Information-theoretic causal inference of lexical flow

Johannes Dellert

Language Variation 4

language science press

Language Variation

Editors: John Nerbonne, Dirk Geeraerts

In this series:

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

ISSN: 2366-7818

Information-theoretic causal inference of lexical flow

Johannes Dellert

language
science
press

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

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

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

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

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

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

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

7 Contact lexical flow inference

Having established PLFI as an exploratory tool for detecting directional contact in the linguistic history of a region, we now turn towards the second task which we set out to tackle within the lexical flow framework. After a summary of where we stand after Chapter 6, and after an overview of what will be different in this chapter, in §7.2 I explain why the contact flow inference problem has the shape of a causal inference problem with hidden common causes.

In §7.3, I explain why the vanilla RFCI algorithm as introduced towards the end of Chapter 3 for causal inference problems of this shape, is difficult to apply on the basis of noisy cognacy overlap data. §7.4 describes my most successful attempt to compensate for these weaknesses, which is to define a significance test for v-structure decisions based on a very close connection with the hypergeometric distribution. The resulting variant of the RFCI algorithm is called Contact Lexical Flow Inference (CLFI), and is presented both in pseudocode and as an informal description in §7.5.

§7.6 evaluates the different CLFI variants resulting from different approaches to skeleton and arrow inference on the same real and synthetic datasets which were already used in the previous chapter. For this purpose, the evaluation metrics have to be adapted to the new problem, and phylum separation is added as a new performance criterion.

An early version of contact flow inference was previously discussed in Dellert (2016a). There, the method was tested on an older version of NorthEuraLex for a language set similar to the current Uralic case study, with promising results which the version presented in this chapter does not significantly improve upon, although it performs better on other case studies.

7.1 The contact flow inference task

To recapitulate the core ideas of lexical flow inference, we systematically compare the cognate overlaps between pairs of languages with other languages in order to find deletable links in a graph which represents paths of lexical transmission. After thinning out the graph structure in this way until no further link

can be deleted, we know which contacts we minimally need to assume in order to explain the observable overlap patterns. In this structure, we compare the overlap patterns among triples of languages in order to extract hints of directionality, with the goal of assigning a directionality to each link in the lexical flow network.

In phylogenetic lexical flow inference, the common ancestors of observed languages were modeled explicitly by reconstructed data, turning every language into a mixture of lexical material transmitted via one of the incoming arrows, with some noise added due to lexical replacement. Such a phylogenetic lexical flow network can be interpreted as a full theory of how the lexicon of each observable language was shaped by inheritance and contact. Since the method is in principle powerful enough to reconstruct contact between proto-languages, PLFI is a fully general method for evolutionary network inference.

Contact network inference can be seen as a synchronic variant of the same basic idea, with a more modest goal. We still attempt to infer directional contact, but only on the level of living languages, without trying to infer when in the history of each language the transfer in question happened. We are thus on the very common and familiar description level of talking e.g. about French loans in English instead of Norman French loans into Middle English, which would be more exact from a diachronic perspective, and the desired outcome of phylogenetic flow inference.

Given the shape of the contact flow inference problem, it is obvious that if we continue to treat languages as variables, and measure dependencies between languages in terms of cognacy overlap, we are now faced with hidden common causes, namely the proto-languages which were modeled explicitly in phylogenetic lexical flow inference, and now create overlap that is not explainable by directed lateral transmission.

While being conceptually simpler, contact flow inference is clearly a less natural problem than phylogenetic flow inference. Since its results do not contain any temporal component such as earlier proto-languages, the resulting graphs cannot be considered evolutionary networks by any definition. Moreover, in contact networks similarities due to common inheritance will appear in the shape of bidirected links, and will be difficult to distinguish from bidirectional contact, which will make the resulting graphs more difficult to interpret and evaluate.

7.2 Advantages and disadvantages of contact flow

The decisive advantage of contact flow inference in comparison to phylogenetic flow inference is that by removing the need for reconstructed proto-languages in

the cognacy overlap data, we will be getting rid of an important source of errors that we have seen appear over and over again in the discussion of the case studies in Chapter 6.

Also, the results will be more grounded in observable facts, as we do not need to build on possibly unrealistic phylogenetic assumptions, and there is no major free parameter like the choice of reconstruction method, which previously influenced result quality so much that it would make or break PLFI as an exploratory tool. In contrast, contact flow inference is a much more data-driven process, and it will not be a surprise that it yields comparatively stable results.

Finally, the absence of proto-languages leads to a smaller problem size for the causal inference methods. This causes significant reductions in running time, which can in the worst case increase exponentially with the number of languages. Depending on the algorithm variant, executing PLFI on the entire NorthEuraLex dataset (107 languages) takes about two to six hours on a single 2 GHz core, whereas the CLFI analysis developed in this chapter never takes more than 20 minutes. Since this difference is bound to become even more pronounced with larger problems, CLFI is clearly a lot more feasible for large-scale exploratory data analysis.

Coming to the disadvantages of CLFI, implementing and tracing the behavior of the algorithm is quite a bit more challenging than it was for PLFI, since we can no longer assume causal sufficiency, and enter the realm of causal inference with latent confounders. As the reader will remember from Chapter 3, this type of causal inference requires a lot more formal machinery, leaves many more detailed choices to the implementation, and comes with a much smaller trove of practical experience gained from applying it to different problem domains.

7.3 Difficulties in applying the RFCI algorithm

What happens if we simply use the existing standard algorithm for causal inference in the presence of latent confounders, and run the RFCI algorithm presented in Chapter 3 on our cognacy-based conditional independence test? It turns out that the absence of additional, reconstructed languages leads to slightly more reliable independence checks, but also that, due to the more comprehensive propagation rules, the consequences of a single wrong result in the v-structure tests can be even more severe than what we have seen in the PLFI case studies.

For instance, consider a run of the RFCI algorithm on the Baltic Sea scenario. Among many correct v-structures such as $fin \multimap olo \multimap rus$ (Olonets Karelian having a large inherited overlap with Finnish, and some Russian loans), the sep-

arating set criterion also creates an erroneous v-structure $olo \circ\!\!\rightarrow rus \leftarrow\!\!\circ lav$, where Russian looks like a mixture of Olonets Karelian (the Russian loans) and Latvian (the inherited stock of shared Balto-Slavic words). Now, the (arguably correct) absence of a different v-structure leads to a first propagation, turning $olo \leftrightarrow rus \circ\!\!- pol$ into $olo \leftrightarrow rus \rightarrow pol$. $rus \rightarrow pol$ is an acceptable arrow (there are indeed some Russian loans in Polish, in addition to the common Slavic material inherited by both languages), but this is more of a lucky coincidence. In the next reasoning step, the new arrow into Polish creates one of the preconditions for one of the RFCI-specific propagation patterns, namely rule $\mathscr{R}4$. The pattern in question is $rus \leftarrow\!\!\circ lit \circ\!\!\rightarrow pol \leftarrow rus$, for which the bidirected erroneous arc provides a discriminating path $olo \leftrightarrow rus \leftarrow\!\!\circ lit \circ\!\!\rightarrow pol$, on which it turns out to be impossible to delete any link, which leads to $rus \leftrightarrow lit \leftrightarrow pol \leftarrow rus$. Finally, the new bidirected link combines with the non-collider $pol \leftrightarrow lit \circ\!\!\rightarrow lav$ to create the wrong arrow $lit \rightarrow lav$. To summarize, it turns out that the root cause for the erroneous arrow between two Baltic languages was a failed v-structure check involving a Uralic minority language in Russia. While the details of these computations might have been difficult to follow, it should now be very clear to how much trouble a single erroneous v-structure test can lead in the RFCI algorithm, and why this means we cannot expect vanilla RFCI to work well on our noisy data. On the plus side, even when the RFCI rules $\mathscr{R}5$ to $\mathscr{R}7$ dealing with selection bias were activated, they were almost never applied in my test runs, showing that at least the absence of selection bias is detected by RFCI.

While my independence tests appear to be good enough for direct application in RFCI, for the v-structure tests I again need to rely on specialized more stable heuristics such as UFR and TSS. In phylogenetic flow inference, every collider had the shape $A \rightarrow B \leftarrow C$, representing the lexicon of B to be a mixture of material from languages A and C. The problem in contact flow inference is that colliders can now be formed by any combination of bidirectional and directional arcs. Since bidirectional links represent the existence of hidden common causes, we would expect every link between related languages to be bidirectional, whereas cross-family contacts should lead to unidirectional links. Consequently, we get colliders that represent very different underlying histories.

For the Baltic Sea scenario, the collider $swe \rightarrow fin \leftrightarrow krl$ arises from a situation where one of two closely related languages borrows material from a language belonging to a different family. In contrast, the collider $deu \rightarrow ekk \leftarrow rus$ represents two cross-family contacts where the donor languages are related. Each of these different scenarios will lead to radically different three-way overlapping patterns, making collider tests much more difficult. For instance, the first collider

will not lead to any overlap between Swedish and Karelian, whereas such overlap exists for the second collider, due to both donor languages being related.

In principle, it would be conceivable to extend TSS to cover the new problem shape, but instead of fitting the three-way overlap to one possible collider scenario $A \longrightarrow B \longleftarrow C$, we would then have to model four different scenarios, and derive predictions for each of these scenarios in order to catch the full range of overlap patterns which can result from local collider scenarios. This leads to a much more difficult problem shape for the already difficult binary classification problem to decide whether a triple of languages forms a collider or not. Instead of going down this not very promising road, I will now develop an alternative test which does not rely on triangle scores, but still performs better than the separating set criterion.

7.4 Significance testing for v-structures

Taking a step back from the RFCI algorithm and considering the problem of inferring v-structures from cognacy data, it turns out that the basic intuition behind the criterion applied by the algorithms can be tested much more strictly on discrete cognacy overlap data. Recall again that the essential idea behind inferring a v-structure $A \longrightarrow B \longleftarrow C$ in the PC and RFCI algorithms was to decide whether B was necessary to separate A and C. What does this mean in terms of overlaps between cognate sets?

The observation I used in deriving UFR was that for B not to be necessary for separation, all cognate sets with reflexes in A and C must also have had reflexes in neighbors forming possible flow paths between A and C not going through B. For instance, to show that English was not necessary for separating Icelandic from French, and therefore establish English as a collider in this triangle, we need to show that there is an alternative path by which all the overlap between Icelandic and French can be explained. The problem in contact flow inference is that these alternative flow paths are not necessarily visible any longer, because they could actually involve proto-languages, as is the case between Icelandic and French, which share some lexical material due to their common Indo-European ancestry. Unless we have other Indo-European languages which form possible flow paths, A and C might therefore share some lexical material which cannot be explained by any path through the network except through B, but the three languages still form a collider $A \longrightarrow B \longleftarrow C$.

These considerations give rise to a possibly more robust way of testing unshielded triples for v-structures. The question is how to test that $c(A, B, C)$, the

count of cognates shared between all three varieties, is significantly smaller than the number we would expect under any of the other causal scenarios. To predict this number, we assume (as before) that when a language borrows lexical material from another, it will sample the lexical material to borrow from the donor language independently from a different language borrowing from the same donor. While this assumption might not be warranted in every individual case (e.g., the name for a newly introduced trade good will often be introduced to many neighboring languages simultaneously), we can still assume this independence of contacts, because there is no obvious mechanism which would coordinate the shape of linguistic influence between two different pairs of languages across their lexicons. In order to violate the independence assumption, such a mechanism would have to make it more likely for words in B which are already borrowings from A to be borrowed further by C from B. On rare occasions, such a preference might occur if e.g. the loans from A fit much better into the phonetic system of C, but this would clearly be an exceptional case that will not be frequent enough to warrant the costs of foregoing a generally applicable heuristic.

The direct consequence of this independence assumption is that under any of the three scenarios $A \longrightarrow B \longrightarrow C$, $A \longleftarrow B \longrightarrow C$, and $A \longleftarrow B \longleftarrow C$, the overall ratio of shared cognates should be roughly equal to the product of the ratio of shared cognates on each of the two links. For instance, if Turkish borrows 30% of its vocabulary from Persian, and Albanian borrows 20% of its vocabulary from Turkish, we would expect 6% of the Albanian lexicon to be of Persian origin. Now assume that the actual amount of lexical overlap between Albanian and Persian was determined to be 5%. How can we decide that the observed ratio δ (5%) is significantly different from the $\hat{\delta}$ (6%) we derived? There is no obvious way to model the distribution of either in a way that would provide a reliable statistical test, and my previous solutions (UFR and TSS directionality inference) both relied on what could be called a local explainability assumption. This assumption that the local scenario completely explains the overlap pattern in each triangle was already a problematic assumption before, even though adding some tolerance through threshold values turned out to work well enough. In the presence of latent confounders, however, the local explainability assumption is violated in most triangles, because in very many cases there will be in overlap due to relationship between at least two out of the three languages.

Sticking closer to the discrete nature of lexical flow as we conceive of it, it turns out that under the null hypothesis that some scenario other than $A \longrightarrow B \longleftarrow C$ holds, $c(A, B, C)$ should follow a hypergeometric distribution. To see this, picture the set $cog(B)$ as an urn containing all the cognate classes with reflexes

in the language B. Picture some of these classes as colored in red, namely the ones shared with A, i.e. all the members of $cog(A, B)$. From this urn, we now randomly pick $c(B, C)$ cognate sets, and ask the question how many of these will be colored red, i.e. have reflexes in A, to predict the count $c(A, B, C)$ of cognate classes shared by all three languages. This immediately gives us a significance test for v-structures, with p-values directly given by the cumulative distribution function of $Hypergeo(c(B), c(A, B), c(B, C))$ at the true value of $c(A, B, C)$.

As an example, take a triple of Russian (*rus*) and two Siberian minority languages which are neither related nor have plausibly been in contact, such as Itelmen (*itl*) and Selkup (*sel*). The cognacy overlaps derived from NorthEuraLex are $c(rus) = 1037$, $c(itl, rus) = 68$, $c(rus, sel) = 100$, and $c(itl, rus, sel) = 27$. Will we reject the null hypothesis that these three languages form a non-collider, i.e. correctly conclude that they do not form a v-structure $itl \circ\!\!\rightarrow rus \leftarrow\!\!\circ sel$? It turns out that we can with surprisingly high confidence, as $chyper(27, 68, 969, 100) = 0.9999999999984805$, i.e. we would not expect to find an overlap pattern like this even if we sampled billions of v-structures. It should be obvious that this is a lot more reliable than building on a separating set criterion.

For an example of a true v-structure, consider another triple of languages consisting of again Russian plus Evenki (*evn*) and Manchu (*mnc*). As determined when discussing the contact languages of Uralic, the true structure here should be $rus \rightarrow evn \leftarrow mnc$. On my automatically inferred cognates, the overlaps are $c(evn) = 1224$, $c(rus, evn) = 66$, $c(evn, mnc) = 134$, and $c(rus, evn, mnc) = 2$. The p-value for the hypergeometric test is $chyper(2, 66, 1158, 134) = 0.01745$, which is below any reasonable significance threshold, allowing us to reject any local causal scenario except the desired v-structure.

In what follows, I will write $vStructTest(A \rightarrow B \leftarrow C)$ for language variables to express a v-structure check. In the FCI directionality inference variant, this will denote the usual check in the first separating set that is found. The shorthand VCI will be used to denote the variant of the algorithm where we check whether $chyper(c(A, B, C), c(A, B), c(B) - c(A, B), c(B, C)) < 0.05$, i.e. the v-structure test developed here at a significance level of 0.05. UFR will continue to be used for the unique correlate flow check as introduced in the previous chapter.

7.5 Contact Lexical Flow Inference (CLFI)

Algorithm 3 shows the adaptations needed to implement the Contact Lexical Flow Inference (CLFI) algorithm. The dependency on a tree and an ancestral state reconstruction method is gone, but the propagation rules have become more nu-

Algorithm 3 CLFI($L_1, \ldots, L_n$)

skeleton inference method $sklM \in \{PC, FS\}$
directionality inference method $dirM \in \{VPC, FCI, VCI, UFR, TSS\}$
$\mathscr{L} := \{L_1, \ldots, L_n\}$, only the input languages
$G := (\mathscr{L}, E) := (\mathscr{L}, \{\{L_i, L_j\} \mid L_i, L_j \in \mathscr{L}'\})$, the complete graph
$S : \mathscr{L} \times \mathscr{L} \longrightarrow \wp(\mathscr{L})$, the separating set storage
$s := 0$
while $s < |\mathscr{L}| - 2$ **do**
  for $\{L_i, L_j\} \in G$ by increasing strength of remaining flow **do**
    if $sklM = PC$ **then**
      for each subset $S \in \wp(N)$ for neighbors N of L_i or L_j **do**
        if $|S| = s$ and $I(L_i; L_j|S) < 0.025$ **then**
          remove $\{L_i, L_j\}$ from G, $S(L_i, L_j) := S(L_i, L_j) \cup \{S\}$
        end if
      end for
    else if $sklM = FS$ **then**
      for each combination $P_1, ..., P_k$ of paths from L_i to L_j of length ≤ 4 **do**
        if $|S| = s$ for $S := \bigcup\{P_1, \ldots, P_k\}$ **then**
          if ratio of $c(L_i, L_j)$ not explainable by flow across S is < 0.025 **then**
            remove $\{L_i, L_j\}$ from G, $S(L_i, L_j) := S(L_i, L_j) \cup \{S\}$
          end if
        end if
      end for
    end if
  end for
  $s := s + 1$
end while
if $dirM = TSS$ **then**
  for $\{L_i, L_j\} \in G$ **do**
    if $sc(L_i \longrightarrow L_j) < 0.72$ **then**
      add arrow $L_i \longrightarrow L_j$ to network
    end if
  end for
else
  for $L_i, L_j, L_k \in \mathscr{L}$ where $\{L_i, L_j\}, \{L_j, L_k\} \in E$ but $\{L_i, L_k\} \notin E$ **do**
    if $(L_i \longrightarrow L_j \longleftarrow L_k)$ is a v-structure according to $dirM$ and $S(L_i, L_k)$ **then**
      add arrow $L_i \circ\!\!\longrightarrow L_j$ to network
      add arrow $L_k \circ\!\!\longrightarrow L_j$ to network
    end if
  end for
  propagate arrows according to $\mathscr{R}_1$ to $\mathscr{R}_3$ (if $dirM = VPC$) or $\mathscr{R}_1$ to $\mathscr{R}_{10}$
end if
return network consisting of G and arrows

merous. This method can only represent the rough structure of the RFCI methods, the way in which the skeleton is revised during the propagation stage cannot be represented in a compact way. The full details of the method need to be taken from §3.2.4, and the literature quoted there.

Like PLFI, the algorithm starts out with a fully connected graph, and attempts to find separating sets of increasing size s. In each iteration for a given size of separating sets, all links which still exist are sorted by the strength of the remaining flow, so that the algorithm first tries to remove the weakest links, and proceeds to the stronger ones later. Depending on the skeleton inference method, separating set candidates of size s for the pair of languages connected by the current link are formed either from the remaining neighbors of both nodes, or only from sets of other nodes that form connection paths between the two languages. If a separating set is found, the current link is removed from the graph. Up to this point, the algorithm is thus identical to CLFI, except that no reconstructed protolanguages are added to the dataset, and no predefined links are added based on the guide tree.

The algorithms mainly differ in the second stage, if directionality inference methods other than vanilla PC or TSS are used. As explained in Chapter 3, a successful v-structure test in the FCI algorithm no longer leads to the addition of fully directed arrows $L_i \rightarrow L_j \leftarrow L_k$, but to underspecified arrows $L_i \circ\!\!\rightarrow L_j \leftarrow\!\!\circ L_k$ that can later become either bidirectional or unidirectional arrows. This happens either through additional successful v-structure tests, or through one of the propagation rules $\mathscr{R}_1$ through $\mathscr{R}_{10}$ as described in Chapter 3. The resulting structure is a contact flow network consisting of both bidirected and directed arcs.

7.6 Evaluation of CLFI

The structure of this section exactly mirrors the order in which PLFI evaluation was performed in the last chapter. I start by discussing the behavior of the evaluation metrics developed there on the contact flow inference problem, and introducing an additional performance measure which captures how well the phylogenetic units are separated in the resulting network. Then, I again decide on one CLFI variant for the case studies by means of global results on the entire NorthEuraLex dataset. The discussion of the case studies refers back to the previous discussion of PLFI performance in each case study, and mainly discusses the differences in behavior, instead of going through each of the problems that persist again. The chapter closes with a validation of the findings about the relative performance of CLFI variants against the simulated data.

7.6.1 Evaluation metrics for contact flow

The results of CLFI can largely be evaluated just like PLFI, given a gold standard graph over the living languages in the dataset. The only difficult question is how a gold standard defined in terms of proto-languages can be flattened into a gold standard on the contact flow level. This ties back to the discussion of the NorthEuraLex gold standard in Chapter 4, where the question was whether we should expect each lexical transfer between proto-languages to be represented as an arrow between one pair of descendant languages in the result.

For contact flow inference, the problem is aggravated by the fact that each such contact can only be visible as an arrow between descendant languages. One could certainly argue that ancient influence of e.g. Proto-Iranian on Proto-Uralic will justify any arrow from an Iranian into a Uralic language, for instance from Persian into Udmurt. From the viewpoint of arrow evaluation, such an arrow would be a true positive. The difficult question is whether the absence of such an arrow also constitutes a false negative. From a local perspective, it is clear that intensive contact between proto-languages should lead to an overlap, detectable as caused by contact, between any pair of descendant languages. However, the lexical flow separation criterion implements a version of Occam's razor when it comes to leaving links in the skeleton, typically leaving only one entry point (e.g. Udmurt) for the borrowed material, and then using the existing network among related languages to distribute the material to the other Uralic languages. From the user's perspective, having a graph that is not cluttered by links between each pair of Uralic and Iranian languages, but still containing the essential information that there was influence of some Iranian on some Uralic language, might be the better solution, especially if the link connects the two languages where the contact is most visible, already indicating a good entry point for closer investigation. Still, relaxing the criterion for false negatives in the skeleton to the point where, say, any influence from some Indo-European on some Uralic language would cover all of the individual contacts we were previously interested in (Swedish on Finnish vs. German on Estonian, for instance), is certainly not the way to go for a quantitative evaluation.

Due to the difficulty in finding a good definition of false negatives, I opted for the local perspective, counting many false negatives for contacts between proto-languages that are actually represented in a satisfactory manner. This means that all the numbers for skeleton recall I will be reporting do not reflect the actual quality of the networks, although they still fulfill their primary purpose of being able to compare the performance of CLFI variants.

Finally, there is one additional level on which contact flow networks can be evaluated. Since this time, we are not putting any phylogenetic information into the procedure, we can evaluate the result in terms of how well it captures the phylogenetic signal in the data. Ideally, the contact network should connect all languages that belong to the same phylum by a subnetwork of bidirected edges (reflecting the common proto-language as the hidden common cause), while at the same time, all links across phyla should not involve hidden common causes, and therefore be monodirectional. This implies a separation of phyla by directed arcs, and can be quantified by a *phylum separation score*, simply defined as the percentage of pairs of languages where the separation induced by the contact flow network (connection or non-connection by a path of bidirected edges) agrees with the separation defined by language family. The phylum separation score will be used as an additional point of evaluation on the simulated data.

7.6.2 Overall quantitative results for NorthEuraLex data

Again, I start by quantitatively evaluating the flow networks produced by different variants of the CLFI algorithm on the entire NorthEuraLex dataset against the gold standard, and we first consider skeleton and arrow performance separately. Table 7.1 compares the skeleton precision and recall. Remember that due to the way false negatives are counted, the skeleton recall suggests a lot more information loss than is actually readable from the output. While the differences between the different methods are much less pronounced than they were for PLFI, the main trend in these results is clearly in favor of flow separation. In both cases, the RFCI skeleton is identical or almost identical to the PC skeleton, indicating that discriminating paths do not form very often in this application if the checks are performed based on flow separation.

Table 7.1: Comparing CLFI variants for contact skeleton performance

	Overlap separation		Flow separation	
	VPC	FCI	VPC	FCI
skPrc	**0.969**	0.961	0.922	0.922
skRec	0.314	0.265	**0.407**	**0.407**
skFsc	0.475	0.416	**0.565**	**0.565**

As before, arrow performance can only be measured on the intersection of links in the inferred skeleton and the gold standard, which will be a smaller or a larger set depending on skeleton performance. Therefore, arrow performance

cannot be reliably compared across reconstructions and skeleton inference variants. Still, we can compare the performance of the four directionality inference methods on the RFCI skeleton. This is done in Table 7.2. We see that vanilla FCI performs very poorly on the better skeleton, clearly motivating the use of more advanced collider tests. Interestingly, the hypergeometric test with FCI propagation is outperformed by TSS-based directionality inference on both skeletons, indicating that TSS is a useful general-purpose method that might also be of help in causal inference on other types of noisy data. Finally, the weakness of UFR, its dependence on correctly inferred unshielded triples, becomes a strength on the thinned-out overlap skeleton, where it outperforms all other methods, whereas it performs worse than VCI on the more dense flow separation skeleton.

Table 7.2: Comparing CLFI variants for arrow performance

	Overlap separation				
	VPC	FCI	VCI	UFR	TSS
arPrc	0.150	**0.234**	0.171	0.231	0.233
arRec	0.400	0.379	0.444	**0.512**	0.400
arFsc	0.219	0.289	0.247	**0.318**	0.294

	Flow separation				
	VPC	FCI	VCI	UFR	TSS
arPrc	0.113	0.167	0.323	0.309	**0.396**
arRec	0.138	0.145	**0.721**	0.691	0.677
arFsc	0.124	0.155	0.446	0.427	**0.500**

Again, the different variants can be ranked by an overall performance score defined as the product of skeleton and arrow F-scores. The resulting ranking in Table 7.3, and the higher arrow precision value for the TSS method, suggest to use the FS-TSS variant for the case studies. Compared to PLFI, the skeleton precision is slightly better in CLFI, although the mentioned problem with the counting of false negatives brings the skeleton F-score into regions lower than the PLFI results. Arrow performance of even the best methods is worse than the values attained for PLFI by some margin, reflecting that the arrow inference task is more difficult without causal sufficiency. As for PLFI, the vanilla variant of the respective standard algorithm (FCI/VPC) does not work well due to the high noise level that needs to be compensated by more robust tests.

Table 7.3: CLFI variants ranked by combined F-score on the NorthEuraLex data

CLFI Variant	skFsc	arFsc	skFsc ∗ arFsc
FS-TSS	0.565	0.500	0.283
FS-VCI	0.565	0.446	0.252
FS-UFR	0.565	0.428	0.242
OS-UFR	0.475	0.318	0.151
OS-TSS	0.475	0.294	0.140
OS-FCI	0.475	0.290	0.137
OS-VPC	0.475	0.219	0.104
OS-VCI	0.416	0.247	0.103
FS-FCI	0.565	0.155	0.088
FS-VPC	0.565	0.124	0.071

7.6.3 Qualitative discussion of NorthEuraLex scenarios

Getting back to the case studies, this time I use the FS-TSS variant of the CLFI algorithm on the same data, and again visualize the difference to the gold standard in the form of evaluation graphs, with the same color coding as before. Due to the false negative issue, we can expect to see many more dotted arrows in light gray this time, indicating how many links in the skeleton were counted as missing, and explaining in a visual way how the low skeleton recall values came about.

7.6.3.1 Case study 1: the Baltic Sea area

Repeating the first experiment on the Baltic Sea data, we see in Figure 7.1 that most of the contacts which were inferred successfully by PLFI appear in the contact flow network as well. As discussed in the PLFI case study, the problem with the influence of German on Livonian disappears under the TSS criterion. However, two other problems have appeared instead.

Firstly, CLFI shares the problem that it is most parsimonious for the model to explain away the influence of *deu* on *lav* by conditioning on *liv*, such that Livonian is inferred as acting as an intermediary for the transport of German lexical material into Latvian. This complements the now correctly detected v-structure pattern $deu \rightarrow liv \leftarrow lav$, and produces an additional arrow $liv \rightarrow lav$ which combines into the erroneous bidirected arrow. Note that Dutch as a proxy for Low German plays a role here again, this time explaining the West Germanic

loans in Latvian that cannot have traveled via Livonian because they are not attested there.

Secondly, in measuring the influences between Slavic languages, Belarusian is seen as influencing Polish instead of either a bidirected arrow representing common descent, or a directed arrow from Polish into Belarusian as according to the gold standard. The very strong score ratio of 2.333 in favor of the wrong direction is mainly due (25.8% of the weight) to an almost perfect fit of the overlap between the two languages and Russian to the v-structure $bel \longrightarrow pol \longleftarrow rus$. Here again, working with data at the representation level of cognacy overlaps shows its weaknesses, as correctly determining the direction of borrowing between these languages can only be done by looking at the actual word forms, and analyzing the sound changes.

7.6.3.2 Case study 2: Uralic and contact languages

The overall results of the Uralic case study, visualized in Figure 7.2, are again quite convincing, especially in terms of phylum separation, with the exception of Kazakh (*kaz*), which becomes separated from the other Turkic languages by erroneous incoming directed arcs, and the wrong bidirected link between Latvian and Livonian, both of which were already explained in the first case study.

The only major problem with this result is a very interesting cluster of inverted arrows into German, which did not appear in the smaller Baltic scenario, although it included all the involved languages as well. For Danish and German, the triangles with Swedish and Norwegian are the only two relevant ones, and the same holds for Norwegian and Danish in reversed roles. The problem now is that all triples fit the v-structure assumption very well. For instance, for $dan - deu - nor$ we have a predicted overlap of 471 cognates according to the formula I derived, at an observed overlap of 482. The counterevidence scores for all triangles are below 0.1, i.e. they all fit the v-structure assumption very well. The problem is that the scenarios with German at the center tend to fit the v-structure assumption slightly better, so that we have a small amount of evidence against $deu \longrightarrow dan$. The TSS score definition only builds on the ratio of scores, not on the actual strength of evidence, which leads to a TSS score ratio of 1.8 in favor of $dan \longrightarrow deu$. In the Baltic sea scenario, this did not happen because Dutch and English provided further sources of high-overlap triples counterbalancing this difference. In general, having more languages in the dataset will always increase the stability of TSS, because there are more weighty triples to factor in. The fewer high-weight triangles are available for a language pair, the more unstable the TSS decision will be. We are going to see this effect very strongly in the Siberian case study.

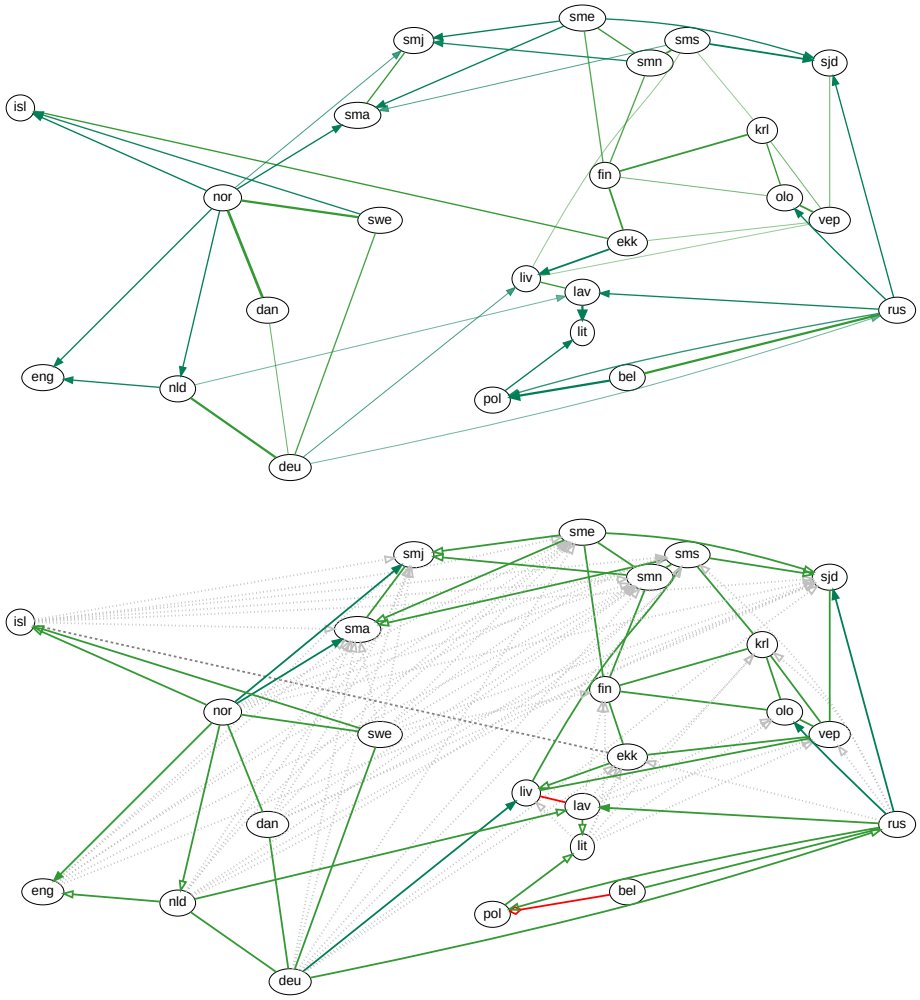

Figure 7.1: Result and evaluation of contact flow on Baltic Sea data

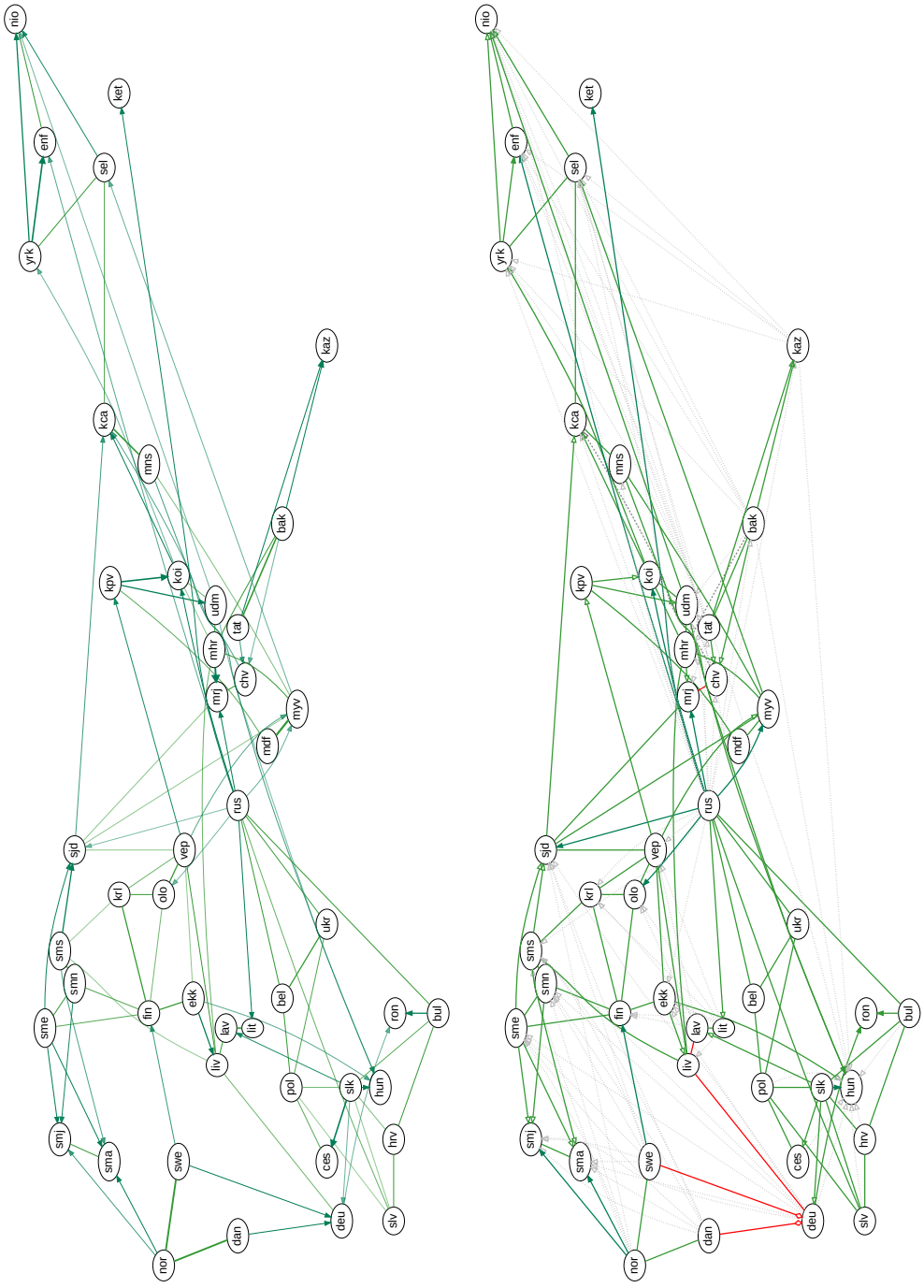

Figure 7.2: Result and evaluation of contact flow on Uralic data

Apart from this cluster of inverted arrows, the inferred contact flow network does not have any serious problems. The empty green arrows in the evaluation graph might serve to highlight a general difficulty of contact flow inference, however. To explain why so much spurious family-internal directionality is inferred, let us consider the Saami languages. Like the other Western Saami languages, Northern Saami (*sme*) has loans from Norwegian, but virtually all of these also exist in the smaller Saami languages that have been in even closer contact with Scandinavian languages. This means that it is most parsimonious for the model to explain away the connection from *sme* to *nor* by conditioning on *sma* and *smj*. Locally, this causes the two languages to look like mixtures of their more easterly relatives with Norwegian, leading to directional arrows from *sme* and *smn* into *sma* and *smj*.

7.6.3.3 Case study 3: the linguistic landscape of Siberia

In this scenario, the results of CLFI are actually worse than those of PLFI. As the results in Figure 7.3 show, the star-shaped influence of Russian on various minority languages is not recognized any more, in most cases leading to bidirectional arcs. This is again due to a lack of high-overlap triples involving the links in question. In the global NorthEuraLex network, the star pattern was inferred just as intended, because there were other Slavic languages in the dataset which could serve to form high-overlap triples involving Russian. To see even more clearly how this problem is ingrained in the mechanics of TSS computation, let us take a look at some details behind the pair $rus - sah$. The following third languages contribute the most to the triangle score sum: Kazakh (30.3%), Itelmen (12%), Buryat and Kalmyk (at 6.7% each). We only have $|cog(rus, sah, kaz)| = 7$, against an overlap of 3.74 predicted for $rus \longrightarrow sah \longleftarrow kaz$, and 13.36 for $sah \longrightarrow rus \longleftarrow kaz$. The fit of both predictions with the true overlap is thus about equal. This pattern repeats for the other triangles, so that the score ratio reaches only 1.029, a signal which is weaker than any reasonable threshold. For other minority languages, the pattern repeats itself, even if some pairs like $rus \longrightarrow ykg$ (TSS ratio 1.357) are much closer to the threshold. So why did everything work much better in the larger scenarios? The reason is that any additional Slavic language such as Ukrainian will provide a high-overlap unshielded triple $ukr - rus - sah$, because Russian will screen off the Russian minority languages from *ukr* during skeleton inference. The TSS criterion yields very high evidence against this being a v-structure, which tips the balance in favor of arrows going out of Russian for all languages Russian separates from Ukrainian. To summarize, TSS helps to aggregate and weight evidence from different triples, but in the absence of

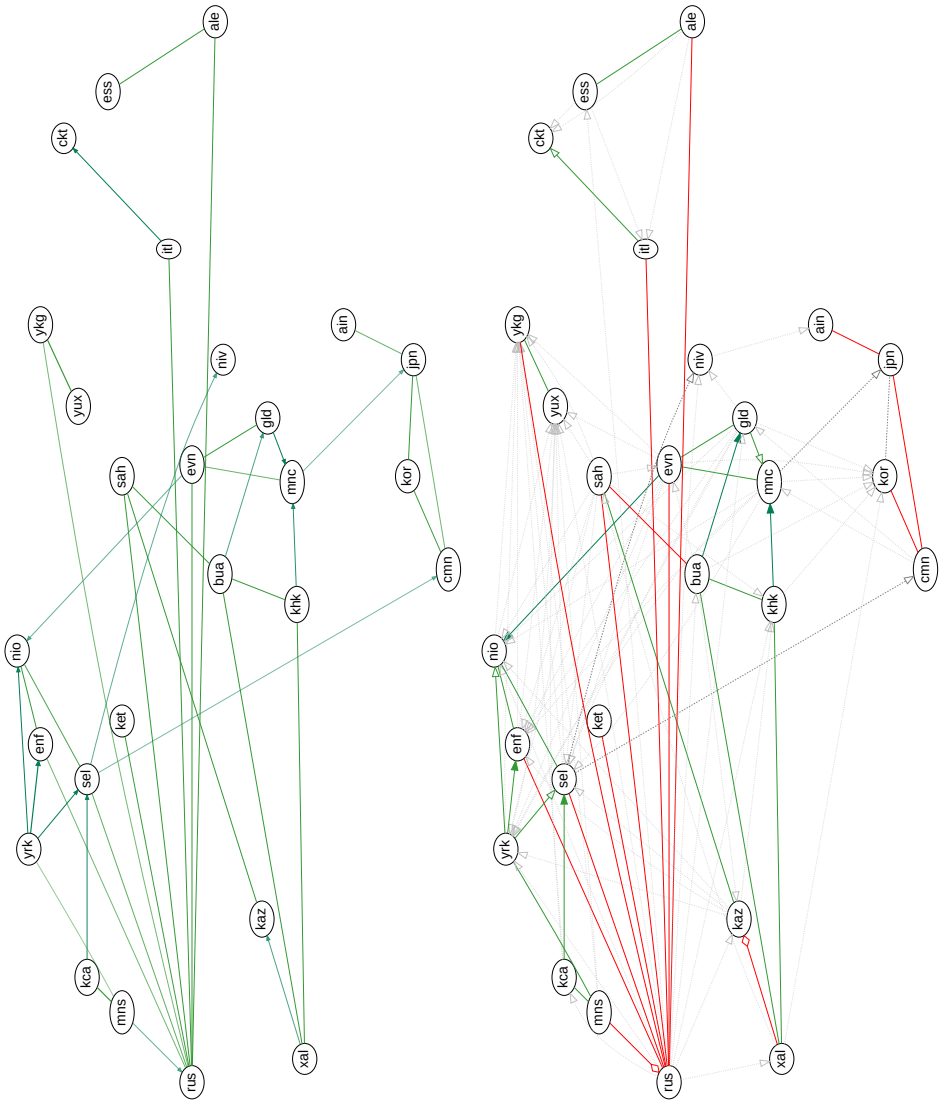

Figure 7.3: Result and evaluation of contact flow on Siberian data

unshielded triples creating strong directional signals, TSS will be unstable or inconclusive. Using a pairwise score like TSS does not provide a means to overcome the theoretical results about causal inference, which tell us that unshielded triples are needed to securely establish the direction of causality.

A further interesting phenomenon is displayed by the two-member Chukotko-Kamchatkan family. Itelmen (*itl*), the language whose lexicon was influenced much more strongly by Russian, is inferred to have been the intermediary for transmitting the Russian loans into Chukchi (*ckt*), yielding a directional signal between the two related languages. This is a problem that is especially virulent in small language families, which is why we have not yet seen it in the other case studies. The wrong internal structure of Tungusic is also created by *evn* as the obvious entry point for all the Russian loans, which then get transmitted within the family on the path $evn \longrightarrow gld \longrightarrow mnc$, although this effect is not strong enough to yield a directional signal above the threshold.

Finally, failure to recognize the directionality of contacts between Chinese, Japanese, and Korean is again due to the absence of high-overlap triples that would yield directional information. The most relevant triple for all connections between these three isolates (in our study) is the one formed by the three languages. But the three-way overlap between the three languages is only very small at $|cog(cmn, jpn, kor)| = 14$, showing that the expected problems with recognizing Chinese loans based on Mandarin Chinese have indeed materialized. In this triangle, there is not enough room for different directions to vastly differ in the fit of their prediction to the true overlap size, leading to hints of equal strength in every direction. In general, lexical flow inference will always run into problems when isolates are involved, and we can only expect it to work well if both languages connected by the link of interest have close relatives in the dataset. This is the reason why contact flow inference worked so much better on the Baltic Sea and Uralic test sets (where larger families meet) than in isolate-ridden Siberia, where even the colonial language is an isolate in our dataset.

7.6.3.4 Case study 4: a visit to the Caucasus

In the results of the Caucasian case study visualized in Figure 7.4, it turns out that the absence of a proto-language does not help us prevent the erroneous arrow from Uzbek into Persian. The reason is that all the triangles formed from two Turkic languages plus Persian look very much like v-structures according to the predicted overlaps. This is not an issue of data sparseness as in previous cases, because the three-way overlap for such triples will typically exceed 60 cognates. Instead, the problem now is the very high overlap between the Turkic languages,

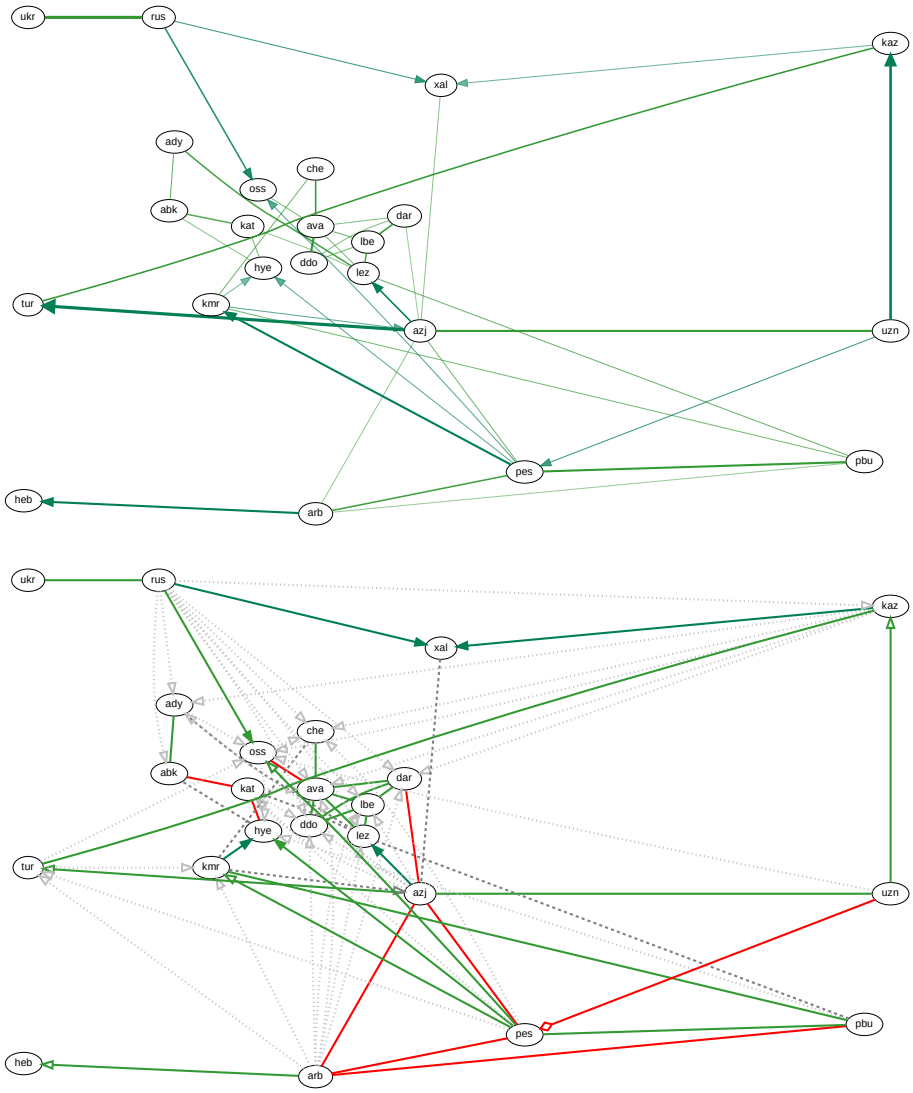

Figure 7.4: Result and evaluation of contact flow on Caucasian data

which would lead to a three-way overlap of the predicted size even if the true story were a v-structure. This is an instance of one of the cases where the TSS criterion is inadequate.

The second interesting question in this case study is why Arabic, which was correctly established to be a major external source of lexical material for the region in UFR-based PLFI, is a source of problems for TSS-based CLFI. Again looking at the TSS score ratios first, we find that the main problem is the triple of Persian, Arabic, and Pashto, with an overlap of $|cog(arb, pes, pbu)| = 70$ which does not fit the story $arb \longrightarrow pbu \longrightarrow pes$. The problem here is the assumption of independent sampling. The Arabic loans in Persian and Pashto overlap a lot more than the assumption of independent contacts would suggest, because they are concentrated in the religious and scientific vocabulary. This non-collider signal counteracts the collider signals coming from triples such as $arb - pes - kmr$. To alleviate this type of problem, one would need a much more explicit flow model which would model the flow beyond the local configuration, enforcing the constraint that there must be a directed path between any pair of languages that share a substantial marker of cognates, and take this into consideration when making local directionality decisions. Unfortunately, it is highly unlikely that inference in such a flow model would be tractable at the scale at which I am operating.

7.6.4 Evaluation on simulated data

As the final part of this chapter, I again check whether we can reproduce our findings about the relative performance of different CLFI variants on the simulated data. This time, a little more thought than before must go into the definition of the gold standard.

7.6.4.1 True and detectable histories

For NorthEuraLex, the inclusion of a contact link into the gold standard presupposed the existence of a discernable layer of loans in the attested part of the language. While this was sometimes difficult to assess based on the available literature, it still provided an external way of getting at the desired information, and the resulting evolutionary network was already relatively flat similar to a contact flow network, because ancient contacts are less clearly known, and less frequently discussed in general descriptions of individual languages.

The main issue in generating such gold standard graphs for the simulated scenarios can be conceptualized as the difference between true and detectable his-

tories. The *true history* is what actually happened during the simulation, including contacts with substrate languages, i.e. languages without living descendants about whose existence we can only know due to loanwords they left in attested languages. This makes the true history easy to define based on the simulation trace.

In contrast, the *detectable history* is only a subset of the events contained in the true history, informally defined as containing all the events of which some trace is still visible in the cognate data for living languages. By means of the detailed protocol of the complete history for each simulated scenario, the visibility of each event can be determined exactly by checking whether it is part of any word trace leading to a cognacy relation in the input data.

7.6.4.2 Summarizing generated contact histories

A true history gold standard will simply contain every contact link through which more than 25 lexical items were transmitted (based on 1,000 simulated concepts, and our CMI threshold of 0.025). However, expecting CLFI to infer this true history will not lead to a fair assessment of the system's performance. Instead, we need a gold standard that is comparable in difficulty to the equivalent task on the NorthEuraLex data. In other words, we need a way to extract a picture of the history of the linguistic region from the simulation protocols that is roughly comparable in shape and abstraction level to the NorthEuraLex gold standard.

This leads me to the following solution for generating gold standard graphs for the simulated data: Exploiting the existing infrastructure for tracing the history of every attested word (the detectable history), we consider each pair of languages in turn, and count the number of current words that were once borrowed from one language or one of its ancestors to the other language or one of its ancestors, stopping once we meet the lowest common ancestor of both languages. After discarding borrowing events which took place within the same cognate class, we arrive at a total number of transferred items in both directions, and put the appropriate arrow into the gold standard network if the number of borrowings in the respective direction exceeds 10 lexical items.

7.6.4.3 Results

Again, we start with the skeleton performance data in Table 7.4. The skeleton recall numbers are globally much better than on the NorthEuraLex data, and the differences between the different separation methods are again rather small, es-

pecially the difference between PC and FCI skeletons. Since the simulated gold standards based on detectable histories are defined in a much more objective way than the NorthEuraLex gold standard, the difference in recall suggests that a large portion of the contacts postulated by the NorthEuraLex data might not actually be detectable from the data, and that quite a few links, especially those reflecting very ancient contacts, should be removed from the gold standard. Apart from this difference, the small divergences in performance between skeleton inference methods show a very similar pattern as on the NorthEuraLex data, although this time, FCI consistently performs better than the vanilla PC algorithm. This difference once more illustrates that FCI can only play out its strengths on very clean datasets such as my simulated data, whereas it appears very prone to be affected by the type of noise seen in automatically inferred cognacy overlap data.

Table 7.4: Comparing CLFI skeleton performance on simulated data

	Overlap separation		Flow separation	
	VPC	FCI	VPC	FCI
skPrc	**0.966**	0.963	0.933	0.934
skRec	0.559	0.621	0.740	**0.750**
skFsc	0.708	0.755	0.825	**0.832**

The numbers for arrow performance in Table 7.5 show that as in PLFI, the vanilla variant of causal inference fares a lot better on simulated data than on NorthEuraLex, very likely again due to the absence of noise, as opposed to the relatively high level of noise resulting from automated cognate clustering. Also, the FCI and VCI methods show some promise on the simulated data, whereas FCI was almost useless on NorthEuraLex. This is different from our observations when evaluating PLFI on simulated data, where the TSS directionality influence was clearly the best. The reason for this might be that the theory behind TSS was not adapted to the possible presence of hidden common causes. In comparison to PLFI, the arrow F-scores of the best method are a bit worse. The best arrow F-score on the MLmlt reconstruction was reached by the TSS method at 0.447, whereas we are now at 0.407 with the FS-VCI variant.

Next, Table 7.6 ranks all the variants of CLFI according to their combined performance score on the simulated data. It is interesting that the performance on NorthEuraLex data is quite consistently much worse than on the simulated data, although as for PLFI, the combination of flow separation with the more robust VCI, TSS and UFR methods does not suffer as much from this. Unlike for CLFI, very different methods end up in the top ranks on simulated and NorthEuraLex

Table 7.5: Comparing CLFI arrow performance on simulated data

	Overlap separation				
	VPC	FCI	VCI	UFR	TSS
arPrc	0.366	**0.529**	0.366	0.405	0.434
arRec	0.486	0.452	0.486	0.409	**0.513**
arFsc	0.417	**0.488**	0.417	0.407	0.470

	Flow separation				
	VPC	FCI	VCI	UFR	TSS
arPrc	0.272	0.283	**0.405**	0.256	0.325
arRec	0.452	0.404	0.409	**0.516**	0.504
arFsc	0.339	0.332	**0.407**	0.342	0.395

data, with only the FS-TSS method staying in one of the top positions across data types. This seems to indicate that the way in which the gold standard was extracted from the simulated data might not be the perfect choice.

Though not the best-performing method on the simulated data, FS-TSS is clearly the best method according to the phylum separation measure. Separating the phyla appears not only to work well in selected example scenarios (such as the Baltic and Uralic case studies), but also across 50 sometimes rather challenging simulated scenarios. This shows that FS-TSS might be a useful tool for discerning different language families in situations which at first sight look rather chaotic.

Table 7.6: CLFI variants ranked by combined F-score on the simulated data

CLFI Variant	simulated	NELex	difference	rank on NELex	phyloSep
`OS-FCI`	0.368	0.137	-0.231	6	0.714
`OS-TSS`	0.355	0.140	-0.215	5	0.737
`FS-TSS`	0.329	0.283	-0.046	1	0.783
`OS-VPC`	0.315	0.104	-0.211	7	0.711
`OS-VCI`	0.315	0.103	-0.212	8	0.711
`FS-VCI`	0.288	0.252	-0.036	2	0.738
`OS-UFR`	0.288	0.151	-0.137	4	0.738
`FS-FCI`	0.283	0.088	-0.195	9	0.642
`FS-UFR`	0.282	0.242	-0.040	3	0.723
`FS-VPC`	0.282	0.071	-0.211	10	0.673

Summing up the results of CLFI, we have seen in the case studies that while the problems previously caused by reconstruction have disappeared, the reliability of v-structure tests appears to have dropped in comparison with PLFI. This is perhaps not too surprising, as only the phylogenetic lexical flow paradigm has clear instances of the lexicon of one language being in a very literal sense a mixture of words from other languages. In contrast, contact flow is frequently faced with situations where parts of the overlaps in the triple are actually due to common inheritance, leading to much more unpredictable overlap patterns, and hence to lower performance of v-structure tests. Still, the overall quality of CLFI results was quite comparable with PLFI, giving us another tool for data exploration that also performs quite well at phylum separation.

References

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

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

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

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

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

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

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

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

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

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

Beckwith, Christopher I. 2005. The ethnolinguistic history of the early Korean peninsula region: Japanese-Koguryŏic and other languages in the Koguryŏ,

Paekche, and Silla kingdoms. *Journal of Inner and East Asian Studies* 2(2). 34–64.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hausenberg, Anu-Reet. 1998. Komi. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 305–326. London: Routledge.
Hawkins, John A. 1990. Germanic languages. In Bernard Comrie (ed.), *The major languages of Western Europe*, 58–66. London: Routledge.
Helimski, Eugene. 1998. Selkup. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 548–579. London: Routledge.
Hewitt, George. 2004. *Introduction to the study of the languages of the Caucasus* (LINCOM handbooks in linguistics 19). Munich: Lincom Europa.
Hewson, John. 1974. Comparative reconstruction on the computer. In John M. Anderson & Charles Jones (eds.), *Proceedings of the 1st International Conference on Historical Linguistics*, 191–197. Amsterdam.
Ho, Trang & Allan Simon. 2016. *Tatoeba: Collection of sentences and translations*. http://tatoeba.org/eng/ (Last accessed 2019-06-10.)
Hochmuth, Mirko, Anke Lüdeling & Ulf Leser. 2008. Simulating and reconstructing language change. (Unpublished manuscript.) https://edoc.hu-berlin.de/handle/18452/3133 (Last accessed 2019-06-10.)
Hock, Hans H. & Brian D. Joseph. 1996. *Language history, language change, and language relationship. An introduction to historical and comparative linguistics*. Berlin: Mouton de Gruyter.
Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London B: Biological Sciences* 269(1493). 793–799.
Holman, Eric W. 2005. Nodes in phylogenetic trees: The relation between imbalance and number of descendent species. *Systematic Biology* 54(6). 895–899.
Hruschka, Daniel J., Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel & Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1). 1–9.
Huelsenbeck, John P. & Jonathan P. Bollback. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology* 50(3). 351–366.
Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267.
Huson, Daniel H. & Celine Scornavacca. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61(6). 1061–1067.
Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2). 245–291.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ramsey, Joseph, Jiji Zhang & Peter L. Spirtes. 2006. Adjacency-faithfulness and conservative causal inference. In Rina Dechter & Thomas Richardson (eds.),

Proceedings of the 22nd annual conference on Uncertainty in Artificial Intelligence (UAI 2006), 401–408. Arlington, VA: AUAI Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vajda, Edward J. & Andrey Nefedov. 2009. Ket. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for 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: Prosveščenie.
Volodin, A. P. & P. J. Skorik. 1997. Čukotskij jazyk. In A. P. Volodin, N. B. Vaxtin & A. A. Kibrik (eds.), *Jazyki mira: Paleoaziatskie jazyki*, 23–39. Moskva: Indrik.
Wells, John C. 1995. *Computer-coding the IPA: A proposed extension of SAMPA*. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm (Last accessed 2019-06-10.)
Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. *The ASJP database (version 17)*. http://asjp.clld.org/ (Accessed 2017-05-22.)
Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2018. *The ASJP database (version 18)*. http://asjp.clld.org/ (Accessed 2019-06-10.)
Wichmann, Søren & Jan Wohlgemuth. 2008. Loan verbs in a typological perspective. In Thomas Stolz, Dik Bakker & Rosa Salas Palomo (eds.), *Aspects of language contact*, 89–122. Berlin: Mouton de Gruyter.
Wiebusch, Thekla. 2009. Mandarin Chinese. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/22 (Last accessed 2019-06-09.)
Willems, Matthieu, Etienne Lord, Louise Laforest, Gilbert Labelle, François-Joseph Lapointe, Anna Maria Di Sciullo & Vladimir Makarenkov. 2016. Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16(1). 180.
Willems, Matthieu, Nadia Tahiri & Vladimir Makarenkov. 2014. A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology* 12(05). 1450024.

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

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

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

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

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

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

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

Name index

Language index

Subject index

Information-theoretic causal inference of lexical flow

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

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

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

ISBN 978-3-96110-143-6