EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

Uralic and its Neighbors as a Test Case for a Lexical Flow Model of Language Contact

Szeged, January 20, 2015

Johannes Dellert

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

Table of Contents

NorthEuraLex

Causal Inference

Lexical Flow Inference

Results

Future Work

NorthEuraLex: Overview

- realizations of 1.016 basic concepts across 103 languages of Northern Eurasia compiled from dictionaries (Dellert, 2015)
- advantage compared to existing datasets (like IDS):
  deep coverage of a large continuous geographic area
- data on 26 Uralic languages has been released and is available for inspection and expert feedback (which is very welcome):
  `http://sfs.uni-tuebingen.de/~jdellert/northeuralex`
- Languages used for this study:
  - ▷ 6 Finnic and 6 Saami languages
  - ▷ 2 Mordvinic, 2 Mari, 3 Permian languages
  - ▷ Khanty, Mansi, Hungarian, 4 Samoyedic languages
  - ▷ 4 Turkic languages, and Ket as a contact isolate
  - ▷ 4 Germanic, 2 Baltic, 6 Slavic languages, Romanian

NorthEuraLex: Conversion to IPA

- for every language, we have converters from orthography to IPA
- only based on the literature and readily available recordings; some errors are unavoidable
- accurate enough for modeling historical developments
- infrastructure:
  - ▷ simple formalism based on greedy matching and replacement in freely definable contexts of arbitrary length
  - ▷ description language for defining transducers on X-SAMPA
  - ▷ cascades of between 1 and 6 transducers for each language; for most languages in Cyrillic orthography, one transducer for consonants and one for vowels works reasonably well
  - ▷ should be compilable into finite-state transducers (currently being done as a student project)

NorthEuraLex: Reduction to ASJP classes

- for cross-linguistic comparisons, we will always need to abstract away from some phonetic detail (e.g. exact representation of diphthongs, palatal vs. palatalized)
- full IPA with all diacritics has too many symbols for reliably estimating alignment scores
- popular in computational historical linguistics: **ASJP classes** as used by the Automated Similarity Judgment Program (Brown et al., 2008) across more than 6.000 languages
- 41 classes, no segment length, no coarticulation!
- very suboptimal for Uralic, but ensures that we are not overfitting the method to a single language family

NorthEuraLex: Data Sample

fin	*korva*	[kɔrʋɑ]	korwa
ekk	*kõrv*	[kɤrv]	korv
liv	*kūora*	[ku:orɑ]	kuora
sme	*beallji*	[pe̯æʎ:i]	peEli
smn	*pelji*	[peʎi]	peli
sms	*pe´llj*	[pɛʎ:ʲə]	pEl3
mrj	*пӹлӹш*	[pɯlɯʃ]	puluS
mhr	*пылыш*	[pɤ·lɤʃ]	poloS
mdf	*пиле*	[pilʲe]	pile
myv	*пиле*	[pilʲe]	pile
udm	*пель*	[pelʲ]	pel
kpv	*пель*	[peʎ]	pel
hun	*fül*	[fyl]	fil
mns	*паль*	[paʎ]	pal
sel	*ӄо*	[qo]	qo
yrk	*xa*	[xɑ]	xa
nio	*коу*	[kou]	kou

Table of Contents

NorthEuraLex

Causal Inference

Lexical Flow Inference

Results

Future Work

Causal Inference: Basic Idea

- statistical techniques to infer causal relationships between variables from observational data alone (Pearl, 2009)
- not possible for two variables: “correlation is not causation”
- but: interaction between more than two variables often provides hints about underlying causal scenario
- model causal scenarios as **causal DAGs** (directed acyclic graphs) over the variables, systematically exploit hints to infer properties of the underlying causal DAG
- possible under conditions of sufficiency and faithfulness

Conditional Independence and Causal Graphs

- core building block: 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 influence of Z"
- PC algorithm: 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 $(X \perp\!\!\!\perp Y \mid Z_1, \ldots, Z_n)$

Unshielded Collider Criterion

- directionality inference on the causal skeleton
- for each pattern of the form $X - Z - Y$ (**unshielded collider**), 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 $X \rightarrow Z \rightarrow Y$
- under certain conditions, we can add additional arrows if they are the only option not to introduce cycles or additional v-structures (stage 3 of PC algorithm)

Table of Contents

NorthEuraLex

Causal Inference

Lexical Flow Inference
- Correlate Detection
- Lexical Flow Independence
- Inferring Directionality
- The Full Algorithm

Results

Future Work

Lexical Flow Inference: Basic Idea

- model **languages as variables**
  - ▷ language is represented by its basic lexicon
  - ▷ for each concept, we get an observation of each language
  - ▷ we will not use phonetic strings directly, but automatically derived correlate judgments
- define a conditional independence measure on languages
  - ▷ leads to an algorithmic test of hypotheses like "does Russian influence explain away all the similarities between Ket and Selkup?"
- apply a custom variant of causal inference
  - ▷ we should get a **graph telling us how languages influenced each other** (e.g. Swedish on Finnish, not in reverse!)
  - ▷ evaluation on Uralic and its contact languages

Lexical Flow Inference: Sound Correspondences

- detection of correlates based on **weighted string distances**
- cognates will have become dissimilar due to regular sound change, but we still want to recognize them
- idea: infer model of ASJP segment correspondences for each language pair, using a variant of the method described by List (2012), and use it to estimate segment distances for each pair
- examples of resulting model:
  - ▷ Finnish `[k]` cheap to align to Hungarian `[h]`
  - ▷ Finnish `[s]` cheap to align to Northern Saami `[C]`
  - ▷ Hungarian `[f]` cheap to align to Northern Saami `[p]`

Lexical Flow Inference: Weighted Alignment

- problem on dictionary forms: string-distance overestimates similarity e.g. between verbs which share an infinitive ending
- solution: for each segment in the ASJP strings, infer information content from trigram models for each language and word class
- use information content as additional weight for alignment (matching of low-information material costs less)
- examples of alignments (transparency = information content):

p	e	E	l	i	SE: beallji "ear"	0.379082
f	i	-	l	-	HU: fül "ear"	

k	o	r	w	a	FI: korva "ear"	0.707607
f	i	-	l	-	HU: fül "ear"	

-	o	t	-	a	FI: ottaa "to take"	0.302580
v	o	t	m	a	ET: võtma "to take"	

Lexical Flow Inference: Correlate Clustering

- an emerging subfield of computational historical linguistics
- Java re-implementation of the LexStat toolchain (List, 2014)
- use UPGMA (Sokal and Michener, 1958) to derive a **hierarchical clustering** of phonetic strings based on their pairwise distances
- cut the tree at a given **threshold to partition** the strings into clusters of similar forms, which we then assume to be correlates
- write $cor(L_1, \ldots, L_n)$ for the correlate sets shared between languages $L_1, \ldots, L_n$, and $c(L_1, \ldots, L_n) := |cor(L_1, \ldots, L_n)|$
- if a dataset has expert cognacy annotation, we can just use it!

Lexical Flow Independence: Problem

- most natural way to define a conditional independence test on correlate sets is based on a mutual information measure:

$$I(L_1, L_2; Z) := \frac{|cor(L_1, L_2) \backslash \{c \mid \exists \{Z_1, \dots, Z_k\} \subseteq Z : c \in cor(Z_1, \dots, Z_k)\}|}{\min\{|cor(L_1)|, |cor(L_2)|\}}$$

- intuitively: ratio of correlates between L_1 and L_2 which cannot be explained by borrowing through any subset of languages in Z
- if $I(L_1, L_2; Z)$ is smaller than chance, assume $(L_1 \perp\!\!\!\perp L_2 \mid Z)$
- assumption of standard PC algorithm (making it tractable): possible separating sets are all subsets of immediate neighbors of L_1 and L_2 in the current skeleton
- true in the fully stochastic case, but problematic in ours (two unconnected neighbors do not constitute a possible transmission path!)

Lexical Flow Independence: Solution

- solution: explicitly model the **lexical flow** for independence testing, i.e. retain a connected subgraph for each correlate set
- only test separating sets which form a union of acyclic paths between L_1 and L_2 in the current skeleton
- implementation uses a depth-first search of the current graph to get all such paths of length ≤ 4, generates all combinations of these paths which lead to separating set candidates of a given cardinality
- longer paths would be necessary in theory, but did not lead to different results on my data, at a much higher computational cost (also, assuming long chains of borrowing events is risky)

Inferring Directionality: Intuition

- causal inference: if Z was not necessary to explain away $X - Y$, this excludes all causal patterns except $X \rightarrow Z \leftarrow Y$
- but: if L_1 borrows from L_2 which in turn borrows from L_3, none of the lexical material from L_3 might appear in L_1!
- idea: for each triple of languages (L_1, L_2, L_3) and a given causal scenario, we can **measure the difference between expected and observed number of correlates**
- do this for each triangle involving L_1 and L_2, derive a **counterevidence score**, weight the scores according the information each triangle contributes
- if counterevidence in one direction is much stronger, add an arrowhead to the link, representing dominant direction of influence

Inferring Directionality: Computation

- $r(L_1, L_2) := c(L_1, L_2) / \min\{c(L_1), c(L_2)\}$
- $r(L_2, L_3) := c(L_2, L_3) / \min\{c(L_2), c(L_3)\}$
- expected number of correlates under assumption $L_1 \leftarrow L_2$: $r(L_1, L_2) \cdot r(L_2, L_3) \cdot \min\{c(L_1), c(L_3)\}$
- amount of information in triangle rises with overlap of L_2 and L_3
- this leads to the **counterevidence score**:

$$sc(L_1 \rightarrow L_2) := \sum_{L_3} c(L_2, L_3)^2 \cdot \frac{c(L_1, L_2, L_3)}{r(L_1, L_2) \cdot r(L_2, L_3) \cdot \min\{c(L_1), c(L_3)\}}$$

- decision depends on ratio of $sc(L_1 \rightarrow L_2)$ and $sc(L_2 \rightarrow L_1)$

The Algorithm

Algorithm 1 infer_network$(L_1, \ldots, L_n)$

$G := (\{L_1, \ldots, L_n\}, \{\{L_i, L_j\} \mid 1 \leq i \neq j \leq n\})$, the complete graph
$s := 0$
while $s < n - 2$ **do**
    for $\{L_i, L_j\} \in G$ by increasing strength of remaining flow **do**
        for each combination $P_1, \ldots, P_k$ of paths from L_i to L_j of length ≤ 4 **do**
            if $|S| = s$ for $S := \bigcup\{P_1, \ldots, P_k\}$ **then**
                if ratio of $c(L_i, L_j)$ not explainable by flow across S is < 0.02 **then**
                    remove $\{L_i, L_j\}$ from G
                end if
            end if
        end for
    end for
    $s := s + 1$
end while
for $\{L_i, L_j\} \in G$ **do**
    if $sc(L_i \to L_j)/sc(L_j \to L_i) < 0.9$ **then**
        add arrow $L_i \to L_j$ to network
    end if
end for
return network consisting of G and arrows

Table of Contents

NorthEuraLex

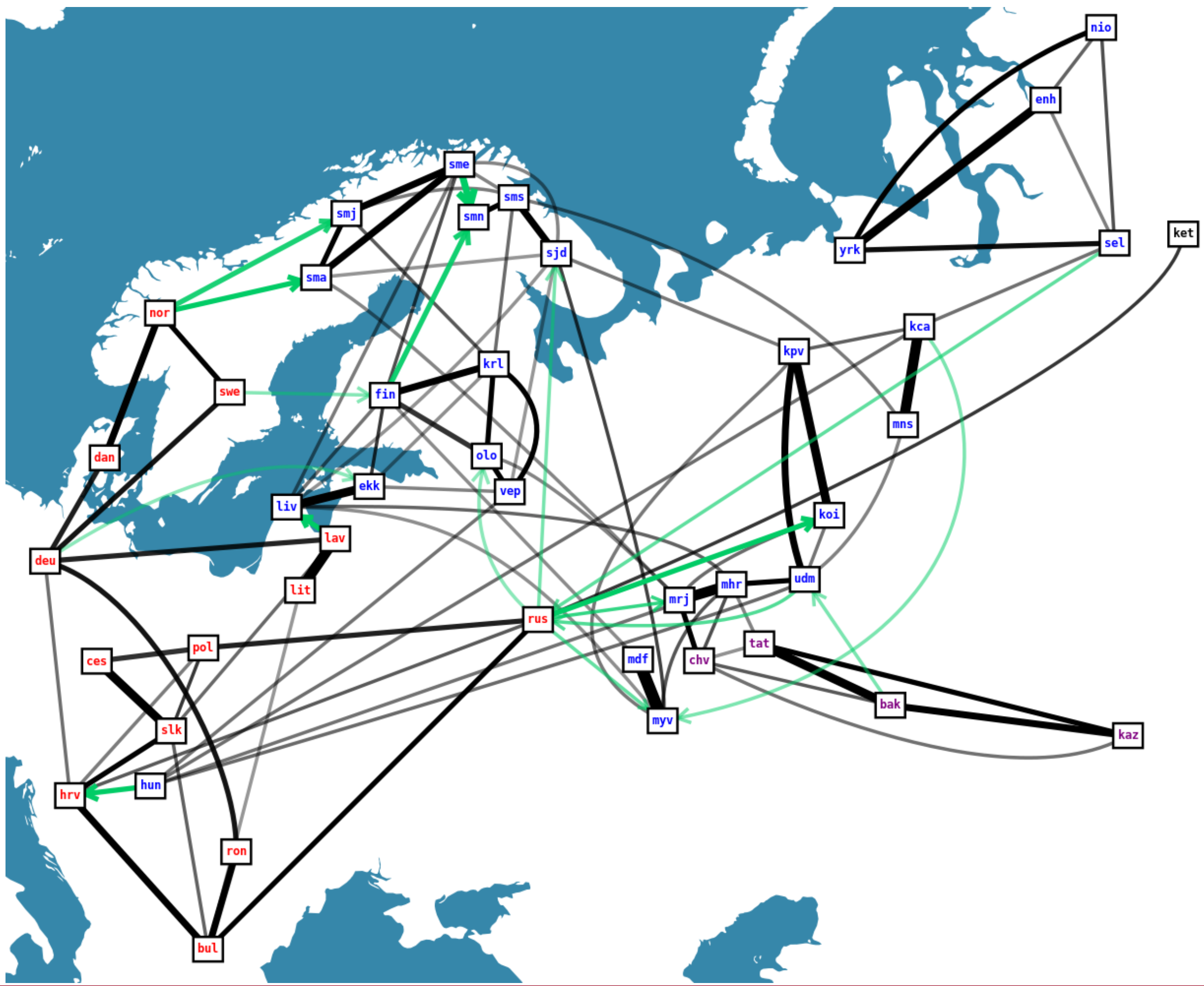
Causal Inference

Lexical Flow Inference

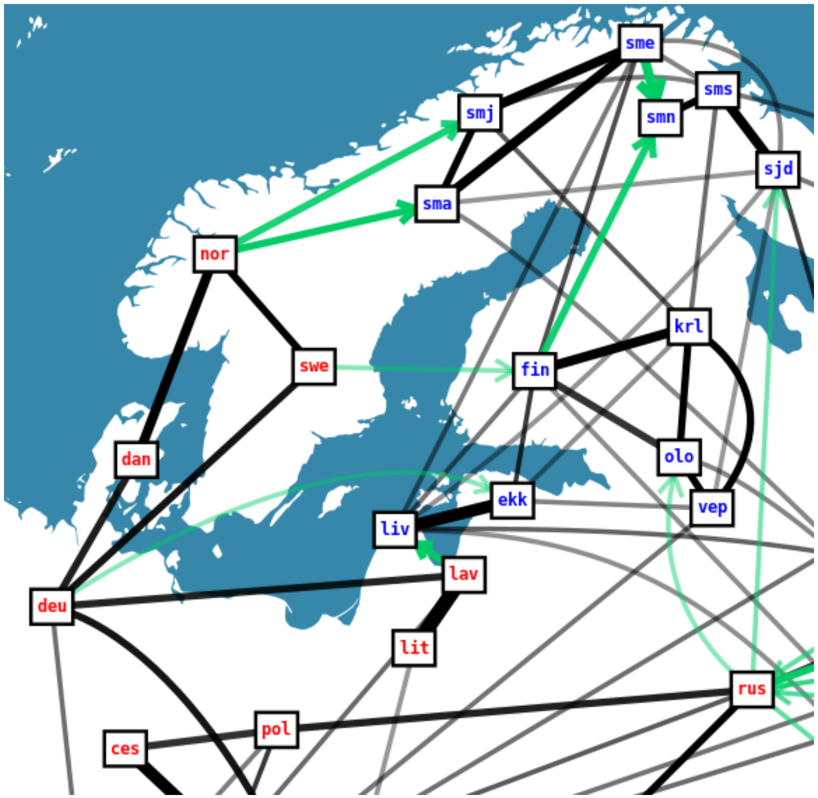
Results

Future Work

Results: Global Picture

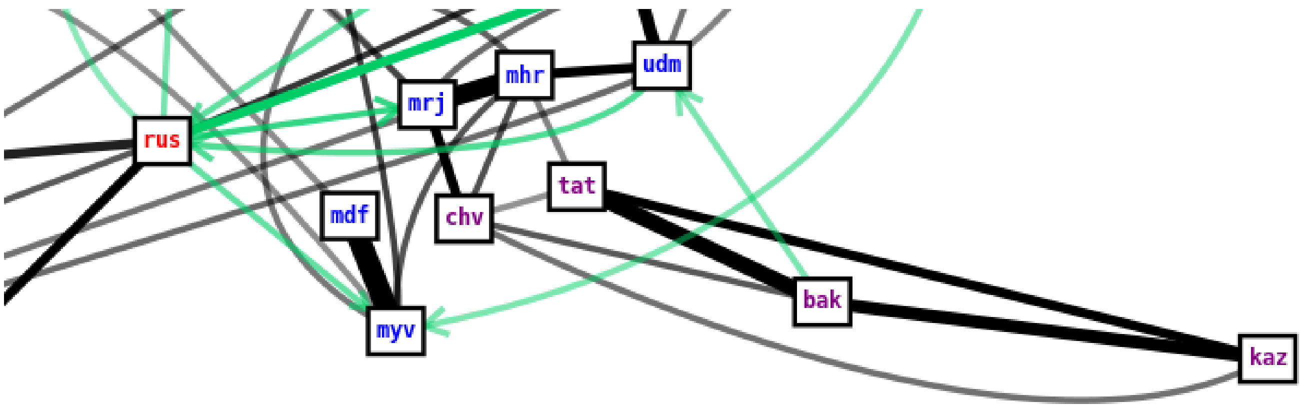


Results: Western Branches



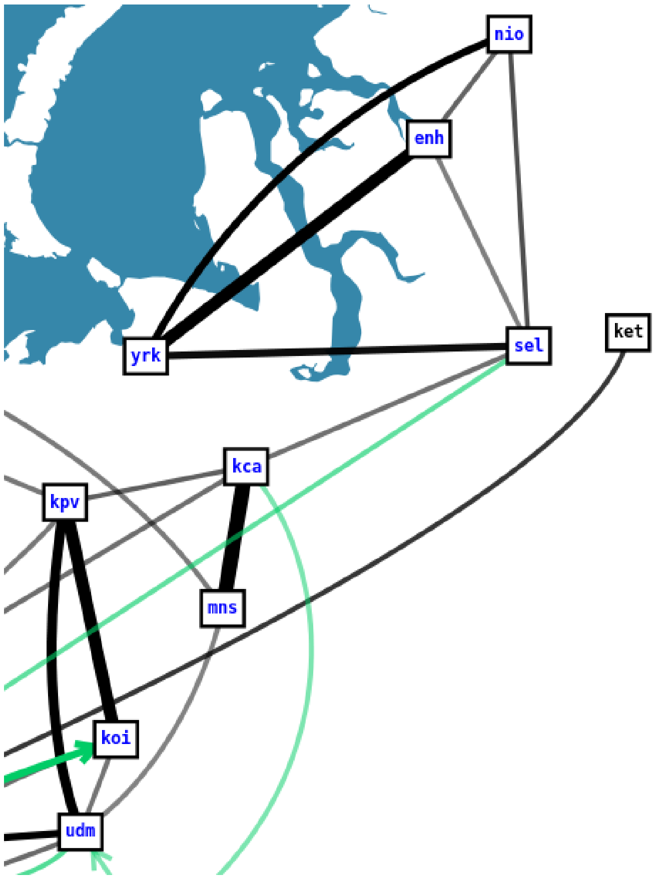
- the causal skeleton is good, shows dialect continua
- all cross-family links are detected as directional, clean split of families
- all inferred arrows point in the correct direction
- slightly questionable: Inari Saami as a mixture of Finnish and Northern Saami

Results: Central Branches



- the causal skeleton is correct (no spurious links)
- wrongly inferred links from Udmurt and Selkup into Russian
- influence of Bashkir on Udmurt is correct
- algorithm does not manage to disentangle contacts of the two Mari languages with the Turkish neighbors Chuvash and Tatar (because most of the contact involved proto-languages?)

Results: Eastern Branches



- the causal skeleton and the strength of links makes sense
- interesting: Selkup is the only Samoyedic language for which an external influence is inferred
- direction of influence between Ket and Russian is not recognized: a problem with isolates, we might need time layering to resolve such cases

Table of Contents

NorthEuraLex

Causal Inference

Lexical Flow Inference

Results

Future Work

Future (and Ongoing) Work

- evaluation on other language families and **simulated data**
- derive explicit models of proto-languages by ancestral state reconstruction on a given phylogeny, use the algorithm to infer **influences between proto-languages**
- explore and evaluate existing techniques for inferring the existence of **hidden common causes**
- move beyond correlate classes, explore conditional independence measures on **realization distances**
- improvements to the data and all its components (IPA transcription, lexical choices, closing gaps)

Acknowledgments

Thanks are due to:

- Alina Ladygina (data for Enets, Hill Mari, and Komi-Permyak)
- Alla Münch (data review for Turkic languages, data for Ket)
- Thora Daneyko (some of the IPA converters, data for Tatar)
- Natalie Clarius (data for Bashkir)
- Ilja Grigorjew (data for Kazakh)
- Roland Mühlenbernd (data for Chuvash)
- Pavel Sofroniev (`sanavirta` visualization tool)

References

Brown, C. H., Holman, E. W., Wichmann, S., and Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals*, 61(4):285–308.

Dellert, J. (2015). Compiling the Uralic Dataset for NorthEuraLex, a Lexicostatistical Database of Northern Eurasia. *Septentrio Conference Series*, 0(2):34–44.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

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

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.