Information-theoretic causal inference of lexical flow

Johannes Dellert

Draft of June 17, 2019, 18:12

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
© 2019, Johannes Dellert
Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):
http://creativecommons.org/licenses/by/4.0/

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

ISSN: 2366-7818
DOI:10.5281/zenodo.3247415
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

Contents

Preface v

Acknowledgments ix

1 Introduction 1

2 Foundations: historical linguistics 7
- 2.1 Language relationship and family trees . . . 7
- 2.2 Language contact and lateral connections . . . 11
- 2.3 Describing linguistic history . . . 12
- 2.4 Classical methods . . . 13
  - 2.4.1 The comparative method . . . 14
  - 2.4.2 Theories of lexical contact . . . 19
- 2.5 Automated methods . . . 25
  - 2.5.1 Lexical databases . . . 26
  - 2.5.2 Phylogenetic inference . . . 30
  - 2.5.3 Phylogeographic inference . . . 34
  - 2.5.4 Automating the comparative method . . . 36
  - 2.5.5 On the road towards network models . . . 40
- 2.6 The lexical flow inference task . . . 45
  - 2.6.1 Phylogenetic lexical flow . . . 45
  - 2.6.2 Contact flow . . . 45
- 2.7 The adequacy of models of language history . . . 46

3 Foundations: causal inference 51
- 3.1 Philosophical and theoretical foundations . . . 51
  - 3.1.1 Correlation and causation . . . 52
  - 3.1.2 Causality without experiment . . . 54
  - 3.1.3 Conditional independence . . . 56
  - 3.1.4 Bayesian networks . . . 61
  - 3.1.5 Causal interpretation of Bayesian networks . . . 63

3.2 Causal inference algorithms . . . 64
3.2.1 Causal graphs . . . 65
3.2.2 Determining conditional independence relations . . . 71
3.2.3 The PC algorithm . . . 76
3.2.4 The FCI algorithm . . . 80
3.2.5 Alternative algorithms . . . 87

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

4.6.2 Case study 1: the Baltic Sea area . . . 139
4.6.3 Case study 2: Uralic and contact languages . . . 144
4.6.4 Case study 3: the linguistic landscape of Siberia . . . 149
4.6.5 Case study 4: a visit to the Caucasus . . . 164

5 Simulating cognate histories **171**
5.1 Simulation and in-silico evaluation . . . 171
5.1.1 Advantages and shortcomings of simulation . . . 171
5.1.2 Principles of in-silico evaluation . . . 173
5.2 Generating phylogenies . . . 174
5.2.1 Models of lexical replacement . . . 175
5.2.2 Simulating how languages split and die . . . 176
5.3 Modeling lexical contact . . . 178
5.3.1 Modeling the preconditions for contact . . . 178
5.3.2 A monodirectional channel model of language contact . 179
5.3.3 Opening and closing channels . . . 179
5.3.4 Simulating channel behavior . . . 181
5.3.5 Overview of the simulation . . . 182
5.4 Analyzing the simulated scenarios . . . 182
5.4.1 Are the scenarios realistic? . . . 186
5.4.2 Are the scenarios interesting? . . . 191
5.5 Potential further uses of simulated scenarios . . . 193

6 Phylogenetic lexical flow inference **195**
6.1 Modeling languages as variables . . . 196
6.1.1 Languages as phoneme sequence generators . . . 196
6.1.2 Languages as cognate set selectors . . . 197
6.2 A cognate-based information measure . . . 198
6.3 Conditional mutual information between languages . . . 201
6.4 Improving skeleton inference . . . 202
6.4.1 Problem: stability on discrete information . . . 202
6.4.2 Flow Separation (FS) independence . . . 203
6.5 Improving directionality inference . . . 204
6.5.1 Problem: monotonic faithfulness and v-structures . . . 204
6.5.2 Unique Flow Ratio (UFR): flow-based v-structure testing 206
6.5.3 Triangle Score Sum (TSS): aggregating directionality hints 208
6.6 The phylogenetic guide tree . . . 213
6.7 Deriving proto-language models . . . 214
6.7.1 Ancestral state reconstruction algorithms . . . 214

6.7.2 Evaluation of ASR algorithms on simulated data 220
6.8 Phylogenetic Lexical Flow Inference (PLFI) 223
6.9 Evaluation of PLFI 225
6.9.1 Evaluation metrics for phylogenetic flow 226
6.9.2 Overall quantitative results for NorthEuraLex data . . . 228
6.9.3 Qualitative discussion of NorthEuraLex scenarios 230
6.9.4 Evaluation on simulated data 242

7 Contact lexical flow inference **249**
7.1 The contact flow inference task 249
7.2 Advantages and disadvantages of contact flow 250
7.3 Difficulties in applying the RFCI algorithm 251
7.4 Significance testing for v-structures 253
7.5 Contact Lexical Flow Inference (CLFI) 255
7.6 Evaluation of CLFI 257
7.6.1 Evaluation metrics for contact flow 258
7.6.2 Overall quantitative results for NorthEuraLex data . . . 259
7.6.3 Qualitative discussion of NorthEuraLex scenarios 261
7.6.4 Evaluation on simulated data 269

8 Conclusion and outlook **275**
8.1 Summary 275
8.2 Future work 277
8.3 Final remarks 281

References **287**

Index **303**
Name index 303
Language index 307
Subject index 313

8 Conclusion and outlook

In this final chapter, I review the results of all the previous chapters, putting some of them into new contexts, and assessing their relationship to the current state of the field. Then, there is a longer section about possible future work building on my results. No research project of any scale is complete without having opened some new avenues for further research, and this book is certainly no exception. For the immediate future, there are many possible improvements to explore, and many steps to take in order to make the new software tools accessible to the wider community of historical linguists who are open to experimenting with computational methods. I therefore describe my current plans about future improvements to data and software, and list my ideas for continuing research in the area of applying causal inference on the level of entire languages. These ideas revolve around possible ways of assigning confidence values to arcs in lexical flow networks, and how more fine-grained methods of estimating conditional mutual information could lead to future improvements. I then comment on the possible place of lexical flow algorithms in the landscape of tools for computational historical linguistics, and a few final remarks express my personal opinion about where the field of computational historical linguistics is headed, and which parts of the uncharted research landscape seem most in need of exploration and development.

8.1 Summary

I start by revisiting all chapters of this book, and informally summarizing the new methods and findings that can be found in each of them. After the introductory chapters revisiting the current state of the fields of computational historical linguistics and of causal inference, Chapter 4 describes the basic infrastructure that I implemented to get from paper dictionaries via raw phoneme sequences for the NorthEuraLex database to cognacy overlap data for most written languages of Northern Eurasia. While adapting and evaluating relatively standard techniques for most subtasks (sound correspondence detection, clustering cognates based on a string distance matrix), the chapter also includes some small innovations.

The most important one is probably information-weighted segment alignment (IWSA), a new alignment method which not only uses sound correspondences to infer lower distance values between cognate words from related languages, but adds an additional weighting by the information content of each segment according to a gappy trigram model. This model automatically disregards frequently occurring morphological material such as infinitive endings when computing phoneme sequence distances by alignment, doing away with the need for stemming when working with dictionary data. The last sections of the chapter justify the decisions behind my gold standard of detectable language contact events in four subareas of Northern Eurasia, representing very different linguistic situations – from intense but monodirectional contact between two large language families (the Baltic sea area) to a chaotic situation with several interacting indigenous families influenced from the outside by imperial languages from four different families (the Caucasus).

Chapter 5 then presents a new simulation model for generating realistic cognate overlap testsets for entire linguistic regions, with up to ten language families interacting on an irregularly shaped continent, spawning new languages that spread to a limited number of locations. Language extinction is modeled as only occurring if a neighboring language splits and expands into an area previously occupied by another language, which then becomes extinct. I show that 50 random scenarios generated by the model are structurally very similar to the NorthEuraLex dataset, agreeing with the real data in many measures of tree structure, cognate set geometry, and word age distribution. These findings make the datasets generated by the model a valuable resource for other research in computational phylogenetics, but especially for evaluating lexical flow inference algorithms like the ones introduced in this book.

The starting point of Chapter 6 was the question how a cognacy-encoded dataset can be used to define a consistent conditional mutual information measure on sets of languages, providing a mathematical model in which similarities of two languages in the form of lexical overlap can be explained away by the influence of other languages. In the causal inference framework, conditional independence tests can be used to answer the question which lateral connections need to be assumed in addition to a given phylogenetic tree to explain how the lexical material covering some set of concepts in a given set of languages ended up in the observable configuration. The coarse-grained nature of cognacy data posed considerable challenges to applying the causal inference paradigm, but a combination of new scoring methods based on the lexical flow metaphor was found to be sufficient for capturing a large number of horizontal connections.

With the causal skeleton in place, the idea of detecting colliders (i.e. places where the lexicon of one language looks like a mixture of two other languages), and distinguishing them from other contact patterns, provides a directional signal for most lateral connections. The resulting directed graph can be interpreted as depicting the process by which the observable languages were generated from a small set of proto-languages. In the phylogenetic flow model explored in this chapter, the common ancestors of observed languages had to be modeled explicitly by some initial phylogenetic theory, and by reconstructing the ancestral states. Experiments on the simulated data confirmed that maximum-likelihood methods for ancestral state reconstruction are superior to other approaches, but also showed the limited reconstructability of contacts between proto-languages. While contacts between observable languages were inferred with sufficient reliability, signals between reconstructed proto-languages turned out to be unstable against different reconstruction strategies.

Informed by these problems, Chapter 7 explored the alternative approach of not trying to infer historical contacts, but a network which indirectly models the presence of hidden common causes in the form of common proto-languages. Such contact flow networks constitute another new way of summarizing and visualizing cognacy data. They provide a clear display of directional signal in the contacts, while not displaying ancestral links on an equal footing, which could cause the interpreter of a phylogenetic flow network to assume that such knowledge could be inferred just as reliably as lexical flow between living languages. In terms of causal inference, the presence of hidden common causes implies we can no longer assume causal sufficiency, which creates a need for one of the variants of the more complex FCI algorithm to be applied. Again, a specialized technique for detecting collider patterns was necessary to achieve acceptable performance, but the resulting networks turned out to contain less severe errors than the results of phylogenetic lexical flow inference. In addition, the best combination of heuristics turned out to lead to networks in which languages with a common ancestor are very likely connected by a chain of bidirectional arcs, whereas languages whose only cognate overlaps are due to contact are linked by monodirectional arcs, very often pointing in the correct direction.

8.2 Future work

To address the most important issue first, lexical flow inference remains in a somewhat unsatisfying position from a philosophical point of view. While lexical flow arguably provides a much more accurate picture of the forces shaping the

lexicons of languages than mathematically simpler models such as phylogenetic trees or galled networks, applying the method to available data still requires some very crude simplifying assumptions to be mathematically tractable, and to lead to satisfactory results. The resulting approach is neither fully empirical, nor fully grounded in linguistic theory. While this general problem concerns many subareas of computational linguistics due to the very complex nature of languages as systems, in the present case this problem is exacerbated by the discrete nature of the underlying data and the rather high noise levels. These problems forced a framework which started out motivated by a well-developed and attractive mathematical theory to undergo so many modifications that it ultimately shifted towards a heuristic method that is mainly of interest for initial data exploration.

The lack of secure mathematical underpinnings is not a severe problem for an exploratory tool which can be used to quickly detect points of interest and to generate hypotheses which can later be tested using traditional methods. However, it detracts from the method's value as a potential way of arriving at reliable new knowledge. As few errors as the methods might make when disentangling the interactions between two language families in contact, the knowledge that it will make some mistakes means that we cannot expect the method to provide us with definitive answers to open questions about historical language contacts.

In principle, this is a problem haunting all mathematical methods, but in contrast fo fully probabilistic evolutionary network models, there is currently no good way to quantify the uncertainty inherent in the results of lexical flow inference. What would already be possible using the current implementation is an ensemble-based approach, as I have already hinted at in several places during the discussion of results for the case studies. Such an approach would perform a selection of different variants of PLFI or CLFI, such as FS-UFR and FS-TSS, and only include in the output structure the directional arrows that all variants agreed on. This could be exploited for a massive increase in precision at the cost of recall, and would therefore allow the linguist user to be much more confident of the results on their lexical dataset.

The next step will be to improve the mathematical underpinnings of the current approach by systematically exploring the possibilities of moving from heuristics and threshold values to sampling and statistical testing. One of the first steps would be to replace the threshold value currently used for conditional independence tests by an actual statistical test for vanishing conditional mutual information. Such a test will require deriving the distribution of overlap patterns under the null hypothesis, using more complex arguments of the type which led to my derivation of the hypergeometric distribution of overlaps for the v-structure test in Chapter 7.

A first simple approach to quantifiable uncertainty will very likely be based on resampling methods. If we run PLFI or CLFI a thousand times on bootstrap resamples of the cognacy data, we can derive an empirical joint distribution over variables representing the presence and directionality of each link, which could be marginalized to provide us with confidence values for each link. Due to the generality of graph models, this procedure would be conceptually simpler than the techniques by which consensus trees are inferred from samples of the posterior distribution in Bayesian phylogenetic tree inference. On the other hand, the resulting graphs will likely be much denser, and less easy to interpret, than consensus trees, because contradictory signals would simply lead to a proliferation of low-confidence links. Rerunning PLFI thousands of times on problems of the size of my case studies could be performed in a few days thanks to the short running time for each analysis, and due to the possibility of performing multiple runs in parallel.

The ultimate goal would of course be a comprehensive Bayesian approach to PLFI and CLFI inspired by the current state of the art in tree inference, which would also model the uncertainty inherent in all results of the pre-processing stages such as detecting cognates, and projecting cognate sets back in time to the proto-languages. First steps in this direction will likely be inspired by the model presented by Murawaki & Yamauchi (2018) for typological data. Their autologistic model jointly infers the contribution of vertical stability, horizontal diffusibility and universality to observable typological feature distributions, all of which have obvious parallels in the lexical flow inference task. Even though Murawaki and Yamauchi emphasize the differences from lexical data, which they perceive to be much less characterized by uncertainty, their focus on overcoming the problems posed by uncertainty and missing values in typological data makes their work very interesting for automatically inferred cognacy data, which are arguably just as uncertain as typological features. Despite all these promising ideas, given the small number of languages for which existing fully probabilistic evolutionary network approaches are still tractable, and the very long convergence times of Bayesian inference even for the much simpler problem of phylogenetic tree inference, it appears highly unlikely for such a comprehensive approach to scale to hundreds of languages in the near future.

In parallel to these endeavors, I am going to revisit other possibilities for conditional independence tests over basic vocabularies, with the goal of extracting an additional meaningful signal beyond pure cognacy from the form differences within each cognacy class. Such tests have the potential to be both more sensitive and more reliable than the current coarse-grained measure of conditional mu-

tual information. In this context, it will also be worthwhile to explore neural approaches, letting the computer decide which features and feature combinations are relevant for classifying sets of languages as independent. This would also make it possible for lexical flow inference to benefit from the recent advances in distributed representations. The input for such methods could consist of a mixture of cognacy data and information-weighted form distances, but adding bag-of-sounds representations or even sets of positional phonetic features would very likely be worth exploring as well.

For all of these future directions of research, large amounts of high-quality training and test data will be of immense help. As mentioned in Chapter 4, the NorthEuraLex project is currently going through its second phase, with the purpose of adding not only another 89 languages, but also an etymological annotation layer based on the most recent etymological dictionaries of the larger families. Etymological coverage will inevitably remain incomplete even if it eventually evolves to model most of the available literature, but even a partial loanword annotation will provide an avenue to evaluating not only the existence of links in lexical flow networks, but also their weights. For this, the evaluation would likely be defined on the level of individual loans, counting how many loanwords were correctly detected as such, and whether the source language was inferred correctly. In a first step, the same analysis could already be performed on the simulated data in the near future, because here the ground truth of each etymology is known by design.

In the immediate future, the most pressing task is to refine the released software in such a way that it allows other linguists to experiment with PLFI and CLFI on their own datasets. The current version is still a typical piece of academic software, with all that entails in terms of sparse documentation, suboptimal error handling, and lacking introductory materials which could provide an entry point for novice users. While these issues are in the process of being addressed, the software is already released under the GPL in a Github repository[1], allowing other researchers to explore the potential of causal inference on language data from an advanced starting point.

It would also be very interesting to simply apply the existing infrastructure to the many small cognacy-encoded datasets that are currently in development, perhaps helping scholars doing research on an underexplored language family to come up with a good first idea of possible contact signals hidden in their data. My intuition is that the ability to infer a phylogenetic network (albeit with a few errors) within hours, could be an attractive prospect for such scholars, but this intuition remains to be tested and substantiated.

[1] https://github.com/jdellert/lexflow

Of course, once they have reached some level of maturity, lexical flow inference methods should be compared against future more probabilistic evolutionary network models, especially as soon as those are performant enough to tackle problems of the size we have been dealing with in this book. It will also be interesting to investigate how much of the signal is lost when using one of the more limited network types, thereby assessing whether the high generality of lexical flow networks is actually an advantage for finding the relevant contact patterns. All of these comparisons should also be performed on large amounts of simulated data as generated by my model, which could perhaps be further improved by tuning the currently fixed replacement rate to a development set.

Finally, coming to the question of what could be improved about the various pre-processing steps in order to maximize the potential of lexical flow inference, the top item on the list would be further progress in automated cognate detection. While ancestral state reconstruction might not improve much beyond the current state as it is unclear where additional information might come from, I am quite certain that cognacy detection could be much improved with the help of hitherto underused signals in the data. Better models of conditional sound changes and other important parts of the comparative method are certainly a worthwhile direction for future efforts, and lexical flow inference methods will immediately benefit from any advances in this field. In my view, the main reason why the field has not yet moved much beyond PMI-based sound correspondences has been the small amount of freely available lexical data with full phonetic encoding, a problem which has started to be addressed only fairly recently. It is likely that NorthEuraLex will be able to serve as a valuable resource in the development of such models.

8.3 Final remarks

The main contribution of my work to the field of computational historical linguistics could be summarized as follows: it provides a previously unexplored framework for evolutionary network inference from lexical data, managing to produce very general networks for problems of unprecedented size at acceptable error rates. The high performance of the method currently comes at the expense of knowledge about the uncertainty inherent in every part of the result, making it more of an exploratory tool than a possible source of proofs about historical language contacts and relationships.

One might perhaps have expected more in light of the very attractive theory behind the causal inference paradigm. This theory tells us that under certain

assumptions, causal inference will provide us with objectively true statements about the causal relationships between statistical variables. In reality, the problem is that many of these assumptions, such as the reliability of higher-order conditional independence tests, typically do not hold, often not even approximately. Causal inference is therefore a lot more difficult to apply to a new problem than the attractive mathematical paradigm would suggest. It took a lot of effort to overcome the difficulties caused by violating the assumptions, and yet we end up with a result about the truths of which no guarantee can be given. Arguably, this is also the case for other types of reasoning, such as the traditional way arguments are made in historical linguistics. Still, human reasoners are much more flexible in the types of knowledge they can take into consideration, and they can actually come close to the ideal of considering all the available data that can be brought to bear on a specific question, such as the question whether Korean and Japanese are related by inheritance. No piece of evidence is "out of scope" in the workflow of historical linguistics, and a single new bit of knowledge can take all plausibility from an entire theory.

From the perspective of many historical linguists, the trend towards answering such questions based on simulation models and probabilistic methods contributes to a tendency to reduce the types and scope of evidence from which conclusions are drawn. This reductionism can be seen as a symptom of an ongoing perhaps unhealthy mathematization of the field. Databases make it easy to abstract over all the minute details which scholars so painstakingly collected over the past centuries, and to instead put one's time and trust into the refinement of mathematical models with the purpose of answering very general questions. The availability of large databases does not necessarily bring about advances in terms of the questions the field was previously interested in, but instead causes a shift in the focus of the field towards those questions which can be answered by statistical analysis of such data, even if the connection to the open questions of the field is tenuous at best. In their very critical assessment of existing mathematical approaches to historical linguistics, Pereltsvaig & Lewis (2015) describe recent developments within the field of geography as a cautionary example. In geography, this has led to a focus on statistical phenomena such as the distributions of city sizes across the world, and away from the development of tools which would allow us to understand in a very specific case how a certain city developed compared to a neighboring city, and for which reasons. If this grand-scheme mentality is combined with glossing over disturbing facts as one would treat measurement errors in physics, mathematical modeling turns into a potentially unhealthy trend, which might cause many sciences to confine themselves

to continually redigesting noisy databases in order to explore this type of very general and abstract questions. A similar trend can already be observed in computational historical linguistics, where the most widely read (and cited) papers make very general claims about languages evolving in bursts (Atkinson et al. 2008), or universals of the human lexicon (Youn et al. 2016). While such results are certainly interesting in their own right, mathematical methods have so far contributed surprisingly little to answering the very complicated questions of detail that are involved in proving language relationship. Publications which explore phylogenetic networks as a means to shed more light on the history of a single language family or historical region, which are much closer to what historical linguists are interested in, seem much less attractive to the computational community. But the existing computational work that tries to answer the old questions that the comparative method was not able to solve conclusively, will not be readily accepted by the historical linguistics community as long as wrong partial results are treated as mere flukes that will not have an impact on the truth of the final result. Strong claims are derived from a mere two hundred words per language, although there are good reasons why historical linguistics has always built on the entire documented lexicon of the relevant languages, along with morphological and typological features that are beyond the scope of this book. In my view, statistical methods, especially the state-of-the-art methods operating in a Bayesian paradigm, could nevertheless gain wide acceptance if they incorporate much more of the available knowledge, as is generally considered good practice among Bayesian statisticians. Just as in other fields where statistics has been applied much longer, leaving out knowledge that would be available to form more informed priors should be considered problematic, and wider coverage of the available knowledge is where the focus of further developments in phylogenetic methods should lie.

The best-performing methods for automating parts of the comparative method display another set of very common problems. As algorithms get more complex, and are typically trained on gold standard data through machine learning, they turn more and more into black boxes, thus called because internal calculations become impossible to interpret for humans. For such systems, it must typically remain unclear whether they really capture some linguistically interpretable signal, or are actually trained to rely on much cruder criteria, as has frequently been the case in the history of machine learning. In my view, much of the problem is caused by the standardized ways in which tools are commonly evaluated in computational linguistics. For instance, methods for automated cognate detection are commonly evaluated only against a selection of test sets covering a single

language family each. This means that there is a very high prior probability for words of the same meaning that sound vaguely familiar to be cognates. A system trained on this type of input will typically produce many false positives when applied to a dataset which spans several language families. Moreover, the focus on attaining ever higher F-scores loses some of its appeal when one becomes aware of the fact that this measure is dominated by the easier cases (cognacy among close siblings), hiding the fact that performance for the difficult cases (cognacy detection across subfamilies) is still a long shot from what human linguists can achieve based on the classical methods.

Partly due to these problems, the future role I see for mathematical models and computational tools in historical linguistics is less in fully computational theories, but more in the paradigm of machine-assisted theory development. Conceptually, a toolbox for machine-assisted historical linguistics would largely automate simple tasks such as dictionary lookup, applying postulated sound changes, phonetic pattern matching to find additional cognates, and finding the optimal sequence of conditional replacement rules, while still relying on human intuition and curiosity to make the high-level decisions, and to receive heuristic hints on which variants to explore next. The human linguist would be able to manipulate parts of the system's initial output at will, e.g. to reject an automatically generated sound law, mark a word that the system was uncertain about as an obvious borrowing, or expand a cognate set proposed by the system with additional forms which the imperfect automated cognate detection component missed. This interaction of course requires the system output to be framed in terms that a historical linguist is used to thinking in. The feedback would then be used by the system in its next round of automated theory refinement, changing bits of the model to accomodate the linguist's ideas, re-applying the model to the parts of the data the analyses of which the user has not yet declared final, and then displaying the results back to the user. This basic feedback loop would potentially lead to much accelerated development of etymological theories, because the human linguist could feed the system with ideas that it found impossible to generate on its own, whereas the computer could tell the human linguist in an instant whether e.g. their new idea for a reconstructed form covers all attested reflexes, and direct their attention to potential unresolved problems. The resulting fully specified theory (including all the decisions that were manually enforced or confirmed by the linguist) could then be shared in a digital format, which would allow other linguists to load the theory into their copies of the system, configuring it according to their knowledge, and inspecting the resulting changes to the automated analysis.

I have recently been given the chance to take the first steps towards building a prototype of such a system. The planned Etymological Inference Engine (EtInEn) will be built on Probabilistic Soft Logic (PSL), a recent framework for relational learning which allows to combine inviolable constraints (which I am using to implement the core logic of the comparative method) with weighted rules (which will allow modeling the heuristic rules which are commonly used when the logic does not yield results). The final system will be built around a backbone of a database of elementary assumptions connected by PSL constraints and rules, which will be used as a common interface to transmit information between a variety of specialized reasoning components. Many smaller parts of the infrastructure developed during the research leading up to this book could be turned into such components. Information-weighted sequence alignment will help in automatically finding good candidates for cognates that no longer overlap in meaning due to semantic shifts. NorthEuraLex, covering a large number of well-researched languages, will provide the starting point for a very rich and accessible testset to play around with. Finally, PLFI and CLFI will find their place among many other tools as a quick way to generate a unique view on a dataset, helping to isolate the contacts which minimally need to be assumed to explain the shared lexical material, and coming up with a good starting hypothesis for this much quicker than a committee of human linguists could.

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

Aikio, Ante, 142, 145, 161, 191
    see also Ánte, Luobbal Sámmol Sámmol
Anikin, A. E., 152
Ánte, Luobbal Sámmol Sámmol, 145
    see also Aikio, Ante
Atkinson, Quentin D., 34, 176, 189, 285

Baba, Kunihiro, 71
Bailey, H. W., 167
Beckwith, Christopher I., 159
Bereczki, Gábor, 146
Bergsland, Knut, 161
Bergstrom, Carl T., 126
Bollback, Jonathan P., 221
Bouchard-Côté, Alexandre, 40
Bouckaert, Remco, 34
Bouma, Gerlof, 115
Bowern, Claire, 13, 28, 34
Brown, Cecil H., 115
Bryant, David, 40
Buch, Armin, 92, 109
Buck, Carl D., 91

Campbell, Lyle, 17
Chaves, Rafael, 75
Chickering, David Maxwell, 87
Claassen, Tom, 87
Collinder, Björn, 160
Colombo, Diego, 80, 85, 86
Comrie, Bernard, 89
Corson, David, 143
Cover, Thomas M., 73
Csató, Éva Ágnes, 153

Dahl, Östen, 140
De Oliveira, Paulo Murilo Castro, 172
De Vaan, Michiel Arnoud Cor, 188
Dellert, Johannes, 3, 28, 92, 94, 95, 109, 251
Dol'gopol'skij, Aron B., 97
Dunn, Michael, 29, 158
Dybo, Anna V., 152
Dyen, Isidore, 29
Décsy, Gyula, 141

Ellison, T. Mark, 40
Embleton, Sheila M., 26, 172
Evans, Bethwyn, 13

Feist, Timothy Richard, 143
Felsenstein, Joseph, 30, 33, 34, 215
Finkenstaedt, Thomas, 11
Fisher, Ronald A., 54
Fortescue, Michael D., 134, 136, 161
François, Alexandre, 12
Friedman, Nir, 62

Geisler, Hans, 29
Goldberg, Yoav, 202
Grant, Anthony, 132
Gray, Russell D., 32, 34
Greenhill, Simon J., 28, 30, 194

Grünthal, Riho, 146
Gruzdeva, Ekaterina, 158
Guy, Jacques B. M., 25, 115

Häkkinen, Jaakko, 145, 146, 157, 161
Halilov, Madžid Šaripovič, 165
Halojmova, K. N., 92
Hammarström, Harald, 91
Haspelmath, Martin, 22, 132
Hauer, Bradley, 131
Hausenberg, Anu-Reet, 148
Hawkins, John A., 191
Helimski, Eugene, 152, 191
Heskes, Tom, 87
Hewitt, George, 164, 165
Hewson, John, 25, 39
Ho, Trang, 95
Hochmuth, Mirko, 171, 173
Hock, Hans H., 19
Holden, Clare Janaki, 32
Holman, Eric W., 177, 189
Holton, Gary, 35
Hörberg, Thomas, 175
Hruschka, Daniel J., 114
Huelsenbeck, John P., 221
Huson, Daniel H., 40, 332

Jäger, Gerhard, 92, 127, 225
Janhunen, Juha, 154, 155, 159, 188
Johanson, Lars, 153
Jordan, Fiona M., 32
Joseph, Brian D., 19

Kalisch, Markus, 72
Kaufman, Terrence, 19, 20, 22
Kessler, Brett, 116
Key, Mary Ritchie, 89
Kobyliński, Zbigniew, 141
Koller, Daphne, 62
Kondrak, Grzegorz, 36, 115, 131
Koptjevskaja-Tamm, Maria, 140
Kroonen, Guus, 12

Ladefoged, Peter, 7, 103
Lehtinen, Jyri, 44
Lehtisalo, Toivo, 188
Lewis, Martin W., 35, 284
Lindén, Krister, 102
List, Johann-Mattis, 29, 36, 44, 98, 102, 115, 116, 123, 126, 133, 225
Lloyd, Stuart, 125

Maathuis, Marloes H., 80
Maddieson, Ian, 7, 103
Martin, Samuel E., 160
Maslova, Elena, 157
Meek, Christopher, 78
Menges, Karl Heinrich, 153, 157
Menovščikov, G. A., 92
Michener, Charles D., 31, 125
Moravcsik, Edith A., 23
Morrison, David A., vii, 40–43
Murawaki, Yugo, 172, 281
Murayama, Shichirō, 160
Muysken, Pieter, 23
Myers-Scotton, Carol, 23

Needleman, Saul B., 37
Nefedov, Andrey, 132
Nei, Masatoshi, 31
Nguyen, Lam-Tung, 216
Nikolaeva, Irina, 136
Nikolayev, Sergei L., 136
Novgorodov, Innokentij, 132

Oakes, Michael P., 39, 115

Pagel, Mark, 175

Pakendorf, Brigitte, 132
Pearl, Judea, vii, 52, 62, 63
Pereltsvaig, Asya, 35, 284
Piispanen, Peter S., 160
Purvis, Andy, 189
Puura, Ulriikka, 142

Radzik, Ryszard, 141
Raghavan, Usha Nandini, 126
Rama, Taraka, 127
Ramsey, Joseph, 80
Reichenbach, Hans, 55
Richardson, Thomas, 68, 70, 81
Rießler, Michael, 132
Roch, Sebastien, 194
Rosvall, Martin, 126
Rot, Sándor, 143
Róna-Tas, András, 148

Saitou, Naruya, 31
Salminen, Tapani, 146
Sammallahti, Pekka, 141, 188
Sankoff, David, 32, 218
Sankoff, Gillian, 179
Schmidt, Christopher K., 132, 159, 160
Schulte, Kim, 132, 149
Schulze, Christian, 171
Scornavacca, Celine, 332
Senn, Alfred, 143
Sergejeva, Jelena, 143
Sicoli, Mark A., 35
Siegl, Florian, 148
Simon, Allan, 95
Skorik, P. J., 158
Smolicz, Jerzy J., 141
Snir, Sagi, 194
Sofroniev, Pavel, 127
Sokal, Robert R., 31, 125
Spirtes, Peter, 63, 68, 70, 72, 76, 79–82
Starostin, Sergei A., 136
Steiner, Lydia, 36
Steudel, Bastian, 207
Suhonen, Seppo, 143
Swadesh, Morris, 176
Syrjänen, Kaj, 29

Taagepera, Rein, 148
Tadmor, Uri, 91, 132, 187
Thomas, Joy A., 73
Thomason, Sarah Grey, 19, 20, 22
Thordarson, Fridrik, 166
Turchin, Peter, 125

Vahtre, Lauri, 157, 158
Vajda, Edward J., 132, 156, 161
Van der Sijs, Nicoline, 132
Van Hout, Roeland, 23
Vejdemo, Susanne, 175
Viires, Ants, 157, 158
Viitso, Tiit-Rein, 141
Volodin, A. P., 92, 158

Wells, John C., 99
Wichmann, Søren, 24, 27, 89
Wiebusch, Thekla, 132
Willems, Matthieu, 43, 44
Wohlgemuth, Jan, 24
Wolff, Dieter, 11
Wunsch, Christian D., 37

Yamauchi, Kenji, 281
Yang, Ziheng, 220
Yeung, Raymond W., 75, 76
Youn, Hyejin, 285

Zachrisson, Inger, 142
Zajceva, N. G., 142
Zhang, Jiji, 81, 82, 85

Language index

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

Baltic languages, 136, 138, 141, 143, 236, 254
Bantu languages, 32, 164
Bashkir, 148, 153, 239, 241
Basque, 61, 91, 122
Belarusian, 141, 213, 264
Bokmål, 140
Breton, 94
Bulgar, 180
Bulgarian, 46, 238
Burushaski, 91, 100
Buryat, 154, 156, 241, 267

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

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

Eastern Iranian languages, 166
Eastern Saami languages, 143
Enets, 137, 148
English, 8, 9, 11–15, 17, 19–22, 24, 37, 38, 43, 46, 58, 69, 99–101, 106, 110, 111, 119, 132, 138, 160, 177, 179, 229, 252, 255, 264
Erzya, 149, 238
Eskimo-Aleut languages, 157, 160, 161, 241
Estonian, 94, 141, 143, 236, 238, 260
Evenki, 152, 156, 157, 241, 257

Faroese, 140
Finnic languages, 138, 141, 143, 236
Finnish, 17, 46, 77, 94, 121, 141–143, 199, 236, 253, 260
Finno-Permic languages, 145, 188
Finno-Saamic languages, 145
Finno-Ugric languages, 145, 188
Frankish, 21, 49
French, 21, 43, 46, 49, 94, 99, 100, 103, 138, 149, 179, 252

Galician, 96
Georgian, 29, 164, 165, 242
German, 9, 14–17, 19–21, 38, 43, 46, 53, 56–58, 93, 100, 101, 106, 108–111, 119, 141, 143, 148, 149, 199, 236, 237, 260, 263, 264
Germanic languages, 9, 11, 43, 58, 77, 90, 100, 149, 191
Gothic, 58, 61, 96
Greek, 29, 137, 149, 167, 200
Greenlandic, 136, 160

Hebrew, 9, 91, 111, 168
Hill Mari, 149
Hindi, 43, 136, 166, 179
Hittite, 96
Hungarian, 21, 91, 94, 96, 121, 134, 146, 147, 149, 180, 238
Hunnic, 153

Icelandic, 11, 53, 56, 59, 140, 200, 236
Inari Saami, 94, 142, 143
Indo-Aryan languages, 58, 136
Indo-European languages, 9, 29, 34, 44, 58, 90, 91, 96, 134, 164
Indo-Iranian languages, 136
Ingush, 164, 166
Inuit languages, 160
Inuktitut, 160
Inupiaq, 160
Iranian languages, 22, 58, 147, 164–167, 244, 260
Irish, 11, 13, 59, 99, 136
Italian, 96, 103
Itelmen, 92, 152, 158, 161, 241, 257, 267, 269

Japanese, 53, 56, 62, 64, 65, 67, 69, 94, 96, 101, 132, 137, 159, 160, 241, 269

Kabardian, 164
Kalmyk, 152, 155, 164, 241, 267
Karachay-Balkar, 167
Karelian, 141, 142, 255
Karluk Turkic languages, 153
Kartvelian languages, 165, 167
Kazakh, 153, 154, 167, 241, 264, 267
Ket, 134, 152, 156
Khaladj, 153
Khalkha Mongolian, 154
Khanty, 146, 148, 238

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

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

Malayalam, 136
Manchu, 94, 152, 155, 257
Mandarin Chinese, 65, 91, 132, 159, 187, 269
Mansi, 94, 146, 148, 238
Mari languages, 146, 148, 238
Meadow Mari, 149, 239
Middle Chinese, 60, 62, 64, 65, 67, 159
Middle English, 24, 252
Middle Mongol, 154
Moghol, 154
Moksha, 149
Mongguor, 154
Mongolian, 154, 179
Mongolic languages, 96, 136, 152, 154, 155, 159, 166
Mordvinic languages, 146

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

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

Old Japanese, 60, 65, 68, 160
Old Korean, 60, 66–68
Old Norse, 43, 140
Old Prussian, 141
Old Turkic, 154
Olonets Karelian, 141, 143, 253
Ossetian, 90, 136, 166
Ottoman Turkish, 168

Paleosiberian languages, 156
Pali, 21, 58
Pama-Nyungan languages, 34
Papuan languages, 28
Pashto, 101, 136, 166, 271
Pecheneg, 167
Permian languages, 146–148, 238
Persian, 11, 24, 101, 154, 165–168, 244, 256, 260, 269, 271
Polish, 141, 143, 213, 254, 264
Portuguese, 160

Romance languages, 11, 21, 61, 134, 238
Romani, 187
Romanian, 46, 132, 149, 238
Russian, 11, 20, 100, 141–143, 148, 152, 154–158, 160, 161, 164, 165, 168, 213, 236, 237, 239, 241, 242, 253, 257, 264, 267
Ryukyuan languages, 160

Saami languages, 141, 142, 191, 236, 267
Sakha, 132, 152–157, 241
Samoyedic languages, 91, 145, 152, 157, 188, 191, 238, 239
Sanskrit, 21, 58, 96, 136, 154
Scytho-Sarmatian languages, 166
Selkup, 148, 156, 241, 257
Semitic languages, 38, 108, 111, 168
Siberian Turkic languages, 91, 153
Siberian Yupik, 92, 134, 160, 161
Skolt Saami, 94, 142, 143, 213
Slavic languages, 46, 58, 90, 136, 141, 149, 164, 238, 264, 267
Slovak, 148
Sorbian, 96
South Caucasian languages, 165
South Slavic languages, 147
Southern Saami, 94, 142, 204
Spanish, 8, 23, 59, 93, 110
Svan, 165
Swedish, 11, 46, 57, 94, 140, 142, 204, 236, 237, 255, 260, 264

Tatar, 117, 148, 153, 241
Telugu, 69, 134
Tocharian, 96
Tok Pisin, 24
Tsez, 136, 165, 166
Tsezic languages, 165
Tundra Yukaghir, 136, 157
Tungusic languages, 96, 152, 155, 159, 241, 269
Turkic languages, 90, 96, 114, 136, 147, 149, 152, 164–166, 180, 238, 241, 244, 264, 271
Turkish, 24, 117, 148, 149, 153, 166, 180, 244, 256
Turkmen, 153

Udmurt, 96, 134, 137, 146, 148, 149, 238, 239, 260
Ukrainian, 96, 267
Uralic languages, 17, 29, 38, 44, 77, 90, 94, 134, 144, 149, 152, 157, 160, 188, 238, 260
Urdu, 166

Uyghur, 153
Uzbek, 24, 153, 168, 244, 269

Veps, 138, 141, 142, 236

Welsh, 13, 136
West Frisian, 96
West Germanic languages, 237
West Slavic languages, 147
Western Iranian languages, 166
Western Saami languages, 94, 143, 234, 267

Xibo, 155

Yaghnobi, 166
Yeniseian languages, 35, 91, 156, 161
Yukaghir languages, 157, 158, 160, 241
Yupik languages, 160

Subject index

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

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

causal DAG, 66
Causal Faithfulness Condition, 68
causal graph, 65
Causal Markov Condition, 68
causal skeleton, 77
causal sufficiency, 66
chain, 66
cognacy class, 13
cognate, 13
collider, 66
combined information content, 109
common cause principle, 55
comparative method, 14
completed partially directed acyclic graph (CPDAG), 70
conditional independence, 57
conditional mutual information, 75
confounder, 52
Conservative PC algorithm, 80
contact flow network, 46
Contact Lexical Flow Inference (CLFI), 257
contraction property, 58
creole, 24

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

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

faithfulness, 68

FCI algorithm, 80
flow separation (FS), 205
fork, 66

galled network, 43
galled tree, 43
GES algorithm, 87
graphoid axioms, 60

hidden common cause, 52
Hungarian, 147
hybridization network, 43

IELex, 29
independence, 56
inducing path, 81
InfoMap algorithm, 126
information content, 107
internal borrowing, 21, 194
intersection property, 58
isolate, 190
IWD (Information-Weighted Distance), 110
IWSA (Information-Weighted Sequence Alignment), 110

joint entropy, 74
joint reconstruction, 220

label propagation algorithm, 126
language contact, 11
language family, 9
Levenshtein distance, 36
lexical item, 7
lexical replacement, 8
loanword, 11

m-separation, 68
majority-based reconstruction, 217
marginal reconstruction, 220
Markov condition, 62
Markov equivalence, 70
maximal ancestral graph (MAG), 68
maximum likelihood, 33
maximum parsimony, 32
median network, 41
minimal lateral network, 44
monotone faithfulness, 207
monotonicity, 76
multi-value ML reconstruction, 221
multi-value MP reconstruction, 218
mutual information, 74

neighbor-joining algorithm, 31
neighbor-net, 42
NorthEuraLex, 28

outlier, 34

parent relation, 65
parsimony network, 42
partial ancestral graph (PAG), 70
partial correlation, 71
path (in graph), 65
PC* algorithm, 79
Pearson correlation, 71
phylogenetic inference, 30
Phylogenetic Lexical Flow Inference (PLFI), 225
phylogenetic network, 40
phylogenetic tree, 9
phylum separation score, 261
pointwise mutual information, 74

recipient language, 11
reticulation cycle, 43
RFCI algorithm, 85
rooting, 34

Samoyedic languages, 147

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

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

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

v-structure, 66

wave model, 12
weak union property, 58
weighted imbalance score, 190
WOLD (World Loanword Database), 132

X-SAMPA encoding, 99

Information-theoretic causal inference of lexical flow

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

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

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

ISBN 978-3-96110-143-6