Information-theoretic causal inference of lexical flow

Johannes Dellert

Language Variation 4

language science press

Language Variation

Editors: John Nerbonne, Dirk Geeraerts

In this series:

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

ISSN: 2366-7818

Information-theoretic causal inference of lexical flow

Johannes Dellert

language science press

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

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

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

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

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

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

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

3 Foundations: causal inference

In this chapter, I give a concise introduction to the basic notions and methods of causal inference. Since this branch of statistics is quickly growing into a large field, the discussion is focused on leading the reader towards an understanding of the theory behind the methods I am developing. This implies that the exposition will only consider methods which are applicable to discrete data, and disregard the many methods which are being developed and improved for continuous observations.

3.1 Philosophical and theoretical foundations

This section starts with a look at the very basic intuitions and established principles of the field. After some general considerations about the well-known issues in linking correlation to causation, we motivate the need for causal thinking even in the absence of experiments, and turn to the central idea of using nature as an experimenter.

I then introduce the crucial notion of conditional independence between variables, which roughly amounts to a criterion for deciding whether the connection between two variables can be explained away by considering the possibility of mediating other variables. For instance, there probably is a measurable correlation between the occurrence of hats and trousers as pieces on the clothing of humans, which might disappear when we condition on gender. The gender of the body suffices to predict how likely we are able to find a hat or trousers on it, and there is no direct connection in the sense that putting a hat on one's head would cause one to also put on trousers, or the other way around.

Sets of conditional independence statements over a set of variables identify a graphical model over these variables, which is a decomposition of the joint distribution into factors given by a neighborhood relation in a graph. I will present some mathematical notions which were developed to make this relationship more precise, allowing us to exploit it for inferring graphical models from data.

Crucially, the Bayesian networks we can infer from data can be given a causal interpretation. If we see such a structure not as a compact way to model and a convenient handle for calculating a joint distribution, but as representing the actual information flow between the variables, we can interpret each directed edge in a graphical model as expressing a causal influence from the starting variable to the variable the edge goes into. Based on this initial idea, one can build an entire theory of intervention, allowing us to predict what would happen if one of the variables was given a certain fixed value by manipulation, as in an experiment. Pearl (2009) summarizes a wealth of previous work in this direction.

In this book, the emphasis is not on the philosophical issues about causality which the intervention calculus touches upon, but on exploring the idea of defining causal graphs over languages, and interpreting the directed arcs as indicating how languages influenced each other. In this section, we start with some general considerations about causality, and the way complex probability distributions are represented by directed graphs in the Bayesian network paradigm. I then introduce the central idea behind causal inference, i.e. giving a causal interpretation to the efficient representations generated by Bayesian network inference. The mathematical details are then the subject of the next section.

3.1.1 Correlation and causation

There are three major reasons why the common warning that correlation is not causation is true. The first is that correlation does not tell us the direction of causality. Secondly and even more importantly, for any observed correlation of two variables it is always possible (even likely) that there is a *hidden common cause* or *confounder* which influences both variables. In this case, the correlation is not due to a true causal connection which would link the two variables directly, but due to *confounding bias*. The third problem is *selection bias*, a second type of bias which can cause non-causal correlations, and can occur whenever two observed variables have an influence on the sampling process. Selection bias is a very difficult problem for clinical studies, because participation in the study frequently depends on both the strength of symptoms and the availability of treatment, which can create a spurious impression for some treatment to have an effect on the symptoms.

To illustrate the general applicability of the causal inference framework, I will work with two running examples throughout this chapter. For our more classical example with actual statistical variables, take the variable R to mean the average room temperature during the winter months, which we can measure in any household of an entire country. Let O be the average outside temperature in the respective city during the same timeframe, H the heating costs per square meter for the household, and I some measure of the household income. In a wealthy welfare state where virtually everyone can afford as much heating as they want, I would assume the average room temperature R to only be based on personal preferences P, and hence independent of both the outside temperature O and household income I. Since people will try to maintain their desired room temperature R by influencing H, there will obviously be a negative correlation between H and O. On some level, this does mean that lower outside temperatures cause heating costs to increase, but this is only the case because people are trying to keep R constant, i.e. the causality is mediated by the room temperature, on which both the amount of heating and the outside temperature have a causal effect.

To give an example of selection bias, assume a politician wanted to argue for a policy that prevents poor people from spending too much money on heating. In anonymized data provided by a consumer counseling service, there does indeed turn out be a strong correlation between I and H, seemingly supporting the policy. The problem is that this is very likely a spurious correlation due to selection bias, because low-income households with high heating costs (e.g. due to high P) are much more likely to get counseling, and thereby becoming part of the study, than other households.

In lexical flow inference, I will model language varieties as variables which influence each other's lexica. The second running example in this chapter serves to establish and illustrate this view, taking the first steps to an abstract understanding of the approach before we turn to the data needed for an implementation in Chapter 4, and the actual mathematical details in Chapter 6. I will often be deliberately vague in this chapter when I write about computing the correlation between languages, or explaining the dependence between two languages based on a third one, in order to demonstrate how general the reasoning patterns are, and that they would apply to many ways of modeling language varieties as mathematical objects. The algorithms in Chapters 6 and 7 will be based on one particular way of fleshing out these reasoning patterns in terms of shared cognates, but the general reasoning patterns could just as well be applied to typological variables, measures on parallel texts, and many other mathematical models of language.

Generally, if we have some measure of determining whether two languages are related or not, we can frame this relation in terms of independence. For instance, the languages Japanese (*jpn*) and Icelandic (*isl*) are commonly assumed to not be demonstrably related, and should therefore be independent according to any reasonable measure. On the other hand, the languages Danish (*dan*) and German (*deu*) are related, and should therefore come out as dependent. Danish and German do, in fact, have a hidden common cause, namely their latest common ancestor Proto-Germanic (*PGer*). In addition, Danish includes many loanwords from German, which a good model would detect as additional correlation in addition to the amount of correlation caused by the common proto-language, and ideally, it would make it possible to determine German as the source of borrowing. We would thus frame the Danish lexicon as being "caused" by the words of Proto-Germanic and German.

With this general idea in place, let us stop and consider the question to what extent we can say that the lexicon of a language is caused by the lexicon of its immediate ancestor, as well as the lexica of any donor languages. To get a more concrete picture of the actual process underlying this way of thinking, we need to move down to the level at which the process which perpetuates a language actually operates. This process boils down to the acquisition of a language by single speakers, which can effectively stretch across throughout much of their lifetimes. When a child acquires words, the languages of the parents or other close relatives will likely be the source of these words. During later development, speakers might be exposed to additional languages (such as a dominant state language) at school, at the workplace, or while traveling, which adds additional speakers to the list of causes for their lexicon. But even within the same language community, speakers will continually acquire new words which they hear from other humans that they are in contact with. On the most elementary level, the language of an individual speaker can thus with some justification be said to be caused by the languages of other speakers. The causal metaphors becomes more problematic if we lift it to the level of entire languages, summarizing thousands or millions of individual speaker histories to the level where we speak of Proto-Germanic causing German, even if not directly. It would even be wrong to say that the language of one generation is the main cause of the language of the next, because on the level of a population, a generation is just as artificial a category as the language stages I mentioned at the end of the previous chapter. Still, with the disconnect between our abstraction and the actual process in mind, conceiving of languages as being caused by their ancestors as well as additional donor languages will turn out to be a very fruitful metaphor.

3.1.2 Causality without experiment

At least since Fisher ([1925] 1934), one of the foundational works of modern statistics, the mainstream view on causality has been that it can only be determined by experiment, i.e. by manipulating one of the variables while observing the effect on the others. When analyzing observational data without the possibility of manipulation, the principle has been to avoid thinking in causal terms, because there was no way of giving a causal interpretation to observed correlations. This inability and prudent refusal to talk about causality has left classical statistics in a rather problematic situation. Most applied statistics is arguably motivated by causal thinking, because a prime motivation of research is to find out why things happen. For this reason, results obtained from observational data will invariably be given causal interpretations, as questionable as that may be from the statistical point of view.

A very promising and comparatively recent approach to alleviating this tension has arisen in the field of causal inference, where causal notions receive a systematic and consistent treatment with the help of graph theory. From the perspective of inference, the core idea is to see nature as an experimenter, in the sense that the random fluctuations in various variables are comparable to manipulations in experiments, even though they are not performed by humans, and certainly not under controlled conditions. For this reason, even more careful thinking is required to avoid the pitfalls which even controlled experiments suffer from. We must ensure that all the possibly relevant variables are taken into consideration, so that there are no confounders leading us to the wrong conclusions. Moreover, this view requires us to put belief into the common cause principle (CCP), most often attributed to Reichenbach (1956: Ch. 19), where the original formulation reads as follows: "If an improbable coincidence has occurred, there must exist a common cause". In statistical terms, this principle implies that variables are typically not correlated by chance. Any significant correlation between two variables must be due either to a direct causal relationship (in whichever direction), or, if they are measured simultaneously, a common cause.

In our heating costs scenario, I would expect to see a correlation between I and H, indicating that more wealthy households will have higher heating costs per square meter. It does not make sense to posit either that I causes H (a low wage makes it easier to keep each square meter of your house warm), or that H causes I (higher heating costs increase your wage), so according to the CCP there must be a common cause influencing both I and H. After some research, we might find that rich people have a tendency to afford larger windows, which decreases their building's heat insulation and makes it more expensive to maintain the desired room temperature. We might therefore include the confounder into our model, as a variable W measuring the window expanse per square meter. If we have no way of measuring this variable, it becomes a hidden confounder that we need to take care of.

For the language scenario, the CCP applies as well, because we would also expect any similarity between languages to be explainable either by chance (in case the correlation is not significant, for example), by a direct causal relationship (i.e. borrowing), or by a common cause (e.g., a common proto-language, or a common source language for shared loans). The case of an additional confounder that needs to be included into our theory may occur in the case of a common substrate layer.

3.1.3 Conditional independence

Two statistical variables X and Y are said to be *independent*, which we write as $(X \perp\!\!\!\perp Y)$, iff $p(x, y) = p(x)p(y)$ for all values x of X and y of Y. Intuitively, independence means that we do not know any more about the value of Y if given the value of X, and vice versa.

In our example scenario, we assumed the room temperature R to be independent of the household income I. The definition implies that if we know the distribution of room temperatures and the distribution of household incomes, we can predict the joint probability $p(R, I)$ for any given values of R and I. Knowing the average room temperature does not allow us to make a more educated guess about whether we are dealing with a high-income household, and vice versa. In contrast, we said that the outside temperature and the heating costs are not independent. Knowing the outside temperature will change our expectations about the heating costs, and vice versa.

There are many ways in which independence could be defined over languages. If two languages like Japanese and Icelandic are independent, the definition implies that if we know, for example, the word for some animal in Japanese, this will not help us in any way to predict how that animal might be called in Icelandic ($jpn \perp\!\!\!\perp isl$). In contrast, if we know that a snake is called *slange* in Danish, this will allow us to make a more educated guess about the German word. Such an educated guess might be successful in some cases (German *Schlange* ‘snake’) and less successful in others (Danish *ræv* vs. German *Fuchs* ‘fox’), but knowledge of the one language helps us to predict the word in the other languages in enough cases to be statistically relevant ($dan \not\!\perp\!\!\!\perp deu$).

Moving one step further, the central notion for causal inference is the *conditional independence* of a pair of (sets of) variables given a set of other variables. If two variables X and Y are dependent, but conditionally independent given a set of variables Z, we say that the dependence disappears when *conditioning on* Z. Intuitively, if we know the values of the variables in Z, knowing the value of X will not tell us anything new about the value of Y, and knowing Y will not add to our knowledge about X.

For the formal definition, let p be a joint probability function over a finite set V of variables. For any sets $X, Y, Z \subseteq V$, we say that X and Y are conditionally independent given Z [in symbols: $(X \perp\!\!\!\perp Y \mid Z)$] if $p(x|y, z) = p(x|z)$ whenever $p(y, z) > 0$.

In the heating example, we will certainly observe a dependence $(P \not\!\perp\!\!\!\perp R)$ between the temperature preference P and the room temperature R. However, it is certainly not the case that the room temperature will drop just because we want it to do so, and a rising temperature in the bedroom will only increase our discomfort, but not the temperature at which we would want to sleep. Instead, the process connecting these two variables is mediated by the heating costs H, and we are going to have $(P \perp\!\!\!\perp R \mid H)$ because if we are not allowed to manipulate the heating, our preferences will not have any impact on the room temperature any longer.

Conditional independence relations hold between many types of knowledge we might have about languages as well. For instance, Swedish (*swe*) is more closely related to Danish than German is, which is why on average, additional knowledge of German will not help us to understand a Swedish word if we already understand the Danish one $(swe \perp\!\!\!\perp deu \mid dan)$. To illustrate, take the words for ‘bird’ (*fågel/fugl/Vogel*), where German would have helped almost as much as Danish, and the words for ‘ant’ (*myra/myre/Ameise*), where Danish is much closer to Swedish. The latter example also shows why we would likely get $(swe \not\!\perp\!\!\!\perp dan \mid deu)$ from any useful criterion of conditional independence.

The conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$ has a number of interesting properties, most of which should be intuitively obvious. For instance, it satisfies *symmetry* in the sense that $(X \perp\!\!\!\perp Y \mid Z)$ implies $(Y \perp\!\!\!\perp X \mid Z)$, which we can interpret to mean that if X does not tell us anything about Y, neither will Y provide us with any information about X. This is a very natural assumption for languages as well, since language relatedness is a symmetric relation.

It also has the *decomposition* property in the sense that jointly irrelevant pieces of information do not become relevant when considered separately, i.e. $(X \perp\!\!\!\perp YW \mid Z) \Rightarrow (X \perp\!\!\!\perp Y \mid Z)$. For instance, given that the knowledge of both a person’s income and his or her temperature preferences does not allow us to draw any conclusions about the climate, it makes sense to assume that neither will income or preference data alone. To understand why the decomposition property holds for languages as well, consider the situation where two groups of languages do not share any features that are not explainable by a third set of languages. If this is the case, this third set will also explain all the overlap between individual languages from the two groups. For instance, if we have established that Proto-

Indo-European can be used to explain all similarities between Indo-Aryan and Germanic languages ($IndoAryan \perp\!\!\!\perp Germanic \mid PIE$), we can conclude that it will also explain the (fewer) similarities between individual language pairs such as Sanskrit and German ($san \perp\!\!\!\perp deu \mid PIE$), or Pali and English ($pli \perp\!\!\!\perp eng \mid PIE$).

Another property is *weak union*, which states that an irrelevant piece of information Y will not suddenly become relevant if we learn another irrelevant piece of information W, or $(X \perp\!\!\!\perp YW \mid Z) \Rightarrow (X \perp\!\!\!\perp Y \mid ZW)$. In the heating cost example, neither the household income nor the outside temperature allow us to say anything about the room temperature, and we would be very surprised if, say, a correlation between the two temperatures would appear once we look at rich households only. For groups of languages whose similarities are explained by a third language, like in the previous example situation ($IndoAryan \perp\!\!\!\perp Germanic \mid PIE$), we would not expect one of the languages from one group (say, Gothic) to provide additional information that makes the remainder of the group appear more similar to Indo-Aryan, i.e. we would also expect ($IndoAryan \perp\!\!\!\perp NWGermanic \mid PIE, Gothic$). The weak union property thus mirrors the informativeness of languages in historical linguistics, where relevant (i.e. non-random) similarities will always become clearer if additional languages are taken into account.

Conversely, a *contraction* property also holds, stating that information W that is irrelevant after learning another piece of irrelevant information Y, must have been irrelevant all along: $(X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp W \mid ZY) \Rightarrow (X \perp\!\!\!\perp YW \mid Z)$. In the heating costs example, we know that the outside temperature O is not connected in any way to household income I. Assume that in two separate studies of richer and poorer people, no connection between people's preferences P and the outside temperature O was found, i.e. it was established that $(P \perp\!\!\!\perp O \mid I)$. In this situation, it would appear nonsensical if it turned out that in the overall population, we had $(P \not\perp\!\!\!\perp O)$. Similarly, considering the relationships between branches of Indo-European, if we used Iranian evidence to explain the similarities between Slavic and Armenian ($Slavic \perp\!\!\!\perp Armenian \mid PIE, Iranian$), but also found that Slavic and Iranian are independent branches that did not remain in similarity-inducing contact after the split-up of Proto-Indo-European ($Slavic \perp\!\!\!\perp Iranian \mid PIE$), we know that we would have needed Iranian to split Slavic from Armenian. On the other hand, contraction does not exclude the possibility that ($Iranian \not\perp\!\!\!\perp Armenian \mid PIE$), which we would actually not be surprised to find given the considerable influence of Iranian languages on Armenian.

Finally, there is an *intersection* property stating that if given some other knowledge Z, two sets of variables are mutually irrelevant to a set of variables X, neither of them will be relevant to X in isolation (nor jointly, by decomposition): $(X \perp\!\!\!\perp W \mid ZY) \wedge (X \perp\!\!\!\perp Y \mid ZW) \Rightarrow (X \perp\!\!\!\perp YW \mid Z)$. Expanding on the heating cost example to get an instance of this reasoning pattern, let us introduce the latitude L into the picture. Obviously L is going to influence O. Moreover, assume that due to historical scarcity of heating materials, people in climates with lower O have developed a culture that reduces the window size W. In this situation, where both $(L \perp\!\!\!\perp R \mid O, W)$ and $(L \perp\!\!\!\perp W \mid O, R)$ hold, we would not expect to find either $(L \not\perp\!\!\!\perp R \mid O)$ or $(L \not\perp\!\!\!\perp W \mid O)$, because the average outside temperature should already provide all the necessary explanation for the connection between latitude and window size. To also motivate the intersection property among sets of languages, we will consider a likely configuration of conditional independence statements involving Irish (*gle*), Icelandic (*isl*), and Spanish (*spa*). Since these three languages come from separate branches of Indo-European, and have not been in contact since PIE split up, we would expect $(gle \perp\!\!\!\perp isl \mid PIE, spa)$ and $(gle \perp\!\!\!\perp spa \mid PIE, isl)$ to hold in the absence of selection bias. These constraints say that provided with background information about the common Indo-European elements, Icelandic does not tell us anything new about Irish if we already know Spanish, but neither does Spanish if we already know Icelandic. In this situation, it would be nonsensical to assume that both languages together would provide any relevant information about Icelandic, which is mirrored by the intersection property telling us that $(gle \perp\!\!\!\perp isl, spa \mid PIE)$. To understand why this reasoning pattern is only valid if the irrelevance is mutual, consider the same situation with Dutch (*nld*) and English instead of Icelandic and Spanish. The situation for Dutch is similar to the one for Icelandic, so that we will have $(gle \perp\!\!\!\perp nld \mid PIE, eng)$, but the reverse does not hold any more, because English and Irish have been in contact $(gle \not\perp\!\!\!\perp eng \mid PIE, nld)$. Here it makes sense that the reasoning pattern does not apply, because otherwise we had $(gle \perp\!\!\!\perp nld, eng \mid PIE)$, which would imply the wrong statement $(gle \perp\!\!\!\perp eng \mid PIE)$ by decomposition.

After stating the different properties of conditional independence, and understanding that they represent very general reasoning patterns that make sense in many domains, we can now take the decisive step connecting conditional independence constraints to graphs. If we picture the different variables as nodes, and the edges as communication channels which allow the transfer of information in both directions, we find that the conditional independence relation can be assigned a straightforward interpretation in terms of paths in a graph. If we define $(X \perp\!\!\!\perp Y \mid Z)$ as meaning that every path from a node in X to a node in Y will be blocked by some node in Z, the relation has the same five properties, which are therefore called the *graphoid axioms*. Together, they have been found to characterize informational relevance very well in many different contexts, and the examples may already have convinced the reader how close the correspondence in fact is. Conditional independence relationships will give us testable elementary statements which we can use to construct graphs over languages, such as the one in Figure 3.1, which will become our running example to illustrate my view on lexical flow networks. In this much simplified picture of the interactions as reconstructed by linguists, the edges represent information flow, and more specifically, the flow of lexical material, between various historical stages of the three major languages of East Asia. For instance, the theory represented by the graph states that the information flow from Old Chinese to Old Japanese was mediated either by Middle Chinese or Old Korean, whereas Middle Chinese directly influenced it. In the next section, we will establish a systematic correspondence of such path constraints with conditional (in)dependence statements such as, in this case, $(OC \perp\!\!\!\perp OJ \mid OK, MC)$ and $(MC \not\perp\!\!\!\perp OJ \mid OK, OC)$.

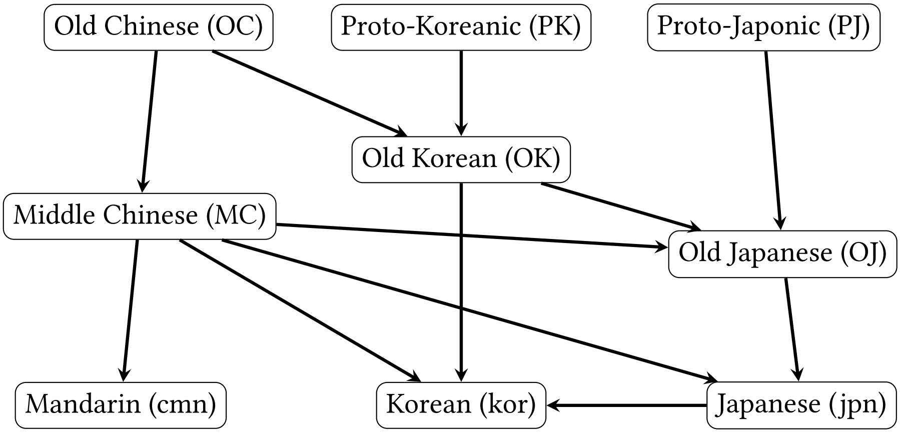

Figure 3.1: Example graph over (selected) languages of East Asia

As a final remark about conditional independence, it is worth emphasizing that conditioning on additional variables can not only remove dependencies, but it can also induce dependencies between otherwise independent variables. As stated earlier for our heating costs example, we are likely to observe an independence $(R \perp\!\!\!\perp O)$ if we do not condition on anything else. However, this is only the case because people will regulate the heating to ensure a constant R. But any fixed investment H into heating will be more or less effective at different outside temperatures. For fixed H, room temperature will therefore begin to depend on the outside temperature, so that we have a conditional dependence ($R \not\perp\!\!\!\perp O \mid H$), whereas we had ($R \perp\!\!\!\perp O$). This is also the pattern underlying selection bias, which can be treated in the framework of conditional independence relations by introducing a hidden selection variable S, and assuming that the data are actually from the conditional distribution given S. This effectively turns every statement of the form ($X \not\perp\!\!\!\perp Y$) into an underlying ($X \not\perp\!\!\!\perp Y \mid S$), which does not contradict ($X \perp\!\!\!\perp Y$) in the underlying truth.

Selection bias is a frequent problem in the application of statistical arguments within historical linguistics. For any linguist trying to prove that two groups of languages are related, it is difficult to avoid a natural tendency to filter the material in some way, e.g. by focusing on certain parts of the vocabulary that seem more promising. This becomes a problem as soon as statistical testing comes into play, where bias-free sampling is essential for estimating the amount of similarity that we would expect by chance, and therefore correctly determining the significance level. To illustrate that selection bias can also have an impact on purely data-driven approaches, assume that we want to assess by large-scale vocabulary comparison whether there is a hidden common ancestor between Basque (*eus*) and Albanian (*sqi*), such as a common Old European substrate. In a balanced sample of our vocabulary, we would very likely find ($eus \perp\!\!\!\perp sqi$). Now assume that we want to build the analysis on parallel word lists which include the oldest attested languages of Europe, such as Gothic (*got*). If the compiler of these word lists had any tendency to tune the selection of concepts towards those for which a Gothic word is known, e.g. in order to create a more fully populated table, the nature of the only available longer Gothic text (a Bible translation) will induce a selection bias. Both Basque and Albanian are heavily influenced in their religious vocabulary by Latin and Romance languages, which could easily lead to a statistically significant similarity on the word list ($eus \not\perp\!\!\!\perp sqi \mid got$), even though we only considered data for the two languages, and might erroneously conclude that ($eus \not\perp\!\!\!\perp sqi$). The same could happen in corpus-based approaches that rely on religious texts such as Bible translations, which are quite frequently the only available written records especially of smaller languages that do not have an indigenous writing tradition.

3.1.4 Bayesian networks

The more variables we consider, the more difficult it becomes to estimate and represent their joint distribution. However, if factorizations of the joint distribution function are possible, they can be used to build a directed acyclic graph

(DAG) where each node contains the conditional probability distribution of one variable given a set of parent variables. The resulting *Bayesian networks* provide a very compact way of representing joint distributions of many variables, and lend themselves very well to a range of important inference tasks. Pearl (1988) is the classical book on Bayesian networks, which explains many of the basic notions and issues in a very detailed and readable fashion, including the motivation of the following definitions.

Let $p(v)$ be the joint probability distribution on an ordered set $V = \{X_1, \dots, X_n\}$ of variables. A set $PA_j \subset V$ is said to be the *Markovian parents* of a variable $X_j \in V$ if it is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In our East Asian example, $\{OJ, OC\}$ is not the set of Markovian parents of Japanese, because according to the graph, this set does not explain all the overlap between Middle Chinese and Japanese. Adding Middle Chinese to the set will still not give us the Markovian parents of Japanese, because the subset $\{MC, OJ\}$ already covers all paths from any other language into Japanese, and Old Chinese is not a Markovian parent, because it is screened off by the other two languages in the set $\{MC, OJ\}$, which is therefore the set of Markovian parents of Japanese. Note that the situation could be different if we had chosen to model, say, Middle Japanese in addition, which would replace Old Japanese in the set of Markovian parents.

If a probability function p admits the factorization $p(x_1, \dots, x_n) = \prod_i p(x_i \mid pa_i)$ relative to a DAG G, we say that p and G are *(Markov) compatible*. In words, each variable must be independent of its non-descendants in the graph if the state of its Markovian parents is known. We say that a joint distribution fulfills the *Markov condition* if there is some graph to which it is compatible. As we have seen, the graph is mirrored by a factorization of the joint probability distribution, which can be used to compactly represent and efficiently perform calculations on joint and marginal probabilities.

Later textbooks such as Koller & Friedman (2009) build on a much more advanced theory and two decades of practical experience in applying graphical models to many problems. However, Bayesian networks have developed into a separate field which is mainly concerned with other tasks, such as the efficient inference of marginal distributions from joint distributions encoded as networks. This newer literature might therefore serve less well to lead the viewer towards an understanding of their importance for causal reasoning, a topic which we turn to in the next section.

3.1.5 Causal interpretation of Bayesian networks

From a philosophical angle, it appears attractive to give a causal interpretation to Bayesian networks, taking the directed arcs in Bayesian networks to represent direct causal influence of one variable on another. Pearl (2009) presents a range of ideas that follow from this viewpoint. Reinterpreting Bayesian networks not as convenient representations of joint probability distributions, but as causal DAGs which actually model the causal processes that generated the data, makes it possible to predict what happens if one of the variables is manipulated from the outside. Pearl develops this idea into a full intervention calculus, providing a framework for calculating answers to counterfactual questions as they arise in jurisdiction when, for instance, responsibility for an accident needs to be decided, and which previously were notoriously difficult to grasp mathematically. As Pearl shows, his fresh view of causality also provides a handle on long-standing statistical paradoxes like Simpson's paradox, where the direction of a correlation between two variables can revert in each individual case when we consider all values of a third value separately. Using the calculus of interventions with its true manipulation operation distinct from conditioning, the seeming paradoxes instantly disappear. As Pearl shows, they only existed due to a confusion of conditioning and manipulation, two operations which the classical statistical methods could not cleanly distinguish. By giving a causal interpretation to Bayesian networks, and thinking about experiments as manipulating the network, we gain a mathematical language for speaking about causality.

Stepping back from predicting the consequences of manipulation, the conditional independence relations we can extract from observational data can be exploited systematically to infer parts of the process which generated the data. This algorithmic side of causal analysis was further developed by Spirtes et al. (2000). Their book provides many of the technical proofs for the theory of inferred causation, and contains the first versions of several central algorithms which I will build upon in this book. It also provides a wealth of impressive examples, including a very in-depth discussion of how causal inference can be used on existing data to prove once and for all that smoking does cause lung cancer. The argument effectively destroys the tobacco industry's last line of defense, which consisted in claiming that the correlation might be due to a genetic predisposition which causes both a taste for cigarettes, and a propensity to develop lung cancer, an explanation which seems absurd, but cannot be ruled out by classical statistical methods.

In this book, I explore how this framework can be applied to languages. If we manage to model language graphs such as the East Asian example in Figure 3.1

as Bayesian networks, we can give a causal interpretation to the arrows, and should be able to use causal inference in order to infer the structure of the network which generated the observable data. If our observations consist of lexical data (as they do in this book), the causal graph can be interpreted as a minimal explanation of how the observed patterns could result from languages influencing each other, either by inheritance (as in the case of Old Chinese and Middle Chinese) or by borrowing (as in the case of Japanese influence on Korean). Each directed arc will represent the transmission of lexical material, and depending on how the underlying conditional independence tests are implemented, each link will correspond to a set of etymologies, detailing which words the model assumes to have been transmitted along which path through the network.

Unfortunately, as we shall see, applying causal inference in practice to a new domain is a lot less straightforward and instantly rewarding than such examples would suggest. The proliferation of additional approximation tricks and processing steps in the literature is a tell-tale sign that one should never expect readily available implementations to yield useful results on a new problem, such as lexical flow modeling in my case. Still, most work on causal inference takes place within the confines of an almost canonical set of basic ideas and algorithmic procedures, and it is this core of the causal toolbox that I will be introducing throughout the rest of this chapter.

3.2 Causal inference algorithms

This section introduces the basics of *causal inference* or *causal discovery*, which is defined as the task to analyse a set of data (observations of three or more variables) to find a *causal structure* (a partially directed acyclic graph) which mirrors the data-generating process as closely as possible. After laying out the basic assumptions and theorems which link conditional independence relations to constraints on the graph structure, I will give an overview of different approaches to testing for conditional independence, and then proceed to motivating and describing the most important causal inference algorithms. Pointers to the relevant literature will enable the reader to find proofs and more general variants of the various mathematical theorems which are needed to explain the motivation behind the algorithms, and why they work.

3.2.1 Causal graphs

I start by defining precisely the mathematical objects that I will be operating on. After defining different types of (partially) directed graphs which can be used to represent causal structures, and basic graph-theoretic notions which will be needed afterwards, I introduce the central notion of d-separation, and its generalization to ancestral graphs. Then, the main assumptions and theorems which link conditional independence relations and possible causal graphs are cited and put into context.

3.2.1.1 Basic definitions

A causal graph $G = (V, E)$ consists of a set of nodes V which represent random variables, and a set of edges $E \subset V \times V$ which will be taken to represent the causal connections between those variables. In the more general variant where the presence of hidden common causes and selection bias cannot be excluded, we have a partition $E = E_{\rightarrow} \cup E_{\leftrightarrow} \cup E_{-}$ into directed arcs which represent direct causal links, bidirected arcs which represent the existence of a hidden common cause for the two variables in question, and undirected arcs to represent the presence of selection bias inducing a dependence between two variables. For each of the three relation types, we will typically just write $X - Y$ for $(X, Y) \in E_{-}$, $X \longrightarrow Y$ for $(X, Y) \in E_{\rightarrow}$, and $X \leftrightarrow Y$ for $(X, Y) \in E_{\leftrightarrow}$.

There is some convenient short-hand terminology which can be used to talk about the relations defined by the different edge types. For instance, the asymmetric directed arcs in $E_{\rightarrow}$ define the *parent* relation, and in subsequent definitions I will write $pa(X) := \{Y \in V : (Y, X) \in E_{\rightarrow}\}$ ro refer to the set of parents of X.

A *path* in a graph G is a sequence $\langle X_0, \dots, X_n \rangle$ of distinct vertices $X_0, \dots, X_n$ where $(X_i, X_{i+1}) \in E$ for $0 \leq i < n$. If in addition, $(X_i, X_{i+1}) \in E_{\rightarrow}$ for all $0 \leq i < n$, we have a *directed path* from X_0 to X_n. If there is directed path from X to Y, or $X = Y$, Y is called a *descendant* of X, and X an *ancestor* of Y. For the set of ancestors of any node $X \in V$, we will write $an(X)$. Applying this terminology to our example network of East Asian languages, there is a directed path from Old Chinese to Mandarin, but no directed path from Japanese to Mandarin. Deviating from the standard terminology in historical linguistics, where every language has a single ancestor, and the recipient language of a layer of borrowings is not called a descendant of the donor language, in graph terminology Old Chinese is an ancestor of both Mandarin and Japanese, whereas Japanese is a descendant of both Old Japanese and Middle Chinese.

A *directed cycle* pattern exists whenever we have a directed path from X to Y, but also a link $Y \longrightarrow X$, allowing us to get back to the beginning of the path. Similarly, an *almost directed cycle* consists of a directed path $\langle X, \dots, Y \rangle$ and a bidirected arc $Y \leftrightarrow X$. In our East Asian language network, we do not have a directed cycle, and due to their temporal nature such cycles will not typically occur in language networks, unless we collapse different stages of three languages between which contacts in all directions have existed. For instance, if we did not treat Old Korean (for which little lexical data is available) as distinct from Modern Korean, we would get a directed cycle $OJ \longrightarrow jpn \longrightarrow kor \longrightarrow OJ$. The existence of directed cycles in a lexical flow network will show us something was wrong with our data, and causal inference algorithms therefore assume that directed cycles will not occur.

If $E_{\leftrightarrow}$ and E_{-} are empty, and there are no directed cycles in $E_{\rightarrow}$, G is called a *causal DAG*. This type of structure is used to model situations in which causal sufficiency holds, i.e. where there are no unobserved common causes which could act as confounders. In the presence of confounders, we will instead rely on *ancestral graphs*, which do allow bidirectional links (nonempty $E_{\leftrightarrow}$), but do not contain any almost directed cycles, and additionally require that nodes connected by undirected edges $X_i - X_j$ do not have any parents, and are not connected to further nodes by ↔ edges. The East Asian language network can be considered a causal DAG because all relevant causes of lexical overlap are explicitly modeled as nodes. If we had left out reconstructed languages such as Old Korean and the two proto-languages, these would act as confounders, and the resulting structure could only be an ancestral graph.

Causal inference relies on certain configurations of directed and bidirected edges. Many of these configurations have special and mnemonic names. For instance, a pattern of the form $X \longrightarrow Y \longrightarrow Z$ is called a *chain* (e.g. $OC \longrightarrow MC \longrightarrow cmn$), and a pattern $X \longleftarrow Y \longrightarrow Z$, such as the pattern $kor \longleftarrow MC \longrightarrow jpn$, is called a *fork*. Most importantly, a *collider* on a path is any node where arrow tips meet. In the directed graph case, colliders can only have the form $X \longrightarrow Y \longleftarrow Z$, such as in the pattern $MC \longrightarrow jpn \longleftarrow OJ$. In the more general case of ancestral graphs where bidirectional edges exist, the patterns $X \leftrightarrow Y \leftrightarrow Z, X \leftrightarrow Y \longleftarrow Z$, and $X \longrightarrow Y \leftrightarrow Z$ count as colliders just as well. An unshielded collider or *v-structure* is a collider $X \longrightarrow Y \longleftarrow Z$ where $(X, Z) \notin E$. The five v-structures in the East Asian language network are $OC \longrightarrow OK \longleftarrow PK$, $OK \longrightarrow OJ \longleftarrow PJ$, $OK \longrightarrow OJ \longleftarrow MC$, $OK \longrightarrow kor \longleftarrow MC$, $OK \longrightarrow kor \longleftarrow jpn$. As we shall see throughout this book, the identifiability of v-structures is the cornerstone of constraint-based causal inference, because they are the source of all evidence of directionality.

3.2.1.2 d-Separation

To precisely capture the links between independence constraints and the graph structure, we will need the notion of *d-separation*. Intuitively, two variables are d-separated by a set of conditioning variables if every path by which information might flow from one node to the other through the graph, is blocked in some way by one of the variables we are conditioning on. In a directed graph, ways in which information flow can be blocked are a bit involved, so that a quite technical definition becomes necessary.

In Pearl's definition, a path p in a DAG G is said to be *d-separated* by a set of nodes $\mathbf{Z}$ iff

1. p contains a noncollider, i.e. a chain $i \longrightarrow m \longrightarrow j$ or a fork $i \longleftarrow m \longrightarrow j$, with $m \in \mathbf{Z}$
2. p contains a collider $i \longrightarrow m \longleftarrow j$ such that $m \notin \mathbf{Z}$ and no descendant of m is in $\mathbf{Z}$

A set $\mathbf{Z}$ is said to d-separate X from Y iff $\mathbf{Z}$ d-separates every path from a node in X to a node in Y. Paths and sets of nodes which are not d-separated are also called *d-connected*.

To illustrate, let us first come back to the heating cost example, and consider the paths between household income I and latitude L. We have assumed a path $I \longrightarrow W \longleftarrow O \longleftarrow L$. This path is d-separated by the empty set $\{\}$ because it contains a collider. $\{W\}$ does not d-separate I and O because it contains the collider, but $\{O\}$ (in a chain) does. We have also assumed a second path $I \longrightarrow W \longrightarrow H \longleftarrow O \longleftarrow L$ which is d-separated by the empty set as well, but d-connected by $\{H\}$. The set $\{H\}$ also d-connects the first path because H is a descendant of W. To summarize, both $\{W\}$ and $\{H\}$ d-connect the variables I and L, whereas the set $\{O\}$ d-separates them because it d-separates both paths.

In the East Asian example network, Middle Chinese and Old Korean are d-separated by Old Chinese because every other path between the two languages contains a collider. If we add Japanese to the set Z, however, the two languages become d-connected, because now we have a path $OK \longrightarrow OJ \longleftarrow MC$ with a collider that gets unblocked because one of its descendants is in K. We thus have one d-connected path, which makes the two languages d-connected. To illustrate how this abstract reasoning corresponds to information flow, note that Old Korean and Middle Chinese will seem completely independent if we discard all the shared material from Old Chinese, e.g. by only looking at the lexical innovations in Middle Chinese, and checking whether they are reflected in Old Korean. However, if we additionally consider the loanwords from both languages in Japanese, we will find that if we determined a word in Japanese as not having existed in Middle Chinese, but we are sure that it already existed in Old Chinese, this will allow us to conclude that it must have been borrowed from Old Korean via Old Japanese, even if the word in question is not attested in any of the few sources in Old Korean. Knowledge of Middle Chinese starts to provide us with information about Old Korean, but only because both of these languages left traces in the lexicon of modern Japanese.

For the case where G is an ancestral graph, Richardson & Spirtes (2002) introduce the more general notion of *m-separation*, which is identical to the definition of d-separation except that the more general definitions of collider and noncollider are used, where bidirected arrows are allowed. A maximal ancestral graph (MAG) for a distribution P then is an ancestral graph for P with the additional property that for any pair of non-adjacent nodes there is a set by which they are m-separated. As we shall see, there is a direct correspondence between m-separation in G and conditional independence relationships in P.

3.2.1.3 Faithfulness

To repeat the definition of a Bayesian network, a distribution p fulfills the Markov condition with respect to a DAG G if it factorizes according to the parent relationship defined by G, i.e. if $p(X_1, \dots, X_n) = \prod_{i=1}^{k} q(X_i \mid pa(X_i, G))$. If there is any such DAG, p fulfills the *Causal Markov Condition*, one of the preconditions for constraint-based causal inference.

A distribution p is called *faithful to* a DAG G if the conditional independence relationships which hold in p are exactly the ones implied by the d-separation criterion on G. We call the distribution p as a whole *faithful* if it is faithful to some DAG. This *Causal Faithfulness Condition* is the second precondition for causal inference. Informally, it ensures that there are no spurious independences which occur just because some numbers happen to cancel out perfectly. For instance, for an independence test based on vanishing partial correlation, this implies that the correlation must never become zero for a pair of dependent variables.

In the heating costs example, the previously determined d-separating sets on our paths $I \longrightarrow W \longleftarrow O \longleftarrow L$ and $I \longrightarrow W \longrightarrow H \longleftarrow O \longleftarrow L$ imply that a distribution that is faithful to our scenario should show the conditional independence relationships $(I \perp\!\!\!\perp L)$ and $(I \perp\!\!\!\perp L \mid \{O\})$, but the conditional dependence $(I \not\!\perp\!\!\!\perp L \mid \{H\})$. Conditioning on H is thus predicted to induce a dependency between I and L, which fits with our previous considerations because in a selection of households with identical heating costs, the richer households will tend to cluster in regions with lower latitudes, because the larger window panes of the rich will cause high heating costs even in less severe winters. As the constraints predicted by d-separation say, the dependence should disappear again if we additionally condition on O, because we then look at each region separately.

3.2.1.4 (In)Dependence constraints and graph patterns

Given faithfulness, a collider $A \longrightarrow C \longleftarrow B$ corresponds to the following two conditional (in)dependence constraints: $(A \perp\!\!\!\perp B)$, but $(A \not\!\perp\!\!\!\perp B \mid C)$. We have seen this in the heating costs example, where the true pattern $W \longrightarrow H \longleftarrow R$ was reflected by the observations that $(R \perp\!\!\!\perp W)$, the room temperature was independent of the window size, but $(R \not\!\perp\!\!\!\perp W \mid H)$, not for fixed heating costs.

In contrast, the fork $A \longleftarrow C \longrightarrow B$ as well as the chains $A \longrightarrow C \longrightarrow B$ and $A \longleftarrow C \longleftarrow B$ all correspond to $(A \not\!\perp\!\!\!\perp B)$, but $(A \perp\!\!\!\perp B \mid C)$. To distinguish between these possibilities, we would need additional variables and additional conditional independencies. This shows that conditional independencies alone do not completely determine causal structure. For instance, if we wanted to use conditional independence tests in order to decide which direction of borrowing is responsible for the shared lexical material between English and Japanese, we will not be able to do this based on data from another heavy recipient of English loanwords such as the Dravidian language Telugu (*tel*). We will find that $(tel \not\!\perp\!\!\!\perp jpn)$ due to shared loanwords from English, but that $(tel \perp\!\!\!\perp jpn \mid eng)$, because these loanwords are the only source of lexical overlap between the two languages. In addition to the underlying fork $(tel \longleftarrow eng \longrightarrow jpn)$, this independence pattern could be due to a chain $(tel \longrightarrow eng \longrightarrow jpn)$ or a chain $(jpn \longrightarrow eng \longrightarrow tel)$, as long as we only take data from these three modern languages into account. This ambiguous configuration of (in)dependence constraints will appear very commonly when language isolates are involved.

So how much about the true graph can we determine from (in)dependence constraints? There are several central theorems in the literature which show that the relationship is rather close. Given causal sufficiency, for each faithful and Markovian probability distribution there is a DAG whose d-separation relationships correspond exactly to the conditional independencies in the distribution. Crucially for the inference task, it further turns out that the v-structures in a DAG G alone fully determine the probability distributions that are compatible with G. If two graphs contain the same v-structures, causal inference cannot distinguish them, and they are Markov equivalent. Markov equivalence therefore partitions DAG structures into Markov equivalence classes, the members of which cannot be distinguished by constrained-based causal inference. Each Markov equivalence class can be represented by a *completed partially directed acyclic graph (CPDAG)*, i.e. an acyclic graph where edges may be undirected, representing the fact that some of the Markov equivalent DAGs have an arrow in one direction on this edge, and some others have an arrow in the reverse direction.

In the absence of causal sufficiency, the correspondence between graph structure and independence constraints gets a little less direct. Again moving to the more complex case where latent common causes and selection bias might be present, we find that each Markov equivalence class of MAGs can be represented by a *partial ancestral graph (PAG)*. For an underlying DAG $G = (X \cup L \cup S, E_{\rightarrow})$ over a set of observed variables X, a set of latent variables L, and a set of selection variables S, a PAG which represents G is a graph $G' = (X, E')$ over X with six edge types $\longrightarrow$, $\circ\!\!\longrightarrow$, $\circ\!\!-\!\!\circ$, $\leftrightarrow$, $-$, and $\circ\!\!-$, if for every distribution P that is faithful to G, we have

- $(X_i, X_j) \notin E' \Longrightarrow \exists Y \subseteq X \backslash \{X_i, X_j\} : \ (X_i \perp\!\!\!\perp X_j \mid Y)_P$
- $(X_i, X_j) \in E \Longrightarrow \forall Y \subseteq X \backslash \{X_i, X_j\} : \ (X_i \not\!\perp\!\!\!\perp X_j \mid Y)_P$
- $X_i \longrightarrow X_j$ or $X_i \circ\!\!\longrightarrow X_j$ or $X_i \leftrightarrow X_j \Longrightarrow X_j \notin an(X_i, G')$
- $X_i - X_j$ or $X_i \longleftarrow X_j$ or $X_i \circ\!\!- X_j \Longrightarrow X_j \in an(X_i, G')$

This rather complex definition captures the type of graph structure we can optimally derive to approximate an underlying true causal graph, based only on conditional independence tests for a subset of observed variables. Whereas the definition of the arrow types $\longrightarrow$ and $\leftrightarrow$ is as before, for the equivalence classes we additionally use the end symbol $\circ$ to designate uncertainty, such that $X_i -\!\!\circ X_j$ means "$X_i - X_j$ or $X_i \longrightarrow X_j$", and $X_i \circ\!\!\longrightarrow X_j$ means "$X_i \longrightarrow X_j$ or $X_i \leftrightarrow X_j$" PAGs will be the output structures of the FCI and RFCI algorithms described in §3.2.4, which I will apply to language data in Chapter 7.

To characterize Markov equivalence classes of MAGs, and therefore the structures represented by PAGs, we need the definition of a special kind of path, which will later also play a role in PAG inference. A *discriminating path* for a vertex V is a path $\langle X, \ldots, W, V, Y \rangle$ of at least three edges, where X and Y are non-adjacent, and every vertex between X and V is a collider as well as a parent of Y.

Informally, a discriminating path provides an environment for a node V which allows us to safely identify it as a collider even within a triangle. Spirtes & Richardson (1997) show that two MAGs are Markov equivalent if and only if they have the same undirected link structure and the same v-structures (i.e. are equivalent as CPDAGs), and furthermore have identical colliders among all nodes for which shared discriminating paths exist. Discriminating paths can therefore be seen as providing the environments in which unshielded colliders can safely be established, even in the presence of confounders.

3.2.2 Determining conditional independence relations

As we have seen in the previous section, any causal inference method which builds on inferring a causal graph from constraints will need a reliable way of deciding for any pair of observed variables X and Y whether they are dependent or not given different subsets $\mathbf{Z}$ of all observed variables. The reliability of these conditional independence tests are the main issue for the reliability of causal inference, because given perfect judgments, the theorems give us certainty that we will arrive at an equivalence class of correct structures. Causal inference algorithms mainly differ in how well they can recover from possible wrong conditional independence decisions.

3.2.2.1 Testing for vanishing partial correlation

The most straightforward statistical tests for conditional independence are based on testing for vanishing partial correlation. As Baba et al. (2004) show, this is only guaranteed to work under the assumption that all involved variables are multivariate Gaussian, and does not provide us with a good test for other distributions, including discrete variables.

The Pearson correlation coefficient ρ_{XY} of two variables X and Y is defined as follows:

$$\rho_{XY} := \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} \tag{3.1}$$

It is thus a normalization of the covariance $Cov(X,Y) := E[(X-E[X])(Y-E[Y)]$, which measures whether the two variables tend to deviate from their means in the same directions. We say that X and Y are *correlated* if and only if $\rho_{XY} \neq 0$.

A conditional variant is defined by the *partial correlation* $\rho_{XY\cdot\mathbf{Z}}$, which is defined as the Pearson correlation $\rho_{R_X R_Y}$ of the residuals R_X and R_Y resulting from the linear regression of X and Y with $\mathbf{Z}$. For instance, to compute the residual R_X for a vector of n regression variables $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ from N observations, we need to find the n-dimensional coefficient vector $\mathbf{w}^*_X$ which optimizes the following minimization problem:

$$\mathbf{w}_X^* = \arg, \min_{\mathbf{w}}; \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z_i} \rangle)^2 \right\} \tag{3.2}$$

The observations of the residual R_X are then $x_i - \langle \mathbf{w}_X^*, \mathbf{z_i} \rangle$ for $1 \leq i \leq N$, from which we can compute the Pearson correlation with the analogous residual R_Y.

A more direct alternative is to compute $\rho_{XY \cdot \mathbf{Z}}$ via a recursive formula, which uses several partial correlations of lower order to compute one partial correlation of higher order, with Pearson correlation as the base case. For any $Z_0 \in \mathbf{Z}$, we have:

$$\rho_{XY \cdot \mathbf{Z}} = \frac{\rho_{XY \cdot \mathbf{Z} \setminus \{Z_0\}} - \rho_{XZ_0 \cdot \mathbf{Z} \setminus \{Z_0\}} \rho_{Z_0 Y \cdot \mathbf{Z} \setminus \{Z_0\}}}{\sqrt{1 - \rho^2_{XZ_0 \cdot \mathbf{Z} \setminus \{Z_0\}}} \sqrt{1 - \rho^2_{Z_0 Y \cdot \mathbf{Z} \setminus \{Z_0\}}}} \tag{3.3}$$

To test for vanishing partial correlation in order to establish conditional independence, Spirtes et al. (2000: 5.5) use Fisher's z-transform of the partial correlation $\hat{\rho}_{XY \cdot \mathbf{Z}}$ in the sample:

$$z(\hat{\rho}_{XY \cdot \mathbf{Z}}) := \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{XY \cdot \mathbf{Z}}}{1 - \hat{\rho}_{XY \cdot \mathbf{Z}}} \right) \tag{3.4}$$

If N is the sample size, $\sqrt{N - |\mathbf{Z}| - 3} \cdot z(\hat{\rho}_{XY \cdot \mathbf{Z}})$ roughly approximates a standard normal distribution if the null hypothesis $\hat{\rho}_{XY \cdot \mathbf{Z}} = 0$ holds. To see whether the vanishing correlation assumption can be rejected, we thus test whether we have $\sqrt{N - |\mathbf{Z}| - 3} \cdot |z(\hat{\rho}_{XY \cdot \mathbf{Z}})| > \Phi^{-1}(1 - \frac{\alpha}{2})$ for the cumulative distribution function Φ of the standard normal distribution. This is the default option for conditional independence tests implemented in the R package `pcalg` by Kalisch et al. (2012).

3.2.2.2 Testing for independence in the discrete case

Spirtes et al. (2000: 5.5) also describe standard procedures for conditional independence tests in the discrete case. If we see each cell count x_{ij} in a table resulting from N samples of two variables X_i and X_j as one multinomially distributed variable, the expected value of x_{ij} under the independence assumption is $E(x_{ij}) = \frac{\sum_j x_{ij} \cdot \sum_i x_{ij}}{N}$. If we add a third variable X_k for which $(X_i \perp\!\!\!\perp X_j \mid X_k)$, the corresponding cell x_{ijk} will have the expected value $E(x_{ijk}) = \frac{\sum_j x_{ijk} \cdot \sum_i x_{ijk}}{\sum_{i,j} x_{ijk}}$.

Analogously, for n conditioning variables $X_{k_1}, \ldots, X_{k_n}$, we get

$$E(x_{ijk_1 \ldots k_n}) = \frac{\sum_j x_{ijk_1 \ldots k_n} \cdot \sum_i x_{ijk_1 \ldots k_n}}{\sum_{i,j} x_{ijk_1 \ldots k_n}} \tag{3.5}$$

These expected cell counts can be tested against the observed values using standard tests. Under the independence assumption, the following two test statistics are both χ^2-distributed for an appropriate number of degrees of freedom df:

$$\chi^2 := \sum_{i,j,k_1,\ldots,k_n} \frac{(x_{ijk_1 \ldots k_n} - E(x_{ijk_1 \ldots k_n}))^2}{E(x_{ijk_1 \ldots k_n})} \tag{3.6}$$

$$G^2 := 2 \cdot \sum_{i,j,k_1,\ldots,k_n} x_{ijk_1 \ldots k_n} \ln\left(\frac{x_{ijk_1 \ldots k_n}}{E(x_{ijk_1 \ldots k_n})}\right) \tag{3.7}$$

In principle, the degrees of freedom for a test of the conditional independence $(X_i \perp\!\!\!\perp X_j \,|X_{k_1}, \ldots, X_{k_n})$ can be computed from the number of categories Cat for each variable as follows:

$$df = (Cat(X_i) - 1) \cdot (Cat(X_j) - 1) \cdot \prod_{i=1}^{n} Cat(X_{k_i}) \tag{3.8}$$

This number is exponential in the number of conditioning variables, which will quickly lead to zero entries in the table that need to be corrected for. In the absence of a general rule, df can be reduced by one for each zero entry as a rough heuristic.

3.2.2.3 Testing for vanishing conditional mutual information

A more general criterion for conditional independence stems from information theory, a branch of mathematics that is concerned with quantifying information and information flow. I will repeat the essential concepts of information theory here to provide some degree of self-containedness. The first chapters of any introductory textbook of information theory will introduce the same concepts with a lot more rigour and detail, also motivating the theory using a wealth of examples. For my purposes in this book, it satisfies to say that information theory will provide us with the mathematical tools for modeling languages as variables, and for defining the conditional independence tests that we will need to apply causal inference. The definitions as I state them in the following are adapted from Cover & Thomas (2006).

The central concept of information theory is called *entropy*, and can be seen as a measure of the expected amount of information provided by a single outcome of (or alternatively, the uncertainty contained in not knowing the outcome of) a random variable X. In the discrete case (to which we will confine ourselves here), the entropy $H(X)$ of a discrete variable X with the set of possible outcomes $\Omega(X)$ is defined as

$$H(X) := -\sum_{x \in \Omega(X)} p(x) \log p(x) \tag{3.9}$$

The entropy of a discrete variable X can be seen as the average information content of a single observation of that variable, or as the expected value of the *self-information* $I(x) := -\log p(x)$ associated with an event $\{X = x\}$. Self-information is also called *surprisal* because it measures the unexpectedness (or amount of surprise) associated with the observation that X has the value x.

If we observe two information sources X and Y at the same time, some of the information we receive might coincide, which means that we cannot just add up the amount of information received by both sources to quantify our overall information. Instead, we generalize entropy to *joint entropy* $H(X, Y)$:

$$H(X, Y) = -\sum_{x}\sum_{y} p(x, y) \log p(x, y) \tag{3.10}$$

The amount of information we receive twice when jointly observing two information sources can be recast as the information that one variable Y provides about the state of the other variable X. This symmetric measure is called the *mutual information* $I(X; Y)$:

$$I(X; Y) := H(X) + H(Y) - H(X, Y) = \sum_{y}\sum_{x} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{3.11}$$

Returning to individual events, we can define mutual information as the expected value of the *pointwise mutual information* $pmi(x; y)$ between two observations:

$$pmi(x; y) := \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{3.12}$$

Pointwise mutual information is very useful for quantifying the strength of associations between pairs of variable values. In §4.4, I will get back to PMI as a standard measure of association between sounds in models of sound correspondences.

Finally, we can measure the information which two variables X and Y provide about each other provided that the values of a set of certain other variables $\mathbf{Z}$ is known. This is called the *conditional mutual information* of X and Y given $\mathbf{Z}$:

$$I(X;Y|\mathbf{Z}) := \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} \sum_{x} p(x,y|\mathbf{z}) \log \left(\frac{p(x,y|\mathbf{z})}{p(x|\mathbf{z})p(y|\mathbf{z})} \right) \tag{3.13}$$

As Yeung (2008) demonstrates, it is easy to derive the following formula for computing conditional mutual information from joint entropies:

$$I(X;Y|\mathbf{Z}) = H(X,\mathbf{Z}) + H(Y,\mathbf{Z}) - H(X,Y,\mathbf{Z}) - H(\mathbf{Z}) \tag{3.14}$$

The decisive property of mutual information for causal inference is that for joint distributions that are faithful to some causal graph, it provides us with a necessary and sufficient criterion for independence:

$$X \perp\!\!\!\perp Y \Longleftrightarrow I(X;Y) = 0. \tag{3.15}$$

More importantly for my application, this also extends to conditional mutual information, giving us the following characterization:

$$(X \perp\!\!\!\perp Y \mid Z) \Longleftrightarrow I(X;Y|Z) = 0. \tag{3.16}$$

Intuitively, this means that two sets of variables are independent given a third set of variables if and only if there is no information flow between the first two sets that could not be mediated by variables from the third set.

Given this equivalence, an obvious idea for implementing a very general independence test now is to check for vanishing mutual information. The problem with mutual information is, however, that it is hard to compute or estimate for any interesting type of variable. This means that to exploit this characterization of independence, we need to rely on other more easily computable measures which in all relevant respects behave just like joint entropy.

Assume we have a set of n discrete random variables $X_1, \ldots, X_n$ with the index set $[n] := \{1, \ldots, n\}$. Then, the criteria for a real-valued function h on subsets of $[n]$ to behave sufficiently like the joint entropy H can be cast into three axioms that are known as the *elemental inequalities*, and are quoted here as in Chaves et al. (2014):

For all $S \subset [n] \backslash \{i, j\}, i \neq j, \ i, j \in [n]$:

- $h([n] \backslash \{i\}) \leq h([n])$ (monotonicity)
- $h(S) + h(S \cup \{i, j\}) \leq h(S \cup \{i\}) + h(S \cup \{j\})$ (sub-modularity)
- $h(\emptyset) = 0$

Intuitively, the monotonicity condition ensures that uncertainty never becomes smaller if we consider a larger set of variables, and the sub-modularity condition ensures that the conditional mutual information derived from the entropy-like measure is always positive.

Together, these inequalities define an outer approximation to the region of vectors in the space of set functions R_n which define some entropy function on all subsets of a set of n discrete random variables. In less technical terms, this means that any set function for which the elemental inequalities hold is close enough in behaviour to entropy that we can use it to derive a consistent measure of conditional mutual information. For more background on this, the reader is referred to Yeung (2008: Ch. 14).

As we will see in Chapter 6, it is relatively straightforward to define measures for which the elemental inequalities hold. Based on a function which can in this regard be seen as a measure of entropy, this will make it possible to establish consistent (if unreliable) independence tests between sets of languages.

3.2.3 The PC algorithm

The first feasible and complete causal inference algorithm based on conditional independence tests is the PC algorithm as presented by Spirtes et al. (2000). This algorithm is a basic building block for many more recent approaches, and is the cornerstone for any understanding of constraint-based causal inference.

3.2.3.1 Preconditions and assumptions

The correctness and completeness of the PC algorithm depends on two very natural conditions, which are however only rarely met in applications to practical problems, and have therefore been weakened for later algorithms which build on the same basic principles.

The first prerequisite is the already mentioned causal sufficiency, i.e. we must assume that there are no unobserved common causes which act as confounders. If we try to circumvent this by operating with DAGs in which some nodes are unobserved, we quickly run into the problem that there are many DAGs over both observed and latent variables for which no equivalent DAG over only the observed variables exist, i.e. the Markov condition breaks down. The minimal example of this is a causal graph of the shape $X_1 \longleftarrow L_1 \longrightarrow X_2 \longleftarrow L_2 \longrightarrow X_3$, where L_1 and L_2 are unobserved. It is not possible to provide a DAG structure over X_1, X_2, and X_3 which corresponds by d-separation to the conditional independence constraints encoded in the larger structure. This will turn out to be a problem in my application, because this pattern is one of the most frequent among languages whenever two language families come into contact, e.g. in the configuration $deu \longleftarrow Germanic \longrightarrow fin \longleftarrow Uralic \longrightarrow yrk$, where Finnish (*fin*) is a Uralic language which was influenced by Germanic, unlike Eastern Uralic languages like Nenets (*yrk*).

The second prerequisite is faithfulness, which we already encountered in the context of defining the correspondence between independence constraints and the graphs they characterize. If the independence tests are too unreliable, and produce a pattern of independence constraints that violates faithfulness, this can be expected to mislead the algorithm, possibly to the point where it derives contradictory constraints which do not correspond to any causal structure.

3.2.3.2 Basic version

The basic architecture of the PC algorithm consists of three phases. In the first phase (Stage I), conditional independence tests are systematically performed to establish an undirected *causal skeleton*. For each pair of variables A and B, we search for a minimal *separating set* S_{AB} with $(A \perp\!\!\!\perp B \mid S_{AB})$ (minimal in the sense that $(A \not\perp\!\!\!\perp B \mid S)$ for any $S \subset S_{AB}$). By doing so for every pair of variables, we construct an undirected graph G with $\{A, B\} \in E$ whenever no such S_{AB} could be found. The one observation which makes the PC algorithm tractable even for dozens of variables is that we do not need to test all possible separating sets S_{AB} when attempting to separate A and B. Instead, we can look for separating sets by increasing size, first removing all links between variables which are unconditionally independent from an initially complete graph, then all links which are independent given separating sets S_{AB} of size 1, and so on.

In addition, two nodes X_i and X_j are d-separated in a DAG G if and only if they are d-separated by either $pa(X_i, G)$ or $pa(X_j, G)$. Therefore, it suffices to check whether two variables are independent given their neighbors in order to check whether they are conditionally independent given any set of variables. As the graph gets sparser by the removal of links, so does the number of neighbors of A and B which the separating set S_{AB} must consist of, which tends to make checking for all possible S_{AB} tractable even for large set sizes.

In the next phase (Stage II), we check for the presence of v-structures. For each triple $A, B, C \in V$ with $\{A, B\} \notin E$, but $\{A, C\}, \{B, C\} \in E$ (i.e. each *unshielded triple*), we add arrowheads pointing at C (leading to a v-structure $A \longrightarrow C \longleftarrow B$) if $C \notin S_{AB}$. This inference is justified by the relationship between graph patterns and (in)dependence constraints established above. There is a v-structure $A \longrightarrow C \longleftarrow B$ if and only if we have $(A \perp\!\!\!\perp B)$, but $(A \not\!\perp\!\!\!\perp B \mid C)$. $(A \perp\!\!\!\perp B)$ is given because we have an unshielded triple, without a direct causal connection between A and B. The absence of C in the separating set S_{AB} implies that $(A \not\!\perp\!\!\!\perp B \mid C)$, because otherwise we would have found $S'_{AB} = \{C\}$ as a separating set before encountering S_{AB}. We infer the existence of a collider $A \longrightarrow C \longleftarrow B$ because all non-collider configurations would have led to $(A \perp\!\!\!\perp B \mid C)$.

If our independence tests were completely reliable (and this is what the basic PC algorithm assumes), we can be sure that we have found exactly the correct v-structures at Stage II. This allows us to orient many of the remaining undirected edges (which were not part of any v-structure) in a third phase (Stage III), by repeatedly applying two simple criteria until no additional arrows can be inferred. If one direction of a link would lead to a new v-structure which was not detected at Stage II, we can add the arrow in the reverse direction (leading to a chain). Moreover, if we assume acyclicity (as implied by the Causal Markov condition), we can also orient any arrow where the reverse orientation would result in a cycle $A \longrightarrow B_1 \longrightarrow B_2 \longrightarrow \cdots \longrightarrow A$. Enforcing the two principles can be achieved by applying four arrow propagation rules, which I will not cite here because they are subsumed by the ruleset of the more complex FCI algorithm in §3.2.4. Thanks to a proof by Meek (1995), it has long been known that application of these four rules until the fixed point (until they do not apply any more) suffices to arrive at the CPDAG representing the Markov equivalence class of the true graph, i.e. containing all the arrows which are common to the causal graphs in the class. This even holds if we add background knowledge in the form of pre-directed arrows, as one would do if e.g. the temporal order makes the directionality of causation obvious.

The result computed by the PC algorithm is the minimal graph structure whose d-separation relationships correspond exactly to the conditional independencies in the distribution, i.e. it will contain all those links which we must assume in order to explain the conditional independence properties of the data, and not a single additional link which would not be necessary. A procedure like the PC algorithm therefore implements Occam's razor, the scientific principle which requires us to pick among adequate explanations the one for which we need to make the smallest number of assumptions. The central role of Occam's razor is

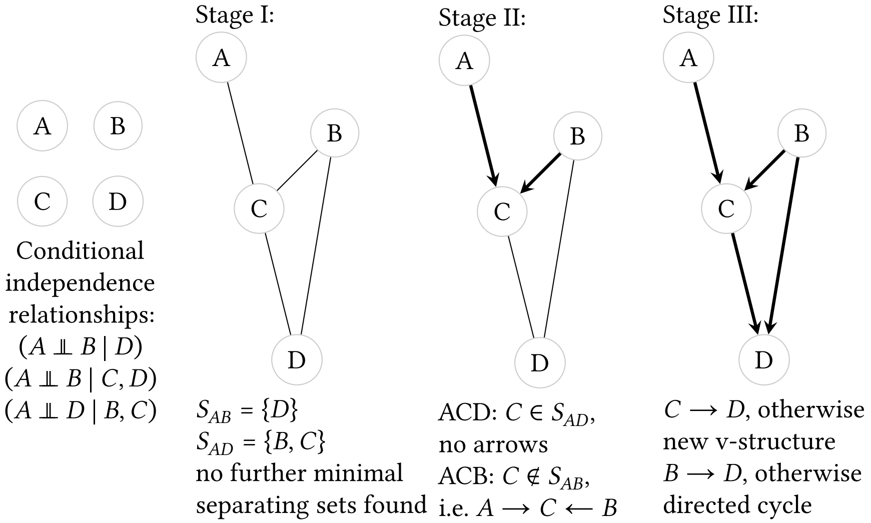
Stage I:
Stage II:
Stage III:
A
B
C
D
Conditional independence relationships:
$(A \perp\!\!\!\perp B \mid D)$
$(A \perp\!\!\!\perp B \mid C, D)$
$(A \perp\!\!\!\perp D \mid B, C)$
$S_{AB} = \{D\}$
$S_{AD} = \{B, C\}$
no further minimal separating sets found
ACD: $C \in S_{AD}$, no arrows
ACB: $C \notin S_{AB}$, i.e. $A \longrightarrow C \longleftarrow B$
$C \longrightarrow D$, otherwise new v-structure
$B \longrightarrow D$, otherwise directed cycle

Figure 3.2: Illustrating the phases of the basic PC algorithm

reflected by the importance of avoiding overfitting in machine learning, where it is typically possible to perfectly model the observable data using a complex model with many parameters (e.g. a fully connected graph), but due to the more difficult inference task, complex models are less likely to classify unseen examples well. It therefore makes sense for inference algorithms to infer the least complex model which still fits the observable data, and this is what the PC algorithm does for joint distributions of causally sufficient sets of variables.

3.2.3.3 More recent variants

The vanilla PC algorithm as I just presented it takes an unnecessary risk in selecting the neighbors out of which separating set candidates are formed. If the distribution is faithful to some DAG, $(A \perp\!\!\!\perp B \mid pa(A))$ implies that A and B are independent given a set of nodes lying on undirected paths between A and B. Conditioning on variables that are not on any connecting path will not cause any blockage of information flow. Therefore, it suffices to include only nodes on connecting paths in separating set candidates. Spirtes et al. (2000: 5.4.2.3) call this modification the *PC* algorithm*, but advise shifting to the more exact separating set candidate selection criterion only after the graph was already thinned out, due to the high memory overload involved in maintaining a list of all connecting paths between any pair of nodes. I will be using a similar idea in Chapter 6, where the shape of my data allows me to adapt an explicit flow criterion for prefiltering the possible separating sets.

A major problem of the vanilla PC algorithm as well as its PC* variant is that the results can vary widely depending on the order in which separating sets are tried out, because the first one will be picked even though there might be many separating sets of the same minimal size. The *Conservative PC* variant by Ramsey et al. (2006) differs in not stopping as soon as a single sepset was found, but checking whether the middle variable is present in all or no separating sets of the current size. Unshielded triples where the relevant variable is contained in some, but not all separating sets, are not oriented as colliders, but marked as ambiguous, and prevented from taking part in the propagation rules. This rule will prevent uncertain directionality information from being propagated, but will often leave many edges unoriented.

The *Stable PC* algorithm by Colombo & Maathuis (2014) uses a majority rule to resolve this problem. This variant decides whether to orient each triple as a collider by counting the ratio of all minimal separating sets which contain B. Both conservativity and the majority rule remove the order-dependence in the presence of conflicting information, but the PC variants defined in this way still tend to yield either unstable or uninformative results, and must be re-run with different thresholds to detect the stable links.

3.2.4 The FCI algorithm

The FCI algorithm can be seen as a generalization of the PC algorithm to the situation where hidden common causes for some of the observed variables might exist. Such connections will be represented by the bidirected arrows (the edge set $E_{\leftrightarrow}$) in ancestral graphs.

3.2.4.1 Basic version

The original version of the *FCI (Fast Causal Inference)* algorithm was given by Spirtes et al. (2000) as a variant of the PC algorithm in the absence of causal sufficiency. Much as the PC algorithm generates a CPDAG with the goal of approximating the underlying true DAG up to Markov equivalence, the FCI algorithm generates a PAG to represent the Markov equivalence class of the underlying true ancestral graph.

The basic procedure of FCI remains to systematically find separating sets for pairs of observed variables, thinning out a fully connected initial graph until we arrive at a skeleton which only connects pairs of variables that cannot be made independent by conditioning on any combination of the other variables. What makes the FCI algorithm much more complicated than the PC algorithm is that due to the possible existence of latent variables, we cannot assume that we can form a separating set for all pairs of m-separated variables from the neighbors in the current skeleton. To see this, consider the following minimal example taken from Spirtes et al. (2000: p.129). We have two pairs of variables A, B and D, E, each of which are dependent due to a hidden common cause. The only direct causal influence on A among observed variables is $D \longrightarrow A$, and the only influence on E is $B \longrightarrow E$. Assume further that neither of the variables A, B, D, E has a direct causal influence on any other variable, observed or unobserved, and that there is an additional causal pattern $B \longleftarrow F \longleftarrow C \longrightarrow H \longrightarrow D$. This pattern induces a dependence $(A \not\perp\!\!\!\perp E)$, but the independence $(A \perp\!\!\!\perp E \mid \{B, C, D\})$. There is a separating set which would allow us to delete the link, but this sets includes a variable which is not directly adjacent to either A and E in the true graph.

The question which sets we need to test in order to ensure that a pair of variables X and Y is not m-separated by any subset of the observed variables gives rise to the notion of an *inducing path*. For two disjoint sets L (latent variables) and S (selection variables) of nodes in an ancestral graph which do not contain X or Y, an inducing path relative to $\langle L, S\rangle$ is a path between X and Y where every intermediate node is either a collider or in L, and every collider on the path is either in S or an ancestor of X or Y. A crucial result by Richardson & Spirtes (2002) shows that X and Y need to be connected in the true ancestral graph (are not m-separated by $Z \cup S$ for any Z disjoint from L and S) if and only if there is an inducing path from X to Y relative to $\langle L, S\rangle$.

For the systematic independence tests that are performed to arrive at a skeleton, the original version of the FCI algorithm relied on the notion of an inducing path graph, which, however, turned out (Zhang 2006) to be less informative than the variant based on ancestral graphs, which is the only one that will be presented here.

After generating an initial skeleton just as in the first phase of the PC algorithm, we cannot yet be sure that all the pairs of still connected variables are actually m-connected in the true ancestral graph, because looking for separating sets only among the neighbors was sufficient to determine d-separation, but does not reliably check for m-separation. Some additional edges might have to be removed, and this is where the inducing path criterion comes in. If we partially orient the links in the initial skeleton by detecting v-structures, many of the paths in the skeleton between each pair of nodes X and Y cannot correspond to inducing paths in the underlying ancestral graph, whereas other might. It therefore suffices to check whether $(X \perp\!\!\!\perp Y \mid Z)$ for each combination Z of nodes Z_k connected to X by what might still represent an inducing path. If we manage to break every possible inducing path between the observed variables X and Y in this way, the inducing path criterion allows us to remove the link between X and Y. In practice, the criterion used to find candidates Z_k in the absence of latent variables checks whether each triple on the path forms either an unshielded collider or a triangle, which is the observable equivalent of an underlying inducing path.

In order to infer the directionality of links in the final skeleton, FCI again relies on the same basic procedure as the PC algorithm, after discarding the directionality information which was used to determine the final skeleton. After redetermining the v-structures as starting points, propagation rules are repeatedly applied, adding partial orientations to additional edges until no further changes occur. The four propagation rules given by Spirtes et al. (2000) have the advantage of still being quite intuitive, but this first version of FCI did not aim to achieve completeness in the sense that it did not necessarily output the most specific maximal ancestral graphs.

Zhang (2008) closes this gap by developing and proving the completeness of a rather complex set of orientation rules, giving rise to the Augmented FCI (AFCI) algorithm. Since this is the version of the rules which is used in my implementation of RFCI (see below), I will provide each rule here, and give an informal explanation of the intuition behind each of them, as well as their status in the overall inference system. For a compact notation of the conditions under which the rules apply, a star is used as an additional wildcard symbol to represent any arrow end state. This is different from the circle in that the circle represents a concrete state with the potential of being turned into an arrow or a line, whereas the star does not correspond to an actual state, and is only used to keep rule notations compact by matching any possible end symbol.

The first four rules ensure arrowhead completeness, i.e. they detect any arrowhead that is present in all members of the Markov equivalence class, based on the assumption that the inferred v-structures are correct. These rules are quite similar to the orientation rules used by the PC algorithm, with one additional rule that looks for discriminating paths which help to distinguish the configurations $A \leftrightarrow B \leftrightarrow C$ and $A \leftrightarrow B \longrightarrow C$ in some cases:

- $\mathscr{R}1$: orient unshielded $A \ast\!\!\longrightarrow B \circ\!\!-\!\!\ast\, C$ as $A \ast\!\!\longrightarrow B \longrightarrow C$
- $\mathscr{R}2$: orient $A \ast\!\!-\!\!\circ\, C$ as $A \ast\!\!\longrightarrow C$ if $A \longrightarrow B \ast\!\!\longrightarrow C$ or $A \ast\!\!\longrightarrow B \longrightarrow C$
- $\mathscr{R}3$: orient $D \ast\!\!-\!\!\circ\, B$ as $D \ast\!\!\longrightarrow B$ if there is a pair of variables A and C with

$(A, C) \notin E$ which is in configurations $A \ast\!\!\longrightarrow B \longleftarrow\!\!\circ\, C$ and $A \ast\!\!-\!\!\circ\, D \circ\!\!\longrightarrow C$

- $\mathscr{R}4$: on a discriminating path $\langle D, \dots, A, B, C\rangle$, orient $B \ast\!\!-\!\!\circ\, C$ as $B \longrightarrow C$ if B is in the separating set found for D and C, and add arrows to form $A \leftrightarrow B \leftrightarrow C$ otherwise

Intuitively, the first rule $\mathscr{R}1$ exploits the assumption that in the previous step of the algorithm, we have found exactly the v-structures which are present in the true ancestral graph. This means that we can exclude any arrow that would lead to an additional collider, giving us additional chains. $\mathscr{R}2$ enforces the absence of almost directed cycles in the ancestral graph.

$\mathscr{R}3$ provides a way to infer additional arrows within shielded triples. If in the configuration it acts upon, we added an arrow from B to D, the second rule would force us to assume $A \ast\!\!\longrightarrow D$ in order to avoid a cycle; in the unshielded triple $A \ast\!\!\longrightarrow D \circ\!\!-\!\!\ast\, C$, the requirement not to introduce additional v-structures would force an additional arrow $D \longrightarrow C$, leading to an (almost) directed cycle $B \longrightarrow D \longrightarrow C \ast\!\!\longrightarrow B$, which cannot exist. Therefore, the arrow $D \ast\!\!\longrightarrow B$ is the only option in this configuration. This inference by contradiction cannot be emulated by propagation rules.

The intuition behind $\mathscr{R}4$ is that discriminating paths show some of the behavior of unshielded triples, because on a discriminating path from X to Y, the colliders are exactly the nodes which do not occur in any m-separating set for X and Y, and the non-colliders are the nodes which occur in every such set. This property allows additional inferences of directionality, much in the same vein as the initial detection of v-structures, but in the presence of bidirectional arcs.

The second block of rules serves to infer the existence of line ends, i.e. they are the overall system's way of detecting selection bias (undirected edges). If we can safely assume that no selection bias is present, these rules will never apply, and can thus safely be ignored:

- $\mathscr{R}5$: orient $A \circ\!\!-\!\!\circ\, B$ and all edges on an uncovered circle path $\langle A, C, \dots, D, B\rangle$ where $(A, D) \notin E$ as well as $(B, C) \notin E$ as undirected $(-)$, if such a path exists

- $\mathscr{R}6$: orient $B \circ\!\!-\!\!\ast\, C$ as $B -\!\!\ast\, C$ if there is an A with $A - B$

- $\mathscr{R}7$: orient $B \circ\!\!-\!\!\ast\, C$ as $B -\!\!\ast\, C$ if there is an A with $A -\!\!\circ\, B$, and $(A, C) \notin E$

A path is called *uncovered* if every subsequence of length 3 on it forms an unshielded triple, i.e. every node is fixed as being either a collider or a non-collider. $\mathscr{R}5$ looks for cycles consisting of unshielded triples, none of which was found to be a v-structure. If we started adding arrows in either direction, $\mathscr{R}1$ would force us to continue adding arrows in the same direction until we arrive at a directed cycle, violating the ancestral graph conditions. Therefore, the only option in such a configuration are undirected links along the entire cycle. $\mathscr{R}6$ directly enforces the ancestral graph property that no arrowhead may point into an undirected edge. The purpose of $\mathscr{R}7$ is similar to $\mathscr{R}3$ in that it covers a reasoning pattern by contradiction that could not be covered by greedy propagation. The reasoning is as follows: If contrary to the rule we assumed $B \leftarrow\!\!\circ\, C$, this would lead to a new v-structure unless $A - B$, in which case we would again violate the ancestral graph conditions.

The third block of rules allows us to turn many partially directed edges $\circ\!\!\rightarrow$ into directed ones, and are therefore essential for the algorithm's ability to distinguish bidirected from directed arcs in the ancestral graph. Two of these rules rely on finding paths that are potentially directed, i.e. contain only links of the shapes $\rightarrow$, $\circ\!\!\rightarrow$, and $\circ\!\!-\!\!\circ$ (with arrows in the direction of the path):

- $\mathscr{R}8$: orient $A \circ\!\!\rightarrow C$ as $A \rightarrow C$ if $A \rightarrow B \rightarrow C$ or $A -\!\!\circ\, B \rightarrow C$
- $\mathscr{R}9$: orient $A \circ\!\!\rightarrow C$ as $A \rightarrow C$ if there is an uncovered potentially directed path from A to C whose second element $B \neq C$ is not adjacent to C
- $\mathscr{R}10$: orient $A \circ\!\!\rightarrow C$ as $A \rightarrow C$ if there is a pattern $B \rightarrow C \leftarrow D$, and two uncovered potentially directed paths from A to B and from A to D, the second elements of which (possibly B or D) do not coincide, and are not adjacent

The rule $\mathscr{R}8$ is again a rule which enforces the non-existence of directed cycles, exploiting the additional conditions imposed on a mixed graph. More specifically, this rule prevents the situation where an arrowhead points into an undirected edge, a condition which was not enforced by $\mathscr{R}2$. The role of $\mathscr{R}9$ is very much analogous to $\mathscr{R}5$, in that it looks for and prevents configurations which would propagate into an almost directed cycle. $\mathscr{R}10$ encodes another instance of reasoning by contradiction. We know that the two potentially directed paths leading away from A form an unshielded non-collider pattern in A (as a collider would have been detected earlier), which implies that the edge from A into at least one of the paths is directed. If we had $A \leftrightarrow C$ instead of $A \rightarrow C$, this initial directed link would propagate by $\mathscr{R}1$ along the entire path, leading to an almost directed cycle via $B \rightarrow C \leftarrow D$ and $A \rightarrow C$.

While the soundness of these propagation rules is comparatively easy to see given the explanations, their joint completeness in the sense that iterative application of these rules will lead to a maximally informative partial ancestral graph is highly non-trivial to prove, and requires large amounts of additional formal machinery. For details on these matters, the reader is referred to the theorems in, and especially the proofs in the appendix of, Zhang (2008).

3.2.4.2 More recent variants

The RFCI (Really Fast Causal Inference) algorithm by Colombo et al. (2012) reconsiders the necessity of the large number of conditional independence tests which typically need to be performed by the FCI algorithm, and manages to reduce the number and order of conditional independence tests by exploiting some additional properties of ancestral graphs. These changes make causal inference without causal sufficiency feasible for dozens of variables, and also make it more stable for small sample sizes, because tests of lower order have more statistical power.

Where FCI tested all subsets of a set of possible m-separators to arrive at the final skeleton, often leading to a combinatorial explosion of tests which needed to be performed especially in sparse graphs, RFCI confines itself to testing only very few sets beyond immediate neighbors, motivated by some important results. For ease of exposition, I will ignore the existence of a set of selection variables S in the original statements, because I will always have $S = \{\}$, equivalent to absence of selecton bias, in my application.

As the first important result, the unshielded triple rule states that a minimal separating set Z for X_i and X_k contains exactly those ancestors X_j of X_i or X_k where both pairs X_i and X_j and X_j and X_k remain dependent given $Z \backslash \{X_j\}$. The RFCI algorithm exploits this by checking all unshielded triples $\langle X_i, X_j, X_k \rangle$ for violations of this pattern, which are then repaired by finding a new minimal separation set for the link ($X_i - X_j$ or $X_j - X_k$) that was found to be inadequate, and removing the offending link, possibly removing or creating new unshielded triples.

Secondly, the discriminating path rule states that if a path $\langle X_i, \ldots, X_l, X_j, X_k \rangle$ is a discriminating path, and no pair of successive vertices on the path can be made independent by conditioning on any subset of the separating set $S_{X_i X_k}$, then if $X_j \in S_{X_i X_k}$, it is an ancestor and not a descendant of X_k, and otherwise it is an ancestor of neither X_l or X_k, nor a descendant of X_k. This fact is used by the RFCI algorithm on triangles of the form $X_k \leftarrow X_l \leftarrow\!\circ\, X_j \circ\!\!\rightarrow\!\!* \, X_k$, where on a minimal discriminating path $\langle X_i, \ldots, X_l, X_j, X_k \rangle$, any edge between pairs of successive vertices violating the rule can be removed. These two rules replace the more straightforward check against all sets of vertices reachable by inducing paths given by the standard FCI algorithm, and manage to lead to many of the edge deletions performed by the refinement stage.

A slight decline in output informativity does persist, however, and can be expressed most concisely as a difference in conditions fulfilled by the PAGs returned by FCI and the ones returned by RFCI. In both variants, absence of an arc between X_1 and X_2 implies the existence of some separating set Y such that $(X_i \perp\!\!\!\perp X_j \mid Y)$, an arrowhead at X_2 expresses that X_2 is not an ancestor to X_1 in any MAG of the equivalence class, and a tail at X_2 that X_2 is an ancestor to X_1 in every such MAG. The difference is in the interpretation of edge existence. In the output of FCI, the existence of an edge between X_1 and X_2 implies that not a single combination of other nodes in the graph constitutes a separating set for X_1 and X_2, whereas in the output of RFCI, this guarantee only extends to separating sets built from adjacents of one of the two nodes. An RFCI-PAG might therefore have some spurious additional edges in comparison to the FCI-PAG, which means that it must be interpreted in a more cautious way.

The resulting PAG is less informative than the output of FCI in some situations, but as Colombo et al. (2012) show, all the causal information it returns is asymptotically correct (i.e. guaranteed to become correct given sufficient amounts of data), and the output provably coincides with the output of FCI on a large class of ancestral graphs. The definition of this class mirrors the difference in edge semantics. The only situation where the RFCI-PAG can have an edge $X_i \ast\!\!-\!\!\ast X_j$ in addition to the FCI-PAG is when there is an inducing path from X_i to X_j relative to the remaining adjacents of X_i as well as another inducing path from X_j to X_i relative to the remaining adjacents of X_j in the initial skeleton, but either there is no inducing path from X_i to X_j relative to the possible d-separation nodes for X_i, or no inducing path from X_j to X_i relative to the possible d-separation nodes for X_j in the refined skeleton of FCI. What this rather involved condition boils down to is that the superfluous edges can only occur between variables that are not connected by ancestry, and will not have line ends in the output of RFCI (i.e. they will not be fully directed).

Since the RFCI algorithm appears to be alone in being able to infer PAGs for dozens of variables without running into severe combinatorial problems, it is the only existing algorithm which is directly applicable to the problem of contact lexical flow inference. I will therefore use my own Java implementation of RFCI to represent the state of the art in causal inference in the absence of causal sufficiency.

3.2.5 Alternative algorithms

While the PC algorithm and its derivatives are mathematically well-motivated and rest on firm theoretical grounds, in practice they suffer from severe error propagation issues if but a single conditional independence test yields an incorrect result. They are thus very unstable, a problem which can become so severe that it is typically necessary to re-run these algorithms with different threshold values for the conditional independence tests, on different variable orderings, and under addition of some random noise, then aggregating the results of the runs into a more stable picture.

However, as we will see in this section, there are good reasons to still focus on constrained-based causal inference for lexical flow inference. To start with, most algorithms in the vast landscape of existing approaches only work on continuous variables, or even only on variables that can be assumed to be normally distributed. Limiting the discussion to approaches which might be relevant for my application, I will only discuss two methods which could in principle be applied to the discrete case, but which I am not exploring any further in this book.

The most important alternative to the constraint-based paradigm can be found in the score-based approaches, which directly model the fit of candidate graphs G to a data representation D as an optimization problem for a score which can be chosen to favor minimal graphs. Using a hill climber or a more advanced general optimization algorithm, a candidate graph G is iteratively modified slightly to test whether the score is improving, until a local maximum is reached. In addition to the advantage of not making any categorically wrong decisions, scores make it possible to quantify the certainty about the result graph. The most popular score-based approach is Greedy Equivalence Search (GES) as described by Chickering (2002). The first phase of GES starts with the empty graph and greedily adds the edges which improve the score most, until a maximum is reached. In the second phase, some edges are removed again as long as this further improves the score. The main disadvantage of score-based methods is the requirement of causal sufficiency, making it impossible to treat hidden common causes correctly. Also, the huge search space tends to make these methods intractable if a structure is to be built over more than a handful of variables.

A different class of causal inference algorithms proceeds by modeling some of the imprecision in the conditional independence judgments based on Bayesian principles. These algorithms are much more stable and less error-prone in practice, but do not have the advantage of theoretical guarantees such as proofs of completeness. Recently, Claassen & Heskes (2012) have proposed to combine both paradigms in order to arrive at a both theoretically sound and computationally stable inference procedure. In their BCCD algorithm, a Bayesian score is assigned to each input statement, quantifying the reliability of each piece of knowledge. By processing the constraints in decreasing order of reliability, the PC algorithm can be guided in such a way that errors due to bad conditional independence judgments happen late during the execution, therefore only causing local errors which are not propagated much further. While the algorithm compares favourably with FCI when unfaithful DAG approximations to MAGs are inferred, the evaluation on test sets spanning only five variables indicates that this approach will not scale to many variables either.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Name index

Language index

Subject index

Information-theoretic causal inference of lexical flow

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

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

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

ISBN 978-3-96110-143-6

9 783961 101436