Information-theoretic causal inference of lexical flow

Johannes Dellert

Language Variation 4

language science press

Language Variation

Editors: John Nerbonne, Dirk Geeraerts

In this series:

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

ISSN: 2366-7818

Information-theoretic causal inference of lexical flow

Johannes Dellert

language
science
press

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

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

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

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

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

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

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Preface v

Acknowledgments ix

1 Introduction 1

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

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

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

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

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

5 Simulating cognate histories **171**
5.1 Simulation and in-silico evaluation . . . 171
5.1.1 Advantages and shortcomings of simulation . . . 171
5.1.2 Principles of in-silico evaluation . . . 173
5.2 Generating phylogenies . . . 174
5.2.1 Models of lexical replacement . . . 175
5.2.2 Simulating how languages split and die . . . 176
5.3 Modeling lexical contact . . . 178
5.3.1 Modeling the preconditions for contact . . . 178
5.3.2 A monodirectional channel model of language contact . 179
5.3.3 Opening and closing channels . . . 179
5.3.4 Simulating channel behavior . . . 181
5.3.5 Overview of the simulation . . . 182
5.4 Analyzing the simulated scenarios . . . 184
5.4.1 Are the scenarios realistic? . . . 187
5.4.2 Are the scenarios interesting? . . . 191
5.5 Potential further uses of simulated scenarios . . . 194

6 Phylogenetic lexical flow inference **197**
6.1 Modeling languages as variables . . . 198
6.1.1 Languages as phoneme sequence generators . . . 198
6.1.2 Languages as cognate set selectors . . . 199
6.2 A cognate-based information measure . . . 200
6.3 Conditional mutual information between languages . . . 203
6.4 Improving skeleton inference . . . 204
6.4.1 Problem: stability on discrete information . . . 204
6.4.2 Flow Separation (FS) independence . . . 205
6.5 Improving directionality inference . . . 206
6.5.1 Problem: monotonic faithfulness and v-structures . . . 206
6.5.2 Unique Flow Ratio (UFR): flow-based v-structure testing 208
6.5.3 Triangle Score Sum (TSS): aggregating directionality hints 210
6.6 The phylogenetic guide tree . . . 215
6.7 Deriving proto-language models . . . 216
6.7.1 Ancestral state reconstruction algorithms . . . 216

6.7.2 Evaluation of ASR algorithms on simulated data 222
6.8 Phylogenetic Lexical Flow Inference (PLFI) 225
6.9 Evaluation of PLFI 226
6.9.1 Evaluation metrics for phylogenetic flow 228
6.9.2 Overall quantitative results for NorthEuraLex data . . . 230
6.9.3 Qualitative discussion of NorthEuraLex scenarios 233
6.9.4 Evaluation on simulated data 244

7 Contact lexical flow inference **251**
7.1 The contact flow inference task 251
7.2 Advantages and disadvantages of contact flow 252
7.3 Difficulties in applying the RFCI algorithm 253
7.4 Significance testing for v-structures 255
7.5 Contact Lexical Flow Inference (CLFI) 257
7.6 Evaluation of CLFI 258
7.6.1 Evaluation metrics for contact flow 260
7.6.2 Overall quantitative results for NorthEuraLex data . . . 261
7.6.3 Qualitative discussion of NorthEuraLex scenarios 263
7.6.4 Evaluation on simulated data 271

8 Conclusion and outlook **277**
8.1 Summary 277
8.2 Future work 279
8.3 Final remarks 283

References **289**

Index **305**
Name index 305
Language index 309
Subject index 315

2 Foundations: historical linguistics

The purpose of this chapter is to give readers with a causal inference background sufficient knowledge of historical linguistics to arrive at a basic understanding of the new application domain. For readers with a linguistics background, it may serve as a quick overview of the relevant core definitions and issues of historical linguistics as I am framing them for the purpose of my work, sometimes deviating a little from the established terminology.

The second half of the chapter is of more interest to the linguist reader. It gives a rough overview of existing computational approaches to modeling language history, and discusses the current state of the art in reference to the methods of classical historical linguistics.

For the exposition, I need to presuppose some basic knowledge of phonology, or the sounds occurring in spoken languages. I will normally represent sounds by means of the IPA (International Phonetic Alphabet), which has become the standard across all branches of linguistics. To learn what these symbols represent, I recommend Ladefoged & Maddieson (1996), the standard textbook of phonology. For readers who are not interested in languages and their pronunciation, but merely want to understand the methods I am developing and describing here, it should also be possible to follow the discussion by treating the IPA as a bag of elementary symbols (an alphabet in the formal sense), and not assigning any meaning or properties to them.

2.1 Language relationship and family trees

While very encompassing definitions of language can be given, at the core, a language such as English or Spanish can be seen as a system of symbols (vocabulary) and combination rules (grammar) used for communication. From this perspective, which is not shared by all traditions within linguistics, a language consists of a collection of symbols (lexical items, such as *them*, *give*, or *renaissance*), and rules how these symbols can be combined (grammar rules, e.g. "to combine an adjective with a noun, put the adjective in front of the noun").

Obviously, individual languages can differ vastly in both the symbols they use and the grammar rules they use to recombine these symbols. While the sequences of sounds used to form the symbols are largely arbitrary and only constrained by limits on pronunciation and distinguishability, grammar rules in languages across the world show much more similarity in structure. The commonalities in the grammars of languages can be described in terms of typological features, such as the very basic fact whether adjectives usually precede the noun (as in English) or follow the noun (as in Spanish).

Just like most complex systems, human languages are constantly undergoing change. While some parts of a language change less quickly than others (e.g. words for body parts vs. slang terms), no part of a language is entirely immune to change. Over the course of millennia, changes will accumulate to the point were two languages which started out as dialects of the same language will end up having no recognizable similarities except the ones dictated by *universals*, constraints on the structure of human language which are ultimately rooted in physiological or cognitive limits.

To give two concrete examples of language change, Old English (OE) from about a thousand years ago, still distinguished different verb forms for the first and second person singular (*ic stele* 'I steal' vs. *þū stilst* 'you steal', cf. *thou stealest* from about 500 years ago), and a boy was called *cnafa*. Crucially, in addition to such instances of grammatical change and lexical replacement, the phonetic shapes of words will invariably change over time due to sound change. For instance, OE *cnafa* [knɑvɑ] became the modern word *knave* [neɪv], where the [k] is not pronounced any more, and the [ɑ] has been lengthened and then become a diphthong [eɪ].

For various external reasons, one (typically more isolated) part of a language community will sometimes not join in a change affecting the rest, or will undergo a change whereas the other speakers of the language do not. This is the prime mechanism by which a language can split into *dialects*, loosely defined as mutually comprehensible, but different variants of the same language. As time goes on, dialects tend to diverge further from each other, up to the point where their speakers do not understand each other any longer, which is when we start to call the former dialects separate languages. Since mutual comprehensibility is a continuum, and the comprehensibility relationship is not transitive (there can be dialect continua where each dialect remains comprehensible to its neighbors, but dialects which are farther apart are different languages), the definition of what we call a dialect and what a language is often arbitrary and not subject to linguistic criteria.

If we trace the development of one language as it recursively splits up into new variants through the ages, we arrive at a tree-shaped pattern which is called a phylogenetic tree. As an illustration, Figure 2.1 visualizes how various Germanic languages are reconstructed as having developed out of a common ancestor language, Proto-Germanic. The height dimension of the tree roughly represents the time dimension, and builds on the estimated time depth of each intermediate stage, i.e. the time at which the respective proto-language is assumed to have split into its daughter languages. Some estimation of time depth is necessary for all advanced methods of phylogenetic inference, but it is a hotly contested topic because none of the methods for estimating time depth has led to full agreement with the known history of families where proto-languages are attested by written records. Today, historical linguists typically avoid talking about chronology ("linguists don't do dates"), due to a long history of premature conclusions which later turned out to be mistaken. Still, in the case of Germanic languages, there are sufficient historical records of many languages that the dates implied by the vertical dimension of the tree are widely considered as very likely.

Languages which are descendants of the same proto-language are said to belong to the same *language family*. In practice, which languages are grouped together actually depends on whether the relationship between them has been proven. A family is thus not different in nature from any of its subgroups defined by a common proto-language, but whether we call it a family depends on the current state of our knowledge. The current partition of the world's languages into about 400 families (about half of these with only a single member) has however turned out to be remarkably stable for decades, indicating that the field might find itself near to a maximum time depth where enough similarites survive to prove genetic relationship. This maximum time depth is commonly assumed to lie between 6,000 and 8,000 years ago, with older relationships provable in the presence of old written records (as in the case of Afro-Asiatic, a family with a time depth of about 10,000 years which includes ancient languages such as Hebrew, Akkadian, and Ancient Egyptian).

Within the Indo-European language family, the Germanic languages form a *taxon*, i.e. a group of more closely related languages which in turn have a common ancestor. Often, families can be decomposed rather cleanly into several such taxa, whereas the question which taxon split off first is often difficult to answer and forms a large part of the debate among experts in the respective language family. The established taxa frequently correspond to a time depth of around 2,000 years, when the similarities between descendant languages are usually still so pervasive that the relationship is obvious to a layman taking a first glance at

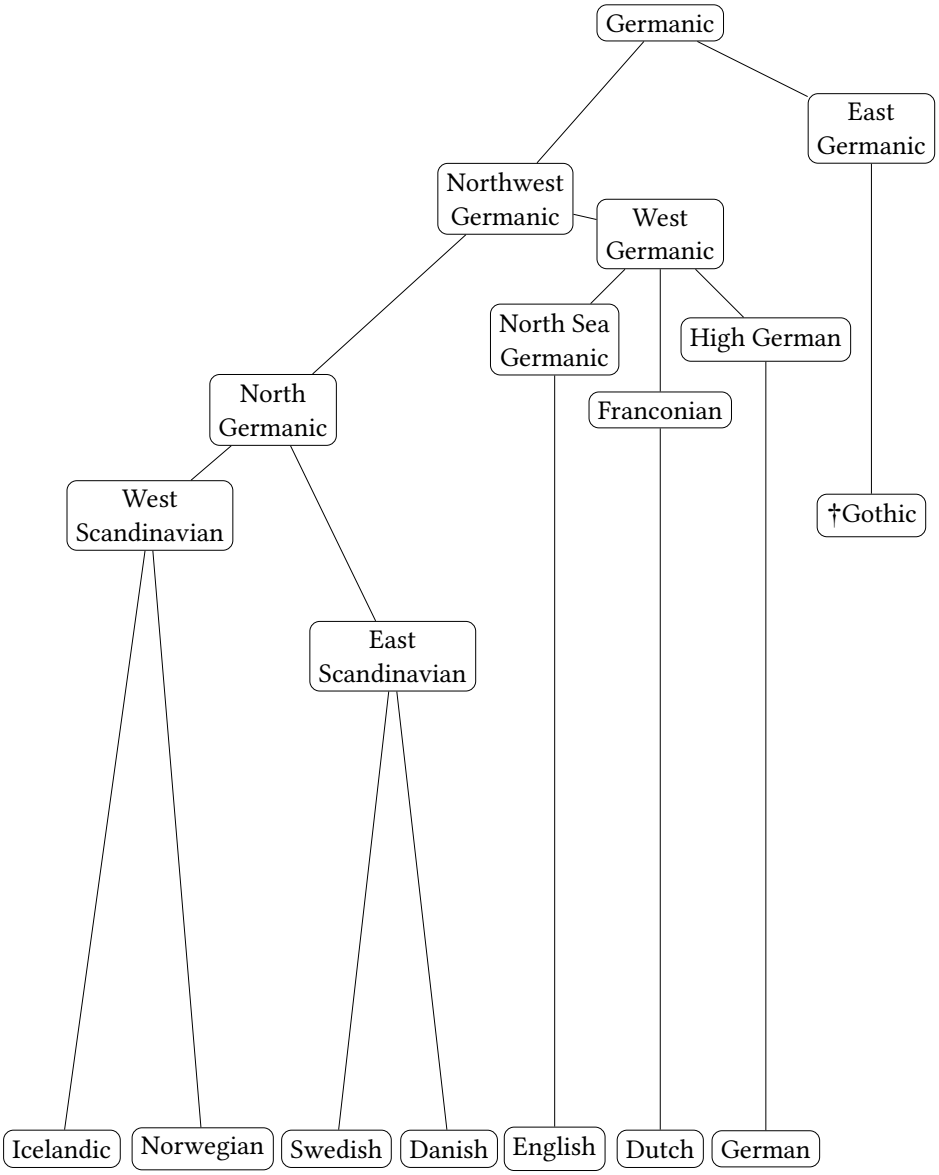

Figure 2.1: A (partial) phylogenetic tree of Germanic languages

the basic vocabulary. For instance, it is quite obvious that English *wife*, German *Weib*, Dutch *wijf*, Swedish *viv*, and Icelandic *víf* are essentially the same word, and a dozen of these close parallels in basic vocabulary would be enough to define Germanic as a taxon. In contrast, proving that Armenian *kin*, Russian *žena*, Irish *bean*, Icelandic *kona*, and Persian *zan* are related in very much the same way, requires a lot more effort and expert knowledge, which is similar to the situation in many other families.

2.2 Language contact and lateral connections

What complicates the picture of neat family trees depicting in which order the languages of a family split off, is that languages are in contact with each other, and that linguistic features do not only result from inheritance or random change, but also from borrowing between languages which are in contact. To the non-linguist reader, the term *borrowing* might seem slightly odd because the element taken from the donor language is never given back in any sense, in which case it helps to mentally equate borrowing with the copying of material. On the lexical level, a loan or loanword is a word which gets copied by one language (the recipient language) from another (the donor language).

English is a very good language for finding examples of contacts, as it represents a very interesting mix of inherited (Germanic) and borrowed (mostly Romance) features. On the level of morphology, English features two competing strategies of forming the comparative degree of an adjective, the Germanic suffix *-er* as in *larger* or *thicker*, and the Romance-style pattern with *more* as in *more interesting* or *more relevant*. The English lexicon is split roughly in half (Finkenstaedt & Wolff 1973), where the first half is dominated by basic vocabulary and words for everyday items and phenomena, which are typically either inherited from Proto-Germanic (*eye*, *rain*, *hammer*) or borrowed from other Germanic languages (*window*, *wing*, *skin*). The other half mainly consists of terms of science and culture (*science*, *pious*, *ignition*), all of which were borrowed from Latin and Romance languages. The convention is to consider the descent of the basic vocabulary and grammatical features as the relation defining the tree (and thereby the family membership), and to treat contact of any intensity as a secondary phenomenon, which makes English a Germanic rather than a Romance language.

Still, the sometimes very visible effects of language contact have always caused an undercurrent of historical linguistics to reject the tree model. Given the ubiquity of language contact, the underlying assumption of the tree model that languages continue to evolve independently after each split, and treating continued

contact as the exceptional case, might seem unnatural. People in this school of thought have tended to adhere to an alternative wave model, which is based on the observation that linguistic innovation tends to spread from a center to the periphery. Sometimes, innovations will sweep across language boundaries in situations of language contact, which then leads to borrowing. A language split occurs when a series of waves does not sweep across an entire language community. François (2014) provides a good recent overview of the theory behind wave models, and argues why they are attractive for describing some patterns of innovation. While generally accepted as well-suited for explaining areal phenomena and dialect continua, the strong assumption of wave-model advocates that any apparently tree-like signal in language evolution arises out of a pattern of overlapping waves is not advocated by many historical linguists any longer. While this view was once very popular due to certain phenomena in the history of Indo-European, it has been weakened by the abundance of quite clearly tree-like patterns in language families since studied. From the perspective of inference, wave models are problematic as well. The problem is that they have little explanatory value in the individual case, since every observable situation can be explained by many more different sequences of waves than sequences of splitting events which generate trees.

2.3 Describing linguistic history

Historical investigations about any given language often amount to proposing etymologies for words. An etymology is a description of a word's history, typically featuring either the information which word in a reconstructed proto-language it evolved from, from which other language it was borrowed, or how it was derived from lexical material which already existed in the language. For instance, the etymology of the English word *house* is given by Kroonen (2013) as inherited from a reconstructed Proto-Germanic word **hūsa-*, which in turn is of obscure origin, although it might derive from a Proto-Indo-European root **kuH-* 'to cover', which would connect the word to e.g. Latin *cutis* 'skin' and Lithuanian *kiautas* 'shell, rind, skin'. The etymologies one can establish thus vary in time depth, and tracing a word's history further into the past is a frequent type of contribution which would in this case advance our knowledge of Germanic. Etymological dictionaries for a single modern language will often include much more information about additional senses and when they developed, as well as the first attestations of some idioms involving the headword. This type of information is not required

for the word to have an etymology, and most etymological dictionaries do not include such information, often due to a lack of historical data.

In classical historical linguistics, words which are derived from the same word in a common proto-language are called *cognates*, whereas borrowed words and their descendants are not counted as belonging to the same cognacy class. While the clean separation of inherited words and loanwords is crucial to the classical method, computational methods have tended to put less emphasis on this distinction. This leads to a somewhat unfortunate difference in terminology between classical and computational historical linguistics, as the latter customarily subsumes both inheritance and borrowing under the cognacy relation. As an alternative term to cover this more liberal notion of cognacy, *correlate* has some currency, but I opt not to use it here because of the otherwise confusing frequent occurrence of the concept of correlates in the statistical sense in the text. Instead, I will use *true cognacy* for the stricter classical sense whenever the distinction is relevant, and otherwise stick to the more liberal usage established in computational historical linguistics.

Expanding on the basic distinction of inherited items and loans in the lexicon of a language lexicon, descriptions of loanwords are usually more fine-grained. Very often, loanwords of roughly the same age from the same donor languages can be grouped into strata or layers. For instance, there is a rather thin stratum of Celtic loans in English, which includes words such as *basket*, *beak*, and *nook*. This Celtic stratum can be further subdivided into an Ancient Brittonic layer (to which the mentioned words belong), and later borrowings from languages such as Welsh (*bard*, *crag*) and Irish (*galore*, *slogan*). In effect, the lexicon of every language can be split into an inherited core vocabulary, and a number of loanword strata which came into the language at different times from different languages.

2.4 Classical methods

This section provides a concise introduction into the mindset and the methods of classical historical linguistics, the discipline to which we owe the bulk of our current knowledge about the history of the world's languages, and against the results of which computational methods are commonly evaluated. Historical linguistics is a much broader field than a short introduction to core principles would suggest, and the non-linguist reader is encouraged to explore the field in its breadth by means of a handbook, such as the excellent recent one by Bowern & Evans (2015).

2.4.1 The comparative method

The primary tool of historical linguistics is the *comparative method*, a well-tested set of principles which has been developing for about two centuries, and has proven its worth as a tool for reconstructing the history of many language families. The key idea is to build on the assumed (and rather robustly attested) regularity of sound changes to reduce the likelihood that observed similarities between words from different languages are only due to chance. For a group of languages, relationship is then proven by reconstructing the sound inventory and the phonetic shape of many words in an assumed common proto-language, and then explaining how the known forms in each descendant language evolved from their equivalent in the proto-language by a series of regular sound changes.

2.4.1.1 Sound correspondences

On the synchronic level, both inheritance and regular sound changes lead to recurring sound correspondences in cognate words. For instance, there is a fairly regular sound correspondence between word-initial English *p* [pʰ] and German *pf* [p͡f], as evidenced by pairs like *pan/Pfanne*, *plum/Pflaume*, and *pluck/pflücken*. The last two examples show that we cannot expect one-to-one correspondences across all comparable segments in cognate pairs: English *u* [ø] can apparently correspond to German *au* [ɑʊ̯] or *ü* [ʏ]. The reason for such a one-to-many correspondence can be either that two different phonemes have merged into English [ø], or that one proto-phoneme diverged into the two German variants due to a conditional sound law, i.e. a change which happened regularly in a specific phonological context, which might have left no traces in the observable forms.

not plain [a]?

not [y]?

2.4.1.2 Sound laws

In the Neogrammarian view, the sound correspondences between related languages are the result of *sound laws*, i.e. regular phonetic changes which occurred while the languages developed from their common proto-language. In the simplest case, a sound change replaces all occurrences of some sound by a different one. However, in reality, sound laws almost never occur unconditionally, i.e. they will typically only apply in a certain phonetic context (such as between vowels, or in stressed syllables). Due to this context dependency, sound changes can increase the number of phonemes in a language, whereas unconditional rules could only keep the number of phonemes constant (if the resulting phoneme was not present in the language before), or decrease it (if two phonemes merge). For in-

stance, the phoneme [t͡ʃ]that is written *ch* in modern English, did not exist in West Germanic, and only developed during the Anglo-Frisian stage in a rather common process called palatalization, where velar plosives become palatals (often affricates) under the influence of adjacent front vowels. In other contexts, ancestral [k] was left untouched. This explains the seemingly irregular sound correspondence between English *choose* (Old English *ċēosan*) and Dutch *kiezen*, whereas we have [k] in both English *cat* and its Dutch cognate *kat*.

if this is a phoneme, it should be //, not []

The crucial idea of the Neogrammarian school of linguists is that such sound laws apply without exception, i.e. they apply to all instances of a sound in a particular context throughout the words of a language. These contexts can be quite complex, as can quickly be demonstrated using English and German. The sound law which produced the mentioned *p/pf* correspondence between English and German is part of the second phase of the High German consonant shift. By comparing the contexts in which this correspondence as well as analogous instances of affrication ([t] ⟶ [t͡s] and [k] ⟶ [k͡x]) occur, one finds that the law must have applied in four distinct contexts: word-initially, when geminated ([p:],[t:],[k:]), after liquids ([l] and [r]), and after nasals ([m] and [n]). For the *p/pf* pair, we have already seen examples of the first context. Instances of the second context are *apple/Apfel* and *copper/Kupfer*. The liquid context is exemplified by *carp/Karpfen*, and as examples of the nasal context we have *swamp/Sumpf* and *cramp/Krampf*. This pattern reliably repeats itself across all the lexical material which both languages inherited from West Germanic.

Since sound changes are historical events which happen during a short timeframe, the laws which shaped the history of a language can be arranged into a sequence in which they occurred. Because sound laws frequently interact (e.g. if one change creates a context where the next law can apply), we can often derive constraints on the possible order in which they must have occurred, leading at least to a partial relative chronology. For instance, we know that the third stage of the High German consonant shift, which in some German dialects turned voiced plosives into voiceless ones (e.g. *[b]* ⟶ *[p]*,must have occurred after affrication, because the voiceless stops would otherwise have become affricatized in turn. We can exlude this order of events based on the fact that the German cognate of English *bread* is not **Pfrot*, but *Brot*, which is pronounced something like [pʀo:t]in the dialects which underwent the third stage of the shift.

no italics for IPA

maybe better [b̥]

While we frequently can derive constraints on the relative order of sound changes in this way, an absolute chronology of sound changes is much more difficult to derive. Typically, it is necessary to rely on historical knowledge or written sources for this. For instance, we know for certain that the High German

sound shift must have happened before the 8th century, because Old High German texts from that time already display the results of that change. For written languages where we have no written sources in scripts which reveal the phonology, a (partial) relative chronology is often the best we can arrive at.

2.4.1.3 Distinguishing inherited words from loans

The work of establishing sound changes and their chronology is necessarily based on reflexes of the same proto-words across descendant languages. The groundwork of historical linguistics has therefore always revolved around the tasks of finding true cognates, distinguishing them from loans, and separating the loanwords in each language neatly into strata. This sometimes very complicated task forms the necessary preparatory work for later higher-level steps such as determining isoglosses, and reconstructing proto-languages in order to establish phylogenetic units.

In addition to reducing the likelihood of words becoming similar due to chance, the regularity of sound change also provides us with the most important source of hints about the etymology of words, especially when deciding whether some word was inherited, or borrowed from a sister language. For instance, the Proto-Germanic shift from [k] to [h] is enough to prove that Latin *cellarium* 'pantry' and German *Keller* 'cellar' cannot be true cognates, because we would expect something like **Heller* in this case. Although interactions between sound laws, and gaps in our knowledge about them, can make this type of argument quite complex, it is typically possible to recognize non-cognates, and estimate the time at which they were borrowed, for a large portion of a language's basic lexicon.

2.4.1.4 Reconstructing ancestral forms

With cognate sets and sound laws established, in theory it should become an almost mechanical task to project the attested words back to reconstructed proto-forms by reverse application of the sound laws. If this back-projection does not lead to the same proto-form if we start from different descendant languages, this is a hint that some of our current hypotheses about sound laws and cognacy relations must be wrong, and provides us with clues about the ways in which our theory needs to be revised.

In practice, there are many phenomena which complicate the picture, and make reconstruction of ancestral forms a non-trivial task. The most pervasive of these is *analogy*, which subsumes a variety of very frequently occurring phenomena by which irregular changes can happen whenever the result becomes

in some way easier to represent and process than the initial form. To quote two of the many examples given by Campbell (1999) in his introduction to historical linguistics, the vowel contrast between the English adjective *old* and its old comparative form *elder* (which still survives in *elder brother*) became leveled to yield current *older*, and German *Natter* is cognate with English `adder`, which lost its initial *n* due to reanalysis of its combination with the indefinite article, such that *a nadder* became *an adder*.

To give another more complex example from a different language family, the Proto-Uralic words **ükte* and **kakta* are one possible reconstruction for the numbers 'one' and 'two'. The latter should have resulted in Finnish **kahda-* by regular sound change, but the actual form is *kahde-*. In contrast, Finnish *yhde-* 'one' is the completely regular result of applying known sound laws to **ükte*. It is generally assumed that the irregular form *kahde-* received its final vowel by analogy, making the words for the two numbers more similar. The decision that **kakta* and not **ükte* is the word that was reshaped is necessarily based on reflexes in other branches of Uralic, which demonstrates why analogy complicates reconstruction. In fact, because irregular changes also appear to have happened to **kakta* in other branches of Uralic, there is still no consensus whether **kakta* is actually the correct reconstruction.

Explanations involving analogy are very common in etymological research, and it seems that analogy is a force just as important as regular sound change in shaping words. For historical linguistics, relying too much on analogy when explaining word forms is quite risky, because allowing almost arbitrary sporadic changes to apply to only one or a few words makes it much easier to fit forms to any reconstructions, detracting from the strength of the method.

2.4.1.5 Establishing phylogenies

By reconstructing ancestral proto-forms for a set of cognates in a set of attested languages, and proving that the regular sound changes and additional assumptions such as analogies correctly generate the different attested forms from the reconstructed form, a historical linguist proves that the attested languages form a phylogenetic unit. By reconstructing older proto-languages as common ancestors of already established younger ones, it should in principle be possible to establish the entire phylogenetic tree of a language family, detailing in which order various genera split off the common proto-language, and how these in turn split into the attested languages.

If we continue this reconstruction process through the ages, shouldn't it be possible to trace the history of each language back to very few large families?

If one assumes *monogenesis*, i.e. that human language has only developed once, and not independently in different places, one could even imagine deriving a world tree, detailing how the modern human languages developed out of a single ancestral language of humankind.

Unsurprisingly, there are limits to the comparative method preventing us from getting this far. The more sound changes accumulate through the ages (especially under complicated conditions), the more indistinguishable the inherited similarities will be from chance similarities. To still isolate individual sound laws and unravel in which contexts and in which order they applied, we would need ever larger numbers of cognacy candidates. The most serious limitation for the method therefore lies in the fact that cognate density actually decreases. Due to semantic change, lexical replacement, and borrowing, the cognates shared between two related languages are bound to get lost with time. Since every language has only a finite number of basic lexemes, already after a few millennia the languages will cease to share enough cognates for regular sound correspondences to be established, and the comparative method ceases to work.

For well-documented language families, the limits of the comparative method in terms of establishing deep ancestry appear to have been reached quite some time ago. What is more, many families which are generally considered established (such as Afroasiatic, and Sino-Tibetan) are not proven as genetic units in the strict sense, as there are no single widely accepted reconstructions of the respective proto-languages. The maximum age of phylogenetic units which can still be safely established using the comparative method seems to lie at between 6,000 and 8,000 years before present, which leads us far into prehistory in most parts of the world, but is a far cry from being able to get back to the times when e.g. the Americas or Australia were settled. Any method which tries to answer questions at higher time depths based on language data will need to resort to statistical arguments, or typological similarities, both of which cannot rule out the possibilities of chance similarity (a risk which is high for typological variables due to universals) and ancient contact.

2.4.1.6 Shared retentions and innovations

But the methodological limits of the comparative method do not only appear at high time depths, but also when making family-internal classification decisions. To reliably separate one taxon from the rest of the family, the amount of lexical overlap in terms of shared cognates is considered an insufficient criterion. Even if we excluded the possibility of borrowing, a larger-than-average lexical overlap between two languages we want to group together can still be either

due to shared retentions (the languages outside the group under consideration changed) or shared innovations (the change happened from the proto-language of the family to the taxon we are trying to establish). The existence of correspondences alone does not yet allow us to decide whether we are dealing with shared retentions or innovations. The main reason why many subgroupings which seem obvious on a lexical level are sometimes not generally accepted is that quite often, the hypothetical common proto-language is so close to the proto-language of the entire family that no regular changes can be detected to define the transition between the two proto-languages. The requirement of demonstrable shared innovations often severely limits the ability of classical historical linguists to clarify the internal structure of a language family beyond the level of securely established branches.

2.4.2 Theories of lexical contact

Whenever speakers of two different languages get in intensive contact with each other, this will invariably leave traces in those languages. According to Hock & Joseph (1996: Section 8.5), the main variable deciding about the shape of lexical influence between two languages is the difference in prestige. This difference is the standard explanation for the fact that languages do not only borrow needed words for new concepts (such as German *Computer* and *Internet*), but also tend to replace perfectly workable and well-established terms for some concepts with those from a prestige language (German *Service* instead of *Dienstleistung*, or *Ticket* instead of *Fahrkarte*).

As Thomason & Kaufman (1988) elaborate in their analysis of contact situations, the decisive factor determining how a contact situation between a high-prestige and a low-prestige language plays out is whether it occurs under conditions of maintaining the low-prestige language, or language shift towards the high-prestige language. The first scenario will typically lead to a situation of widespread bilingualism, where even words for basic concepts tend to be replaced by their borrowed equivalents, as the familarity of the bilinguals with the higher-prestige language increases at the expense of the lower-prestige recipient language. The situation where the prestige gradient is not too high, and both languages continue to coexist for many generations, is the one where one would expect the largest amount of borrowings even of basic vocabulary items. The words borrowed from a language of comparable prestige are said to form an adstratum. For instance, English has a North Germanic adstratum from the time when the Vikings settled large parts of England, and started to intermarry with the local population. Words borrowed by English during this time include very basic vocabulary items such as *to take* and *they*.

In the second scenario, the target language is learned imperfectly by shifting speakers, which tend to retain many phonological and syntactic features of the original language, but typically not much lexical material except terms for local plants and animals. If the number of shifting speakers is demographically relevant, the structural substrate influence will result in a changed variant of the target language. A case in point is the Western Uralic substrate in Russian which shows itself in certain syntactic features which set Russian apart from its Slavic sister languages, such as the lack of a copula in the present tense, and the extensive use of the partitive genitive.

In the case where both languages are maintained (i.e. in the absence of language shift), Thomason & Kaufman (1988) distinguish five degrees of contact intensity, each with characteristic manifestations in the extent of lexical and structural borrowing. The first three stages represent different degrees of lexical borrowing, with accompanying weak structural borrowing that does not cause any shifts in the typological profile. Under circumstances of casual contact, we would only expect content words to be borrowed, and typically non-basic vocabulary. Under slightly more intense contact, conjunctions and adverbial particles will be among the first structural elements which are taken over. Only under very intense contact will we observe borrowing of other function words such as adpositions, pronouns, and low numerals. Bound morphemes such as derivational affixes may also be transferred at this stage, and they can stay functional in the borrowing language. The last two stages describe situations of strong and very strong cultural pressure, where the structural influence is so strong that typological changes can occur. On the lexical level, contact of this intensity will lead to massive borrowing which can even replace large parts of the basic vocabulary.

2.4.2.1 Types of borrowing

In the bilingual environment where most lexical borrowings occur, a loanword is initially borrowed in its original phonetic shape. With time, loans tend to get nativized by sound changes, often up to a point where they are not recognizably foreign any longer. An instance of this is German *Fenster* 'window', which was borrowed from Latin *fenestra* into Old High German, i.e. more than a thousand years ago. Without knowledge of Latin, no native speaker of German would suspect that this word was not inherited from the parent language. By contrast, more recent loans in German, such as *E-Mail* from English, tend to contain sounds foreign to German (the diphthong [eɪ]), or to deviate from the usual orthographic rules (the long vowel [iː] written as *e*, instead of *ie* or *ih*), and are therefore instantly recognized as loanwords.

While from the perspective of cognacy, borrowing across families will tend to be detectable as words from a different cognate class, the quite common situation of internal borrowing within the same language family often leads to a word getting replaced by a cognate word from a sister language. For instance, this is what happened with the English word *guard*, which replaced older *ward* by a cognate borrowed from Frankish via French, where a regular sound change from [w] to [g] had occurred (cf. *William* vs. *Guillaume*). Both cross-family and family-internal borrowing will commonly happen whenever living languages come into close contact.

A lot less frequently, words can also be borrowed from ancient languages, which might even be direct ancestors of the recipient language. This tends to happen with high-prestige written languages that persist as languages of religion or science. The most well-known examples are Latin in Western and Southern Europe, from which Romance languages did not only inherit, but also borrow words, and Sanskrit, which plays the same role across India. Less well-known examples of the same pattern include Old Church Slavonic, from which many words were borrowed into later East Slavic languages such as Russian, and the influence of Pali, the Middle Indo-Aryan language of Theravada Buddhism, on many languages of Southeast Asia. If a language borrows from its own ancestor or a close relative, this very frequently leads to doublets like the *guard/ward* example, which are also called etymological twins. For instance, the Latin word *dīrectus* developed into French *droit* under regular sound change, but the word *direct* was later borrowed into French in addition, where the two words now coexist.

In addition to loanwords, *calques* are the second important type of lexical borrowing. A calque, also called loan translation, is a derived word which is composed of native lexical material after the model of a derived word in another language. For instance, the Hungarian *összefüggés* 'correlation' is composed of the native lexemes *össze* 'together' and *függ* 'to hang' after the model of German *Zusammenhang* 'correlation', literally 'hanging-together'. Massive calquing tends to occur when the vocabulary of a language needs to be expanded rapidly to areas of life that it was not previously applied to (e.g. science or technology), and is especially pervasive when compounding is a preferred word formation strategy in both the model language and the newly expanded one, as has been the case for German and Hungarian. The term 'borrowing' as used in this book does not include calques.

2.4.2.2 Constraints on borrowing

As Haspelmath (2008) states in his summary of loanword typology, an essential step towards a theory of language contact is to determine possible constraints governing which elements of a language are more likely to be borrowed, and in which order elements will be borrowed under intensive contact. From the viewpoint of historical linguistics, understanding these constraints can help to decide open questions in language classification. In the context of the present volume, this knowledge will be of some use in interpreting results, and feeds into the design of a simulation model of some aspects of actually occurring lexical transfer.

The most striking initial observation about borrowability is that the number of content words which can be transferred during intensive contacts seems almost unconstrained. Less than half of the vocabulary of modern English is of Germanic descent, and Armenian has borrowed so many words from neighboring Iranian languages that its nature as a separate branch of Indo-European was only recognized very late in the history of Indo-European linguistics. However, we have already seen that words for the most basic vocabulary are typically exchanged only among languages of roughly equal prestige in long-term contact. This kind of contact is historically quite rare at least in the regions of the world that I will be concerned with here, meaning that basic vocabulary will be a very good predictor of genetic affiliation.

Beyond such general statements about basic and non-basic vocabulary, scholars have established some non-trivial constraints on the borrowability of different parts of the lexicon which seem worth mentioning. For instance, an important factor to which much influence has been attributed is the typological distance between the donor and recipient languages, because very different grammars make it harder to copy words, let alone grammatical features, without causing major changes to the recipient language's system. This helps to explain why conjunctions and adverbial particles are borrowed more often than other functional items. These elements belong to smaller subsystems which tend to be less integrated with the rest of the grammatical system, and are thus more likely to be integrable into the structural fabric of the borrowing language.

Calling into question the predictive power of such theories, Thomason & Kaufman (1988) attack the central role attributed to structural incompatibility as an explanation of resistance to lexical borrowing. Based on some very interesting extreme cases, they argue that any prediction about which parts of the lexicon and the grammatical structure are susceptible to borrowing will mainly need to build on sociolinguistic factors. Under social circumstances which are conducive

to moderate borrowing, however, typological compatibility still appears to influence the extent of structural borrowing, sometimes leading to more intensive interference than one would expect at the given intensity of contact.

Beyond compatibility, an important inhibiting factor for the borrowability of a feature appears to be its overall typological markedness. For instance, morphemes which express more than one function (such as the combined case and number markers of Indo-European languages) are less likely to be borrowed than the typologically more common clearly separable and single-function morphemes (such as case endings of agglutinating languages). Beyond such individual cases, if we consider morphological means to express functions as generally more marked than syntactic means, this general principle can also explain the tendency for morphological complexity to reduce in contact situations.

On the lexical layer, there are differences in borrowability between different types of content words. Most prominently, nouns are borrowed more easily than verbs. This long-held view was substantiated by van Hout & Muysken (1994), who statistically analysed texts for different factors which predict the borrowability of lexical items from Spanish into Quechua. Their explanation for finding many more borrowed nouns than verbs is the motivation of extending referential potential, i.e. giving words to new things. Since new things which need a name are much more common than new actions, this explains the higher borrowability of nouns. But they also find a signal in favor of borrowing lexemes which show little inflection in the donor language. The latter finding ties in well with the theory of language contact developed by Myers-Scotton (2002: Ch. 6), who argues that the main reason for the higher borrowability of nouns as opposed to verbs is that introducing foreign noun phrases tends to be less disruptive to predicate-argument structure.

A well-known phenomenon that can be interpreted as reinforcing this theory was first observed by Moravcsik (1975), who claimed that words for verbal concepts are never borrowed as verbs, and only become borrowable as nominalizations. The part of this extreme claim which still remains valid today in the presence of much more evidence is that languages with complex verbal morphology do not tend to borrow verb stems from other languages, nor act as donors of verbal stems. Instead, verbal concepts are much more likely to be borrowed in the shape of nouns, typically in the form of a source-language nominalization which is then combined with a light verb meaning 'to do'. For instance, this pattern appears very strongly in the Arabic influence on the languages of many Islamic cultures. The Semitic root-pattern morphology is so alien to languages from other families, that they will only borrow verbs in a nominalized form. For instance,

the Arabic verb *daʿāʾ* 'to summon' was borrowed as a verbal noun (*duʿāʾ*) into languages from other families, where it was combined with native light verbs to express the concept of praying. In Persian, this gives us *do'â kardan* 'to pray', whereas the Turkish and Uzbek equivalents are *dua etmek* and *duo qilmoq*, respectively. This strategy of integrating Arabic loans is extremely common in all major Iranian and Turkic languages. Instances of the same strategy are observed many times across the globe by Wichmann & Wohlgemuth (2008), who place it at the lower end of a tentative loan verb integration hierarchy. The partial cognacy relations which result from this type of borrowing become a problem for any attempt to automatically partition the words for a given concept across many languages into cognate classes.

2.4.2.3 Mixed languages

Some languages have interacted with other languages to such a degree that their genetic affiliation becomes difficult to define. The most common type of such mixed languages are the *creoles*, fully developed languages which come into being when a *pidgin*, a simplified auxiliary language as it tends to arise when speakers of very different languages need to communicate, gets nativized by children growing up with the pidgin as their primary language.

The prototypical creole languages all arose from colonization, where the colonial language invariably operates as the *lexifier* of the creole language, i.e. virtually the entire lexicon is inherited from the colonial language, albeit undergoing sometimes significant semantic change. The substrate influence of the other language is seen in the grammatical structure (which often retains little similarity with the lexifier), and often in collocations and idioms. For instance, Tok Pisin, the national language of Papua New Guinea, is an English-based creole where the word *gras* 'grass' has taken on the primary meaning 'hair', via the indigenous conceptualization of hair as *gras bilong het* 'grass belonging to the head'. But apart from the prototypical colonial situation, other languages are sometimes discussed as possibly being creoles as well, especially when massive shifts within the grammatical systems can be shown to have occurred within few generations. The most famous example of this is English itself, which was heavily restructured during the Middle English period, losing almost all inflected forms and becoming extremely simplified in the remaining inflections such as plural formation. What makes this case less prototypical is that the two involved languages were related (making structural borrowing much easier), and that there was no clear developmental gap between the two cultures which would have ensured dominance. This also explains why in this case, the lower-prestige language would have to be treated as the lexifier.

While mixed languages can be difficult to classify in terms of phylogeny if our desire is to trace the development of the entire language system, on the description level of the lexicon, which the work described here is confining itself to, it is entirely unproblematic to just model creoles as immediate descendants of their lexifiers. Therefore, we do not need to be too concerned here with languages that might not have a clear position in a phylogenetic tree, and we can always assume an underlying tree-shaped skeleton to exist in our networks. On the lexical level, one could summarize the position I am taking as follows: there are no equal mixtures of languages, there are only admixtures. In biological terms, we have no hybridization, but potentially massive horizontal gene transfer.

2.5 Automated methods

Looking up many words in dictionaries, cross-referencing them and constantly re-performing these steps when revising earlier findings while solving the puzzle of a language family's development, can be a very time-consuming and even tedious task. Not surprisingly, the potential advantages of being able to automate subtasks in historical linguistics were seen as soon as computing technology became performant enough to operate on large quantities of string data.

The earliest example of applying computers to a problem of historical linguists I was able to find is Hewson (1974), who uses predefined correspondences between Algonquian languages and simple sequences of substitutions to generate all possible projections from attested words into possible Proto-Algonquian forms, and filters out all candidate forms which are reconstructable by some sequence of substitutions from each modern form to arrive at a consistent reconstruction hypothesis. According to the author, this procedure resulted in the detection of 250 previously unknown cognate sets, and was then used as a core for a computer-generated etymological dictionary. From the description it is clear that the system exploits much previous knowledge, both in the representation and preprocessing of the input data, which will not be easily transferable to other language families.

One step closer to modern statistical methods, the COGNATE system first presented in Guy (1984) estimated the probability of sound correspondences using chi-square tests on a sound co-occurrence table based on string positions. The system was evaluated on 300-word lists from 75 languages of Vanuatu, and is reported to have yielded satisfactory results for closely related languages. Unfortunately, neither the system nor the test data appear to remain available.

Embleton (1986) summarizes early developments in lexicostatistics, but also foreshadows many of the approaches and concepts which still figure centrally

in phylogenetic inference. For instance, Embleton proposes the use of clustering algorithms for deriving phylogenetic trees (including branch lengths) from cognate data, and uses a simulation model to analyze the amount of skew in tree inference introduced by borrowing. The discussion also addresses many of the major issues that the field is still struggling to solve, such as the lack of truly independent linguistic features, or the problems caused by selection bias in lists of shared roots or grammatical features that are extracted from the specialist literature.

A factor which hampered progress in this and many other computational fields was the lack of sufficient computing power for testing the already quite advanced algorithmic ideas of these pioneers of computational historical linguistics on substantial amounts of data. When it became clear that these limitations made the early tools too inflexible und unwieldy to attain general acceptance and widespread use among historical linguists, the field did not see any work for about a decade. It was only in the late 1990s that the successes of computational methods in biology inspired a second wave of attention for introducing automatization into other branches of science where the gene metaphor seemed fruitful. Among others, these included literary studies (tracing how works were derived from each other), anthropology (attempting to reconstruct ancient systems of kinship), and linguistics.

This section gives a rough overview of recent developments in applying computational methods to answering questions of relevance to historical linguistics. The discussion is restricted to methods which attempt to find answers to concrete questions about the past of words and languages, and does not include more general results which can be derived from large databases, such as computational proofs of claimed typological universals like sound symbolism, or global correlations involving extralinguistic features such as altitude, climate, and population size.

2.5.1 Lexical databases

The most basic prerequisite for any computational study in historical linguistics is an electronic database which contains the information a linguist would look up in dictionaries or other sources in a standardized format which can be processed by a computer. The absence of such databases has been one of the limiting factors in the expansion of the field, but some very useful resources have become available during the past decade, and the pace at which new resources appear seems to be accelerating.

While databases of typological features have only recently started to receive broad attention, most work so far has been performed on representations of the basic lexicon across a relevant set of languages. Such lexical databases either contain phonetic forms representing the realizations of a concept across the relevant languages in a unified format, or, especially when they cover data within well-known families, the realizations are cognacy-coded. The advantages and disadvantages of these types of databases, as well as examples of both types, are discussed in this section.

2.5.1.1 Databases of phonetic forms

The easiest way to generate some computationally tractable data about a set of languages is to take a list of basic concepts (the words for which still tend to be cognate among more distantly related languages), and dictionaries, and then digitalize the relevant entries, transcribing them from the orthography or the format used in the source into some cross-linguistically applicable string format, usually over some phonetic alphabet which allows to represent all the phonemes of the language family of interest. Many factors complicate this basic procedure, such as the need to bridge different gloss languages in different sources, the low availability of unpublished resources like fieldnotes, inadequate phonetic descriptions which make it impossible to reconstruct the pronunciation at the desired level of detail, grammatical properties which make expert knowledge necessary to isolate the relevant parts of dictionary forms, and imprecise glosses which leave the compiler without certainty that the intended concept was matched. Still, with some experience, very little is needed to compile a database of phonetic forms corresponding to a list of basic concepts. This is the main advantage of settling for phonetic forms, as opposed to more high-level data.

The earliest major effort to create a computer-readable database of basic vocabulary was part of the Automated Similarity Judgment Program (ASJP). The ASJP database aims to cover the words for 40 basic concepts across all documented languages, in a rather rough, but unified phonetic transcription. Version 18 (Wichmann et al. 2018), the most recent version available at the time of writing, includes 7,655 wordlists. While some of these wordlists do not correspond to different languages, but variants of the same language, the number of languages still approaches about two thirds of the global estimate of currently spoken languages. Altogether, the database approaches a size of 310,000 entries, making it by far the largest currently available resource in one consistent format. Over the years, previous versions of ASJP have been used to investigate many linguistic questions like the stability of concepts against borrowing and semantic change,

the question whether sound symbolism creates problematic amounts of lexical similarity between unrelated languages, and correlations between phoneme inventories and extralinguistic factors such as population size or geographic isolation.

More recently, Greenhill (2015) presented TransNewGuinea.org, a database covering more than 1,000 languages and dialects of New Guinea. The database represents a massive effort to make lexical data on the basic vocabulary of Papuan languages, the least well-studied linguistic region of the world, readily available to a wider public on the web. Building on a list of 1027 lexical meanings, various types of published and unpublished resources were processed to build a database in a unified phonetic format that can be processed by computational tools. Due to the very sparse documentation of many languages, at the time of publication the total size of the database had only reached about 145,000 entries, or an average of just over 140 words per language. Work to expand the database by cognacy judgments is under way, but since the bulk of available material has already been processed, it will not be possible for this database to become much larger.

The Chirila database of Australian languages by Bowern (2016) is another good example of a database spanning an entire linguistic region, with the goal to eventually make all known lexical data available. Due to the complicated legal situation when publishing full resources, and a cultural bias of many linguistic groups against giving outsiders access to their languages, only 230,000 of about a million database entries are freely available at the moment, but even this lower number puts Chirila among the largest available databases. In addition to documenting the word forms in the original sources, much effort is put into clarifying or reconstructing the most likely pronunciation in order to arrive at standardized phonemic representations.

The NorthEuraLex (North Eurasian Lexicon) database first presented in Dellert (2015) is similar to the previous two databases in its aim to cover an entire linguistic area, but has the advantage of containing only very few gaps despite covering 1,016 concepts, which is only possible due to the much better documentation of minority languages in Europe and Russia. The version of NorthEuraLex which I will be using for evaluation covers 107 languages, making it comparable in size to the released parts of the Papuan and Australian databases. Since the compilation of NorthEuraLex was a substantial part of the necessary preparatory work for my experiments, as it provides the gold standard for evaluating my lexical flow methods, it will be discussed in much more detail in Chapter 4.

The examples of databases just listed are only the tip of an iceberg of smaller often unpublished databases which cover a single language family or the minority

languages of one country. As can be seen from the recent dates of most publications, there has been an explosion in the number of large-scale lexical database projects during the last two years, a trend which can be expected to gain traction as the field continues to grow.

2.5.1.2 Cognate databases

The other type of lexical database does not consider the phonetic forms of primary importance, but encodes the presence or absence of cognate classes in each individual language. A phonetic database would focus on the information that the words for *hand* in Armenian, Albanian, Greek, and Georgian are [dʒ̂]ɛrkʰ], [dɔrə], [çeɾi], and [χɛlɪ], allowing a program to compare these strings in order to figure out whether they are related. In contrast, a cognate database would not provide the three words in a unified phonetic format, but instead encode the information that the first three words are cognates, while the fourth is unrelated, by assigning a `1` to the first three languages and a `0` to Georgian in a column encoding the absence or presence of this cognate set.

The advantages of cognacy encoding are that binary characters are easier to handle computationally, and that many disturbing factors such as loanwords or morphology are already filtered out during data preparation, leading to much cleaner data. The disadvantage is that cognacy-encoded databases need to be compiled either by experts in the history of the respective language family, or by going through the published etymological literature in language families where such work exists. Both approaches require an enormous amount of work, which makes typical cognacy-annotated databases much smaller than phonetic form databases, and also limits the number of language families for which they are available.

A very early cognate database is the Dyen database which formed the basis of Dyen et al. (1992), an early lexicostatistical study of Indo-European. The database is a small and rather unreliable resource (Geisler & List 2010) which covers 200 concept across 84 Indo-European languages in cognacy-encoded form. After substantial revisions, it today forms the core of IELex (Dunn 2015), a database of increased quality which is continuously being updated, and will soon be released in a major revision. The most recent publicly available version groups about 35,000 words into 5,000 cognate sets.

An equivalent of IELex for the Uralic language family is collected under the name UraLex. The latest available version of UraLex was published together with a phylogenetic analysis of the data by Syrjänen et al. (2013), when the database covered 226 concepts across 17 languages. Given the small size but high time

depth of the language family, the very distributed state of etymological information, and considerable disagreement between different authors, even compilation of this small database has certainly been a substantial effort.

By far the largest effort so far is the Austronesian Basic Vocabulary Database (ABVD) compiled by Greenhill et al. (2008), which covers 210 concepts across more than 1,400 languages, providing virtually complete coverage of the world's largest language family. ABVD partially relies on orthographic forms instead of a fully unified transcription, otherwise it would provide another phonetic form database of a size comparable to the ASJP database, due to its deeper coverage of individual languages. What makes this database unique, however, is that words from a sample of 400 languages (an earlier version of the database) are grouped into more than 34,000 cognate sets. The low time depth of many genetic subunits tends to make these cognate sets a little less interesting than the long-distance cognates from the other databases, but the cognacy-annotated part of ABVD is poised to remain the largest database with expert cognacy annotations for quite some time.

2.5.2 Phylogenetic inference

A major focus of computational historical linguistics has been phylogenetic inference, i.e. the task of inferring phylogenetic trees from language data. The bulk of work in phylogenetic inference has been character-based, typically building on cognacy data encoded in such a way that the presence of each cognate class is treated as a binary character. The older distance-based methods, where a single distance matrix between languages (which can be computed from string data in many different ways) is used to extract tree-like signals, have recently regained some popularity, especially for investigating language families where cognacy-encoded databases do not exist.

Phylogenetic inference already was a well-developed branch of bioinformatics when it started to be applied to large amounts of language data. Computing an optimal tree is inherently a very demanding problem because already the number of possible tree topologies over k languages rises super-exponentially with k, and this does not yet include the inference of branch lengths. Exhaustive optimization according to some optimality criterion is therefore not an option. Instead, heuristic methods are employed, with the risk of hitting a local instead of the global optimum. The following summary is based on Felsenstein (2004), a very popular book-length introduction to phylogenetic tree inference which is also recommended to the reader as an entry point to the field.

Phylogenetic inference methods can be classified as either distance-based or character-based. The distance-based case is the more general one, because any character-encoded dataset can be reduced to a distance matrix in a number of ways. However, the loss of information caused by reducing a character matrix into a simple language distance matrix will typically lead to lower-quality results, so that distance-based method will not typically be used if character-encoded data is available.

Assume we want to infer the best tree from a distance matrix. An ideal distance matrix would correspond directly to some tree by having the property that for any triple A, B, C with structure $((A, B), C)$ [i.e A and B are closer, and C more distantly related], the distance measure fulfills the conditions $d(A, B) < d(A, C)$ and $d(A, B) < d(B, C)$. However, a distance matrix which unambiguously encodes a single tree topology is rarely observed in practice. The reason can be noise introduced by errors in the underlying data, an inadequate distance measure, or a real non-tree-like signal resulting e.g. from loanwords in linguistics, or from horizontal gene transfer in biology. Multiple standard algorithms exist for quickly extracting a plausible tree from a distance matrix which does not consistently represent a tree. These approaches differ in complexity, and in the criterion they optimize.

The oldest and still frequently used distance matrix algorithm is UPGMA, first defined by Sokal & Michener (1958) as an approach to hierarchical clustering. UPGMA (Unweighted Pair Group Method with Arithmetic Mean) progressively fuses clusters into larger phylogenetic units, starting with every node in its own cluster. The algorithm maintains a table of current distances between all clusters. At each step, the two clusters with the minimal distance are fused into a new cluster, and a corresponding node is introduced to the phylogeny. The distance between the new cluster and all existing clusters is defined as a weighted average of the distances to the two clusters, with the weights defined by their relative sizes. Branch lengths are simply defined by the distances among the clusters. UPGMA works quite well on data for which a clock assumption holds, i.e. when we can assume that the changes which increased the distance occurred at roughly equal rates throughout the tree. Under this condition, the UPGMA tree provides the optimal least-square fit between the branch lengths and the distance matrix.

If the length of tree branches is to be minimized wihout a clock assumption, the most popular quick approach is the neighbor-joining algorithm by Saitou & Nei (1987), which is not linked to a simple optimality criterion. Neighbor joining maintains a measure of isolatedness $u(i)$ for each node i, derived from its average distance to all other nodes. At each step, it connects the nodes i and j

which have the smallest $d(i, j) - u(i) - u(j)$, i.e. distance corrected for isolatedness. The distance of the new node ij to each existing node k is then defined as $d(ij, k) := (d(i, k) + d(j, k) - d(i, j))/2$. i.e. the average of the distances to each of the two nodes, with a discount which increases as clusters become less closely connected. This procedure maintains a good balance between internal consistency of clusters, and quick inclusion of isolated nodes which are not particulary close to any other cluster.

If character-encoded data is available, the most straightforward optimality criterion is to build the tree which minimizes the number of assumed evolutionary events. This leads to the *maximum parsimony* paradigm, where the primary design decision is how to count the evolutionary events the number of which we want to minimize. The most straightforward definition is based on the minimal number of character-state changes we have to assume to fit the data to a given tree. Computing this number is typically done according to Sankoff (1975), a dynamic programming algorithm which reconstructs the optimal character states at ancestral nodes as a byproduct. We will therefore take a look at the Sankoff algorithm in detail when reconstructing the presence of cognate sets at protolanguages in §6.8. Parsimony scores for any given tree can be computed very efficiently, but the challenge remains how to traverse the tree space in order to find a tree of maximum parsimony. The most advanced methods for doing that are based on the branch & bound paradigm, where the entire search space is indexed by a decision tree. For each of the alternatives at the current decision node, lower and upper bounds for the still attainable parsimony scores are computed given the decisions already made. If the lower bound of one alternative is higher than the upper bound of the other, an entire (possibly huge) branch of the search tree can be ignored in our search for the maximum. With a good decision tree that maximizes the chances of large differences between alternatives, branch & bound optimization can be quite efficient.

Gray & Jordan (2000) were the first to apply maximum-parsimony phylogenetic inference on substantial amounts of language data to answer an open question of historical linguistics. Using maximum-parsimony trees over 77 languages inferred from about 5,000 cognacy characters, Gray and Jordan compare two competing theories about the Austronesian settlement of the Pacific. While the resulting trees clearly suggest a rapid expansion from Taiwan into Polynesia, they find that the overall signal does not fit the tree model extremely well, suggesting substantial interaction between populations even after the initial settlement. In contrast, Holden (2002) shows that the cognacy pattern for 92 concepts across 73 Bantu languages fits a tree-like pattern rather well. The subgroup-

ings with the highest support (i.e. which consistently appear across a range of maximum-parsimony trees on subsets of the data) were found to closely mirror the earliest farming traditions of sub-Saharan Africa, leading to the conclusion that the modern subgroups have largely remained in place since they became separated, without intensive contact or further large-scale migrations after the initial expansion.

The most modern and most successful approaches to phylogenetic inference are all *probabilistic*, i.e. they build on a model specifying the probability of generating the character-encoded dataset from any hypothesized tree. This requires the generation of trees to be modeled explicitly by an evolutionary model, which can e.g. include separate mutation rates for each branch. However we define our evolutionary model, it will assign a probability $p(D|T)$ to the dataset D given any tree hypothesis T. Now, if our goal is to find the tree T which maximizes $p(D|T)$, we do not actually need to normalize $p(D|T)$ into a probability distribution by considering all other datasets, but it is enough to have some function which ranks trees in the same way as $p(D|T)$ on our fixed dataset D. A function $L(T|D)$ with this property is called a *likelihood function*, and the resulting paradigm is therefore called *maximum likelihood*. As in the case of maximum parsimony, the challenge is to efficiently traverse the tree space in order to arrive at high likelihood values. Typically, the topology will be modified first, and the branch lengths then optimized given the topology. While naive optimization algorithms often work surprisingly well, a lot of technical machinery is needed to efficiently come up with good tree hypotheses on larger datasets. Felsenstein (2004: Ch. 16) is still a good entry point into the ever-growing landscape of algorithms and heuristic techniques trying to solve this challenge.

Further exploiting the advantages of fully probabilistic models, the state of the art in phylogenetic inference relies on Bayesian methods. If we have prior knowledge $p(T)$ about likely tree topologies, mutation rates and branch lengths (e.g. due to historical constraints), we can do better than a simple maximum-likelihood estimate by inverting $p(D|T)$ using Bayes' formula. This formula implies that $p(T|D)$ is proportional to $p(D|T) \cdot p(T)$, i.e. we can maximize the posterior probability $p(T|D)$ of the tree hypothesis given our prior knowledge $p(T)$ about possible tree structures and rates of change. What is more, if we normalize $p(T|D)$ across all possible trees, we actually get an explicit posterior probability distribution. This makes it possible to quantify our certainty about any given solution, and we can sum over the probability mass assigned to entire classes of trees in order to derive confidence windows in the tree space.

The problem is that the normalization factor for $p(T|D)$ will typically not be computable, because already the enumeration of all possible tree topologies becomes intractable very quickly. As soon as continuous branch lengths are involved, the normalization factor becomes an integral that can only be estimated. This is where the main challenge of implementing Bayesian methods lies, and again I point the reader to Felsenstein (2004: Ch. 18) for an introduction and overview. The crucial finding is that if we use specialized sampling techniques, knowledge of the likelihood function suffices to sample trees from the posterior distribution in a way that converges towards the true distribution. Based on large numbers of samples generated in this way, all relevant properties of the posterior distribution can be estimated just as well as if we had access to the full distribution. Since many trees have to be generated and discarded to emulate independent sampling, Bayesian methods are very demanding computationally, and have only recently become feasible to run in acceptable runtimes due to the advancement of computing technology.

All current probabilistic models only infer an *unrooted tree*, in which the branch lengths are defined for the path between every pair of nodes, but the position of the root in the tree remains unspecified. The reason for this is that the most widespread models for $p(T|D)$ are agnostic to the position of the root in the tree, implying that there is no mathematical criterion that can be used to decide between possible root positions. Therefore, after inference the root must be placed in some position on some branch of the tree in a *rooting* step based on external evidence. The simplest approach (and the only one which will feature in this book) is to define one of the languages as an *outlier*, which will be assumed to form a taxon separate from all other languages in the resulting rooted tree.

The application of Bayesian phylogenetic inference to linguistic data was pioneered by Gray & Atkinson (2003), who derive a controversial very early date estimate for Proto-Indo-European. As Bowern & Atkinson (2012) exemplify for Pama-Nyungan, the largest Australian language family, Bayesian phylogenetic inference is very useful for clarifying the higher-order structure of less well-researched language families.

2.5.3 Phylogeographic inference

Within the Bayesian paradigm, it becomes possible to also include other elements of language history into models, and then sample from the joint posterior distribution to generate the most likely scenario. This is the framework within which Bouckaert et al. (2012) modeled the expansion of Indo-European. They defined a phylogeographic model which not only includes the tree topology and time

depths for each split, but also assigns a geographical location to every language at every point in time. As observations, the model was given the current geographical ranges of living Indo-European languages in addition to the cognacy overlaps in basic vocabulary already used previously for time depth estimates. The surprising result of their model is that the most likely ancestral homeland of Indo-European is Anatolia, as opposed to the much more widely accepted homeland in the Pontic steppes. This result is found to be stable under various conditions, including the inclusion of cognate data for ancient languages.

A different approach is exemplified by Sicoli & Holton (2014). They attempt to determine the most likely homeland of the recently substantiated Dené-Yeniseian language family proposal, which would connect the Na-Dene languages of Northwest America to the Yeniseian languages of Central Siberia. Working on only 90 typological characters (lexical cognacy is hard to determine due to the high time depth), they compare different tree constraints in a Bayesian phylogenetic tree inference framework, and determine whether one topological constraint leads to a much better fit to the data than others. Finding no support in favor of Yeniseian splitting off before the diversification of Na-Dene, they conclude that the most likely historical scenario is an ancestral homeland of the family in Beringia, from where Na-Dene speakers migrated into North America in two separate waves, whereas the Yenisean languages are a result of a back-migration from the same area. This Out-of-Beringia dispersal is in conflict with previously dominating views assuming that the Yenisean languages branched off during the migration of Dene-Yenisean speakers from Central Asia into the American continent, but fits together well with recent findings of population genetics.

Much more than the results of phylogenetic inference, this type of work is faced with massive criticism by historical linguists, who remain skeptical about the possibility of deciding such difficult questions by mathematical means based on a very small number of datapoints. A major factor is that homeland questions were previously found to be almost inanswerable even given all the available information about the position of ancient languages, grammar, the full lexicon, and many historical facts. This skepticism is enhanced by the fact that the scenarios computed as most likely by such methods naturally tend to deviate from securely established facts in many details, suggesting that these methods might be too optimistic about the reconstructability of events which are essentially historical in nature. Pereltsvaig & Lewis (2015) give a book-length reply to Bouckaert et al.'s claim that the Indo-European homeland debate can be considered settled in favor of the Anatolian hypothesis due to their result. While their criticism against phylogeographic methods might jump to conclusions a little too easily, the book

can still be recommended to any reader who would like to understand why phylogeographic inference, and the way its results were advertised as resolving a century-old question, was so badly received by historical linguists.

2.5.4 Automating the comparative method

Much closer to the heart of mainstream historical linguists, a small subtrend within computational historical linguistics has consisted in attempts to automate parts of the comparative method. Since each stage of the method has its very own heuristics and rules, the development of software tools for these purposes has become a very specialized area where the usual paradigm of applying off-the-shelf bioinformatics software does not lead very far. The decisive advantage of this approach is that it yields results which can be interpreted and evaluated in the trusted framework of historical linguistics. Soon, the field might lead to convenient helper tools which take over much tedious routine work involved in the comparative method, such as looking for possible cognates with non-identical meaning, or mechanically checking whether a hypothesized sound law covers all examples. Good overviews of the current state of the field are provided by Steiner et al. (2011), who present a very ambitious full pipeline for computational historical linguistics which has apparently not been completely realized, and by List (2014), a dissertation which describes the motivation and the design decisions behind LingPy, the most advanced publicly available workbench for computational historical linguistics.

2.5.4.1 Phonetic distance measures

From the computer's perspective, phoneme sequences, whether encoded in a language's orthography or in a unified phonetic format, are initially just sequences made of distinct symbols, none of which is inherently similar to any other. At least as a prefilter for anything that follows, a program for automating the comparative method will need some capability to decide whether two phoneme sequences are broadly similar. In computational systems, this basic intuition is invariably modeled by some string distance measure. This can be as simple as the number of shared bigrams (two-segment substrings), the longest common subsequence, or just a binary distinction where strings are judged as similar if they share the first letter, and as dissimilar otherwise. Kondrak (2005) systematically evaluates how far one can get using some of these simple measures, and achieves surprisingly good results on information retrieval and cognate detection tasks.

One of the most widely used non-trivial string distance measures is the *Levenshtein distance* or *edit distance*, which counts the minimal number of elemen-

tary editing operations (deletions, insertions, or replacements) needed to transform the one string into the other. In its vanilla definition over an alphabet of distinct symbols, this measure is very efficient to compute using dynamic programming. The Levenshtein distance on either the orthography or some coarse-grained sound-class model tends to lead to a workable first approximation to phonetic form distance. Still, the fact that according to the Levenshtein distance, *gown* is as far away from *owl* as from *gun*, might indicate that using the Levenshtein distance will lead to unsatisfactory results in many specific cases.

Typically, the solution is to estimate symbol similarity matrices, and count replacement of similar symbols by only a fraction of a full replacement when computing the edit distance. For instance, when assessing the similarity of English orthographic strings, changing an *o* to a *u* should be much better than changing an *l* to an *n*. This natural extension to the Levenshtein distance leads to the algorithm first presented by Needleman & Wunsch (1970), which maximizes the similarity score between strings by introducing gaps, and filling a dynamic programming table. Variants of the Needleman-Wunsch algorithm are still the preferred method for computing string distances in distance-based phylogenetics. In gene sequence alignment, there are standardized and well-tested similarity matrices which encode current knowledge about the different probabilities for each nucleotide base to turn into a different one due to mutation. Unfortunately, no such standard matrices exist for phonemes, due to the absence of a global inventory of attested sound changes. In practice, this makes the estimation of a new symbol distance matrix necessary for each dataset, and turns the process into a bit of a dark art. Methods which estimate the distance matrix from large amounts of data consistently fare better than attempts to manually encode the intuitions of historical linguists into a matrix, due to the impossibility for a human to assign an intuitive meaning to the distance weights.

2.5.4.2 Phoneme sequence alignment

Any method which uses dynamic programming to compute some minimal edit distance implicitly constructs an alignment, i.e. a separation of the two or more aligned strings into columns of equivalent segments. A binary alignment specifies which phonemes are cognate in a pair of cognate words. For automated methods, the columns of each alignment provide sound correspondence candidates, which can be counted and correlated to build models of phoneme distances. Binary alignments can also be joined into multiple alignments of entire cognate sets in order to extract multi-way sound correspondences.

The optimal way to align phoneme sequences is still an active area of research, and no single approach has so far materialized as being the best across

datasets. Depending on the language family, the trivial alignment (of identical positions, without assuming any gaps) might work just as well as gappy alignments, and whether vowels match might be almost irrelevant (in Semitic) or crucial (in Uralic).

As part of the data preprocessing, a variant of my previously published binary alignment method called Information-Weighted Sequence Alignment (IWSA) is introduced in §4.3. The information weighting uses language-specific trigram models to weight the phonemes by relevance, assigning a higher penalty to mismatching segments in high-information phonemes. This helps to detect partial cognacy, and avoids some of the skew introduced when comparing dictionary forms e.g. due to shared infinitive endings.

2.5.4.3 Sound correspondence models

In the same ways that global sound distance matrices are estimated, it is possible to infer sound distances for any pair of languages. These will tend to assign low costs to sound pairs which are equivalent across many alignments, and can therefore be interpreted as encoding some of the sound correspondences the comparative method operates with. For instance, given enough examples such as *water*/*Wasser*, *street*/*Straße*, and *foot*/*Fuß*, the alignment costs of English [t] and German [s] will be rather low, encoding the consequence of a part of the High German consonant shift. Since programs for inferring and modeling sound correspondences are part of the toolchain I am using for the evaluation, existing methods for performing this task are covered in §4.4.

2.5.4.4 Cognate and loanword detection

In principle, it would be possible to use inferred sound correspondences in a symbolic way, and to implement the logical criteria as they are applied in historical linguistics to separate cognates from loanwords. The problem with this approach is that it requires very clean decisions on the valid sound correspondences, so that word pairs which violate some of these correspondences could be sifted out as loans. Also, the computational models of sound correspondences are not sufficiently fine-grained to model contexts in ways that are explicit enough to represent sound correspondences as exceptionless.

Another huge problem for better models is data sparseness. Even in classical historical linguistics, where detail questions can often be resolved by means of many other sources of knowledge such as early attestations or historical and cultural knowledge, sound correspondences between more distant languages can

often barely be established, even if the entire lexicons of two languages are taken as raw material for possible cognate pairs.

These reasons make it unfeasible in principle to apply these criteria to lexicostatistical databases, even the largest of which will only cover a portion of the relevant basic vocabulary. Current systems therefore remain in the probabilistic framework, typically applying some clustering algorithm to the distances of language-specific realizations in order to group them into cognate classes. Since an automated cognate detection module again forms a major part of the work presented in this book, an overview of possible approaches to this task is given in §4.5.

2.5.4.5 Automated reconstruction

To have a computer prove language relationship according to the standards of historical linguistics, we would need software which reconstructs the unattested common ancestor language of the claimed genetic unit, and demonstrates how the attested forms can be derived from them as the result of a series of sound laws. This is the task which the earliest work in computational historical linguistics has attempted to automatize, and having tools for reliably inferring ancestral strings remains a highly attractive goal. If an automated tool successfully reconstructed a language attested in ancient texts, or a proto-language which all experts in the respective language family find convincing, based on its modern descendant languages, this would be a very convincing argument in favor of computational methods.

Given the centrality of reconstruction to the comparative method, it is surprising how little work has been done in the area. What makes the task very difficult is the conditional nature of sound changes, and that it is difficult to model which sound changes are plausible and which are not. Moreover, the combinatorial problems involved in determining in which order sound changes applied, do not disappear when attempting to automate the task. Rule-based systems such as the ones presented by Hewson (1974) and Oakes (2000) tried to stick as closely as possible to the way in which sound laws are tested in historical linguistics. These approaches ran into the problem that for reconstruction tasks involving more than a handful of languages, there will frequently be no consistent solution, because some language will invariably display some changes which cannot be captured by the system's logic. There will always be some exceptional behavior like analogy modifying some words, which a historical linguist will be aware of and can decide to make part of the explanation, but which an automated method will not detect.

Like for the other tasks, progress towards workable systems has only been possible by modeling the problem probabilistically. As the endpoint of a series of papers working towards such a model, Bouchard-Côté et al. (2013) present a fully Bayesian framework for reconstructing ancestral wordforms, and evaluate it on the ABVD database. In their eyes, an average normalized Levenshtein distance of 0.125 per wordform (i.e. getting about seven out of eight phonemes right) between the automated reconstruction and expert reconstructions of Proto-Austronesian is a great success. The model they are describing is certainly much more linguistically informed than an earlier Bayesian model by Ellison (2007), and a separate evaluation on the Oceanic subgroup shows that the system's reconstructions differ only twice as much from two expert reconstructions than these differ among each other. However, given that Austronesian is widely considered one of the easier cases for reconstruction due to a very simple phonology, and that a wrongly reconstructed vowel in a single word is sometimes problematic enough in classical historical linguistics to make or break an entire theory, the state of the art in automated reconstruction is still very far from a convincing implementation of the last step of the comparative method. By and large, automated reconstruction can still be considered an open problem.

2.5.5 On the road towards network models

The context of the approach I am exploring in this book takes us back to the phylogenetic inference problem. The crucial last step is that instead of treating conflicting signals in distance or character data as mere noise that makes it more difficult to infer the true tree, it is also possible to see some of the conflicting signal as caused by legitimate secondary connections. To get access to these connections, an obvious idea is to expand the tree model in order to explicitly represent relevant lateral signals. The idea of adding additional reticulation edges to a tree in order to do so, leads to the concept of a phylogenetic network. Huson & Bryant (2006) give an overview of the most common types of phylogenetic networks, also performing a much-needed clarification of terminology which resolves some of the conceptual confusion plaguing this area. A slower-paced introduction to the topic is the more recent book by Morrison (2011), which also serves as my main source for the following overview of the most important types of networks.

The notion of a phylogenetic network actually subsumes two very different types of networks which have not always been kept separate due to the use of a single vague term. Morrison advocates the usage of the term *data-display network* for undirected networks which simply visualize the conflicting signal in

phylogenetic tree inference by means of additional virtual nodes and undirected edges which represent multiple alternative subgroupings. In contrast, the term *evolutionary network* is reserved for types of directed acyclic graphs that are true generalizations of directed rooted trees, where reconstructions of lateral signals (such as lexical admixtures) are represented explicitly by directed secondary arrows connecting nodes in the tree.

2.5.5.1 Data-display networks

Most popular types of data-display networks can be subsumed under the term of a *splits graph*. The defining property of a splits graph is that each edge represents a bipartition of the leaves according to some criterion. Intuitively, edges therefore visualize separability between clusters of nodes. An incompatible pair of bipartitions defines a reticulation, which is visualized by copying both edges, and drawing them into a parallelogram. Drawing these parallelograms will introduce some additional nodes which do not reflect any reconstructed common ancestor, which means that edges cannot be interpreted as representing actual evolutionary events. If there is too much conflicting information, or the data are insufficient to decide on a binary structure, some nodes will be unresolved and visualized as polychotomies (more than binary branching). Split graphs make it easy to get a first impression of the amount and location of conflicting signal for the phylogenetic tree inference task, but do not lend themselves well to any purpose beyond exploratory data analysis.

Many algorithms for computing splits graphs from various types of data have been proposed in the literature, which differ in the number of conflicting splits that will be visualized at the same time, and in the way possible splits are defined. I will only mention the two most popular types here, and refer to Morrison (2011: Section 3.3.1) for more variants.

For data that comes in the shape of binary characters, the prototypical type of data-display network is the *median network*. In the most basic variant, every pair of conflicting splits leads to a duplication of the edges crossing that split, which can lead to exponentially many virtual nodes, and a very high-dimensional structure that is difficult to visualize. Various modifications have been developed with the goal of systematically excluding some less well-attested splits from the visualization. The most commonly used variant, the *reduced median network*, isolates the highly conflicting characters in a preprocessing step, and generates more compatible replacement data as the basis for the median network computation.

Among the network types which can be extracted from distance data, the most popular type of splits graph is called a *neighbor-net*. The process for building a

neighbor-net is similar to tree construction by neighbor joining. Going through pairwise distance scores from the shortest distance to the largest, a pair of nodes is linked at each step, and represented by the subsequent steps as a single virtual node whose distance to the other nodes is computed as the average of the individual distances. Unlike in neighbor-joining, each linked node remains eligible for additional linking until it is linked a second time to a different node. This procedure makes a planar representation possible, however complex the conflicting signal may be. Always selecting the closest pair of unlinked nodes ensures that the strongest conflicts are the ones that get visualized. While neighbor-nets are becoming the most popular type of data display networks due to their readability, one must be aware of the fact that some reticulations might not be supported by actual conflicts in character data, but result from the information loss incurred while summarizing the data in the form of a distance matrix. In contrast, reticulations in a median network always reflect conflicting information from at least one pair of characters.

The second group of data-display networks apart from splits graphs are *parsimony networks* , where each edges represent a link in some maximum-parsimony tree, and the edge length is directly interpretable as the number of character changes according to the maximum parsimony criterion. Parsimony networks are constructed from collections of maximum-parsimony trees, which makes them very costly to compute for larger datasets, and very prone to inaccuracies in the data. This makes them less adequate for often quite noisy linguistic data, and I am only mentioning them here to indicate that types of data-display networks other than splits networks exist. For an overview of different parsimony network inference methods, and pointers to the relevant literature, the reader is referred to Morrison (2011: Section 3.3.1). Morrison also discusses other types of data-display networks which aggregate trees generated from different parts of the data, and are not generated directly from entire character-encoded or distance-matrix data sets.

2.5.5.2 Evolutionary networks

In contrast to data-display networks, internal nodes in evolutionary networks actually correspond to inferred ancestor species, and lateral connections in a good network should correspond to actual instances of lateral contact. These networks obviously contain much more specific information than data-display networks, and inferring them is a very worthwhile, but much more difficult task. Many ways of inferring such networks have been proposed in the literature, but according to Morrison (2011), most algorithms are not performant enough to be

applied to interesting datasets, and not a single one exists in a readily usable implementation that could make it a standard tool in bioinformatics. What is more, existing methods have tended to make unrealistic simplifying assumptions in order to keep algorithms tractable, without much emphasis on asking whether the assumptions reflect properties of the respective problem.

Existing methods have generally prioritized computability by restricting the search space to classes of evolutionary networks with some additional constraints. One obvious idea is to limit the number of reticulations in the network, for example by allowing a maximum number of horizontal connections in the entire tree. This makes very fast inference possible, but is of limited use because it introduces knowledge of the same type which one would actually hope to retrieve from a network. Less ad-hoc constraints involve the notion of a *reticulation cycle*, which is defined as a configuration of two separate directed paths which start in a single node and meet again in another node. An instance of this would consist of directed paths from Proto-Germanic into Old Norse, from Proto-Germanic into West Germanic, from West Germanic into English (all due to inheritance), and a directed arc from Old Norse into English (via horizontal transfer).

To put a bound on the amount of reticulation, a very basic approach is to prevent any nodes from being shared between cycles. Among other things, this implies that every node can only be involved in a single horizontal transfer event. This rather strong constraint defines *galled trees*, which have much nicer mathematical properties than general evolutionary networks, and inference of which is tractable for substantial numbers of nodes. However, this condition makes galled trees inadequate for our linguistic application, as they cannot model e.g. the transfer of English loanwords into both German and Hindi.

Slightly more generally, *galled networks* allow reticulation cycles to share reticulation nodes, i.e. allow one node to serve as a source of horizontal information flow into more than two other nodes, but still prevent any overlap in other nodes. Even though it is weaker, this constraint still implies that each node can only receive information flow from at most two nodes, i.e. from its ancestor and one additional node. Thus, a galled network cannot model facts such as that English received loanwords from both Norman French and Old Norse.

Willems et al. (2014) describe an algorithm for inferring the slightly more general class of hybridization networks from a distance matrix according to a generalization of the Neighbor-Joining principle. The method detects some nodes in the neighbor-joining tree as containing conflicting affiliations, and allows each such conflict node to be modeled as a hybrid of two contributors of lexical material, either by horizontal transfer or inheritance. In the case where two sources of

horizontal transfer are inferred (true hybridization), one of them can receive an alternative interpretation as the influence of a close common ancestor, leading to a maximum of three incoming connections being visualized. The possibility for leaves to become source languages moves the generality of representable evolutionary histories only slightly beyond the level of galled trees. The advantage is that the models yields a quantitative estimate of the contribution by each of the two source taxa for each hybrid, and can designate one of the two sources as the likely ancestor, and the other as the contributor of a lateral signal.

2.5.5.3 Network models in historical linguistics

Unlike phylogenetic trees, phylogenetic networks have only started to be applied to linguistic data. Most existing work is based on data-display networks, which are mainly used to visualize the places in language trees where conflicting phylogenetic signals exist, and typically interpreting these as representing dialect continua as opposed to tree-like dispersal. A case in point is Lehtinen et al. (2014), who compute splits trees over the UraLex database, and interpret the results as showing dialect continua within the westernmost branches of Uralic.

List et al. (2014) have some success in inferring minimal lateral networks from cognacy data. Minimal lateral networks are a less mainstream type of recombination network where all lateral connections are implictly assumed to have their endpoints in living languages, whereas the startpoints can be any internal node in a guide tree. This makes the computational problem a lot easier, and highlights the parts of the tree among which there is a strong lateral signal, but the results cannot be interpreted directly as encoding contact events.

Recently, Willems et al. (2016) compared how splits graph, galled network, and their own hybridization network algorithms perform on a single set of cognacy data. They argue that the network models successfully identify donors and recipients of lexical material, and make it possible to quantify the degree of influence between languages. Unfortunately, their only evaluation is on a heavily modified version of the IELex database, albeit with a lot of interesting detail. Discussing and comparing the results of the three algorithms on several genera, the hybridization networks consistently delivered a clearer picture, often deciding correctly which of the two sources contributed the horizontal signal, and fewer erroneous hybrids than the galled network. This difference might be due to an unfair comparison, because the galled networks frequently represented a language to be a mixture of more than two languages, which is not surprising given their higher generality. On larger datasets from other linguistic regions, contributions from more than two source languages might actually be desirable. Unsurprisingly

given their status as mere data display networks, the splits networks turned out to be barely interpretable, as they included far too many lateral connections.

2.6 The lexical flow inference task

In this section, I wrap up my overview of existing models in computational historical linguistics by defining the two tasks I am attempting to solve in this book, and putting them into the context of the overall methodological landscape. The first task, phylogenetic lexical flow inference, can be conceived as fully general evolutionary network inference without branch lengths. The second task, contact flow inference, focuses on determining historical contacts between attested languages, and only treats proto-languages as hidden common causes for subsets of the observed languages.

2.6.1 Phylogenetic lexical flow

Every rooted phylogenetic network can be analyzed as containing a tree topology and additional lateral connections cross-cutting the tree structure, which cause some nodes to have multiple incoming arrows. If we assume that major lexical influences will typically be monodirectional (as justified by my overview of current knowledge about language contact), assigning a direction to each lateral connection, as is the case in a directed acyclic graph, is an obvious modeling decision. Moreover, as we have seen in the discussion of existing phylogenetic network methods, it is desirable for a network to support any number of incoming and outgoing edges for leaves as well as ancestral nodes, making it possible to represent the very complex evolutionary scenarios which actually occurred in the history of human language.

Quite naturally, this leads to the task of inferring a completely unconstrained directed acylic graph (DAG) over the attested languages of a region, and their reconstructed ancestors according to some inferred tree. The phylogenetic lexical flow inference task which I am trying to solve is thus a lot more challenging than any type of evolutionary network that has so far been inferred over a moderately large number of languages. I hope to show that causal inference at least provides a starting point for addressing this task.

2.6.2 Contact flow

In order to make the problem a little simpler for possible inference algorithms, and not to build too much on unreliable estimates of cognate set presence in an-

cestral languages, we can confine ourselves to just inferring which languages are related by inheritance, and just inferring a model of lexical flow among the observed languages. This is mirrored in some sense by the way in which the history of languages is commonly described. For instance, if we want to summarize the history of the English lexicon in a few sentences, we would typically not refer to previous stages of the languages involved, speaking of Norman French influences on Old English. Instead, the much more common way of expressing this is to simply talk about French influence on English, although the actual process happened between ancestors of the two languages, which are quite arbitrarily called older versions of the modern national languages, as if Latin were called Old Italian.

In this vein, we would expect an automated method which analyses the basic lexicon of modern European languages to infer a relation pair *fra* $\longrightarrow$ *eng*, i.e. largely monodirectional influence of French on English. Similar well-known pairs in Europe would include *swe* $\longrightarrow$ *fin* (Swedish influence on Finnish) and *deu* $\longrightarrow$ *lav* (German influence on Latvian).

So how can we define a correct and complete contact flow network among a set of living languages, in which the earlier stages at which the relevant contacts actually occurred are not explicitly represented? One possibility is to define the contact flow network in such a way that a flow involving at least one proto-language can be represented as the corresponding flow involving any of the descendant languages. For instance, Slavic influence on Romanian might be represented by *bul* $\longrightarrow$ *ron* (Bulgarian on Romanian), by *hrv* $\longrightarrow$ *ron* (Croatian on Romanian), or both. Given the quantity of Slavic loans even in the basic vocabulary of Romanian, we might also want to consider as an error the absence of any such incoming arrow from a Slavic language. In contrast, if a pair of languages shares lexical material only due to common ancestry, this may be represented by a second type of edge, which will be representing by undirected arrows in my contact flow visualizations.

2.7 The adequacy of models of language history

After the short overview of computational models of language history and the place of my own models in this landscape, a good way to conclude the chapter might be to reconsider the relation of these models with linguistic reality from a more distant point of view.

In some respect, every computational model is necessarily at odds with the ways in which language change is known to occur in reality, even if we only concern ourselves with the contents of their lexica. All the models I described

quite literally assume that languages are "bags of words" to which new words get added, and from which other words are removed, as the language transitions across its different historical stages. Even if we buy into this model, and are ready to abstract over the fact that the disappearance of a word will never be a sudden event, but an abstraction over a process of declining usage, we are faced with an additional need for abstraction when mapping the history of a language. The quite common separation into discrete stages such as Old High German or Middle English already constitutes a very far-reaching and unnatural simplification given that the changes between the variants have obviously not all occurred at the same time. Some of the more advanced methods do yield probability distributions over reconstructed states at arbitrary points of time, but these distributions often only reveal just how uncertain we are about the exact order in which words appeared in or disappeared from an attested language. Computationally more tractable methods will tend to restrict themselves to inferring a limited number of intermediate historical states, most commonly the stages just before each split occurred in a tree model.

Even in the rare case where we can rely on written sources to get a glimpse of a historical variety, deciding whether some word existed in the language at the given point in time can be a difficult question. An attested occurrence of a word does not necessarily mean that it had currency within the language at the time of writing, it might well be that a it was merely used to evoke an impression of ancientness, or even only for humorous effect (e.g. when a Shakespearean phrase like 'methinks' is used in contemporary English). On the other hand, a word that is missing might simply not be attested due the small corpus size, and we should not infer anything from its absence. For instance, the Hittites quite plausibly had a word for 'to sneeze', even if none of their surviving texts might use it.

Ultimately, already the concept of a language is an abstraction over the manifold variants and subsets actually represented in the brains of and used by its speakers. For professional reasons, my personal variant of German includes quite a few English words that should very likely not be considered loanwords in 'German', although in my personal variant, they certainly are. It is important to keep in mind that reducing the languages involved to discrete and uniform units in order to model the connections between them in terms of a discrete graph structure, be it a phylogenetic tree or a more general structure like a lexical flow network, amounts to a significant leap of abstraction.

While these abstractions over speaker variation as well as geographical and temporal variants might be seen as illegitimate simplifications by some, it is important to bear in mind that classical historical linguistics routinely builds on very similar abstractions. Generalizing over dialects in order to focus on the

deeper questions of ancestry is an obvious choice given the field's primary interest, but even temporal developments occurring across centuries are typically only discussed if necessary for an argument. A typical etymological dictionary of a European language will describe most Latin elements in the dictionary just as borrowings from "Latin" instead of differentiating between Medieval and Classical Latin, and only mention that the source was Medieval Latin if the word in question is not attested in Latin texts from classical antiquity. In many cases, the other words were technically borrowed from Medieval Latin as well, but this is not specified if the word already existed in the classical language. Therefore, the fiction of uniform languages that can be treated as elementary units does not only form the basis of computational models, but has always been one of the central abstractions of historical linguistics.

The situation is quite different in dialectology and philology, where the differences between the variants used in different places or by different authors tends to be the focus of interest. Especially in dialectology, wave models have always been much more popular than in historical linguistics, because on the microlevel, the wave-like nature of spreading innovations becomes much more clearly visible. In the timeframe of historical linguistics, thousands of years after a change, when only one of the dialects might be attested, or a reconstruction of a single proto-language abstracts over dialectal differences, the result of a complex pattern of shared or isolated innovations will tend to look like a split, giving credibility to the tree metaphor. But even here, phylogenetic tree inference is quite commonly faced with the problem that the phylogenetic signal becomes less tree-like if the language sample includes dialects which were recently in contact, and in quite many language families, the difficulty in deciding which daughter language split off first may be due to the fact that the changes distinguishing the subfamilies were wave-like.

While some level of abstraction is necessary to see the major trends, models differ in their foci on different aspects of linguistic reality that they are primarily interested in. In phylogenetic inference, lateral transmission due to language contact is still mostly seen as disturbing noise that makes it more difficult to infer reliable trees. I would argue that models such as lexical flow models in which contact is not seen as irrelevant noise, but as a complementary force shaping language history that holds interest in its own right, as has been the perspective of classical historical linguistics, are inherently more interesting than even the most advanced tree models.

Phylogenetic flow networks and contact flow networks differ in the importance they assign to the idea of common inheritance. While a phylogenetic flow

network can explicitly model inheritance and borrowing, i.e. both of the primary forces shaping the lexicon, a contact flow network is much more similar in spirit to a wave model. Given what we know about the strengths and weaknesses of tree and wave models, if a contact flow model fits a dataset particulary well, we are very likely dealing with a set of neighboring languages through which innovations have tended to spread in waves. On the other hand, a phylogenetic flow model that only adds some directed lateral connections to a tree model will be a better fit for data covering a set of languages which became differentiated due to geographic separation, with only sporadic interaction at some points in history.

Finally, while I have tried to argue in this chapter that we can expect most instances of language contact to have a strong directional component, I should not conclude without acknowledging that this is another simplifying assumption which might be difficult to defend in some situations. Especially when we collapse the contacts which have happened during different historical stages, like in a contact flow network, it will sometimes have been the case that the main direction of borrowing between a pair of languages has shifted. An actual example of this would be the interaction between French and Dutch. French is differentiated from other Romance languages by a layer of Frankish loanwords which entered the language during the Old French period. In a contact flow network over living languages, where Dutch is likely to be the closest living relative of Frankish, this implies that we would want to infer an arrow from Dutch to French. But of course, much later Dutch then borrowed a substantial amount of cultural vocabulary from French, which should be reflected by an arrow in the opposite direction. Since we do not distinguish between historical stages in a contact flow network, this would be equivalent to connecting both languages by a bidirectional arrow.

Even if as in this book, we build on a framework that is able to infer such bidirectional links, it seems quite reductionist to be content with this summary when the true story is much more complex and interesting. Still, such a lexical flow model would arguably model the linguistic reality much better than a phylogenetic tree where both languages should be inferred to belong to different branches, and their lateral interaction will typically only lead to uncertainty about the internal structure of both branches. Tree models and network models occupy different positions in the trade-off between model adequacy and ease of inference, and as we are going to see in this book, striving for more adequacy inevitably comes at a cost to computational efficiency and reliability of results. Adding more detail, and coming even closer to inferring a realistic model of language history, will lead to even more challenging inference problems.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vajda, Edward J. & Andrey Nefedov. 2009. Ket. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database.* Leipzig: Max Planck Institute for Evolu-

tionary Anthropology. http://wold.clld.org/vocabulary/18 (Last accessed 2019-06-09.)
van der Sijs, Nicoline. 2009. Dutch. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/12 (Last accessed 2019-06-09.)
van Hout, Roeland & Pieter Muysken. 1994. Modeling lexical borrowability. *Language Variation and Change* 6(1). 39–62.
Vejdemo, Susanne & Thomas Hörberg. 2016. Semantic factors predict the rate of lexical replacement of content words. *PLOS ONE* 11(1). 1–15.
Viires, Ants & Lauri Vahtre. 1993. *The red book of the peoples of the Russian empire*. Tallinn. http://www.eki.ee/books/redbook (Last accessed 2019-06-10.)
Viitso, Tiit-Rein. 1998. Fennic. In Daniel M. Abondolo (ed.), *The Uralic languages* (Language Family Descriptions Series), 96–114. London: Routledge.
Volodin, A. P. & K. N. Halojmova. 1989. *Slovar' itel'mensko-russkij i russko-itel'menskij*. Leningrad: Prosveščenie.
Volodin, A. P. & P. J. Skorik. 1997. Čukotskij jazyk. In A. P. Volodin, N. B. Vaxtin & A. A. Kibrik (eds.), *Jazyki mira: Paleoaziatskie jazyki*, 23–39. Moskva: Indrik.
Wells, John C. 1995. *Computer-coding the IPA: A proposed extension of SAMPA*. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm (Last accessed 2019-06-10.)
Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. *The ASJP database (version 17)*. http://asjp.clld.org/ (Accessed 2017-05-22.)
Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2018. *The ASJP database (version 18)*. http://asjp.clld.org/ (Accessed 2019-06-10.)
Wichmann, Søren & Jan Wohlgemuth. 2008. Loan verbs in a typological perspective. In Thomas Stolz, Dik Bakker & Rosa Salas Palomo (eds.), *Aspects of language contact*, 89–122. Berlin: Mouton de Gruyter.
Wiebusch, Thekla. 2009. Mandarin Chinese. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/vocabulary/22 (Last accessed 2019-06-09.)
Willems, Matthieu, Etienne Lord, Louise Laforest, Gilbert Labelle, François-Joseph Lapointe, Anna Maria Di Sciullo & Vladimir Makarenkov. 2016. Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16(1). 180.
Willems, Matthieu, Nadia Tahiri & Vladimir Makarenkov. 2014. A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology* 12(05). 1450024.

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

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

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

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

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

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

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

Name index

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Oakes, Michael P., 39, 115

Pagel, Mark, 175

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

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

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

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

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

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

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

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

Language index

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Veps, 138, 141, 142, 236

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

Xibo, 155

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

Subject index

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

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

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

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

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

faithfulness, 68

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

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

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

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

joint entropy, 74
joint reconstruction, 220

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

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

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

outlier, 34

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

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

Samoyedic languages, 147

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

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

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

v-structure, 66

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

X-SAMPA encoding, 99

Information-theoretic causal inference of lexical flow

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

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

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

ISBN 978-3-96110-143-6