EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
LANGUAGE EVOLUTION:
THE EMPIRICAL TURN

Information-Theoretic Causal Inference of Lexical Flow

Frankfurt, May 3, 2021

Johannes Dellert

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

Table of Contents

Historical Linguistics

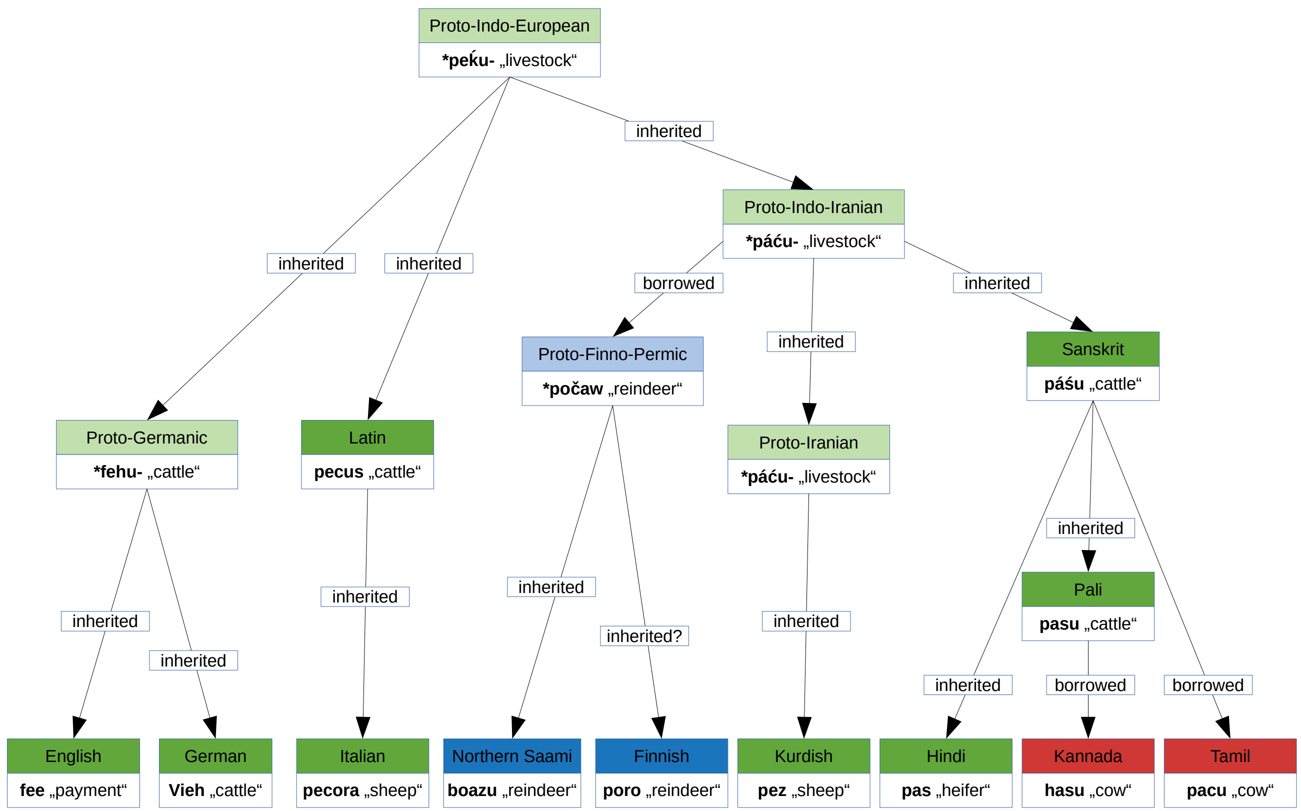
Subject of **historical linguistics**:

- understanding the historical development of languages
  - ▷ Which groups of languages have a common ancestor?
  - ▷ Within such families, which ones are more closely related?
- one of three important windows into the prehistoric past, complementing archaeology and genetics

Goals of classical historical linguistics:

- determine the origin of as many words as possible (**etymology**)
- reconstruct the lexicon and the grammar of unattested common ancestors of attested languages (e.g. Proto-Germanic, Proto-Indo-European)
- provide parsimonious explanations of how the attested languages developed from these reconstructed ancestors

Etymologies: Example



Computational vs. Classical Approaches

Classical methods:

- very successful tradition since early 19th century
- take all the available evidence into account
- ideally results in consistent theories which explain large parts of each language's lexicon and grammar
- not fully formalizable
- problems if there is conflicting evidence
- many interesting questions (e.g. dating) beyond scope

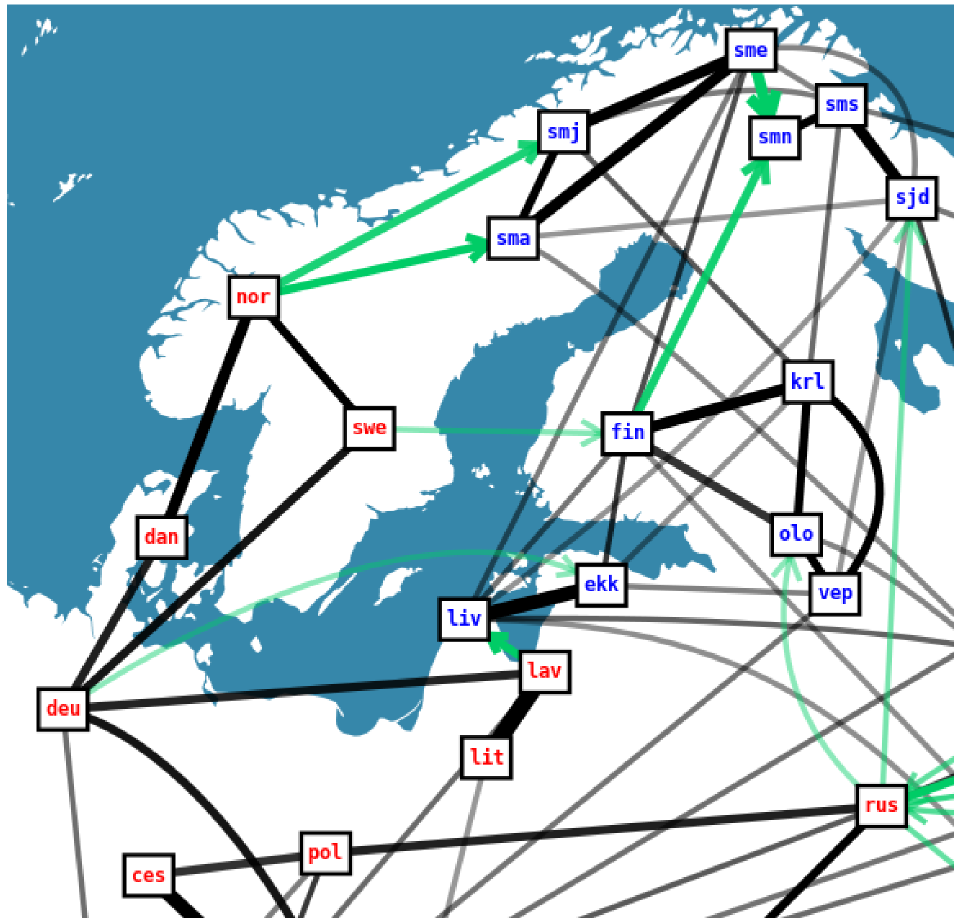
Computational methods:

- new field, largely based on bioinformatics (since 1990)
- work with a small, carefully sampled subset of the data
- ideally help to decide long-standing open questions by providing a framework for dealing with uncertainty
- based on mathematical models of evolution
- evidence is quantifiable, but difficult to interpret
- studies often contradictory

Lexical Flow Inference (LFI)

from this ($\times$ 1,016), algorithmically derive this:

ces	*husa*	[ɦusa]
dan	*gås*	[gɒːˀs]
deu	*Gans*	[gans]
ekk	*hani*	[hɑnʲi]
fin	*hanhi*	[hɑnhi]
krl	*hanhi*	[hɑnhi]
lav	*zoss*	[zuɔ̯sː]
lit	*žąsis*	[ʒɑːsʲɪ̂s]
liv	*gūogõz*	[guːogɨz]
nor	*gås*	[goːs]
olo	*hanhi*	[hɑnhi]
pol	*gęś*	[gɛ̃ɕ]
rus	*гусь*	[guˑsʲ]
sjd	*чуэнь*	[t͡ʃueɲ]
sma	*gaase*	[kɑːsɛ]
sme	*čuonji*	[t͡ʃu̯ɔɲi]
smj	*gássa*	[gasːɑ]
smn	*čuá'njá*	[t͡ʃu̯æɲæ]
sms	*čue´nj*	[t͡ʃuɛɲʲə]
swe	*gås*	[goːs]
vep	*hanh'*	[hɑnhʲ]



Phylogenetic Lexical Flow Inference

A map of the linguistic history of a region should include

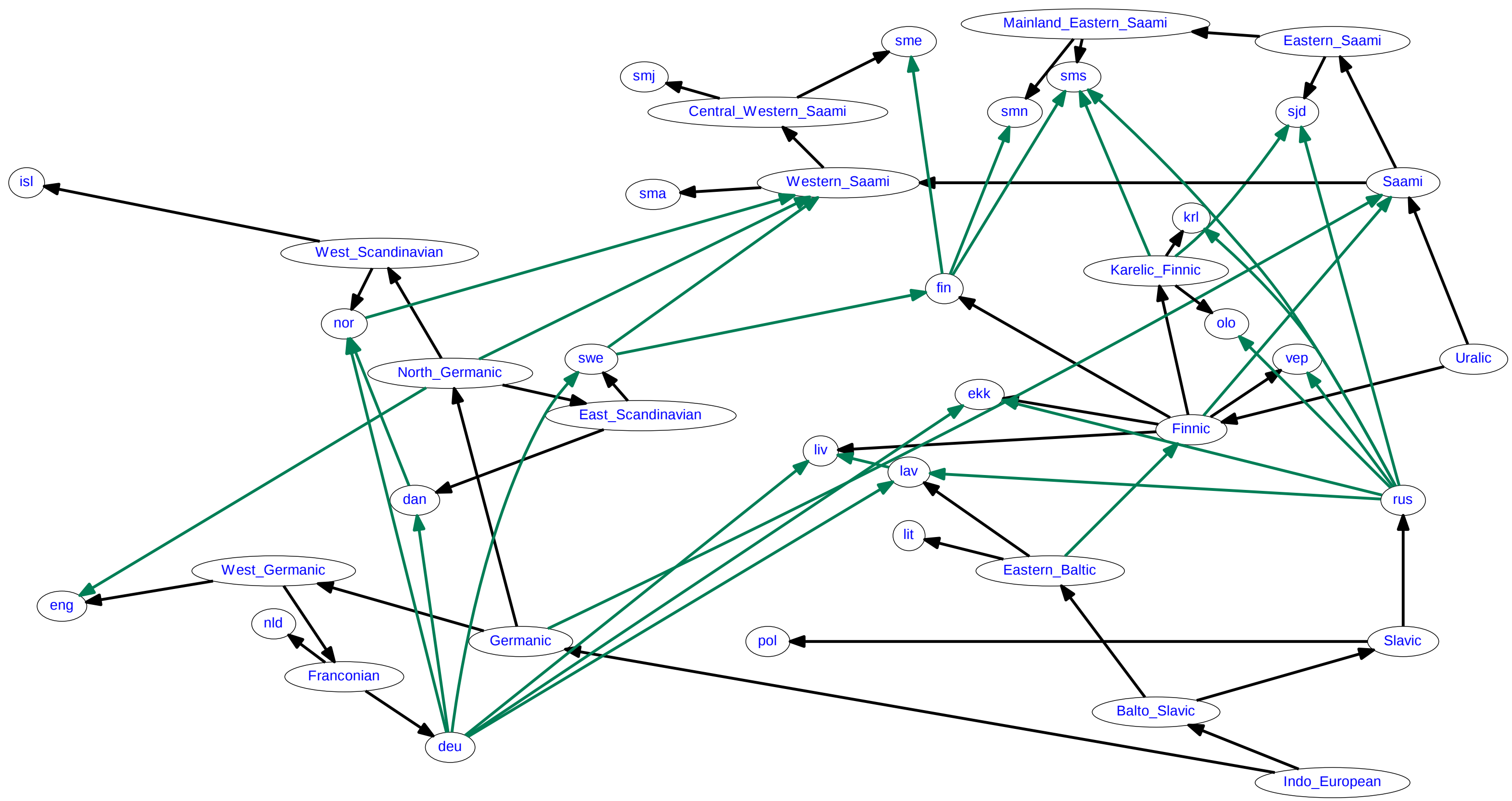
- the paths on which lexical material was inherited (i.e. a phylogenetic tree)
- the paths on which lexical material was borrowed (both among ancestral and living languages)
- taken together, all the paths on which lexical material has "flown" to produce the observable situation (**lexical flow**)

Simplifying assumptions taken in my approach:

- some phylogenetic tree is known (good inference methods exist)
- we rely on reconstructions of the homologue sets present at each proto-language (derived by historical linguists, or using some automated reconstruction method), and treat them as if we observed the data

Phylogenetic Lexical Flow Inference: Example

Desired result for the basic lexicon around the Baltic Sea:



Steps of my PLFI workflow

- infer information models from word forms
- infer sound correspondence models from word forms
- compute optimal pairwise alignments of word forms
- compute form distances
- cluster forms into homologue sets (“cognates”)
- use phylogenetic tree inference to build tree skeleton
- reconstruct status of homologue sets for proto-languages
- make it possible to run conditional independence tests between sets of languages based on homologue overlaps
- infer the causal skeleton (parsimonious contact model)
- infer dominant directionality of lexical flow on the skeleton

Existing Phylogenetic Network Methods

Morrison (2011): two main types of phylogenetic network

- **data-display networks**
  - ▷ generalize unrooted trees
  - ▷ use additional virtual nodes to visualize conflicting signals
  - ▷ examples: median network, neighbor-net
- **evolutionary networks**
  - ▷ generalize rooted trees
  - ▷ all nodes represent some (ancestral) language
  - ▷ lateral connections are directed
  - ▷ examples: galled tree, galled network, hybridization network

Existing Phylogenetic Network Methods

Evolutionary network inference is still in its infancy:

- **probabilistic models** are very complex and need a lot of strong modeling assumptions; inference methods do not scale well to large networks, 7 species is the limit hit by Wen et al. (2016)
- models for more languages restrict the search space rather heavily, usually in terms of reticulation cycles
- **galled trees** do not allow node sharing between reticulation cycles ($\Rightarrow$ multiple donor languages not possible)
- **galled networks** allow reticulation cycles to share nodes, but only reticulation nodes, i.e. multi-way colliders are possible (BUT deu $\leftarrow$ eng $\rightarrow$ hin still not representable)
- **hybridization networks** are only slightly more general (they allow leaves as source languages)

Table of Contents

Information Content and Importance Weighting

Let c_{abc}, c_{abX}, c_{Xbc}, c_{aXc} be trigram and extended bigram counts, then **information content** of a segment *c* in its context *abcde* is

$$I(abcde) := 1 - \max\left\{\frac{c_{abc}}{c_{abX}}, \frac{c_{bcd}}{c_{bXd}}, \frac{c_{cde}}{c_{Xde}}\right\} \quad (1)$$

A (smoothed) information content model derived from 1,000 dictionary forms allows us to focus on the distinctive phonemes:

FRA	BARK	aboyer	/abwaje/	/a b w a j e /
SPA	DRINK	beber	/beβer/	/b e β e r /
TUR	COVER	kaplamak	/kaplamak/	/k a p l a m a k /

This is mainly useful as a way of reducing words to their stems!

Information-Weighted Sequence Alignment

Modify alignment to using the following distance measure:

$$M(i,j) := M(i-1,j-1) + d(a_i,b_j) \cdot s(a_i,b_j), \tag{2}$$

where $d(a_i, b_j)$ is the phoneme distance inferred from the data by PMI, and the **combined information content** $s(a_i, b_j)$ is the quadratic mean of both information content scores:

$$s(a_i,b_j) := \sqrt{\frac{I(a_{i-2}\ldots a_{i+2})^2 + I(b_{j-2}\ldots b_{j+2})^2}{2}} \tag{3}$$

IWSA: Examples

g	ə	f	ʁ	i	i	ʁ	ə	n	DEU: gefrieren FREEZE	0.361784
-	-	f	ɹ	i	i	z	-	-	ENG: freeze FREEZE	

θ	a	l	-	d͡ʒ	ARB: θaldž SNOW	0.312547
ʃ	ɛ	l	ɛ	g	HEB: šeleg SNOW	

Figure: Visualizations of IWSA for two word pairs

Phoneme Distances

- phoneme distances are inferred via pointwise mutual information (PMI), where the expected distribution is estimated via resampling in the tradition of Kessler (2001):

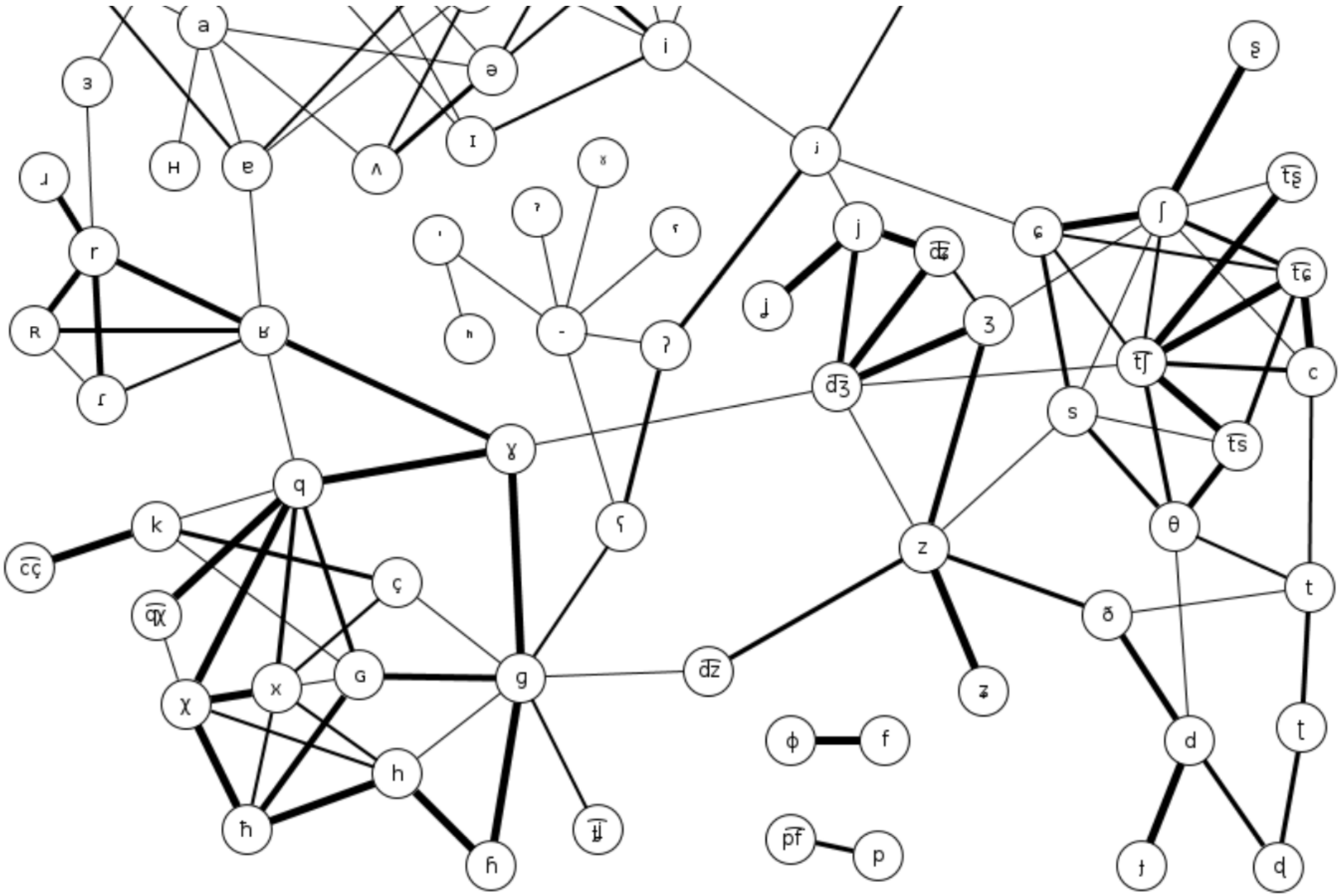
$$w_{glo}(x, y) := \log \frac{p(x, y)}{\hat{p}(x, y)}$$

- in the information-weighted case, both $p(x, y)$ and $\hat{p}(x, y)$ are based on **weighted counts**:

$$c(x, y) := \sum_{L_1, L_2 \in \mathcal{L}} \sum_{\substack{(a,b) \in lex(L_a, L_b), \\ sc(a,b) < 1.2}} \sum_{\substack{1 \leq i \leq \max\{m,n\}, \\ al(a,b).a_i = x, \\ al(a,b).b_i = y}} I^2_{L_a, L_b}(a_i, b_i)$$

- global PMI scores based on 1.3M potential homologue pairs from NorthEuraLex 0.9, and equal number of random word pairs

Global Phoneme Distances: Visualization



Sound Correspondences

- local PMI scores (inferred from the data for a single language pair) to represent some of the sound correspondences:

$$w_{L_1,L_2}(x,y) := \frac{w_{glo}(x,y) + \log \frac{p_{L_1,L_2}(x,y)}{\hat{p}_{L_1,L_2}(x,y)}}{2}$$

- $p_{L_1,L_2}(x,y)$ and $\hat{p}_{L_1,L_2}(x,y)$ are estimated like in the global case: five alternations of re-estimation and re-filtering of candidates

Sound Correspondences: Example

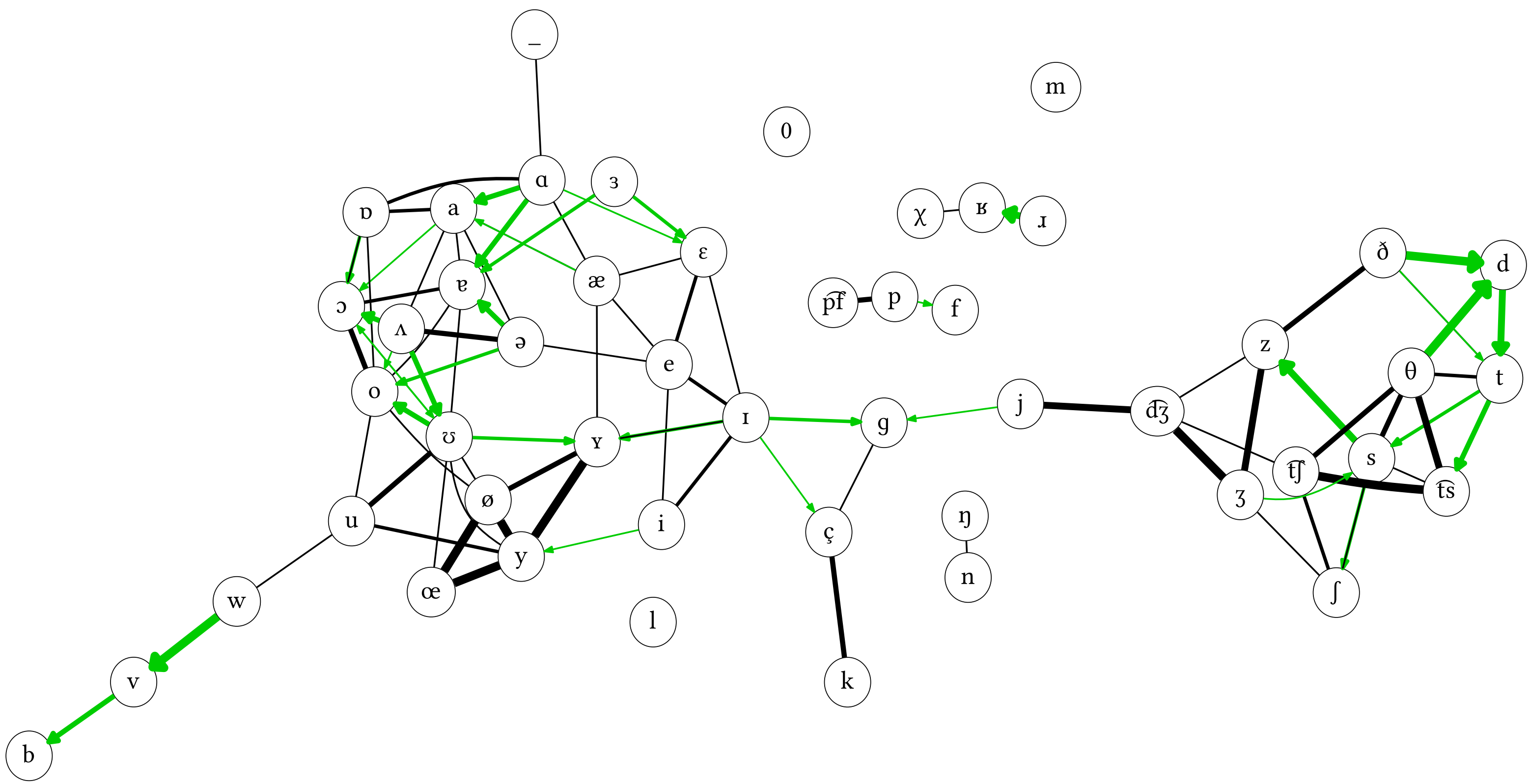


Figure: Drift graph of inferred correspondences from English to German.

Homologue Clustering

- to infer homologue sets for a concept, apply a clustering algorithm (UPGMA in my case) to the pairwise distance matrix between all relevant forms, and cut off the resulting tree at an empirically chosen threshold value
- impression of the results for FISH:

26	niv: [c^{h}o]	cmn: [y]			
27	mns: [χul]	hun: [hɒl]	sme: [kʊɔlli]	sjd: [kuuʎʎ]	sma: [kʉɛliɛ]
	mrj: [kol]	mdf: [kal]	nio: [kolɨ]	krl: [kɑlɑ]	olo: [kɑlɑ]
	fin: [kɑlɑ]	sel: [qælɨ]	ekk: [kɑlɑ]	smj: [gʊuɔllɛ]	yrk: [xɑʎɑ]
	myv: [kal]	vep: [kɑlɑ]	mhr: [kol]	liv: [kɑlɑɑ]	smn: [kyeli]
	kca: [χuɬ]	ale: [qɑχ]	sms: [kuɛll^{j}ɘ]		
28	pbu: [kab]				
29	kan: [miinu]	tam: [miin]			
30	bua: [zagahaŋ]	khk: [t͡saɢas]	xal: [t͡saɦsɐn]		
31	udm: [t͡ɕorɨg]	abk: [ɑp^{h}sɨd͡z]			
32	itl: [əɲt͡ʃ]				
33	deu: [fɪʃ]	eng: [fɪʃ]	nld: [vɪs]	ket: [jiɕ]	
34	por: [pejʃə]	cat: [pɛʃ]	ita: [peʃʃe]	ron: [peʃte]	spa: [peθ]

Table of Contents

Ancestral State Reconstruction

Possible methods for ancestral state reconstruction, part 1:

- **Majority (Mjrty)**: just reconstruct the set that exists in the largest number of children, no reconstruction in case of draw
- **Maximum Parsimony, Single Value (MPSgl)**: reconstructs exactly one cognate set for each node, based on minimizing the number of replacement events that need to be assumed
- **Maximum Parsimony, Multiple Values (MPMlt)**: decide for each homologue set separately, based on minimizing the number of presence/absence switches assumed

For maximum parsimony, I use my own implementation of the standard algorithm by Sankoff (1975).

Ancestral State Reconstruction

Possible methods for ancestral state reconstruction, part 2:

- **Maximum Likelihood, Single Value (MLSgl)**: based on explicit parameterized evolutionary model which fully describes how each state is likely to evolve along a given phylogenetic tree (with branch lengths), select the most likely homologue set for each ancestral node in the maximum-likelihood estimate
- **Maximum Likelihood, Multiple Values (MLMlt)**: like MLSgl, but estimating binary presence/absence values, reconstructing the homologue set if presence is more likely than absence

For maximum likelihood reconstruction, I use the R package `phangorn` by Schliep (2011).

Performance of Reconstruction Methods

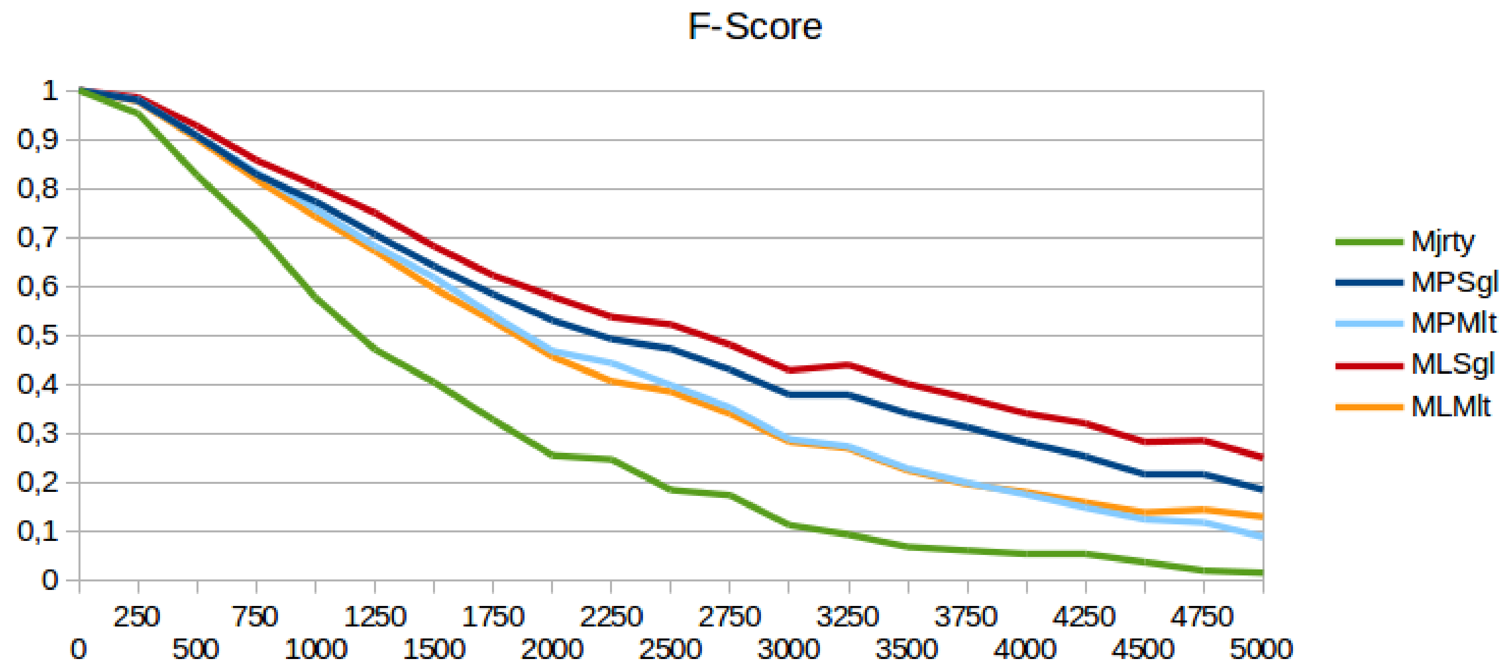


Figure: Development of ASR performance with age of reconstructed language, based on simulated data where the ancestral states were fully known.

Table of Contents

Conditional Independence

- causal inference needs a **conditional independence** relation
- $(X \perp\!\!\!\perp Y \mid Z)$ intuitively means:
  "any dependence between the variables X and Y can be explained by the joint influence of a set of variables Z"
- task: enable conditional independence tests between languages
- challenge: not obvious how to model entire languages as statistical variables, and conditional independence relations need to follow a quite complex set of axioms

Axioms of Conditional Independence

- **symmetry**:
  $(X \perp\!\!\!\perp Y \mid Z) \Rightarrow (Y \perp\!\!\!\perp X \mid Z)$
- **decomposition**:
  $(X \perp\!\!\!\perp YW \mid Z) \Rightarrow (X \perp\!\!\!\perp Y \mid Z)$
- **weak union**:
  $(X \perp\!\!\!\perp YW \mid Z) \Rightarrow (X \perp\!\!\!\perp Y \mid ZW)$
- **contraction**:
  $(X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp W \mid ZY) \Rightarrow (X \perp\!\!\!\perp YW \mid Z)$
- **intersection**:
  $(X \perp\!\!\!\perp W \mid ZY) \wedge (X \perp\!\!\!\perp Y \mid ZW) \Rightarrow (X \perp\!\!\!\perp YW \mid Z)$

Information-Theoretic Treatment

Basic notions of information theory (informally):

- **entropy** $H(X)$: amount of information provided by X
- **joint entropy** $H(X, Y)$:
  amount of information provided by X and Y together
  (not $H(X) + H(Y)$, some information might be redundant!)
- **mutual information** $I(X; Y)$: amount of information that knowing the result of X or Y provides about the other

Important relation: $I(X; Y) = H(X) + H(Y) - H(X, Y)$
(given joint entropy, we can derive mutual information)

Important criterion: $X \perp\!\!\!\perp Y \Leftrightarrow I(X; Y) = 0$
("vanishing mutual information is independence")

Information-Theoretic Treatment

Conditional variants of the notions (again informally):

- **conditional entropy** $H(X|Y)$: amount of additional information provided by X if Y is already known
- **conditional mutual information** $I(X;Y|Z)$: amount of additional information that knowing the result of X or Y provides about the other, provided that we already know Z

Important relation:
$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$
(only (joint) entropies are needed to derive conditional MI)

Important criterion: $(X \perp\!\!\!\perp Y \mid Z) \Leftrightarrow I(X;Y|Z) = 0$
(the all-important equivalence stays valid in conditional case)

Joint Information Measure for Languages

Variables in my model:

- lexical variables: Lex_i: Ω to phonetic strings
- $Hom(Lex_1, \ldots, Lex_n)$: result of homologue detection
- $Hom_i(Hom)$: homologue sets touched by language L_i

Information measure h for sets of languages:

$$h(L_i, \ldots, L_n) := h(Hom_1, \ldots, Hom_n) := \left| \bigcup_{i=1}^{n} Hom_i \right| \tag{4}$$

$\Rightarrow$ We count the **number of homologue sets touched** by the lexicon of any of the languages! (cf. *descriptive complexity*)

Cognate-Based Information Measure

Chaves ea. (2014): Three axioms (**elementary inequalities**) suffice for a measure h to "behave sufficiently like entropy" (i.e. vanishing CMI defines a conditional independence relation):

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

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

My measure $h(L_i, \ldots, L_n)$ can be proven to meet all of these conditions, i.e. we can use it to derive consistent conditional independence tests!

Conditional Independence between Languages

- from h we derive **conditional mutual information between languages** L_1 and L_2 given a set of languages $\mathbf{S} := \{S_1, \ldots, S_n\}$:

$$i(L_i, L_j; \mathbf{S}) := h(L_i, S_1, \ldots, S_n) + h(L_j, S_1, \ldots, S_n) - h(L_i, L_j, S_1, \ldots, S_n) - h(S_1, \ldots, S_n)$$

- intuitively: how many homologues between L_i and L_j cannot be explained away by also being homologous to a word in one of the languages in **S**?
- we can now compute answers for questions like: how much of the lexical overlap between Hungarian and Albanian can be explained by shared influence (= borrowing) from Turkish?
- if CMI vanishes, we have conditional independence, which will allow us to remove a direct contact link between e.g. Hungarian and Albanian from our network

Interpretation and Normalization of CMI

- Example: $i(X_{sqi}; X_{hun}|X_{tur}) =$
  $h(X_{sqi}, X_{tur}) + h(X_{hun}, X_{tur}) - h(X_{sqi}, X_{hun}, X_{tur}) - H(X_{tur})$
  $= 4299\ (|Hom_{sqi}| + |Hom_{tur}| - |Hom_{sqi} \cap Hom_{tur}|)$
  $+\ 3827\ (|Hom_{hun}| + |Hom_{tur}| - |Hom_{hun} \cap Hom_{tur}|)$
  $-\ 5892\ (|Hom_{sqi}| + |Hom_{hun}| + |Hom_{tur}| - |Hom_{sqi} \cap Hom_{hun}|$
  $\quad -|Hom_{sqi} \cap Hom_{tur}| - |Hom_{hun} \cap Hom_{tur}|$
  $\quad +|Hom_{sqi} \cap Hom_{hun} \cap Hom_{tur}|)$
  $-\ 2195\ (|Hom_{tur}|)$
  $=$ 39 (correlates of `sqi` and `hun` that are not from `tur`)
- set with cardinality $i(X_{sqi}; X_{hun}|X_{tur})$ is **interpretable**!
- need comparability $\Rightarrow$ normalize: $\hat{i}(X; Y|Z) := \frac{i(X;Y|Z)}{h(X,Y)}$
  $\hat{i}(X_{sqi}; X_{hun}) = 0.0143$, and $\hat{i}(X_{sqi}; X_{hun}|X_{tur}) = 0.0101$
- $\hat{i}(X; Y|Z) \in [0, 1]$, test against global threshold

Table of Contents

Causal Inference: Basic Idea

- techniques to infer causal relationships between variables from observational data alone (Pearl, 2009)
- not possible for two variables: "correlation is not causation"
- interaction between three or more variables often provides hints
- we need to assume Reichenbach's **Common Cause Principle**: "no correlation without causation"
- systematically use tests to extract hints about underlying causal structure, summarize findings as directed acyclic graphs over the variables (**causal DAGs**)

d-Separation

A path p in a DAG G is **d-separated** by a set of nodes **Z** iff

- p contains a noncollider, i.e. a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$, with $m \in \mathbf{Z}$
- p contains a collider $i \rightarrow m \leftarrow j$ such that $m \notin \mathbf{Z}$ and no descendant of m is in **Z**

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

Faithfulness

A distribution p fulfills the **Markov condition** with respect to a DAG G if it factorizes according to the parent relationship defined by G, i.e. if $p(X_1, \ldots, X_n) = \prod_{i=1}^{k} q(X_i \mid pa(X_i, G))$.

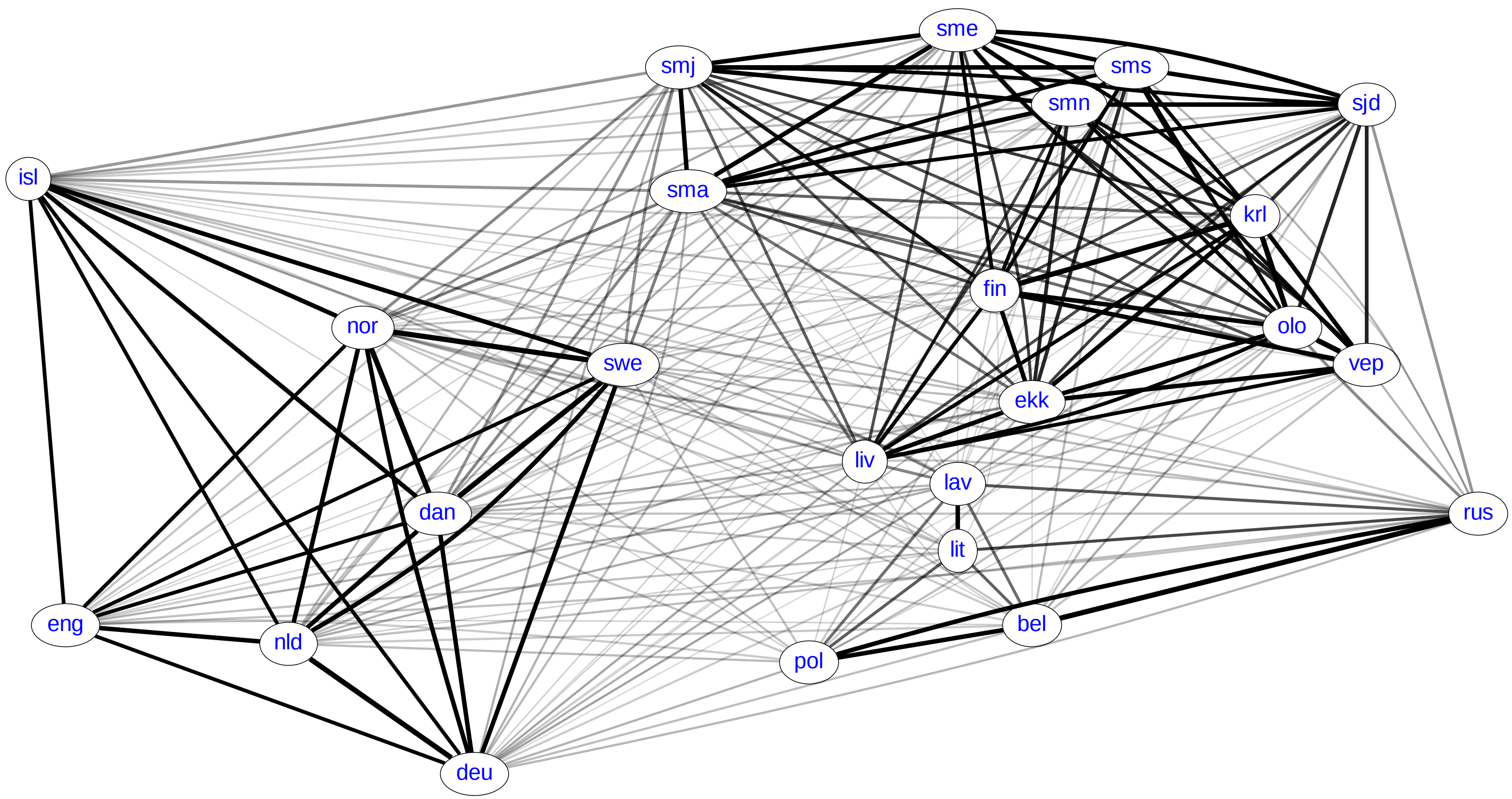
A distribution P is **faithful to a DAG** G if the conditional independence relationships which hold in P are exactly the ones implied by the d-separation criterion on G. We call the distribution P **faithful** if it is faithful to some DAG.

Causal Inference: Inferring the Skeleton

- PC algorithm by Spirtes et al. (2000): sequence of conditional independence tests reduces a complete graph to a **causal skeleton**, where no link can be explained away by conditioning on other variables
- removal of link $X - Y$ relies on finding a **separating set**, i.e. a set of variables $\{Z_1, \ldots, Z_n\}$ such that $(X \perp\!\!\!\perp Y \mid Z_1, \ldots, Z_n)$
- example: $(\textit{sma} \perp\!\!\!\perp \textit{fin} \mid \textit{swe}, \textit{Uralic})$
- conditional independence tests are performed with increasing separating set size, and links are removed after each such stage
- if all confounders are observed, and there is a DAG which is faithful to the underlying distribution, the PC algorithm provably results in an equivalence class containing that DAG

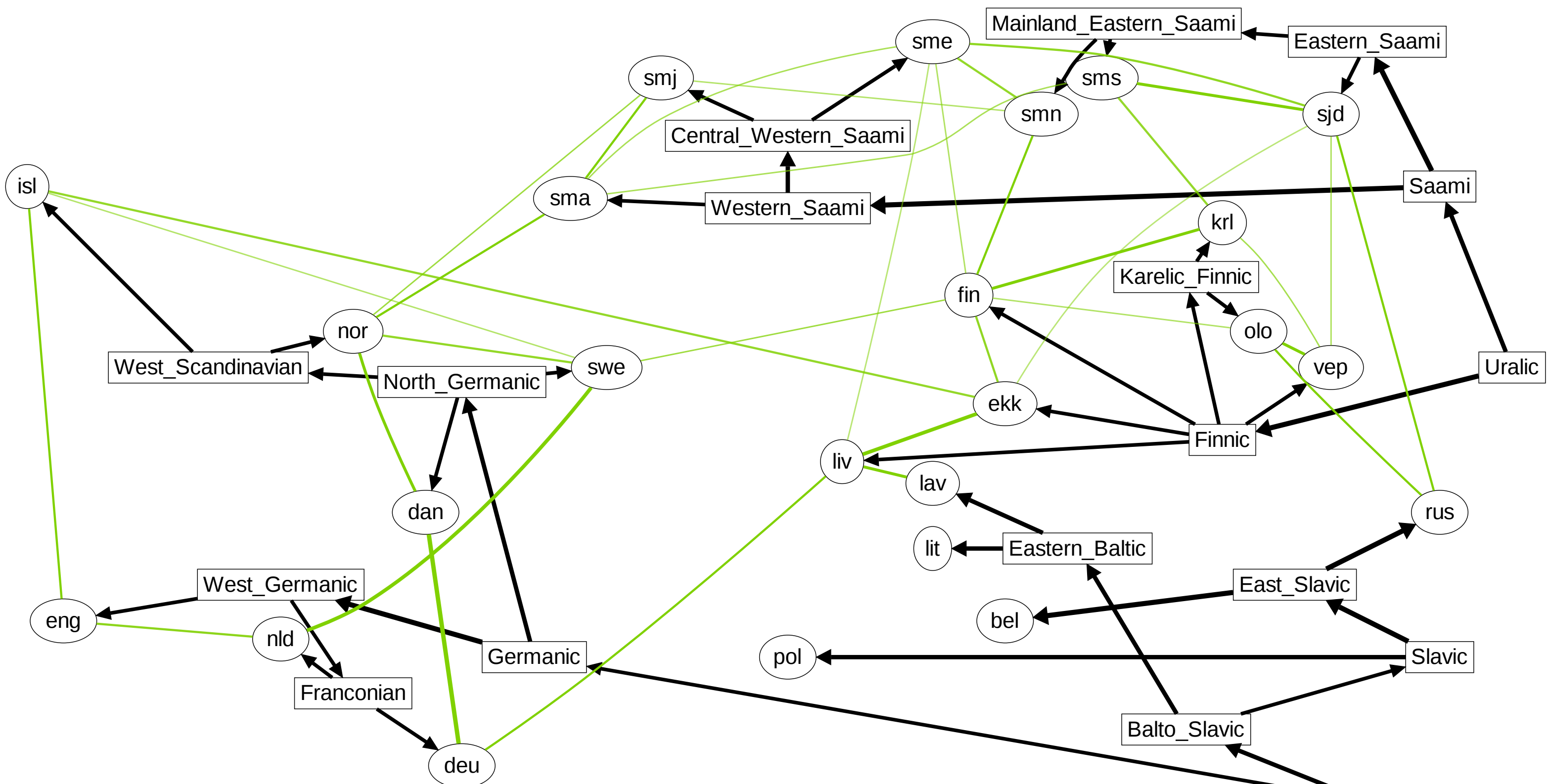
Phylogenetic Lexical Flow Inference: Skeleton

Example input, visualzing pairwise homologue overlaps:



Phylogenetic Lexical Flow Inference: Skeleton

Example result of PC algorithm using vanishing CMI as test:



Causal Inference: Directionality Inference

- for each pattern of the form $X - Z - Y$ (**unshielded triple**), ask whether the central variable was part of the separating set that was used for explaining away the link $X - Y$
- underlying idea: if Z was not necessary to explain away $X - Y$, this excludes all patterns except $X \rightarrow Z \leftarrow Y$ (a **v-structure**)
- reason: we would expect some information flow in all three scenarios $X \leftarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$, and $X \rightarrow Z \rightarrow Y$
- this relies on a causal **faithfulness** assumption: we can measure $(X \perp\!\!\!\perp Y \mid Z)$ iff this is implied by the true causal graph
- example: $swe - fin - Fennic$, $(swe \perp\!\!\!\perp Fennic)$, i.e. Finnish not necessary to separate Swedish from Fennic, therefore $swe \rightarrow fin \leftarrow Fennic$

Causal Inference: Propagating Directionality

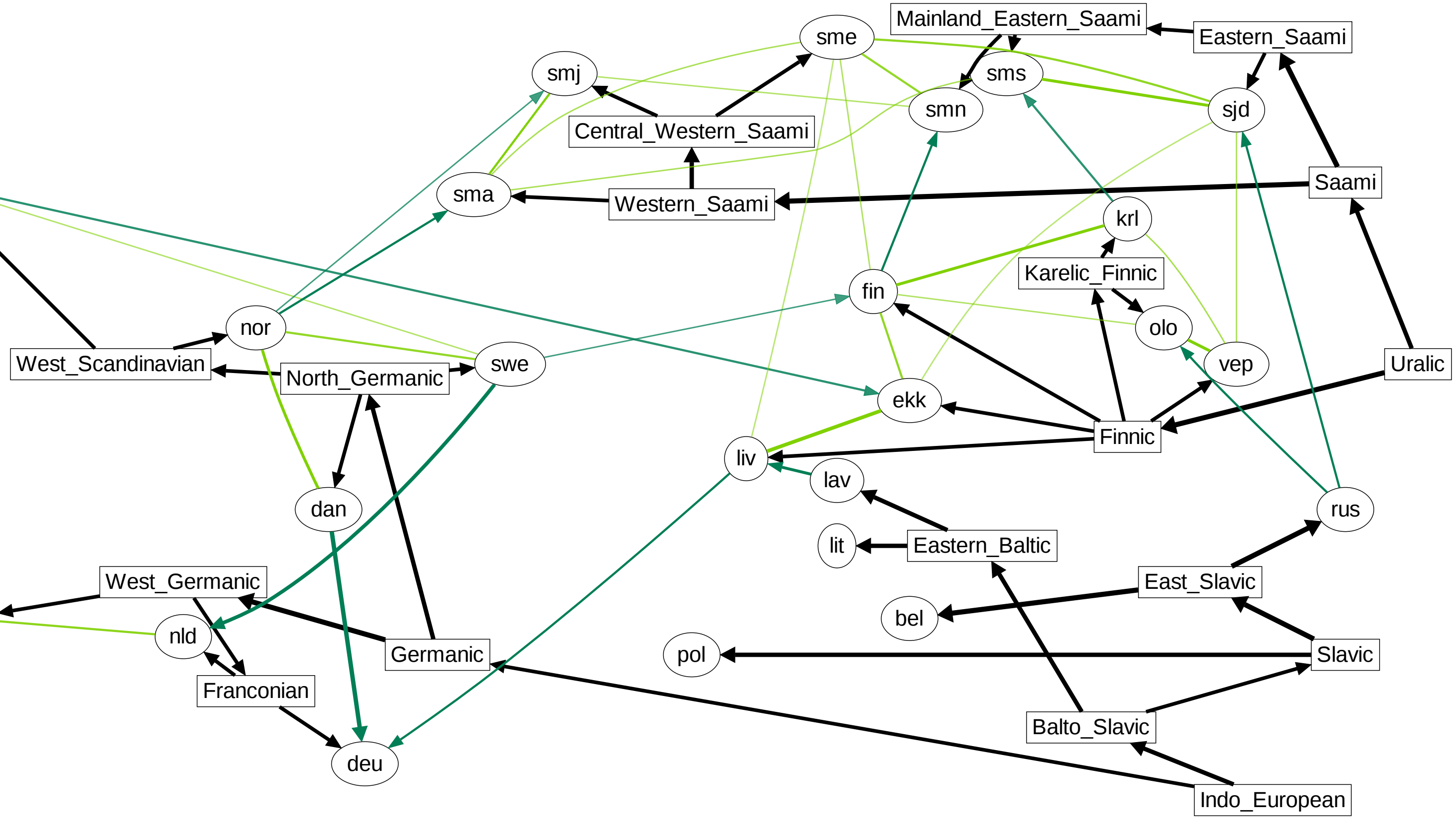
- if all possible common causes are measured, the faithfulness assumption implies we can be sure to have detected exactly the true v-structures
- this provides an inference rule $X \rightarrow Z — Y \Rightarrow X \rightarrow Z \rightarrow Y$
- the PC algorithm uses this rule to **propagate directionality information** through the graph, in many case assigning a direction to each node in the causal skeleton
- example: Glottolog gives us *Franconian* $\rightarrow$ *deu*, we found it impossible to separate *deu* – *liv*, but (*Franconian* $\not\perp\!\!\!\perp$ *liv*) and (*Franconian* $\perp\!\!\!\perp$ *liv* | *deu*), no v-structure, therefore *deu* $\rightarrow$ *liv*

Directionality Inference for Languages

- big problem: on our coarse-grained information measure, conditional independence test are less reliable than needed
- good solution for skeleton inference: **Flow Separation (FS)**
- for directionality inference: **Unique Flow Ratio (UFR)**, direct collider test instead of separation set criterion
- best in experiments: **Triangle Score Sum (TSS)**, a heuristic aggregate of fit scores across triples involving each link

Phylogenetic Lexical Flow Inference: Directionality

Example result of FS plus UFR:



Evaluation Measures

	arrow in result	**no arrow in result**
arrow in standard	*true positive*	*false negative*
no arrow in standard	*false positive*	*true negative*

Table: Table of elementary definitions for skeleton evaluation.

	$\rightarrow$ **in result**	$\leftarrow$ **in result**	— **in result**
$\rightarrow$ **in standard**	*true positive* + *true negative*	*false positive* + *false negative*	*false negative*
$\circ\!\!\rightarrow$ **in standard**	*true positive*	*false positive*	*true negative*
$\leftrightarrow$ **in standard**	*false negative*	*false negative*	*true negative*

Table: Table of elementary definitions for arrow evaluation.

Results for PLFI on NorthEuraLex

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

Table: Comparing skeleton performance on MLsgl and MLmlt reconstructions.

	FS on MLsgl reconstruction				FS on MLmlt reconstruction			
	VPC	SPC	UFR	TSS	VPC	SPC	UFR	TSS
arPrc	0.185	0.154	**0.615**	0.546	0.240	0.000	0.410	**0.500**
arRec	0.114	0.050	**0.585**	0.585	0.122	0.000	0.695	**0.689**
arFsc	0.141	0.076	**0.600**	0.565	0.162	(0.0)	0.516	**0.579**

Table: Comparing arrow performance on MLsgl and MLmlt reconstructions.

Results for PLFI on Simulated Data

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

#	MLsPC	MLsPS	MLsFS	MLmPC	MLmPS	MLmFS
skPrc	**0.851**	0.798	0.711	**0.855**	0.797	0.710
skRec	0.539	**0.722**	0.659	0.527	**0.720**	0.658
skFsc	0.660	**0.758**	0.684	0.652	**0.757**	0.683

Table: Skeleton performance for perfect and reconstructed ancestors.

Results for PLFI on Simulated Data

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

	FS on MLsgl reconstruction				FS on MLmlt reconstruction			
	VPC	SPC	UFR	TSS	VPC	SPC	UFR	TSS
arPrc	0.490	0.512	0.432	**0.555**	0.485	0.508	0.435	**0.561**
arRec	0.362	0.290	**0.423**	0.343	0.354	0.288	**0.422**	0.347
arFsc	0.417	0.370	**0.428**	0.424	0.409	0.368	0.428	**0.428**

Table: Arrow performance for perfect and reconstructed ancestors.

Table of Contents

Open Questions: Ongoing Work

- performance improvements in flow separation in order to derive confidence values on each link via bootstrapping
- revisit information weighting and improve its mathematical foundations, with potential gains for the entire pipeline
- determine dependence of performance on available number of independent characters (can we do better with 2,000 concepts, how much worse with 500?)
- investigate impact of erroneous elementary decisions (are they just noise, or do they propagate?)
- release code as a well-organized package, making the methods accessible not only to very experienced computational linguists

Open Questions: Application to Other Types of Data

- can we apply this to other levels of linguistic description, e.g. typological variables or paradigm structures?
  (likely difficult due to lack of universally measurable features)
- will this work on the level of dialects, where cognacy is uninformative, but we can measure values for a considerable number of phonological and other features?
- how about other dimensions of variation? measure word usage patterns to imply causal graphs between authors or documents?
- the transpose of the problem (inferring influences between concepts, with languages as observations) is highly interesting as well (causal graph will be a semantic map!)
- always interested in suggestions and possible collaborations!

References I

Bentz, C., Alikaniotis, D., Samardžić, T., and Buttery, P. (2017). Variation in Word Frequency Distributions: Definitions, Measures and Implications for a Corpus-Based Language Typology. *Journal of Quantitative Linguistics*, 24(2-3):128–162.

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, I., Baysarova, Z., Mühlenbernd, R., Wahle, J., and Jäger, G. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia.

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages.* CSLI Publications, Stanford.

Morrison, D. A. (2011). *An introduction to phylogenetic networks*. RJR Productions.

Pearl, J. (2009). *Causality*. Cambridge University Press.

References II

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

Schliep, K. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Thomason, S. G. and Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley and Los Angeles.

Wen, D., Yu, Y., and Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet*, 12(5):e1006006.

The NorthEuraLex database

NorthEuraLex database as published in Dellert et al. (2020):

- list of 1,016 cross-linguistically applicable concepts
- goal: realizations of these concepts across all sufficiently documented languages of Northern Eurasia (currently 107 languages from 20 different families, expanding to 196 languages until the end of the year)
- (mostly) automated transcription of collected words into the International Phonetic Alphabet to make the data comparable
- cannot be automated, a lot of manual work is needed!
- web interface (and releases) at `www.northeuralex.org`

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution: The Empirical Turn

Data Collection Sources

рыкIв — 326 —

колóть, продырявливать, проды-рявить *что*.
рыкIвашара *перех. побуд. от* **кIвашара** II 1) заставлять (заста-вить) *кого* танцевать танец «куа-шарá» (*см.* **кIвашара**); 2) *перен.* избивáть, избить, бить, побить *кого*; ◇ **заджвы мса кьымта агIа-рыкIвашара** отколотить когó--либо (*букв.* заставить плясать, как роговóй волчóк).
рыкIвырпкIвырра *перех. побуд. от* **кIвырпкIвырра** щекотать, по-щекотать *кого*.
рыласра *перех. побуд. от* **лас-хара** 1) облегчать, облегчить *что*, делать (сделать) лёгким *что* (*в ве-се*); **ахIатла рыласра** облегчить нóшу; 2) ускорять, ускорить *что*; **анхара алгара рыласра** ускорить окончание работы.
рылаххра *перех.* вбегать, вбе-жать (*в гущу кого-чего-л.*); **ауаIа рылаххра** вбежать в толпу.
рылашарара *перех. побуд. от* **лашарахара** освещать, осветить *что*; **апещ рылашарара** осветить кóмнату.
рылащцарара *перех. побуд. от* **лащцарахара** затемнять, затемнить *что*, замаскирóвывать (замаски-ровáть) свет.
рымайрара *перех. побуд. от* **майрахара** делать (сделать) лёг-ким, легкó выполнимым, облег-чать, облегчить *что*.
рымдза стул; **арымдза аквчIва-ра** сесть на стул; *ср. тж.* **сакъвы**.
рымцырара 1. *перех.* 1) опорож-нять, опорожнить *что*; **амашакв рымцырара** опорожнить мешóк; 2) опустошать, опустошить *что*; **2.** *в знач. сущ.* 1) опорáжнивание; 2) опустошéние (*действие*).
рымчра *перех.* усиливать, уси-лить *что*; **акъару рымчра** удвоить энéргию; ◇ **ачей рымчра** крéпко заварить чай.
рынашхыйара *перех. побуд. от* **нашхыйахара** удручать, удручить, расстраивать, расстрóить *кого*.
рынага 1. дрóжжи; **2.** *в роли*
опр. дрожжевóй; **рынага хъварп** дрожжевые грибкú.
рыпара: къвыд р[…]
гадать на фасóли.
рыпхащара *пере[…]*
пхащара стыдить, п[…]
рыпхдзыхIвара *п[…]*
пхдзыхIвара 1) зас[…]
вить) *кого* потéть, […]
перен. вгонять (в[…]
крáску, заставля[…]
кого краснеть, пок[…]
рыпшара *с.-х.* в[…]
рыпшара *перех.*
вать, провéять *что*
рыпшдзара орна[…]
орнамент.
рыпшдзара 1. *пе[…]*
укрáсить *что*; **2.**
украшение (*действ[…]*)
пшдзауа жьыпIта,
дзауа чгIвычанI *по[…]*
шéние для костéй, о[…]
шéние для тéла.
рыпшкара *перех.*
измельчить *что*;
вспушить, взбивá[…]
3) разрыхлять, р[…]
агрыа рыпшкар[…]
пóчву.
рыпIатIаура *пер[…]*
пIатIаура сдвигáть
с мéста.
рыратра *перех.*
растворить *что*; **а[…]**
йрыратра раствор[…]
2) расплавлять, р[…]
рысасира *перех.* […]
ра оттáивать, оттá[…]
рысасира оттáять […]
рытаразга *тех.* […]
рытаразра *перех.*
исправлять, испра[…]
тья рытаразра пра[…]
уточнять, уточни[…]
уап рытаразра ут[…]
рытшвара *пере[…]*
тшвахара сýживат[…]
живать, заузить *что*; **ахъвда ры-тшвара** сýзить вóрот.
рытыбгIара *перех. побуд. от* **тыбгIахара** расширять, расширить *что*.

ndab	〈名〉	藏：འདབས ＊ ndap		下面的
	※emdeŋ ～ 裤腿口			
ndab ka		藏：ཁ ＊ kwa		下摆
ndaka 参见 warma ndaka				
ndaGgə-	〈动〉	藏-：བདག- ＊ ɦdak-		看管；
ndamrə	〈名〉	藏：འདམ་དྲི ＊ ndamrə		腥味
ndamrə ʂda-				出腥味
ndar-	〈动〉			渴
ndara	〈名〉	藏：མདའ་ར ＊ ndara		战壕
ndərgə-	〈动〉	藏-：འདུ- ＊ ndə-		包括
nde-	〈动〉			吃
ndeʥi u:-				吃喝
ndeɢa-	〈动〉			喂
ndegu	〈名〉			食物
ndegu u:gu				饮食
ndegu jamala				食品
ndemʂgə-	〈动〉	藏-：འདེམ- ＊ ndem-		选举
ndewa	〈名〉	藏：སྡེ་བ ＊ ɦdewa		村子，
ndoŋ	〈名〉	藏：མདུང ＊ ndoŋ		矛
ndəba	〈名〉	藏：སྡོམ་པ ＊ ɦdomba		鉴戒
	※tɕanletɕa ŋlada～əlɢadʑə 闻者足戒。			
ndəbda	〈名〉	藏：ɦdopta		烟草
ndəɢ	〈名〉	藏：མདོག ＊ ndok		颜色
ndəɢ gi				素的；

duge, gå an dajpedh IV (ij dajph-det duger ikke) (dajpa juhtedh-det går an å flytte nå)
duge, klare seg dåhkasjidh, dåhkesje- *v.*
duge, merkes dååjredh (mohte dej bienji baenieh goh edtjin ennje dååjrh gænnah)
dugelig, brukbar, kompetent, kvalifisert, anse som- gaagnadehtedh I-
dugelig, passende, brukbar gyönegs, gyönege, gyönehke (manne dellie leam gyönege-jeg er vel passende til det? er jeg da duganes til det? (litt foraktelig uttrykk))
dugg, rim, tynt islag jyjsege
dukke opp komme plutselig, vise seg (om ulv) delhkiehtidh, delhkehte*komme plutselig* - *v.*
dum gåffoe- *adj./adv.*
dum kjempe staaloe *i eventyr og sagn* - *s.*
dum, toskete jåasoeh- *adj./adv.*
dumheter gjøre- jebjiedidh
dun aegkie- *s.*
dusin dusijne
dusk duahpa- *s.*
duskregne, sildre, begynne å- sijregåetedh I
dvs. dihte jeahta=d.j.
dykke tjarnedh, tjarna- *v.*
dyktig, flink væjkeles (veartenen væjkalommes tjoejkijh-verdens beste skiløpere)
dynke, fukte lietsedh I
dyp giengeles- *adj./adv.*
dypeste på det-, midten av vannet voernge
dypet (ut på-), ut på havet, ut mot kysten dåvvese
dyppe og spise njåalodh, njååle- *v.*
dyr 1) kreehke 2) juvre 3) vijre *udyr vilt*- *s.*
dyrka mark ientje- *s.*
dyrkamark, eng, innmark ientje
dyrt 1) dovrehke 2) dovres- *adj./adv.*
dø (om menneske) sealadidh
dø 1) jaemedh I, jaama 2) jaamedidh, jaamede 3) sealadidh, sealede- *v.den ene etter den andre om mennesker*
dø av alderdom, gå bort, eldes aalterostedh V
dø ut få til å-, utrydde, kutte (et tre) slik at det ikke skyter skudd gïerehtehtedh I
død jaame- *adj./adv.*
død, avdød jaemehke
død, det å dø sealadimmie
døden jaemede
døden jaemede- *s.*
døden, det å dø jaememe
dødsgudinnen Rovhte
dødssyk, halvdød (av sykdom) aasmeles
døgn dygne- *s.*
dølgsmål gjøre noe i-, være anonym, gå i skjul, spille på runebomme tsiemedh I
dømmes, idømmes døøpmesovvedh
dønning dïelme, haaven dïelm
døpe kristedh
døpe, bade laavkodh
døper, en som døper (eg: en som bader) laavkoje (Jåhha laavkoje-døperen Johannes)
døpes, bli døpt, bli badet lååvkesovvedh
dør, port okse- *s.*
dørene en som går ut og inn i- skoerkedæjja- *s.*
dørene gå ut og inn i- skoerkedidh, skoerkede- *v.*
dåehkie gruppe
dåp, bad laavkome
dårlig bli -(om vær) dormenidh
dårlig forfatning i-, ute av drift, i ustand smalhtjan
dårlig jobb utføre-, slurve slaerviedestedh I
dårlig passende, passelig madtege- *adj./adv.*
dårlig vis på- nåake-laakan- *adv.*
dårlig, elendig gååre
dårlig, elendig, udugelig skraape, skraapoe (skraape bienje-dårlig hund) (skraape bijle-dårlig bil) (skraapoe kaarre-dårlig kar)
dårlig, elendig, udugelig, skit, avføring bæjhke (bæjhke bienje-elendig hund) (bæjhke kaarre-udugelig fyr)
dårlig, passelig, passende madtege *adv*
dårlig, passelig, passende, lite av, svært lite av madtege (madtege graesie-lite gress) (madtege beapmoeh)
dårlige den-, svake siden (av noe) haaltje-bielie

E

effektiv radtjoes (attr), radtjohke (pred)
egen mening, eget synspunkt jijtsh vuajnoe
egen, opposisjonell, egenrådig, sur, tverr, sta beeke
egen/egne sin/sine- jijtjese, jijtse (aerebi lea jijtse boelvem laarhkenamme) (aerebi lea jijtsh soermh ryøkneme)
egeninteresse jijts-buerie
egenrådig jietjeraarehke
egenrådig, sur, tverr, sta, egen, opposisjonell beeke
egens hans/hennes- altemse (det er hans egen skrift- altemse tjaeleme dihte)
egentlig darhke- *adv.*
eget synspunkt, egen mening jijtsh vuajnoe
egg munnie- *s.*
eggkjele munnie-giebnie- *s.*
eiendom jeeluve *i form av reinflokk* - *s.*
eiendom, eiendel, gods, vare eeke- *s.*
eier av -buerie,-burrie- (jis aaj naan røøvrebuerie,nov gujht dle meehti vaedtsedh vijredh-om det var noen eier av gevær, da kunne han jo gå på jakt)
eier rein båatsoe-buerie
eik, gammel furu haajhke- *s.*
einer gasngese- *s.*
ekkel (å se til el. av smak), illeluktende, snusket dielies

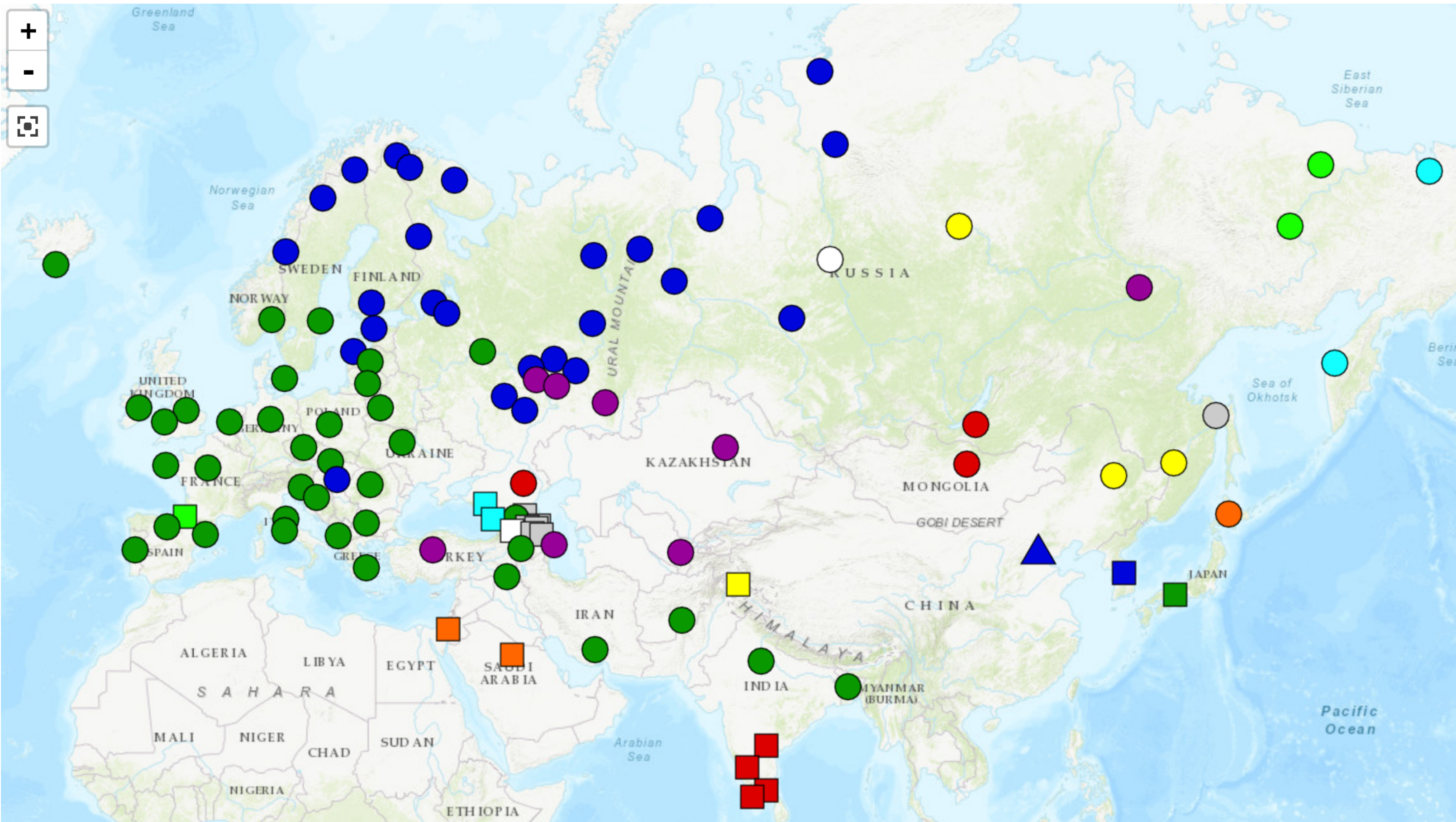
Side 16

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

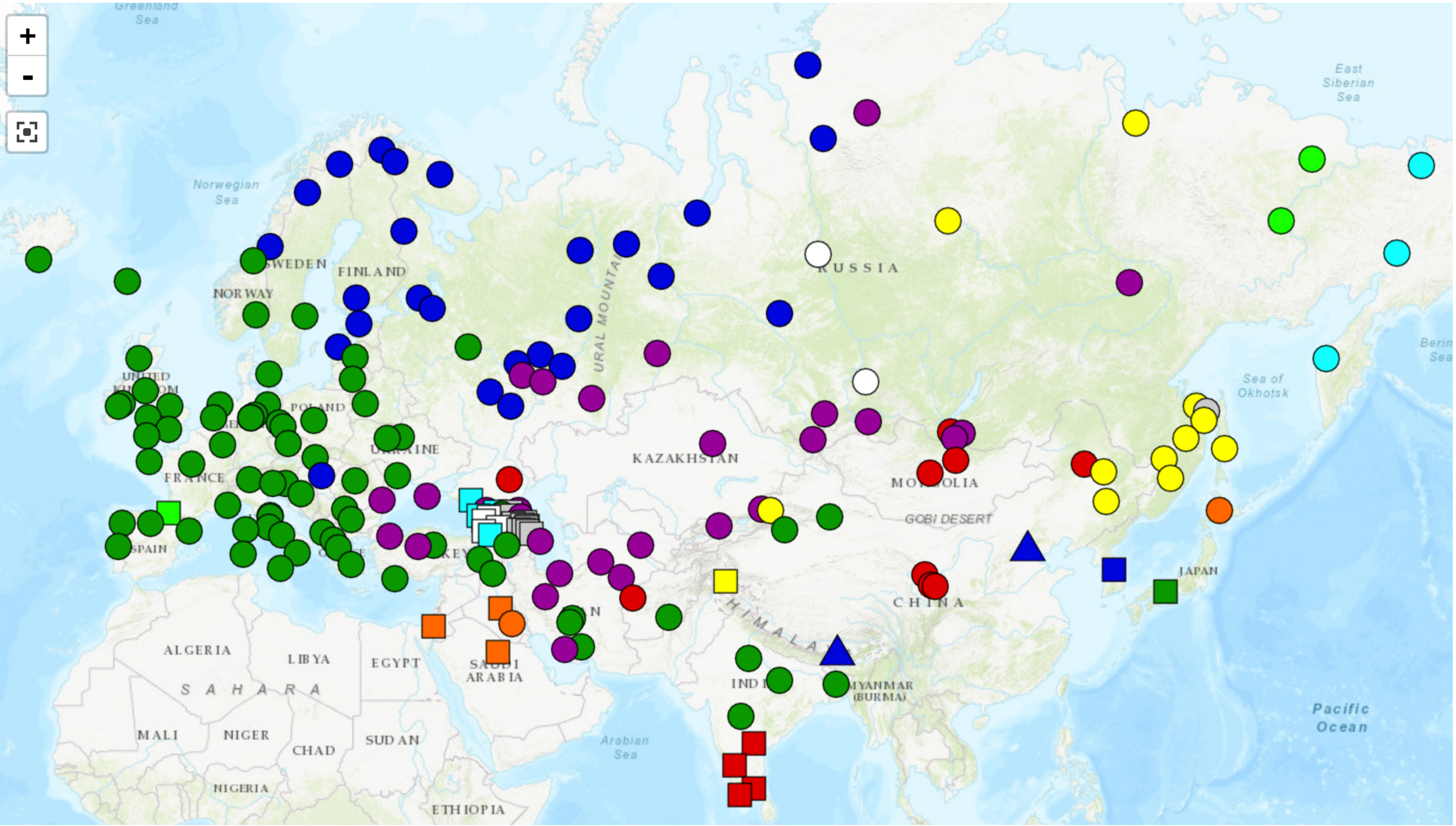
Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

NorthEuraLex 0.9 (situation in 2020)



NorthEuraLex 1.0 (planned for December 2021)



Generating Testset Data by Simulation

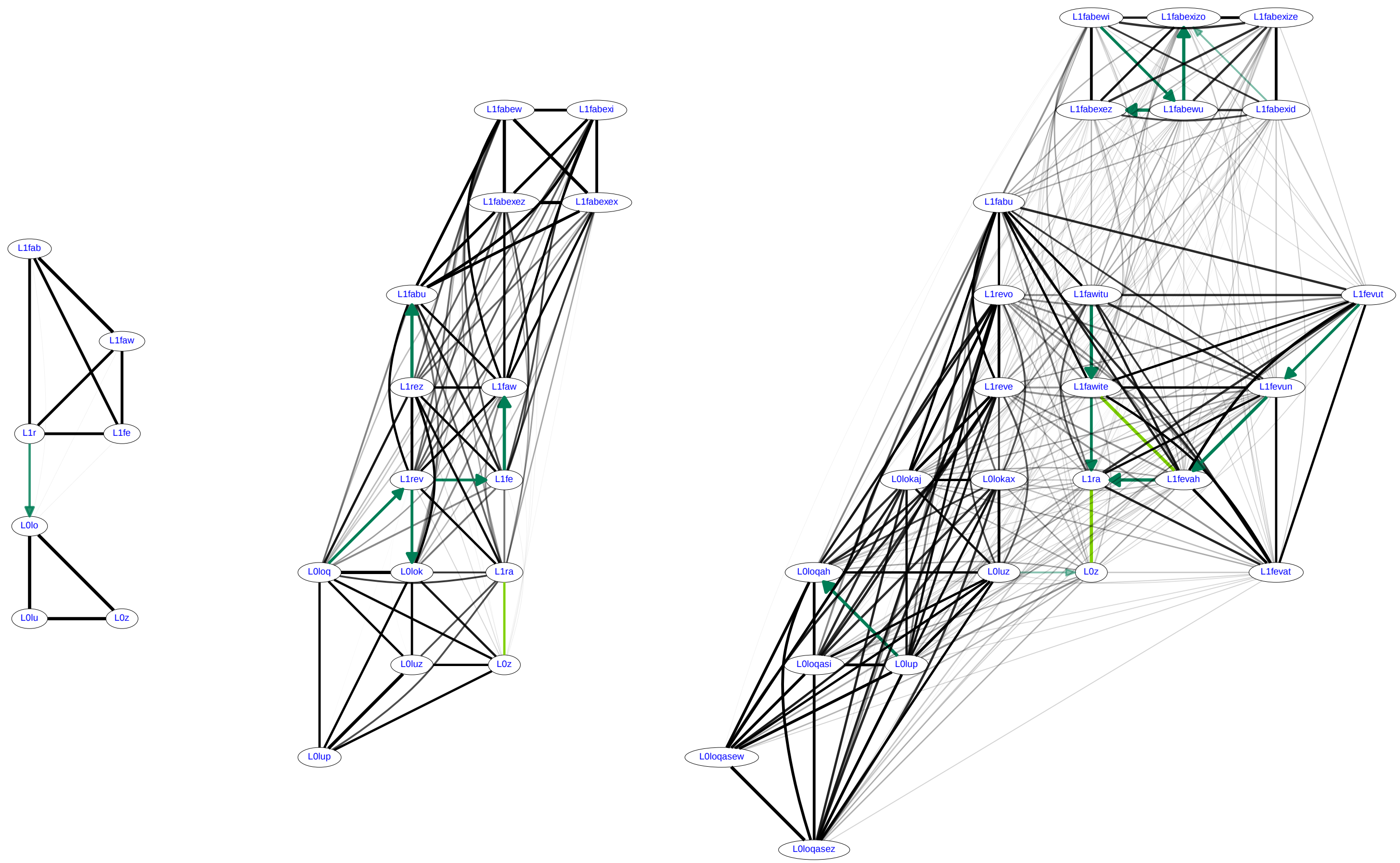
Advantages of using simulations:

- arbitrary amount of test data
- abstract away from problems caused by error-prone cognate detection, tree inference, and ancestral state reconstruction

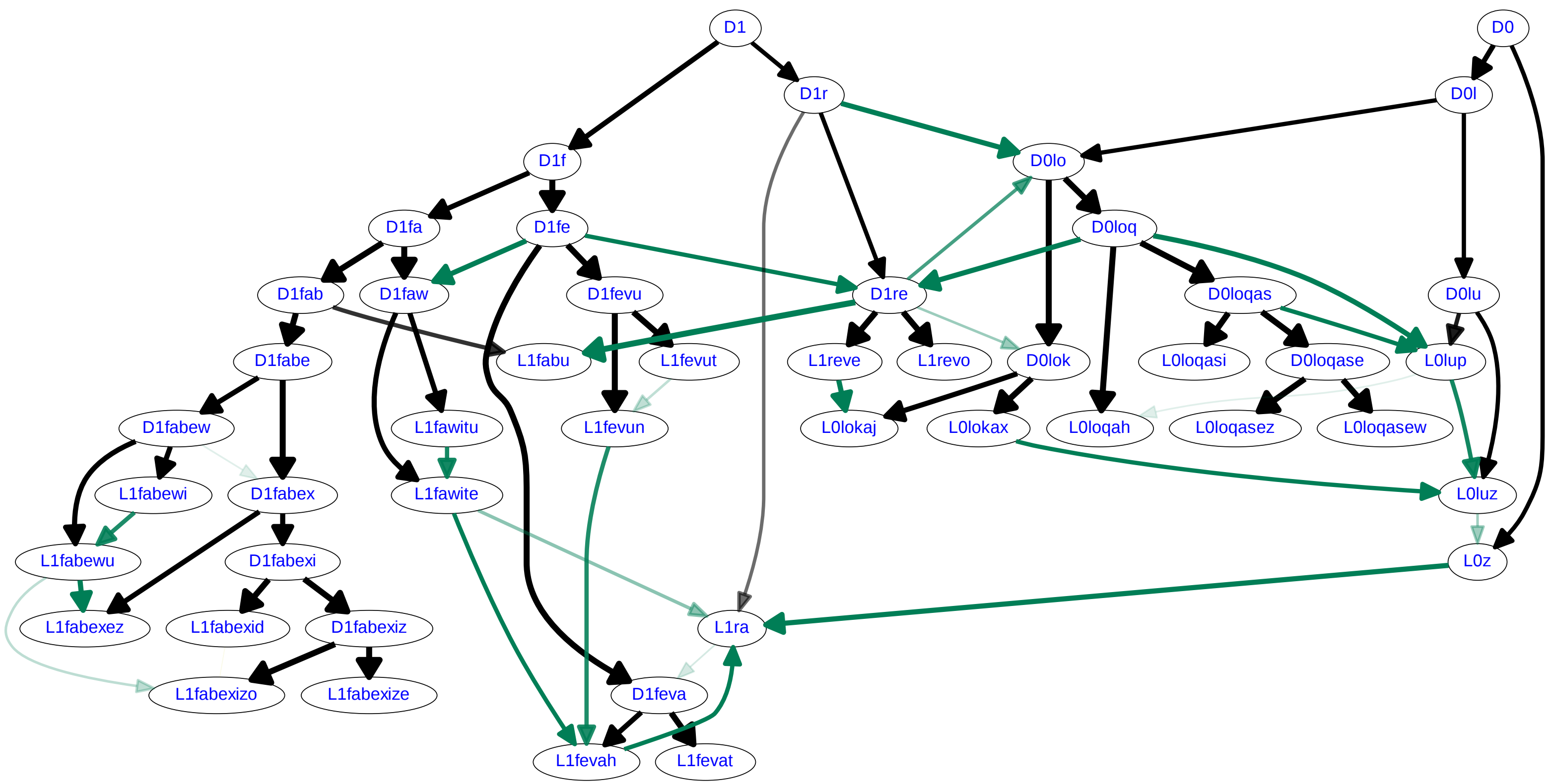
Core design decisions of my simulation model:

- languages split at random intervals, filling a continent
- a language does not become extinct without reason, it only gets replaced if a neighboring language splits into its territory
- we explicitly model lexical replacement in each language (longer splits will lead to less cognate set overlap)
- monodirectional contact channel can open at any time between neighbors, on which cognate IDs are randomly copied over
- every single event modifying the data is tracked, we retain access to complete knowledge

Example: The Simulation Process



Example: A Simulated Flow Network



Directionality Inference: Unique Flow Ratio (UFR)

The second solution I explored:

- define a score for unshielded triples for making the collider decisions, based on the same intuitions plus a flow criterion
- propagate the decisions by the PC propagation rules

Details of the **Unique Flow Ratio (UFR)** score:

- idea: quantify the notion of "Z needed to remove X — Y"
- let cog_{XYZ} be the cognates shared between between X, Y, Z
- cog_{XYZ*}: the cognates which no path excluding Z could have transported between X and Y (**unique flow**)
- $ufr_1 := \frac{\frac{|cog_{XYZ*}|}{\min(|cog_X|,|cog_Y|,|cog_Z|)}}{\frac{|cog_{XZ}|}{\min(|cog_X|,|cog_Z|)} \cdot \frac{|cog_{YZ}|}{\min(|cog_Y|,|cog_Z|)}}$ ("as much UF as expected?")
- $ufr_2 := cog_{XYZ*}/cog_{XYZ}$ ("how relevant is flow through Z?")
- $ufr := ufr_1 \cdot ufr_2$, v-structures will typically have $ufr < 0.02$

Example of PC Algorithm

A B

C D

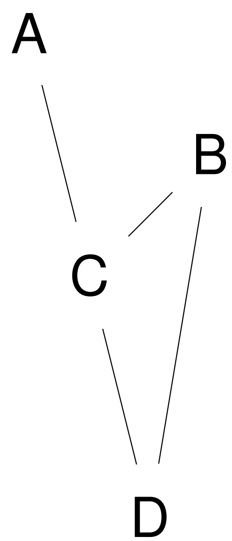
Conditional independence relationships:

$(A \perp\!\!\!\perp B \mid D)$

$(A \perp\!\!\!\perp B \mid C, D)$

$(A \perp\!\!\!\perp D \mid B, C)$

Stage I:

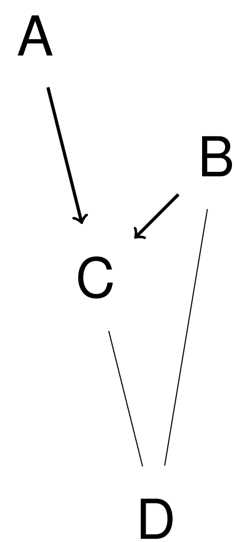


$S_{AB} = \{D\}$

$S_{AD} = \{B, C\}$

no further minimal separating sets found

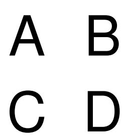
Stage II:



ACD: $C \in S_{AD}$, no arrows

ACB: $C \notin S_{AB}$, i.e. $A \to C \leftarrow B$

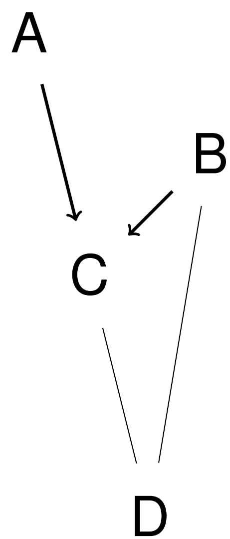
Example of PC Algorithm



Conditional independence relationships:

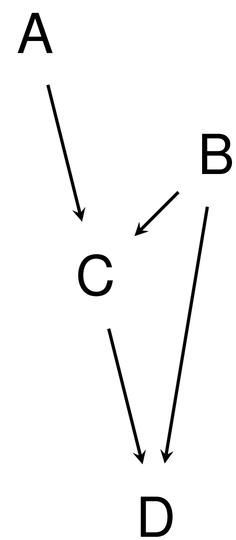
$(A \perp\!\!\!\perp B \mid D)$
$(A \perp\!\!\!\perp B \mid C, D)$
$(A \perp\!\!\!\perp D \mid B, C)$

Stage II:



ACD: $C \in S_{AD}$, no arrows
ACB: $C \notin S_{AB}$, i.e. $A \rightarrow C \leftarrow B$

Stage III:



$C \rightarrow D$, otherwise new v-structure
$B \rightarrow D$, otherwise directed cycle

Directionality Inference: Triangle Score Sum (TSS)

- consider each unshielded triple $l_1 \rightarrow l_2 \leftarrow l_3$
- define $w(l_1 \rightarrow l_2; l_3) := \frac{|cog(l_1) \cap cog(l_2)| \cdot |cog(l_2) \cap cog(l_3)|}{|cog(l_2)|}$,
  i.e. the cognate overlap between l_1 and l_3 we would have expected if the true pattern had been $l_1 \leftarrow l_2 \rightarrow l_3$ or $l_1 \leftarrow l_2 \leftarrow l_3$
- aggregate from all triples into $sc(l_1 \rightarrow l_2) := \sum_{l_3} w(l_1 \rightarrow l_2; l_3)$, use threshold on $sc(l_1 \rightarrow l_2)/sc(l_2 \rightarrow l_1)$ to make decision

Directionality Inference: Improved v-structure test

Another variant for CLFI (without reconstructed proto-languages)

- a simple test based on the hypergeometric distribution
- in a v-structure $X \rightarrow Z \leftarrow Y$, we would expect the number k of isolectic sets covering X, Y, Z to be low
- we want to model the distribution of k under null hypothesis that it is not a v-structure
- we get probability of getting k sets covering all three variables if we randomly draw sets for covering Z and Y from all sets covering Z, some of which also cover X
- $k \sim Hypergeo(N, K, n)$ with N (red balls) the number of sets covering X and Z, K (black balls) the number of sets covering Z, but not X, and n (sample size) the sets covering Y and Z
- we can simply check whether $chyper(k, N, K, n) < p$ for a p-value of our choice (in the experiments: p = 0.1)

Joint Entropy

For discrete variables $X_1, ..., X_n$ with joint distribution $P(x_1, ..., x_n)$, the **joint entropy** is defined as

$$H(X_1, ..., X_n) = -\sum_{x_1} ... \sum_{x_n} P(x_1, ..., x_n) \log_2[P(x_1, ..., x_n)]$$

where $P(x_1, ..., x_n) \log_2[P(x_1, ..., x_n)] := 0$ if $P(x_1, ..., x_n) = 0$.

The joint entropy

- is the standard way of measuring the uncertainty associated with (or the information contained in knowing the outcomes of) a set of variables taken together
- is larger than the maximum of single variable entropies
- never shrinks when additional variables are added

Joint Entropy vs. Sum of Entropies

It can be proven that the joint entropy is always smaller or equal to the sum of the individual entropies:

$$H(X_1, ..., X_n) \leq H(X_1) + ... + H(X_n)$$

If the two sides are equal, the variables are independent.

The difference (called **total correlation**) can be conceived as capturing “synergy”, or the information shared between the involved variables.

Core idea of the Extended Common Cause Principle:

- measure the strength of total correlations
  for subsets of the observed variables
- use this to establish the existence of common causes

Issue 1: Selection Bias due to Choice of Concepts?

- concepts are more likely to be included in NorthEuraLex if their form distances reflect language distances well
- language distances inferred from independently determined set of 50 stable concepts
- Q: Could this cause selection bias in lexical flow inference?
- A:
  - ▷ it causes a bias for more stable words to be sampled
  - ▷ but stability is not a variable in the model
    $\Rightarrow$ no selection bias within the set of variables
  - ▷ bias could induce dependence between related languages, strengthening dependence induced by proto-language

Issue 2: Information Content without Context?

- first step in toolchain:
  inference of information weights for sequence alignment
- goal: remove necessity of stemming, compensate for combinatorics when dealing with sound systems of different size
- Q: Could this method be improved by integration of syntactic context and language use?
- A:
  - ▷ corpus-derived information measures very useful on the word level, e.g. Bentz et al. (2017) for morphological complexity
  - ▷ on the segment level, relevance for historical linguistics should not depend on the commonness of the word in question
  - ▷ data sparseness problem for most languages

Issue 3: Exclusive Focus on the Lexical Level

- P: Flow contact model exclusively builds on the lexicon.
- Q: Do the results generalize to the more complete picture one would get by taking other dimensions into account?
- A:
  - ▷ in the network summary, other dimensions will not add much
  - ▷ in the hierarchy of contact intensity by Thomason and Kaufman (1988), lexical influence becomes visible first, i.e. visible contact without lexical borrowing is very rare
  - ▷ approaches on other types of data suffer from differences in expert judgments, and lack of independent samples
  - ▷ these domains could add some borrowable features, but their number will be far smaller than the number of etyma

Issue 4: Abstracting over Speakers in Contact

- P: “Language change happens through people.”
- Q: Could the social processes actually taking place during contact be interfaced or integrated with such models?
- A:
  - ▷ methods could be applied to individual speakers (or authors)
  - ▷ measuring the language of individual speakers is difficult (especially for marginalized smaller languages)
  - ▷ to test feasability, simulation could be extended
  - ▷ problem: many parameters can only very roughly be estimated already, parametrization of social processes?

Issue 5: Dictionary Data vs. Functional Needs

- P: "Actual language use depends on functional need."
- Q: Does this limit the usefulness of conclusions drawn from the lexical inventory alone?
- A:
  - ▷ dictionary data for small languages do not reflect usage
  - ▷ reason: no functional need for minority language
  - ▷ language contact and borrowing might be much more pervasive in reality than documentation suggests
  - ▷ also applies beyond lexical data (grammars are remolded)
  - ▷ crucially: actual contact strength never weaker than detectable in the sources (no false positives)