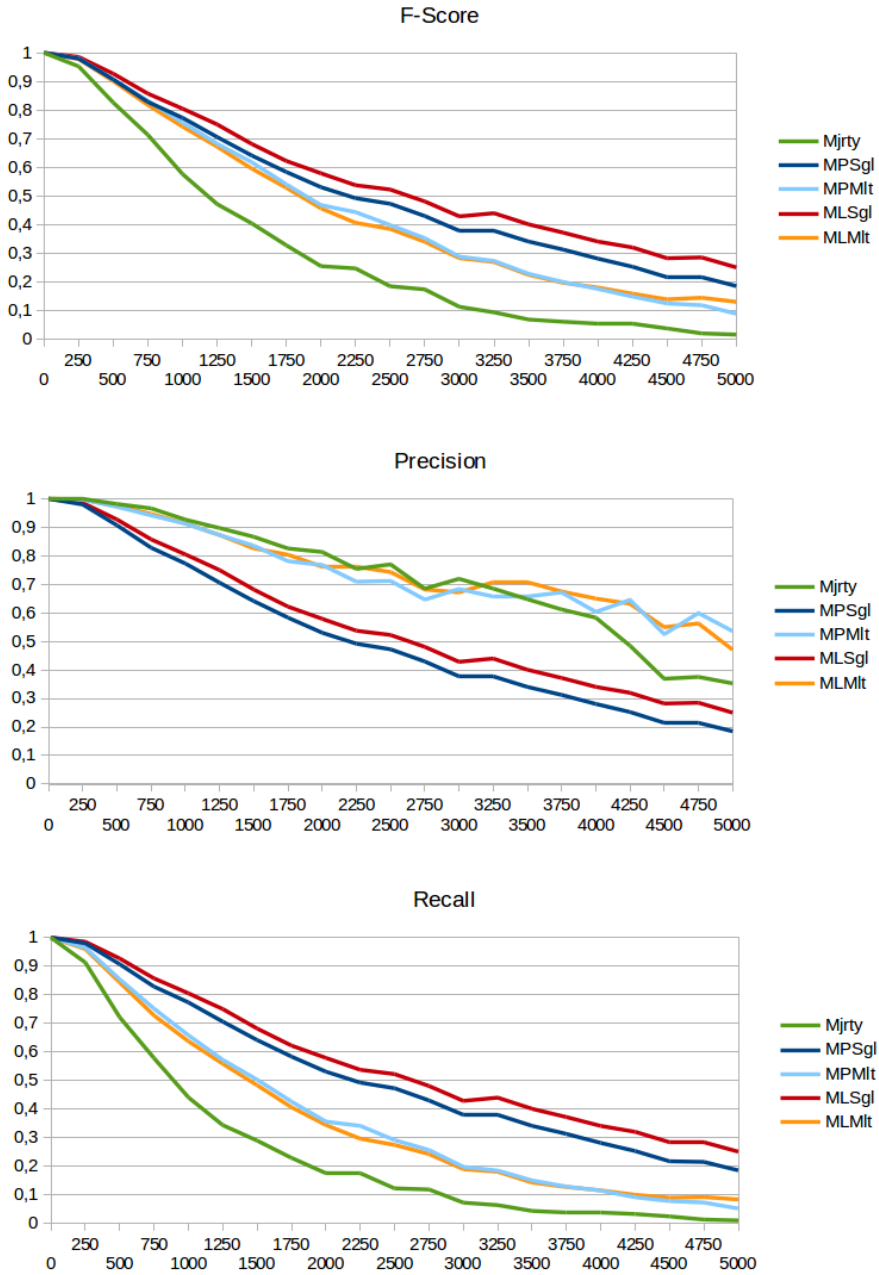


Figure 6.3: Development of ASR performance with age of reconstructed language

ML reconstruction over MP reconstruction was recently confirmed by Jäger & List (2017) on a different test set comprising real data from Indo-European, Austronesian, and Sinitic languages. In contrast to my simulated data, the performance of the methods is only evaluated against reconstructions at the level of the respective proto-language, i.e. at a high time depth. Just as in my experiment, the ML-single variant (ML-multi in their terminology) wins by a large margin.

6.8 Phylogenetic Lexical Flow Inference (PLFI)

Given a tree skeleton predetermined either by previous knowledge or inferred by means of a phylogenetic method, we can now apply ASR methods to derive cognate sets for the tree's internal nodes, and proceed to apply causal inference to the resulting dataset. Of course, this means that the performance of the method will hinge very much upon the quality of the reconstruction.

Very optimistically assuming that the output of our ASR method approximates the true history very closely, we treat all the nodes in our phylogenetic tree as observable languages, and apply lexical flow inference to a mixture of attested languages and reconstructed proto-languages. This is the algorithm which I propose to call Phylogenetic Lexical Flow Inference (PLFI). Algorithm 2 gives a description of all PLFI variants in pseudocode.

The PLFI algorithm requires either an expert tree, or a tree inferred by some phylogenetic tree inference algorithm, as input in addition to the cognacy data. After a preprocessing stage where the tree is reduced to the leaves for which data are available, some ancestral state reconstruction method (MLsgl by default) is applied to the tree and the data at the leaves in order to infer the presence or absence of each cognacy class at every non-leaf node in the reduced tree. The causal graph is then built over all the tree nodes.

The PC algorithm starts out with the fully connected graph, i.e. a network in which every pair of nodes is connected by a link. This graph is progressively thinned out to yield the skeleton by means of conditional independence tests based on separating set candidates. In each iteration, the algorithm increases the size s of separating sets it considers. At each stage, it iterates through all the links remaining in the graph, and tries to build a separating set of size s from the neighbors of the two languages it tries to separate.

Unlike in the vanilla PC algorithm, there is a defined order in which the links are tested for deletability. The PLFI algorithm always starts with the weakest remaining link, on the grounds that such links are more likely to arise due to random fluctuation in the noisy cognacy judgments. The links which represent the highest overlap are always checked last.

Algorithm 2 PLFI(L_1, \dots, L_n)

```

1: ASR method  $asrM \in \{MjrtY, MPsgl, MPmlt, MLsgl, MLmlt\}$ 
2: skeleton inference method  $sklM \in \{PC, PS, FS\}$ 
3: directionality inference method  $dirM \in \{VPC, SPC, UFR, TSS\}$ 
4:  $T := phyloInference(\{L_1, \dots, L_n\})$ , or an expert tree
5:  $T := reduce(T, \{L_1, \dots, L_n\})$ , the phylogenetic tree reduced to attested leaves
6:  $T := asr(T, asrM)$ , add cognate classes to ancestral nodes by reconstruction
7:  $\mathcal{L} := nodes(T)$ 
8:  $G := (\mathcal{L}, E) := (\mathcal{L}, \{\{L_i, L_j\} \mid L_i, L_j \in \mathcal{L}'\})$ , the complete graph
9:  $S : \mathcal{L} \times \mathcal{L} \rightarrow \wp(\mathcal{L})$ , the separating set storage
10:  $s := 0$ 
11: while  $s < |\mathcal{L}| - 2$  do
12:   for  $\{L_i, L_j\} \in G$  by increasing strength of remaining flow do
13:     if  $sklM \in \{PC, PS\}$  then
14:       for each subset  $S \in \wp(N)$  for neighbors  $N$  of  $L_i$  or  $L_j$  do
15:         if  $sklM = PC$  or all elements of  $N$  are on paths from  $L_i$  to  $L_j$  then
16:           if  $|S| = s$  and  $I(L_i; L_j|S) < 0.025$  then
17:             remove  $\{L_i, L_j\}$  from  $G$ ,  $S(L_i, L_j) := S(L_i, L_j) \cup \{S\}$ 
18:           end if
19:         end if
20:       end for
21:     else if  $sklM = FS$  then
22:       for each combination  $P_1, \dots, P_k$  of paths from  $L_i$  to  $L_j$  of length  $\leq 4$  do
23:         if  $|S| = s$  for  $S := \bigcup \{P_1, \dots, P_k\}$  then
24:           if ratio of  $c(L_i, L_j)$  not explainable by flow across  $S$  is  $< 0.025$  then
25:             remove  $\{L_i, L_j\}$  from  $G$ ,  $S(L_i, L_j) := S(L_i, L_j) \cup \{S\}$ 
26:           end if
27:         end if
28:       end for
29:     end if
30:   end for
31:    $s := s + 1$ 
32: end while
33: if  $dirM = TSS$  then
34:   for  $\{L_i, L_j\} \in G$  do
35:     if  $sc(L_i \rightarrow L_j) < 0.72$  then
36:       add arrow  $L_i \rightarrow L_j$  to network
37:     end if
38:   end for
39: else
40:   for  $L_i, L_j, L_k \in \mathcal{L}$  where  $\{L_i, L_j\}, \{L_j, L_k\} \in E$  but  $\{L_i, L_k\} \notin E$  do
41:     if  $(L_i \rightarrow L_j \leftarrow L_k)$  is a v-structure according to  $dirM$  and  $S(L_i, L_k)$  then
42:       add arrows  $L_i \rightarrow L_j$  and  $L_k \rightarrow L_j$  to network
43:     end if
44:   end for
45:   propagate arrows according to rules  $\mathcal{R}_1$  to  $\mathcal{R}_3$ 
46: end if
47: return network consisting of  $G$  and arrows

```

The separating set candidates which are tried out for the conditional independence tests depend on the skeleton inference method. If the skeleton inference method is set to *FS*, the flow-separation criterion introduced in §6.4 is applied, so that every separating set candidate is composed of paths connecting the two languages in the current skeleton. Using one of the other skeleton inference methods, this first stage of the algorithm can also be configured to behave just like the vanilla PC algorithm or like PC*.

For the directionality inference stage, the user has the choice between four variants. The vanilla PC variant only differs from stable PC and the UFR criterion in the way in which v-structures are detected. In all three cases, the three standard directionality propagation rules of the PC algorithm are applied until all links are directed or none of the rules applies any longer. The only directionality inference method which works differently is the TSS-based variant, which infers the directionality on each arc separately by checking the TSS ratio against the threshold determined in §6.5.

6.9 Evaluation of PLFI

There are two main ways to evaluate phylogenetic lexical flow inference which promise to be of interest. First, we can evaluate on perfect proto-data to determine the theoretical maximum performance the method could achieve if we had access to a perfect reconstruction. This gives us an upper bound on performance, because any real reconstruction will deviate from this perfect picture. To generate the input data for PLFI, we simply take the final state of the simulation for all languages, whether living or dead. This implies we include data from entire unattested lineages, including what we have earlier called para-languages and substrates, so that in this scenario, we have actual causal sufficiency, and the PC algorithm should be applicable without restrictions.

The more realistic evaluation of the method builds on reconstructed proto-data. Here, we reduce the known tree to ancestors of living languages, leaving only the lowest common ancestor in the cases where internal nodes become unary because one of two branches is deleted. Then, we apply one of the ASR algorithms to produce the data for the internal nodes of the reduced tree. To limit the number of cases to consider, we only evaluate on the two ML reconstruction methods which performed best on the simulated data, in order to be certain that our findings on simulated data carry over to the NorthEuraLex dataset.

6.9.1 Evaluation metrics for phylogenetic flow

Since causal inference as we employ it consists of the two stages of skeleton inference and directionality detection, for both of which we have multiple options at our disposal, it makes sense to first evaluate performance at the skeleton inference task, and then evaluate the different methods for directionality detection on the results of the best skeleton inference method.

For each connection $L_1 - L_2$ which was found by the phylogenetic flow algorithm in the reconstructed network G_{res} , we can ask whether it corresponds to a lateral connection $L_1 - L_2$ in G_{true} . If this is the case, we call the inferred connection a true positive (tp), otherwise a false positive (fp). If a lateral connection in G_{true} does not have an equivalent in G , we count it as a false negative (fn). If for a pair of languages L_1 and L_2 , neither graph has a connection $L_1 - L_2$, we count it as a true negative (tn). From these four numbers tp , fp , fn , and tn , we can compute precision and recall, the standard measures of performance on binary classification tasks. The *skeleton recall* (SkRc) is then defined as $\frac{tp}{tp+fn}$, i.e. the ratio of links in the true skeleton which the algorithm managed to reconstruct. Analogously, the *skeleton precision* (SkPr) can be written as $\frac{tp}{tp+fp}$, i.e. the ratio of links in the reconstructed skeleton which are correct. Both measures can be combined in a standard way via $2 \cdot \frac{SkPr \cdot SkRc}{SkPr + SkRc}$ to the *skeleton F-score* (SkFs), a combined performance measure which reaches high values if precision and recall are well-balanced.

For the evaluation on perfect proto-data, it is easy to adapt these standard performance measures, because there are no complications due to gaps in our knowledge which ASR cannot close. When evaluating on reconstructed proto-data, the situation is a little more complicated because some of the true connections cannot conceivably be found in the absence of substrates, and even in a perfect result, only the links between ancestors of related languages will be represented. For simulated contacts between pairs of languages where either only the donor or recipient of lexical material is in the input data (again, cases like substrate languages), we need to define which structures in the result graph would count as correctly reflecting reality, and which structures we would not accept as equivalent to the true story. In terms of precision, we will accept a link where either the reconstructed donor or the reconstructed recipient is the lowest ancestor of the true donor or recipient in the reduced tree, but not if both donor and recipient are wrongly detected, or contact is inferred between descendants. For each connection $L_1 - L_2$ in the inferred network G_{res} , we therefore ask whether it is compatible with some connection in G_{true} , in the sense that it reflects a lateral

connection $A_1 - A_2$ for two languages $A_1 \in \text{anc}(L_1)$ and $A_2 \in \text{anc}(L_2)$ in G_{true} . For example, if a link $isl \rightarrow eng$ is found, we would accept it on the grounds of $\text{North Germanic} \rightarrow eng$.

While the definition of true positives and false positives remain rather straightforward in this way, the definition of false negatives becomes a bit more involved. Is $\text{North Germanic} \rightarrow eng$ captured by the inferred skeleton if it features a connection from any North Germanic language to English? Or should we require that all North Germanic languages should be connected to English by a lateral connection? Recall that the last option would require separate exclusive lexical flows of detectable size from every single North Germanic languages into English, which will typically not be possible. For this reason, I choose to relax the condition and only require the weaker representation. More formally, a link $L_1 - L_2$ in G_{true} which is not present in G_{res} does not count as a false negative if there are descendants $D_1 \in \text{des}(L_1)$ and $D_2 \in \text{des}(L_2)$ in the phylogenetic tree such that $D_1 - D_2$ in G_{res} .

With the skeleton in place, we can proceed to measure the quality of directionality inference on the links. The idea is to consider all correct links in the skeleton for which a directionality can be derived from the gold standard, and then analyse for which of these links the correct orientation was inferred. The fact that we actually have three possibilities for the gold standard (\rightarrow , \rightrightarrows , and \leftrightarrow), relative to which we have three possibilities for the result (\rightarrow , \leftarrow , $-$), makes it a little less natural to define positives and negatives than for the skeleton measures. However, if we decide to count an arrow which points in the wrong direction as a false positive, and take the equivalence of $-$ and \leftrightarrow as well as the compatibility of \rightrightarrows in the gold standard with both \rightarrow and $-$ in the result into account, we arrive at a plausible solution, which is defined by Table 6.2.

Table 6.2: Table of elementary definitions for arrow evaluation

	\rightarrow in result	\leftarrow in result	$-$ in result
\rightarrow in standard	<i>true positive + true negative</i>	<i>false positive + false negative</i>	<i>false negative</i>
\rightrightarrows in standard	<i>true positive</i>	<i>false positive</i>	<i>true negative</i>
\leftrightarrow in standard	<i>false negative</i>	<i>false negative</i>	<i>true negative</i>

Based on these elementary definitions, we can again define precision and recall measures in the standard way. Informally, the *arrow recall* (ArRc) then measures how many of the arrows in the gold standard on links in the derived skeleton also occur in the inferred network with the correct directionality. To complement this measure, *arrow precision* (ArPr) quantifies how many of the arrows

in the reconstruction are justified by the gold standard. The trade-off between these two measures is of the same nature which one would typically capture in the precision-recall paradigm. If a directionality inference algorithm aggressively infers arrows even in the face of conflicting or weak evidence, this will increase arrow recall at the expense of arrow precision. A very cautious directionality inference scheme which assumes bidirectionality by default, will lead to a higher arrow precision at the cost of arrow recall. To handle this trade-off, we again mix both measures into the *arrow F-score* (ArFs) defined as $2 \cdot \frac{ArPr \cdot ArRc}{ArPr + ArRc}$, which will be our primary measure for comparing the performance of the different variants.

6.9.2 Overall quantitative results for NorthEuraLex data

Our first step for the evaluation is to compare the different methods in terms of skeleton precision and recall as well as arrow precision and recall on the entire NorthEuraLex dataset. This will allow us to choose the best method for the case studies in the next section.

Table 6.3 compares the skeleton precision and recall obtainable by the different conditional independence checks on our two maximum-likelihood reconstructions. While the single-value ML reconstruction led to the highest overall F-scores in the reconstruction experiments on simulated data, we find that the multi-value reconstruction consistently leads to better performance in all measures, especially in recall. These are the consequences of using the reconstruction with the highest precision, as this reconstruction will introduce the least noise, letting the patterns appear more clearly. The noise introduced by single-value reconstruction is so strong that it not only decreases the precision (leading to spurious lateral connections), but also the recall (letting weaker conditions disappear into the noise).

Table 6.3: Comparing skeleton performance of MLsgl and MLmlt reconstructions on the NorthEuraLex data

	MLsgl reconstruction			MLmlt reconstruction		
	PC	PS	FS	PC	PS	FS
skPrc	0.970	0.907	0.856	0.965	0.914	0.859
skRec	0.265	0.376	0.431	0.404	0.502	0.557
skFsc	0.416	0.532	0.574	0.570	0.648	0.676

The arrow performance measures are only defined for the intersection of links in the inferred skeleton and the gold standard, which will be a smaller or a larger set depending on the skeleton performance. Therefore, arrow performance can-

not be reliably compared across reconstructions and skeleton inference variants. Still, we can compare the performance of the four directionality inference methods on the best skeleton. This is done in Table 6.4. As the results show, the two standard directionality inference methods used in the causal inference literature do not work at all, due to the non-exact conditional independence tests and the resulting difficulty to detect v-structures based on separating sets. On the multi-value reconstruction, the stable PC variant does not even manage to infer a single correct arrow. The two directionality inference methods introduced in this chapter fare a lot better, but there is an interesting contrast in their behavior on the different reconstructions. The single-value reconstruction gives an advantage to UFR, whereas TSS is clearly superior on the multi-value reconstruction.

Table 6.4: Comparing arrow performance of MLsgl and MLmlt reconstructions on the NorthEuraLex data

	FS on MLsgl reconstruction			
	VPC	SPC	UFR	TSS
arPrc	0.185	0.154	0.615	0.546
arRec	0.114	0.050	0.585	0.585
arFsc	0.141	0.076	0.600	0.565

	FS on MLmlt reconstruction			
	VPC	SPC	UFR	TSS
arPrc	0.240	0.000	0.410	0.500
arRec	0.122	0.000	0.695	0.689
arFsc	0.162	(0.0)	0.516	0.579

Finally, to rank the different variants for overall performance, we can multiply the skeleton and arrow F-scores, capturing the intuition that the best approach should result in both a good skeleton and correct directionality information. The resulting numbers are given in Table 6.5, motivating our use of the MLmlt-FS-UFR variant for the case studies. Depending on the application, a different variant which is tuned towards more reliability at the expense of only finding the most prominent patterns, might be preferable. For applications where the focus is on precision (e.g. if computational means for deciding a research question are needed), the numbers on NorthEuraLex suggest that the MLsgl-FS-TSS variant might be the best option.

Table 6.5: PLFI variants ranked by combined F-score on the NorthEuraLex data

PLFI Variant	skFsc	arFsc	skFsc * arFsc
MLmlt-FS-UFR	0.6759	0.5793	0.3916
MLmlt-PS-UFR	0.6477	0.5405	0.3501
MLmlt-FS-TSS	0.6759	0.5157	0.3486
MLsgl-FS-TSS	0.5736	0.6000	0.3442
MLsgl-FS-UFR	0.5736	0.5647	0.3239
MLmlt-PS-TSS	0.6477	0.4737	0.3068
MLsgl-PS-UFR	0.5315	0.5455	0.2899
MLsgl-PS-TSS	0.5315	0.5075	0.2697
MLmlt-PC-UFR	0.5699	0.4333	0.2470
MLmlt-PC-TSS	0.5699	0.3175	0.1809
MLsgl-PC-UFR	0.4156	0.4242	0.1763
MLmlt-PS-VPC	0.6477	0.2069	0.1340
MLmlt-PC-SPC	0.5699	0.2105	0.1200
MLmlt-FS-VPC	0.6759	0.1622	0.1096
MLsgl-PC-TSS	0.4156	0.2424	0.1007
MLsgl-PS-VPC	0.5315	0.1852	0.0984
MLmlt-PC-VPC	0.5699	0.1714	0.0977
MLsgl-PC-SPC	0.4156	0.2308	0.0959
MLsgl-FS-VPC	0.5736	0.1408	0.0808
MLmlt-PS-SPC	0.6477	0.1132	0.0733
MLsgl-PS-SPC	0.5315	0.0889	0.0472
MLsgl-PC-VPC	0.4156	0.1081	0.0449
MLsgl-FS-SPC	0.5736	0.0755	0.0433
MLmlt-FS-SPC	0.6759	0.0000	0.0000

6.9.3 Qualitative discussion of NorthEuraLex scenarios

To put some flesh on the performance measures on all of NorthEuraLex, we now turn back to the case studies discussed in Chapter 4. In the four following subsections, the lexical flow network inferred by the MLmlt-FS-UFR variant of the PLFI algorithm is given along with a second visualization which is color-coded for the difference to the gold standard.

In the result graphs as in the gold standard, green arrows represent lateral connections for which directionality information could be inferred, and green lines mark lateral connections with conflicting evidence of directionality. In addition,

yellow is used for links for which no causal evidence is available (typically isolated groups of two languages), and black arrows show the Glottolog tree which was part of the input, and remained an immutable part of the skeleton during all computations. The thickness of the lines symbolizes the unexplained cognacy overlap at the end of skeleton inference, i.e. the ratio of shared lexical material for which no other paths through the remaining graph exist. Evaluating these link weights in the lexical flow networks inferred by my algorithms would require a much more fine-grained gold standard which includes not only the contacts discussed in the literature, but a quantification of each contact's strength in addition. Some of this work was already done when compiling the gold standard, where some links will be left out because they did not seem to concern the subset of the lexicon covered by the database, but as we are going to see in the following discussion, even this simple pre-selection was not very successful. An evaluation of inferred contact strengths would therefore require a much more refined gold standard building on intersecting available lists of loanwords for many language pairs with the NorthEuraLex data, which will have to be postponed to future work.

In the evaluation scheme I will use here, the evaluation graph symbolizes the fit of the inferred lexical flow network to the gold standard. It does not include link weights, but otherwise uses the same layout as the result graph, to which it adds red as well as additional dashed and dotted grey lines and arcs to highlight all types of errors in both the inferred skeleton and the inferred directionality information. The predefined phylogenetic tree backbone remains in black. Filled green arrows now stand for directional influences between related and unrelated languages that are compatible with the gold standard, whereas green lines symbolize lateral connections between related languages, which are always acceptable as possible artifacts because of imperfect phylogenetic trees or non-tree-like signals caused by dialect continua and similar phenomena leading to overlapping isoglosses. Empty green arrows represent links among related languages for which no directional contact is in the gold standard, but which are directional in the result.

Other edge types encode various types of errors. Dotted light gray is used for false negatives in skeleton inference, i.e. lateral connections that were part of the gold standard, but are not found in the result graph. These arrows are kept rather unobtrusive because the type of error they symbolize is arguably less problematic than the other categories. Spurious links (false positives in the skeleton inference) are dashed lines in dark gray, a pattern which is used for both arrows and undirected links between unrelated languages that are not justified

by the gold standard. Finally, red is the color of the most problematic types of errors. Red lines with the empty diamond have the inverse direction to the one they should have according to the gold standard, whether among a pair of related or unrelated languages. Red lines are correctly inferred links which have a directionality in the gold standard, but are either bidirectional or undirected in the result. Figure 6.4 summarizes the different edge types used in evaluation graphs for easier reference. Intuitively, evaluation graphs with large numbers of red connections indicate low performance, and a perfect network would only contain black and green (plus perhaps some dotted) connections.

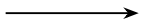










	predefined arrow (defined by underlying tree)
	correct directed link across families (skeleton TP, arrow TP, arrow TN)
	correct directed link within family (skeleton TP, arrow TN)
	correct undirected link (skeleton TP, arrow TN)
	spurious arrow within family (skeleton TP, arrow FP)
	inverted arrow on correct link (skeleton TP, arrow FP, arrow FN)
	missing arrow on correct cross-family link (skeleton TP, arrow FN)
	spurious directed link (skeleton FP)
	spurious undirected link (skeleton FP)
	missing directed link (skeleton FN)
	missing undirected link (skeleton FN)

Figure 6.4: Summary of combined color, line style and shape coding used in evaluation graphs

After a quick summary of the result for each case study, I will mostly focus on individual cases of red and spurious links, and go into the details of the computation to elucidate why this variant of PLFI failed on these links. These investigations will help to get a full picture of why PLFI is not a perfect method, and lead to some ideas for possible future improvements beyond the current state.

6.9.3.1 Case study 1: the Baltic Sea area

Already at first glance, the visualization of results for the Baltic sea case in Figure 6.5 displays only very few errors. The major contacts in the area are all inferred successfully: North Germanic influence on Western Saami (here imperfectly represented as influence from Norwegian on the individual languages), Russian influence on all the minority languages on its territory (invisible in the cases of Veps and North Karelian, due to the mentioned purism of the sources), Swedish influence on Finnish, and Latvian loans in Livonian.



Figure 6.5: Result graph (top panel) and evaluation graph (bottom panel) of phylogenetic flow on Baltic Sea data

Some other influences in the gold standard, such as the Baltic influence on Finnic and the Finnic influence on Saami, were not visible in the data, most likely because the multi-value reconstruction these results are based on did not project sufficiently many cognate sets to the level of these proto-languages. This is something we will observe in many cases, due to the cautious reconstructions in this paradigm. Only in graphs derived from single-value reconstructions do the numbers of cognate sets in the ancestral languages become so high that influences between them cannot be explained away completely by their descendants.

Other contacts in the gold standard are imperfectly represented by undirected lateral signals, such as the German influence on the continental Scandinavian languages. This is represented by an inverted link from German to Danish, whose close relationship with Norwegian is recognized, but not as directional, and a second lateral connection involving Dutch and Swedish. This part of the skeleton could hint at a problem with the language sample in this case study, since much of the German material in Danish and Swedish was actually borrowed from Low German, an unobserved language the closest relative to which in our dataset is Dutch. A different problem causes the spurious link from Icelandic into Estonian. This should actually be another link from German, the Germanic language which by far had the largest lexical influence on Estonian. Now the problem is that some of the material shared with German can be inferred as having flowed through Livonian, which contains an even larger share of German loanwords. Some other Germanic words which cannot have travelled via Livonian are present in rather archaic forms in Icelandic, causing most cognates for the remaining overlap to be detected for that language instead of other Germanic languages.

The wrong directionality of the arrow from Danish into German is simply due to the fact that the unshielded triple *dan* – *deu* – *Franconian* is detected to be a v-structure due to a very low UFR score of 0.0013. This erroneous low score is due to high overlap of 461 words between Danish and reconstructed Franconian, of which not a single item can be explained only by paths going through German. This is due to the existence of alternative routes, one through inheritance from Germanic, and the other through the Dutch-Swedish connection. We thus have a case where the logic of UFR breaks down due to the complex interplay of path configurations.

Coming to the final inverted arrow *liv* → *deu*, here it is the unshielded triple *lav* – *liv* – *deu* which does not look like a v-structure at all. This triple has a very high UFR score of 0.6364, caused by the fact that out of the 22 items shared exclusively by Latvian and German (mostly German loans in Latvian), 21 are also shared by Livonian. From the pattern *lav* → *liv* ∘ ∘ *deu* which arises after

Latvian influence on Livonian was successfully detected, the propagation rule infers the erroneous link *liv* → *deu* because otherwise a previously rejected v-structure would have to be assumed. To interpret the result, to the flow model Livonian very much looks like a transmitter of Latvian words into German, after having explained away the contact link from German into Latvian demanded by the gold standard.

To sum up, the two serious mistakes that were produced for this scenario are caused by the fact that the UFR-based v-structure test is not as reliable as it would have to be to guarantee a correct result. As always in constraint-based causal inference, even a single erroneous v-structure test can have strong effects due to propagation. Interestingly, the alternative method TSS has no problem at all to assign the correct directionality to the arrows involving German, once more showcasing the motivation for the alternative method. Since the TSS method makes other mistakes, a way towards avoiding inverted arrows (the worst type of mistake) could be to aggregate the results of both methods, only returning the arrows on which all directionality inference methods agree.

6.9.3.2 Case study 2: Uralic and contact languages

Moving on to the second case study, we see in Figure 6.6 that while there are a few more problematic arrows, the overall results are still rather convincing. Since the Western part of this case study was already covered by the previous experiment, I will not comment further on the Baltic Sea area here, except for one interesting point. In the absence of Dutch from the language set, the West Germanic material present in Swedish is now inferred as being shared with Standard German, the only West Germanic language remaining in the dataset. This highlights one property of the inference method: if relevant languages are not part of the dataset, the method will find the most plausible explanation involving only the attested languages and their reconstructed ancestors. The addition or removal of one language can have consequences beyond the immediate vicinity of the language in question, due to alternative routing of lexical flow and changed propagation patterns for directionality information.

Moving to European Russia, we see that the most dominant trend of the region, the pervasive influence of Russian on many of the minority languages of the Russian Federation is inferred correctly (dark green arrows pointing outwards from *rus*). The spurious arrows mainly concern inferred secondary influences between branches of Uralic, on which there is often no consensus among scholars, and which are therefore not represented in my gold standard. For instance, the inferred influence of Komi (*koi*) on Khanty (*kca*) is not implausible at all, and

neither is Khanty influence (*kca*) on Mansi (*mns*). On the other hand, most of the long-distance arrows are clearly spurious, such as influence of Estonian (*ekk*) on Hungarian (*hun*), and of Erzya (*myv*) on Mansi (*mns*). Let us inspect a third example, the reconstructed secondary influence of Udmurt (*udm*) on Nganasan (*nio*), more closely. The reason why the link remains during skeleton inference is legitimate: there is some material shared between Samoyedic and the Permian languages, but not the rest of Uralic. The reason why Nganasan and Udmurt were selected to model these lexical isoglosses is again due to the imperfect nature of the cognate detection. All the other Samoyedic languages have undergone much more disruptive sound change than Nganasan, causing the system to find more cognates between the more conservative Nganasan and the other Uralic languages. To a lesser extent, the same pattern applies on the Permian side, where Udmurt has undergone fewer sound changes than Komi. Finally, the erroneous arrow is caused by the fact that according to the UFR criterion, the Erzya (*myv*) lexicon does not look like a mixture of Russian and Udmurt, and neither does Udmurt form a v-structure with Erzya and Nganasan, causing the arrow from Russian into Erzya to be propagated by the principle of avoiding additional v-structures.

While a connection of Romanian (*ron*) with Bulgarian (*bul*) is inferred correctly by skeleton inference, the directionality of influence between the two languages is inferred to be the opposite of the real situation, where the Romanian lexicon is an obvious mixture of Slavic and Romance elements. The problem here is that Romanian is the only Romance language in the dataset, meaning that on the reduced skeleton, the mixed character of the Romanian lexicon would have to be detected from a v-structure *Indo-European* \rightarrow *ron* \leftarrow *bul* with Bulgarian or some other Slavic language. The UFR score for this triangle is rather close to zero at 0.0340, but not close enough for our empirically determined threshold. An additional Romance language in the dataset would yield a much cleaner v-structure *Romance* \rightarrow *ron* \leftarrow *bul*, whereas too little of the Romance material in *ron* can be reconstructed for *Indo-European*.

A second interesting area where some problems of the method become visible is the interaction between the Turkic and Uralic languages of the Volga region. The inferred network displays some shared material between Chuvash (*chv*) and the two Mari languages (*mrj* and *mhr*), but cannot decide on the directionality of either connection. According to the gold standard, there should be arrows from Chuvash into both Mari languages, but this presupposes that *mrj* \rightarrow *chv* \leftarrow *mhr* is not a v-structure. Unfortunately, a v-structure is exactly what we get by the UFR criterion, since on the rather dense skeleton, transmission via *chv* is not

needed to explain even a single item shared between *chv* and both variants of Mari, as all of these are projected up to Proto-Mari. Here, the fact that local criteria are used is really showing its negative consequences, because nothing forces the model to explain how this Turkic material ended up in Proto-Mari, which is independent of any Turkic influence conditioned on its two descendants.

A larger part of the Turkic element in Meadow Mari (*mhr*) is wrongly attributed to influence from Bashkir (*bak*), as it is correctly inferred for Udmurt. As the orange color indicates, this is a problem of skeleton inference. Here, the reasons can be traced back to the fact that cognacy data are too coarse-grained to distinguish between different closely related donor languages. While Chuvash and Bashkir are not very closely related, their divergence has primarily happened on the phonetic level, which is not visible in the inferred cognacy data. Taking a look at the actual forms, it instantly becomes clear that Chuvash has been the main source of Turkic material in Mari, but this fact is hidden by the cognate set abstraction. A future improved measure of conditional mutual information which is computed from phonetic distances (see §6.1.1) could prove superior here.

While skeleton inference performed with good overall precision, the one truly inexplicable link remaining in the inferred skeleton is the connection between the Balto-Slavic and Samoyedic proto-languages. Inspecting the cognate sets the distribution of which is explained by flow on the spurious link, we see that it mainly consists of Balto-Slavic cognate classes for concepts like *APPLE* and *CAT*. As Russian loanwords, these cognate classes are also present in a majority of Samoyedic languages, causing them to be reconstructed for the proto-language. It is unclear how this effect can be avoided in general, and how the system could successfully infer that a widespread cognate class in a minority language family might be due to separate borrowings from the majority language, without pre-determining the desired result by explicitly modeling which languages are majority and minority languages. This problem is going to become even more visible in the Siberian case study.

6.9.3.3 Case study 3: the linguistic landscape of Siberia

The Siberian data display the same large-scale pattern which we already saw in European Russia: all the minority languages have borrowed much of the vocabulary for modern life from Russian. As desired, this pattern also appears as the dominant feature of the linguistic area in the inferred lexical flow network visualized in Figure 6.7. However, two of the arrows, the ones from Russian to Sakha (*sah*) and Evenki (*evn*), do not have the desired direction. Again, the reason lies in incorrect results of v-structure tests. The triangle *bua* – *sah* – *rus* does not look

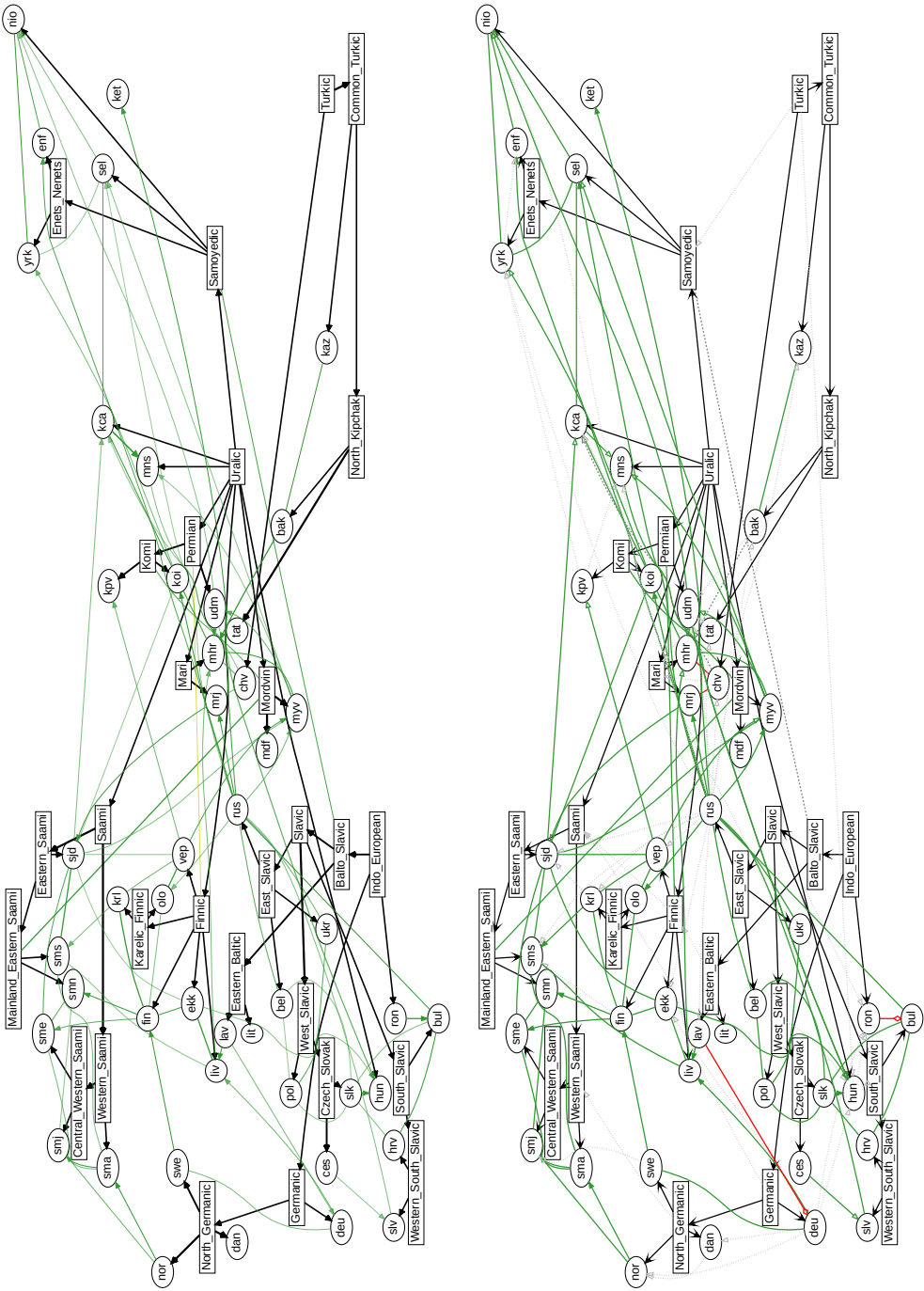


Figure 6.6: Result and evaluation of phylogenetic flow on Uralic data

at all like a v-structure, because a very large part of the material shared between Russian and Buryat (*bua*) is also shared with Sakha, to the point where in a similar pattern to the German-Livonian problem discussed before, conditioning on Sakha actually screens off Russian from Buryat. Due to the failed v-structure test, the correct arrow from Buryat into Sakha is propagated into Russian. The other failed test is for the v-structure $Tungusic \rightarrow evn \leftarrow rus$, which again misses the UFR criterion, albeit at a rather low UFR score of 0.057. Since the inheritance from Tungusic is fixed, the only way to resolve this triple in such a way that no v-structure arises, is again by inferring an arrow into Russian.

Coming to the spurious connections, the influence of Russian on the two Yukaghir languages was detected as going into Proto-Yukaghir, again because of the impossibility for the reconstruction algorithm to decide that a cognate class appearing in both daughter languages should not be projected to the proto-language. Assuming two separate arrows into the two individual languages is simply not the parsimonious solution if we do not include the knowledge that Proto-Yukaghir had already ceased to exist when the Yukaghirs were colonized. Exactly the same problem also leads to the spurious connections between Russian and Proto-Chukotko-Kamchatkan as well as between Russian and Eskimo-Aleut.

An erroneous v-structure $sah \leftrightarrow xal \leftarrow kaz$ is inferred due to zero unique flow between the Turkic languages across the Mongolic language *xal*, indicating that the true pattern $sah \leftarrow xal \leftarrow kaz$ is very unlikely. In fact, the inferred configuration is not as incorrect as the evaluation criteria imply. Arrows between *xal* and *kaz* in both directions can be justified based on the gold standard, as it includes arrows $Mongolic \rightarrow kaz$ and $Kipchak \rightarrow xal$. The problem can thus be reduced to the fact that *Kipchak* is not a node in the reduced tree for this scenario, because other Kipchak languages like Bashkir and Tatar are not part of the language sample.

The spurious connections of Selkup (*sel*) to Chinese and Itelmen are due to a slightly too high noise level in cognate detection. For instance, Selkup and Chinese have an overlap of 28 cognate classes according to the inferred cognacy relation, and the 0.025 threshold would have kicked in at 25 cognate classes. This scenario also provides a nice example for why spurious connections are very problematic, because the Selkup-Chinese connection is also responsible for the inverted arrow from Japanese into Chinese, due to an inferred v-structure $sel \rightarrow cmn \leftarrow jpn$. Still, by focusing only on all of these problems one must not forget that in many other cases, PLFI works just as intended, and that many of these errors will disappear under the TSS directionality criterion, again making the case for a combined approach to enhance stability.

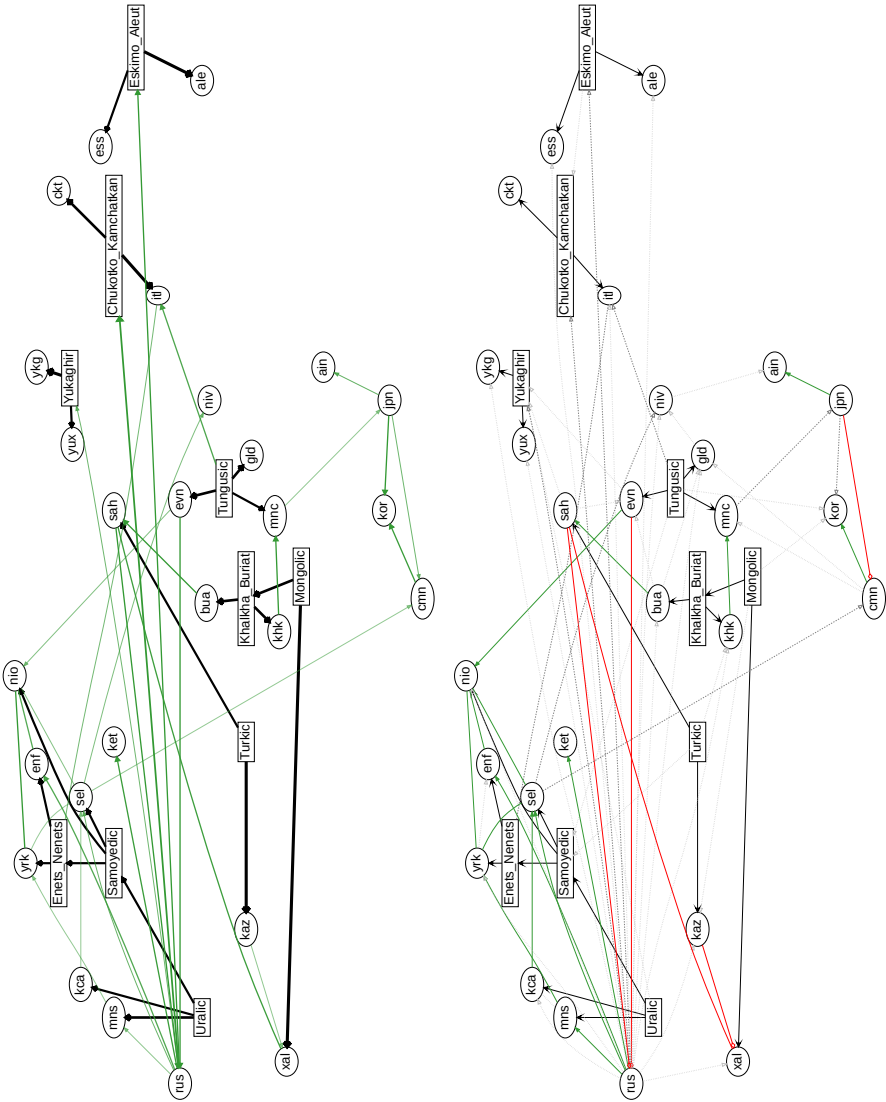


Figure 6.7: Result and evaluation of phylogenetic flow on Siberian data

6.9.3.4 Case study 4: a visit to the Caucasus

With the Caucasus scenario, we finally encounter the most complex test case. As can be seen on visual inspection of the evaluation graph in Figure 6.8, the more peripheral influences from Arabic and Russian into the Caucasus area are inferred correctly, but in the chaotic situation among Caucasian languages, there is just too much interacting and contradictory signal for PLFI to perform well.

The problems within Daghestanian might also be due to imperfections in the gold standard (the Caucasus being the only region where it was almost impossible to find literature on language contacts for some languages), but the case of Georgian (*kat*), an isolate in this dataset, points to a further issue that requires some consideration.

The underlying signal indicating language contact always comes in the shape of cognate classes present in some child language of an ancestral language which does not contain them, but a more distant language does. This is the pattern which makes the data non-tree-like, and causes lateral connections that cannot be explained away. Now, the directionality inference can recognize that the recipient language is a mixture of its own ancestor and the donor language. This entire mechanism cannot work reliably for isolated languages, because there is no proto-language in the model, and if we assumed an additional proto-language representing an earlier stage of the isolate, the reconstruction algorithm could not decide based on the single descendant language which cognate classes must have existed at that node, and would project all the inherited material up into the proto-language. The only reason why the arrow from Russian into the Siberian isolate Ket was inferred successfully in the previous case study was the other erroneous arrow from Sakha into Russian. This problem only occurs among isolates, however. An additional Slavic language in the dataset would have shared most of the Russian loans as well, and Proto-Slavic would have provided the necessary third language for a negative collider test involving the link from Ket into Russian. In a sense, the method faces the same limitations which historical linguists face when trying to infer the directionality of loans between isolates. This is notoriously difficult to do, and can only be done when loanwords are recognizable due to language-internal reasons, e.g. because they do not adhere to some phonological constraints governing the rest of the lexicon.

Finally, let us investigate why the important connections $tur \rightarrow kmr$ and $pes \rightarrow tur$ could not become part of the skeleton. The reason is that with Azeri (*azj*), there is another Turkic language that is lexically very close to Turkish, but has interacted with both Iranian languages even more, because of lexical contact on a more equal footing with Kurdish as minority languages of Iran. The overlap

of Azeri with Persian and Kurdish therefore subsumes the overlaps of both languages with Kurdish, leaving only the links with Azeri in the skeleton. The signal behind the spurious arrow from Azeri into Kurdish is actually the one which should have created the missing arrow from Turkish into Kurdish.

Finally, let us explore why the nature of Azeri and Uzbek as Turkic languages with many Persian loans is not understood by the algorithm, which instead produces the wrongly directed arrows $azj \rightarrow pes$ and $uzn \rightarrow pes$. Here, Persian is the language which looks like a mixture of elements of Azeri and Uzbek, the assumed transmitted material being exactly the Persian loans the two Turkic languages share, because these are projected into and then explained by Turkic. On a more abstract level, this phenomenon is another example of how the indistinguishability of inherited words and widespread loans can result in erroneous arrows. Unlike other erroneous v-structures, this is purely a reconstruction problem which also hits the TSS criterion, where the arrow $uzn \rightarrow pes$ has an evidence ratio of 3.575. In the global NorthEuraLex network, this problem does not occur, because the many other Turkic languages untouched by Persian produce a Turkic reconstruction that does not contain any of these. The situation could therefore be improved by considering more Turkic languages. It is an interesting question whether to a historical linguist, all the Persian elements in Turkish and Uzbek would actually be recognizably foreign if only three Turkic languages were attested.

6.9.4 Evaluation on simulated data

As the final analysis in this chapter, we now return to the simulated data. There are two important questions answers to which the simulated data will help us find. First, we want to quantify how much potential performance we lost by using ancestral state reconstruction methods and acting as if the reconstructed data were actually observed. Second, we want to know whether the performance on the simulated data is comparable to what we observed on NorthEuraLex, and whether our previous findings about the relative advantages of the different skeleton and directionality inference methods generalize beyond my dataset.

We start by comparing the skeleton performance measures for the three skeleton inference methods on the perfect data (i.e. the picture we get if we take the actual states of the simulation when the proto-languages split) to the results on the two best reconstructions. The results are given in Table 6.6. If the best skeleton inference method is picked on both types of data, we see an only moderate decrease in F-score from about 87% to 79% for the more exact single-value reconstruction, and to 74% for the more generous multi-value reconstruction. Precision and recall

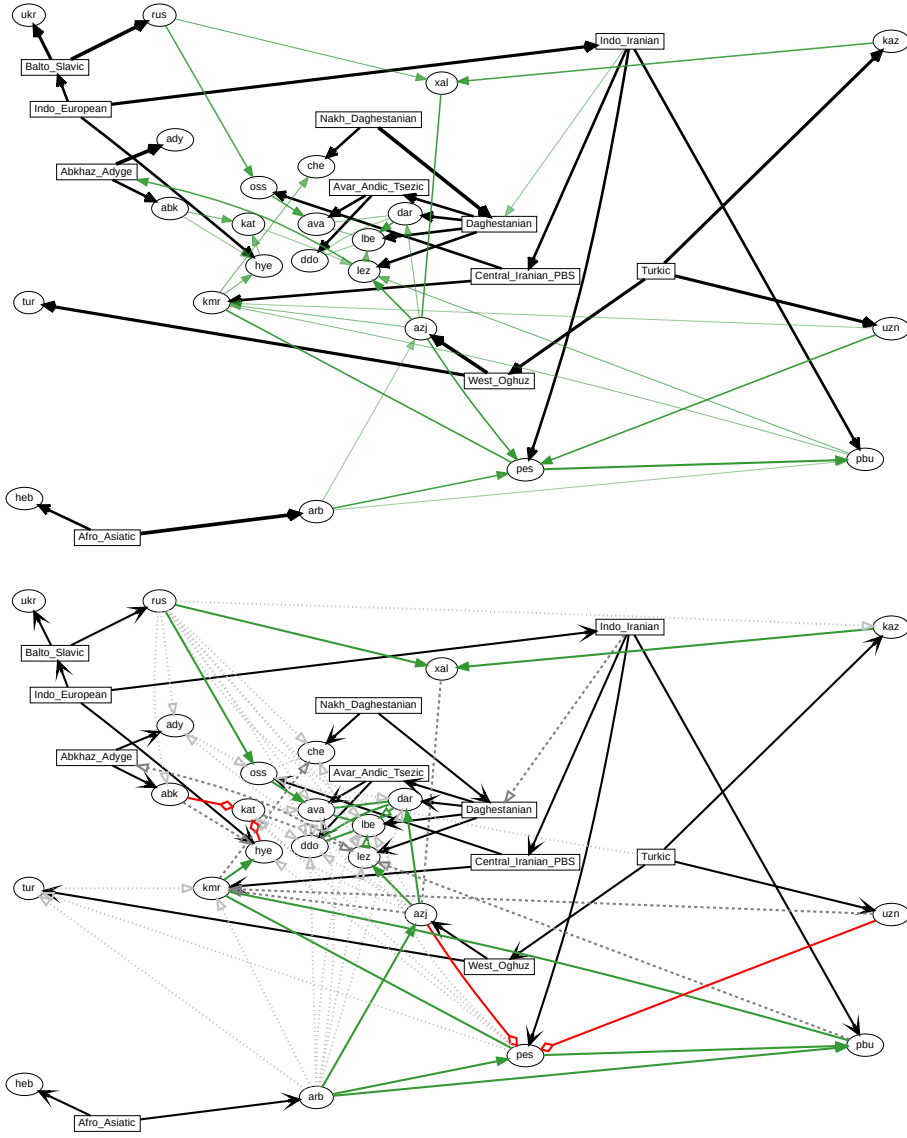


Figure 6.8: Result and evaluation of phylogenetic flow on Caucasian data

suffer about equally, showing that the reconstructed data contain a more noisy version of the same signal. We see that a good reconstruction method can help us a long way towards results comparable to what we would get on perfect data, confirming our impression that it is possible in principle to extract information about historical language contacts from a cognacy-encoded dataset covering only their living descendants. Interestingly, while flow-separation methods worked best for the NorthEuraLex data, the performance of PS is comparable on perfect data, and even superior on the reconstructed data, especially due to a much higher recall at comparable precision. Interpreting this result, erroneous reconstructions appear to have a rather strong effect on the reliability of connecting paths, indicating that while not clearly superior on perfect data, on noisy reconstructions the PS variant is surprisingly robust.

Table 6.6: Comparing skeleton performance for perfect ancestral data and the two best reconstructions

#	PrfPC	PrfPS	PrfFS
skPrc	0.901	0.870	0.829
skRec	0.780	0.915	0.915
skFsc	0.837	0.892	0.870

#	MLsPC	MLsPS	MLsFS
skPrc	0.851	0.798	0.711
skRec	0.539	0.722	0.659
skFsc	0.660	0.758	0.684

#	MLsPC	MLsPS	MLsFS
skPrc	0.855	0.797	0.710
skRec	0.527	0.720	0.658
skFsc	0.652	0.757	0.683

The consequences of reconstruction vs. observed data for arrow performance are again not easily quantifiable in our framework, because they result in different skeletons. Still, within each variant we can compare the arrow performance resulting from the different directionality inference methods. To maintain comparability with the NorthEuraLex results, only the numbers for the FS method are given in Table 6.7. The apparently low arrow performance for the perfect ancestral data is due to the higher skeleton recall, which leaves many weak links

in the skeleton where overlaps are small and directionality evidence therefore uncertain.

Instead, these numbers provide us with a further piece of the answer to the second question. Contrary to what we saw on the NorthEuraLex dataset, TSS directionality can now compete with the UFR criterion across reconstructions. The theoretical considerations leading to the TSS method seem to apply much better on this dataset. So what is the underlying reason? The only obvious difference between the two types of scenarios is that the simulated data have perfect cognate clustering, whereas the automated cognate clustering that I performed to derive cognacy overlap data from NorthEuraLex is quite noisy. Essentially, this noise causes non-zero values for δ even in unshielded colliders, making these more difficult to detect, and preventing the TSS method from reaching its full potential. In contrast, the UFR method can handle noise much better, but as we have seen in the case studies, it tends to run into problems in dense graphs where unshielded triples are rare, which is the case in the Caucasus as well as in many of the simulated scenarios. The fact that the statistical assumptions behind causal inference hold much better for perfect cognacy judgments also shows in the much better performance of the VPC and SPC directionality inference methods on the simulated data. While both were completely useless on noisy cognacy data, on clean cognacy data they do manage to capture some of the signal, although their results remain very unreliable, and the specialized directionality inference methods perform better by a significant margin.

Again, we can combine skeleton and arrow F-scores by multiplication to derive an overall performance figure for each of the compared methods, which makes it possible to quantify approximately how much overall performance we lose due to reconstruction. The resulting numbers are ranked by the best result on perfect ancestral data in Table 6.8, which compares the performance on the two reconstructions against the perfect data, also quantifying the losses or gains in percentages. Overall, the FS-UFR variant is clearly the best-performing on the perfect data, whether FS-TSS is the best on reconstructed data. Moreover, most methods perform between 20% and 30% worse on both single-value and multi-value ML reconstruction on the perfect data. Encouragingly, the specialized directionality inference method TSS does not follow the general pattern, but provides a method that at less than 10% decrease is stable against the negative consequences of reconstruction, and this at moderate performance.

Finally, we can compare the combined scores to elucidate to what extent the methods behave similarly on the simulated and the NorthEuraLex data. In Table 6.9, the different variants are ranked by their overall performance on the

Table 6.7: Comparing skeleton performance for perfect ancestral data and the two best reconstructions

FS on perfect ancestral data				
	VPC	SPC	UFR	TSS
arPrc	0.414	0.362	0.438	0.371
arRec	0.415	0.313	0.585	0.366
arFsc	0.414	0.336	0.501	0.368

FS on MLsgl reconstruction				
	VPC	SPC	UFR	TSS
arPrc	0.490	0.512	0.432	0.555
arRec	0.362	0.290	0.423	0.343
arFsc	0.417	0.370	0.428	0.424

FS on MLmlt reconstruction				
	VPC	SPC	UFR	TSS
arPrc	0.485	0.508	0.435	0.561
arRec	0.354	0.288	0.422	0.347
arFsc	0.409	0.368	0.428	0.428

Table 6.8: Analysis of consequences of reconstructed data for selected PLFI variants

PLFI Variant	perfect data	MLsgl	(diff)	MLmlt	(diff)
FS - UFR	0.436	0.293	(-32.8%)	0.293	(-32.8%)
FS - VPC	0.360	0.285	(-20.8%)	0.279	(-22.5%)
PS - VPC	0.346	0.273	(-21.1%)	0.271	(-21.7%)
FS - TSS	0.320	0.290	(- 9.4%)	0.293	(- 8.4%)
PC - VPC	0.313	0.232	(-25.9%)	0.237	(-24.3%)
FS - SPC	0.292	0.253	(-13.4%)	0.251	(-14.0%)

simulated data, and the equivalent figure on the NorthEuraLex data is given for comparison. The difference in percent, plus the rank of each method in the ranking by performance on the NorthEuraLex data, are given in addition to facilitate interpretation of results. Apart from the already mentioned advantage of TSS on simulated data, and of UFR on noisy-cognate data, the methods agree on four of the top-five methods, and the advantage of specialized skeleton and directionality inference techniques persists for the simulated data. This shows that my findings for NorthEuraLex generalize well to the simulated data, validating both the simulation model and the PLFI paradigm.

Table 6.9: Comparison of PLFI variants between datasets

PLFI Variant	simulated	NorthEuraLex	difference	rank on NELex
MLm1t-FS-UFR	0.293	0.392	+0.099	1
MLm1t-FS-TSS	0.293	0.349	+0.056	2
MLsgl-FS-TSS	0.293	0.344	+0.051	3
MLsgl-FS-UFR	0.290	0.324	+0.034	4
MLsgl-FS-VPC	0.285	0.081	-0.204	9
MLm1t-FS-VPC	0.279	0.110	-0.169	6
MLsgl-PS-VPC	0.273	0.098	-0.175	8
MLm1t-PS-VPC	0.271	0.134	-0.137	5
MLsgl-FS-SPC	0.253	0.043	-0.210	11
MLm1t-FS-SPC	0.251	0.000	-0.251	12
MLm1t-PC-VPC	0.237	0.098	-0.139	7
MLsgl-PC-VPC	0.232	0.045	-0.187	10

To summarize my findings, the PLFI paradigm of reconstructing ancestral cognates and treating them as additional observations in a causal inference paradigm has turned out to work reasonably well, although the quality of results depends a lot on the quality of the reconstruction as well as the choices of skeleton and directionality influence algorithms in the causal framework. The most important positive result of the evaluation is that the recall values for the skeleton are high, indicating that if lateral connections exist in the data, they will generally be present in the lexical flow graph. Also, the skeleton precision values indicate that about three quarters of the lateral connections in the graph will turn out to correspond to actual contacts on closer examination. This shows that PLFI lives up to its promise as a promising exploratory tool for historical linguists in the initial stages of clarifying the linguistic history of a region. Directionality inference is less reliable, but we have seen that in scenarios involving the contacts

between larger families (and not isolates), about four out of five inferred arrows will have the correct direction. On the downside, deciding whether influences occurred between extant languages or their ancestors turned out to be very difficult and unstable across different reconstructions, which guides us towards the less ambitious contact flow task that will be tackled in the next chapter.

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-Koguryōic and other languages in the Koguryō,

References

- 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 lexico-statistical 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, Lluís Barceló-Coblijn, Olga Feher & Tessa Verhoef (eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVLANG11)*. 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'gopoli'skij, Aron B. 1964. Gipoteza drevnejšego rodstva jazykov Severnoj Evrazii. Problemy fonetičeskikh sootvetstvij. In Sergej P. Tolstov (ed.), *VII meždunarodnyj kongress antropologičeskikh i ètnografičeskikh 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 praprot'urskij 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 Bowerman & 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 murteutumisen 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-dagestanskije jazykovye kontakty: (na materiale avarsko-čežskih 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.

References

- 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. Taipei: 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 Axelsson, Sam Hardwick, Tommi A. Pirinen & Miikka Silverberg. 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 Baptiste. 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.

References

- 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.),

References

- 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 Permian, 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 Evolu-

- tionary 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: Prosveshchenie.
- 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 Makarencov. 2016. Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16(1). 180.
- Willems, Matthieu, Nadia Tahiri & Vladimir Makarencov. 2014. A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology* 12(05). 1450024.

References

- 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

- Aikio, Ante, 142, 145, 161, 191
see also Ánte, Luobbal Sámmol Sámmol
- Anikin, A. E., 152
- Ánte, Luobbal Sámmol Sámmol, 145
see also Aikio, Ante
- Atkinson, Quentin D., 34, 176, 189, 285
- Baba, Kunihiro, 71
- Bailey, H. W., 167
- Beckwith, Christopher I., 159
- Bereczki, Gábor, 146
- Bergsland, Knut, 161
- Bergstrom, Carl T., 126
- Bollback, Jonathan P., 221
- Bouchard-Côté, Alexandre, 40
- Bouckaert, Remco, 34
- Bouma, Gerlof, 115
- Bowern, Claire, 13, 28, 34
- Brown, Cecil H., 115
- Bryant, David, 40
- Buch, Armin, 92, 109
- Buck, Carl D., 91
- Campbell, Lyle, 17
- Chaves, Rafael, 75
- Chickering, David Maxwell, 87
- Claassen, Tom, 87
- Collinder, Björn, 160
- Colombo, Diego, 80, 85, 86
- Comrie, Bernard, 89
- Corson, David, 143
- Cover, Thomas M., 73
- Csató, Éva Ágnes, 153
- Dahl, Östen, 140
- De Oliveira, Paulo Murilo Castro, 172
- De Vaan, Michiel Arnoud Cor, 188
- Dellert, Johannes, 3, 28, 92, 94, 95, 109, 251
- Dol'gopoli'skij, Aron B., 97
- Dunn, Michael, 29, 158
- Dybo, Anna V., 152
- Dyen, Isidore, 29
- Décsy, Gyula, 141
- Ellison, T. Mark, 40
- Embleton, Sheila M., 26, 172
- Evans, Bethwyn, 13
- Feist, Timothy Richard, 143
- Felsenstein, Joseph, 30, 33, 34, 215
- Finkenstaedt, Thomas, 11
- Fisher, Ronald A., 54
- Fortescue, Michael D., 134, 136, 161
- François, Alexandre, 12
- Friedman, Nir, 62
- Geisler, Hans, 29
- Goldberg, Yoav, 202
- Grant, Anthony, 132
- Gray, Russell D., 32, 34
- Greenhill, Simon J., 28, 30, 194

- Grünthal, Riho, 146
Gruzdeva, Ekaterina, 158
Guy, Jacques B. M., 25, 115
- Häkkinen, Jaakko, 145, 146, 157, 161
Halilov, Madžid Šaripovič, 165
Halojmov, K. N., 92
Hammarström, Harald, 91
Haspelmath, Martin, 22, 132
Hauer, Bradley, 131
Hausenberg, Anu-Reet, 148
Hawkins, John A., 191
Helinski, Eugene, 152, 191
Heskes, Tom, 87
Hewitt, George, 164, 165
Hewson, John, 25, 39
Ho, Trang, 95
Hochmuth, Mirko, 171, 173
Hock, Hans H., 19
Holden, Clare Janaki, 32
Holman, Eric W., 177, 189
Holton, Gary, 35
Hörberg, Thomas, 175
Hruschka, Daniel J., 114
Huelsenbeck, John P., 221
Huson, Daniel H., 40, 332
- Jäger, Gerhard, 92, 127, 225
Janhunen, Juha, 154, 155, 159, 188
Johanson, Lars, 153
Jordan, Fiona M., 32
Joseph, Brian D., 19
- Kalisch, Markus, 72
Kaufman, Terrence, 19, 20, 22
Kessler, Brett, 116
Key, Mary Ritchie, 89
Kobyliński, Zbigniew, 141
Koller, Daphne, 62
- Kondrak, Grzegorz, 36, 115, 131
Koptjevskaja-Tamm, Maria, 140
Kroonen, Guus, 12
- Ladefoged, Peter, 7, 103
Lehtinen, Jyri, 44
Lehtisalo, Toivo, 188
Lewis, Martin W., 35, 284
Lindén, Krister, 102
List, Johann-Mattis, 29, 36, 44, 98, 102, 115, 116, 123, 126, 133, 225
Lloyd, Stuart, 125
- Maathuis, Marloes H., 80
Maddieson, Ian, 7, 103
Martin, Samuel E., 160
Maslova, Elena, 157
Meek, Christopher, 78
Menges, Karl Heinrich, 153, 157
Menovščikov, G. A., 92
Michener, Charles D., 31, 125
Moravcsik, Edith A., 23
Morrison, David A., vii, 40–43
Murawaki, Yugo, 172, 281
Murayama, Shichirō, 160
Muysken, Pieter, 23
Myers-Scotton, Carol, 23
- Needleman, Saul B., 37
Nefedov, Andrey, 132
Nei, Masatoshi, 31
Nguyen, Lam-Tung, 216
Nikolaeva, Irina, 136
Nikolayev, Sergei L., 136
Novgorodov, Innokentij, 132
- Oakes, Michael P., 39, 115
- Pagel, Mark, 175

- Pakendorf, Brigitte, 132
 Pearl, Judea, vii, 52, 62, 63
 Pereltsvaig, Asya, 35, 284
 Piispanen, Peter S., 160
 Purvis, Andy, 189
 Puura, Ulriikka, 142
- Radzik, Ryszard, 141
 Raghavan, Usha Nandini, 126
 Rama, Taraka, 127
 Ramsey, Joseph, 80
 Reichenbach, Hans, 55
 Richardson, Thomas, 68, 70, 81
 Rießler, Michael, 132
 Roch, Sebastien, 194
 Rosvall, Martin, 126
 Rot, Sándor, 143
 Róna-Tas, András, 148
- Saitou, Naruya, 31
 Salminen, Tapani, 146
 Sammallahti, Pekka, 141, 188
 Sankoff, David, 32, 218
 Sankoff, Gillian, 179
 Schmidt, Christopher K., 132, 159, 160
 Schulte, Kim, 132, 149
 Schulze, Christian, 171
 Scornavacca, Celine, 332
 Senn, Alfred, 143
 Sergejeva, Jelena, 143
 Sicoli, Mark A., 35
 Siegl, Florian, 148
 Simon, Allan, 95
 Skorik, P. J., 158
 Smolicz, Jerzy J., 141
 Snir, Sagi, 194
 Sofroniev, Pavel, 127
 Sokal, Robert R., 31, 125
- Spirtes, Peter, 63, 68, 70, 72, 76, 79–82
 Starostin, Sergei A., 136
 Steiner, Lydia, 36
 Steudel, Bastian, 207
 Suhonen, Seppo, 143
 Swadesh, Morris, 176
 Syrjänen, Kaj, 29
- Taagepera, Rein, 148
 Tadmor, Uri, 91, 132, 187
 Thomas, Joy A., 73
 Thomason, Sarah Grey, 19, 20, 22
 Thordarson, Fridrik, 166
 Turchin, Peter, 125
- Vahre, Lauri, 157, 158
 Vajda, Edward J., 132, 156, 161
 Van der Sijs, Nicoline, 132
 Van Hout, Roeland, 23
 Vejdemo, Susanne, 175
 Viires, Ants, 157, 158
 Viitso, Tiit-Rein, 141
 Volodin, A. P., 92, 158
- Wells, John C., 99
 Wichmann, Søren, 24, 27, 89
 Wiebusch, Thekla, 132
 Willems, Matthieu, 43, 44
 Wohlgemuth, Jan, 24
 Wolff, Dieter, 11
 Wunsch, Christian D., 37
- Yamauchi, Kenji, 281
 Yang, Ziheng, 220
 Yeung, Raymond W., 75, 76
 Youn, Hyejin, 285
- Zachrisson, Inger, 142
 Zajceva, N. G., 142
 Zhang, Jiji, 81, 82, 85

Language index

- Abaza, 164
Abkhaz, 103, 136, 137, 164
Abkhaz-Abaza languages, 164
Adyghe, 136, 137, 164
Afro-Asiatic languages, 9
Ainu, 94, 100, 134, 158
Akkadian, 9
Albanian, 29, 61, 149, 256
Aleut, 134, 160, 161
Algonquian languages, 25
Altaic languages, 152, 159
Ancient Brittonic, 13
Ancient Egyptian, 9
Arabic, 23, 91, 103, 108, 111, 137, 154, 166–168, 242, 271
Arghu Turkic languages, 153
Armenian, 11, 22, 29, 58, 136, 164, 167
Australian languages, 28
Austronesian languages, 30, 32, 40, 160
Avar, 165–167
Avar-Andic languages, 165
Azeri, 153, 165, 167, 244

Baltic languages, 136, 138, 141, 143, 236, 254
Bantu languages, 32, 164
Bashkir, 148, 153, 239, 241
Basque, 61, 91, 122
Belarusian, 141, 213, 264
Bokmål, 140

Breton, 94
Bulgar, 180
Bulgarian, 46, 238
Burushaski, 91, 100
Buryat, 154, 156, 241, 267

Celtic languages, 13
Chechen, 134, 164, 166, 167
Chinese, 94, 101, 155, 159, 160, 179, 241, 269
Chukchi, 136, 152, 158, 269
Chukotko-Kamchatkan languages, 136, 157, 158, 161, 241, 269
Chuvash, 148, 153, 238
Circassian languages, 164, 166, 167
Classical Armenian, 96
Common Turkic, 153
Croatian, 46, 148

Daghestanian languages, 164, 165, 242
Danish, 53, 56, 57, 99, 140, 160, 199, 236, 264
Dargin languages, 165
Dargwa, 165, 167
Daur, 154
Dené-Yeniseian languages, 35, 161
Dongxiang, 154
Dravidian languages, 91, 136
Dutch, 11, 15, 49, 59, 132, 160, 236, 237, 263, 264

- Eastern Iranian languages, 166
Eastern Saami languages, 143
Enets, 137, 148
English, 8, 9, 11–15, 17, 19–22, 24, 37,
38, 43, 46, 58, 69, 99–101,
106, 110, 111, 119, 132, 138,
160, 177, 179, 229, 252, 255,
264
Erzya, 149, 238
Eskimo-Aleut languages, 157, 160,
161, 241
Estonian, 94, 141, 143, 236, 238, 260
Evenki, 152, 156, 157, 241, 257
Faroese, 140
Finnic languages, 138, 141, 143, 236
Finnish, 17, 46, 77, 94, 121, 141–143,
199, 236, 253, 260
Finno-Permic languages, 145, 188
Finno-Saamic languages, 145
Finno-Ugric languages, 145, 188
Frankish, 21, 49
French, 21, 43, 46, 49, 94, 99, 100, 103,
138, 149, 179, 252
Galician, 96
Georgian, 29, 164, 165, 242
German, 9, 14–17, 19–21, 38, 43, 46,
53, 56–58, 93, 100, 101, 106,
108–111, 119, 141, 143, 148,
149, 199, 236, 237, 260, 263,
264
Germanic languages, 9, 11, 43, 58, 77,
90, 100, 149, 191
Gothic, 58, 61, 96
Greek, 29, 137, 149, 167, 200
Greenlandic, 136, 160
Hebrew, 9, 91, 111, 168
Hill Mari, 149
Hindi, 43, 136, 166, 179
Hittite, 96
Hungarian, 21, 91, 94, 96, 121, 134, 146,
147, 149, 180, 238
Hunnish, 153
Icelandic, 11, 53, 56, 59, 140, 200, 236
Inari Saami, 94, 142, 143
Indo-Aryan languages, 58, 136
Indo-European languages, 9, 29, 34,
44, 58, 90, 91, 96, 134, 164
Indo-Iranian languages, 136
Ingush, 164, 166
Inuit languages, 160
Inuktitut, 160
Inupiaq, 160
Iranian languages, 22, 58, 147, 164–
167, 244, 260
Irish, 11, 13, 59, 99, 136
Italian, 96, 103
Itelmen, 92, 152, 158, 161, 241, 257,
267, 269
Japanese, 53, 56, 62, 64, 65, 67, 69,
94, 96, 101, 132, 137, 159, 160,
241, 269
Kabardian, 164
Kalmyk, 152, 155, 164, 241, 267
Karachay-Balkar, 167
Karelian, 141, 142, 255
Karluk Turkic languages, 153
Kartvelian languages, 165, 167
Kazakh, 153, 154, 167, 241, 264, 267
Ket, 134, 152, 156
Khaladj, 153
Khalkha Mongolian, 154
Khanty, 146, 148, 238

- Khinalugh, 165
Kildin Saami, 132, 142, 149, 213
Kipchak Turkic languages, 153, 155, 167, 241
Kolyma Yukaghir, 157
Komi, 149, 179, 238
Komi-Permyak, 146
Komi-Zyrian, 134, 146, 148
Korean, 64, 66, 101, 159, 160, 269
Koreanic languages, 160
Kumyk, 165, 167
Kurdish, 137, 167, 244
Kurmanji, 167
Kyrgyz, 153

Lak, 165
Latin, 11, 12, 16, 20, 46, 61, 148, 149, 188
Latvian, 46, 94, 100, 120, 138, 141, 143, 177, 236, 237, 254, 263
Lezgian, 165
Lezgic languages, 165
Lithuanian, 12, 120, 141, 143
Livonian, 94, 141, 143, 177, 236, 237, 263
Low German, 141, 236, 263
Lule Saami, 142

Malayalam, 136
Manchu, 94, 152, 155, 257
Mandarin Chinese, 65, 91, 132, 159, 187, 269
Mansi, 94, 146, 148, 238
Mari languages, 146, 148, 238
Meadow Mari, 149, 239
Middle Chinese, 60, 62, 64, 65, 67, 159
Middle English, 24, 252
Middle Mongol, 154
Moghul, 154

Moksha, 149
Mongguor, 154
Mongolian, 154, 179
Mongolic languages, 96, 136, 152, 154, 155, 159, 166
Mordvinic languages, 146

Na-Dené languages, 161
Nakh languages, 164, 166
Nakho-Daghestanian languages, 164
Nanai, 155, 158
Navajo, 161
Nenets, 77, 148, 179, 188
Nganasan, 94, 238
Nivkh, 122, 158, 159, 161
Nogai, 165, 167
North Germanic languages, 19, 140, 143, 204, 229, 234
North Karelian, 141, 236
Northeast Caucasian languages, 164, 165, 167
Northern Saami, 121, 142, 143, 204, 267
Northwest Caucasian languages, 136, 156, 164
Norwegian, 94, 140, 204, 236, 264, 267
Nynorsk, 140

Ob-Ugric languages, 146, 149
Oghur Turkic languages, 153
Oghuz Turkic languages, 153, 165, 167
Oirat, 155
Old Chinese, 60, 62, 64, 65, 67
Old Church Slavonic, 21
Old English, 8, 15, 96
Old French, 49
Old High German, 16, 20, 120

- Old Japanese, 60, 65, 68, 160
Old Korean, 60, 66–68
Old Norse, 43, 140
Old Prussian, 141
Old Turkic, 154
Olonets Karelian, 141, 143, 253
Ossetian, 90, 136, 166
Ottoman Turkish, 168
- Paleosiberian languages, 156
Pali, 21, 58
Pama-Nyungan languages, 34
Papuan languages, 28
Pashto, 101, 136, 166, 271
Pecheneg, 167
Permian languages, 146–148, 238
Persian, 11, 24, 101, 154, 165–168, 244, 256, 260, 269, 271
Polish, 141, 143, 213, 254, 264
Portuguese, 160
- Romance languages, 11, 21, 61, 134, 238
Romani, 187
Romanian, 46, 132, 149, 238
Russian, 11, 20, 100, 141–143, 148, 152, 154–158, 160, 161, 164, 165, 168, 213, 236, 237, 239, 241, 242, 253, 257, 264, 267
Ryukyuan languages, 160
- Saami languages, 141, 142, 191, 236, 267
Sakha, 132, 152–157, 241
Samoyedic languages, 91, 145, 152, 157, 188, 191, 238, 239
Sanskrit, 21, 58, 96, 136, 154
Scytho-Sarmatian languages, 166
Selkup, 148, 156, 241, 257
- Semitic languages, 38, 108, 111, 168
Siberian Turkic languages, 91, 153
Siberian Yupik, 92, 134, 160, 161
Skolt Saami, 94, 142, 143, 213
Slavic languages, 46, 58, 90, 136, 141, 149, 164, 238, 264, 267
Slovak, 148
Sorbian, 96
South Caucasian languages, 165
South Slavic languages, 147
Southern Saami, 94, 142, 204
Spanish, 8, 23, 59, 93, 110
Svan, 165
Swedish, 11, 46, 57, 94, 140, 142, 204, 236, 237, 255, 260, 264
- Tatar, 117, 148, 153, 241
Telugu, 69, 134
Tocharian, 96
Tok Pisin, 24
Tsez, 136, 165, 166
Tsezic languages, 165
Tundra Yukaghir, 136, 157
Tungusic languages, 96, 152, 155, 159, 241, 269
Turkic languages, 90, 96, 114, 136, 147, 149, 152, 164–166, 180, 238, 241, 244, 264, 271
Turkish, 24, 117, 148, 149, 153, 166, 180, 244, 256
Turkmen, 153
- Udmurt, 96, 134, 137, 146, 148, 149, 238, 239, 260
Ukrainian, 96, 267
Uralic languages, 17, 29, 38, 44, 77, 90, 94, 134, 144, 149, 152, 157, 160, 188, 238, 260
Urdu, 166

Uyghur, 153

Uzbek, 24, 153, 168, 244, 269

Veps, 138, 141, 142, 236

Welsh, 13, 136

West Frisian, 96

West Germanic languages, 237

West Slavic languages, 147

Western Iranian languages, 166

Western Saami languages, 94, 143,
234, 267

Xibo, 155

Yaghnobi, 166

Yeniseian languages, 35, 91, 156, 161

Yukaghir languages, 157, 158, 160,
241

Yupik languages, 160

Subject index

- ABVD database, 30
- alignment, 37
- almost directed cycle, 66
- ancestor (in graph), 65
- ancestral graph, 66
- arrow F-score, 230
- arrow precision, 230
- arrow recall, 230
- ASJP database, 27
- ASJP encoding, 98
- Augmented FCI (AFCI) algorithm, 82

- B-Cubed measures, 131
- Bayesian methods, 33
- Bayesian network, 62
- BCCD algorithm, 88
- borrowing, 11
- branching process, 176

- causal DAG, 66
- Causal Faithfulness Condition, 68
- causal graph, 65
- Causal Markov Condition, 68
- causal skeleton, 77
- causal sufficiency, 66
- chain, 66
- cognacy class, 13
- cognate, 13
- collider, 66
- combined information content, 109
- common cause principle, 55

- comparative method, 14
- completed partially directed acyclic graph (CPDAG), 70
- conditional independence, 57
- conditional mutual information, 75
- confounder, 52
- Conservative PC algorithm, 80
- contact flow network, 46
- Contact Lexical Flow Inference (CLFI), 257
- contraction property, 58
- creole, 24

- d-separation, 67
- data-display network, 40
- decomposition property, 57
- descendant (in graph), 65
- dialect, 8
- dialect continuum, 8
- directed cycle, 65
- directed path (in graph), 65
- discriminating path, 70
- Dolgopolsky encoding, 97
- donor language, 11
- drift graph, 119

- elemental inequalities, 75
- entropy, 74
- etymology, 12
- evolutionary network, 41

- faithfulness, 68

Subject index

- FCI algorithm, 80
- flow separation (FS), 205
- fork, 66
- galled network, 43
- galled tree, 43
- GES algorithm, 87
- graphoid axioms, 60
- hidden common cause, 52
- Hungarian, 147
- hybridization network, 43
- IELex, 29
- independence, 56
- inducing path, 81
- InfoMap algorithm, 126
- information content, 107
- internal borrowing, 21, 194
- intersection property, 58
- isolate, 190
- IWD (Information-Weighted Distance), 110
- IWSA (Information-Weighted Sequence Alignment), 110
- joint entropy, 74
- joint reconstruction, 220
- label propagation algorithm, 126
- language contact, 11
- language family, 9
- Levenshtein distance, 36
- lexical item, 7
- lexical replacement, 8
- loanword, 11
- m-separation, 68
- majority-based reconstruction, 217
- marginal reconstruction, 220
- Markov condition, 62
- Markov equivalence, 70
- maximal ancestral graph (MAG), 68
- maximum likelihood, 33
- maximum parsimony, 32
- median network, 41
- minimal lateral network, 44
- monotone faithfulness, 207
- monotonicity, 76
- multi-value ML reconstruction, 221
- multi-value MP reconstruction, 218
- mutual information, 74
- neighbor-joining algorithm, 31
- neighbor-net, 42
- NorthEuraLex, 28
- outlier, 34
- parent relation, 65
- parsimony network, 42
- partial ancestral graph (PAG), 70
- partial correlation, 71
- path (in graph), 65
- PC* algorithm, 79
- Pearson correlation, 71
- phylogenetic inference, 30
- Phylogenetic Lexical Flow Inference (PLFI), 225
- phylogenetic network, 40
- phylogenetic tree, 9
- phylum separation score, 261
- pointwise mutual information, 74
- recipient language, 11
- reticulation cycle, 43
- RFCI algorithm, 85
- rooting, 34
- Samoyedic languages, 147

- Sankoff algorithm, 218
- SCI encoding, 98
- selection bias, 52
- self-information, 74
- separating set, 77
- single-value ML reconstruction, 221
- single-value MP reconstruction, 219
- skeleton F-score, 228
- skeleton precision, 228
- skeleton recall, 228
- sound change, 8
- sound law, 14
- splits graph, 41
- Stable PC algorithm, 80
- stratum, 13
- sub-modularity, 76
- substrate, 20
- substrate language, 191
- symmetry property, 57

- taxon, 9
- time depth, 9
- Triangle Score Sum (TSS), 213
- true cognacy, 13
- typological feature, 8

- unique flow, 209
- Unique Flow Ratio (UFR), 209
- universal, 8
- unrooted tree, 34
- unshielded collider, 66
- unshielded triple, 78
- UPGMA, 31
- UraLex, 29

- v-structure, 66

- wave model, 12
- weak union property, 58

- weighted imbalance score, 190
- WOLD (World Loanword Database), 132

- X-SAMPA encoding, 99

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