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

Johannes Dellert
Universität Tübingen
Seminar für Sprachwissenschaft
jdellert@sfs.uni-tuebingen.de

December 31, 2015

Abstract

This paper introduces a new method for inducing a language contact model from lexical data. Based on sets of etymologically related words which can be either automatically inferred or expert-annotated, the method analyses possible paths of borrowing in terms of lexical flow. The criterion of vanishing lexical flow gives rise to a conditional independence relation between languages, allowing a variant of the PC algorithm for causal inference to be applied.

The resulting partially directed network represents a parsimonious model of common ancestry and directional contact between the languages in the dataset. In an evaluation on a large lexical database comprising 1,016 concepts across 26 Uralic languages and 18 neighboring languages, the method is shown to detect and correctly infer the directionality of many instances of cross-family language contact which had a large impact on the basic lexicon.

1 Introduction

Recent computational methods for historical linguistics have the disadvantage of being imprecise due to abstraction over relevant details, but the advantage of weighing more evidence than a human brain can process in principled and reproducible ways. While methods for estimating phylogenies from cognacy judgments are already highly developed and in widespread use (e.g. [1, 2, 3]), the network models

used to account for language contact are still in their infancy [4]. Simple network methods such as Neighbor-Net [5] offer a visual summary of contradictory signals in the underlying character data in the form of reticulations, but do not give any hint about their interpretation. A reticulation in such a network may be the effect of anything from a dialect continuum to massive lexical borrowing.

The Uralic family is an ideal test case for evaluating automated methods because it is relatively small and quite well-understood. Previous computational work on Uralic by the BEDLAN group applies standard methods to infer trees [6] and networks [7], with interesting results concerning the reality of subgroupings above the level of primary branches. For these purposes, the group collected expert cognacy judgments for the realizations of around 300 concepts across 18 Uralic languages.

The method presented in this paper needs more than a few hundred concepts to yield good results, but has the advantage of inferring an explicit directional model of lexical influence between languages. Because expert cognacy judgments for a larger number of concepts are not readily available, I resort to automatically inferred sets of **correlates** (i.e. etymologically related words, not necessarily cognates) over a larger lexical database which covers Uralic and its neighbors.

Building only on correlate sets, the method uses ideas from causal inference to infer a partially directed network between attested languages, where each link represents either common inheritance or contact, and directed links can be taken to represent the dominant direction of borrowing. An initial evaluation on the Uralic dataset shows that the method correctly detects all the major cross-family contact events, and infers the right directionality for all but a few of them.

2 Conditional Independence and Causal Inference

Causal inference [8] is a relatively new subfield of statistics which attempts to infer causal relationships between variables from observational data alone. While the fact that correlation is not causation prevents us from inferring the direction of causality between an isolated pair of variables, the interaction between more than two variables often provides hints about the possible causal scenarios underlying the data.

The core building block of causal inference is a **conditional independence** relation between the variables involved. Intuitively, the conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$ expresses that any dependence between the variables X and Y can be explained by the influence of a third variable Z.

The inference techniques I will be using are inspired by the PC algorithm [9]. The first stage of the PC algorithm uses a sequence of conditional independence tests to reduce a complete graph to a **causal skeleton**, an undirected graph over the variables

where each link expresses an interaction which cannot be explained away by conditioning on other variables. Each removal of a link $X - Y$ relies on finding a **separating set**, i.e. a set of variables $\{Z_1, \ldots, Z_n\}$ such that $(X \perp\!\!\!\perp Y \mid Z_1, \ldots, Z_n)$.

In the second stage, the separating sets which were used to explain away each link are used to detect v-structures among triples of variables, i.e. causal patterns of the shape $X \rightarrow Z \leftarrow Y$. The presence of v-structures typically allows the PC algorithm to infer the directionality of causal influence along many links in the skeleton.

3 The Data

I am testing my inference procedure on a development version of NorthEuraLex [10], a lexical database of Northern Eurasia which aims to cover the realizations of 1,016 concepts across 100 languages of Northern Eurasia. The version I am using contains near-complete coverage of this concept list for 26 Uralic languages, and 18 languages which have historically been in close contact with Uralic languages.

On the Uralic side, the database includes six Finnic languages (Finnish, North Karelian, Livvi-Karelian, Veps, Standard Estonian, Livonian), six Saami languages (Southern Saami, Lule Saami, Northern Saami, Inari Saami, Skolt Saami, and Kildin Saami), the two written variants of both Mordvinian (Erzya and Moksha) and Mari (Meadow Mari and Hill Mari), three Permic languages (Komi-Zyrian, Komi-Permyak, Udmurt), Northern Khanty, Northern Mansi, Hungarian, and four Samoyed languages (Nothern Selkup, Tundra Nenets, Tundra Enets, Nganasan).

The contact languages consist of four Turkic languages (Chuvash, Tatar, Bashkir, Kazakh) and 13 Indo-European languages, including four Germanic (German, Danish, Swedish, Norwegian), two Baltic (Latvian and Lithuanian), and six Slavic languages (Russian, Polish, Czech, Slovak, Croatian, Bulgarian), as well as a single Romance language (Romanian). In addition, the Yeniseian language Ket was included as an important contact language in central Siberia.

3.1 Sound Correspondence Model and Phonetic String Distance

All the data was normalized to the ASJP format [11], a de-facto standard in distance-based approaches which reduces IPA to 41 equivalence classes. The encoding is designed to make long-distance comparison easier, but ignores some features (such as vowel length) that are highly relevant for Uralic. To give a few examples, Northern Saami *čalbmi* is represented as `[CalEbmi]`, and Hungarian *egyedül* as `[ECEdil]`.

For each segment in the ASJP strings, the information content was inferred from trigram models for each language and word class. This is necessary when operating on

dictionary forms, since a string-distance based method would otherwise overestimate the similarity e.g. between verbs which share an infinitive ending.

To enable the detection of cognates which have become dissimilar due to sound change, a model of segment correspondences was inferred separately for each language pair using a variant of the method described by List [12], which results in a segment distance matrix for each language pair. For instance, the distance matrix for Finnish and Hungarian makes it cheap to align a `[k]` to an `[h]`, and the matrix for Hungarian and Northern Saami assigns a low cost to aligning `[s]` and `[C]`. Using the language-specific segment distances and the information content as weights, normalized edit distances were computed for all pairs of realizations of the same concept.

3.2 Correlate Inference

The automated inference of correlate sets (under the name of cognacy detection) is an emerging subfield of computational historical linguistics [13, 12, 14]. My implementation of the LexStat toolchain [15], like the original, uses the UPGMA algorithm [16] to derive a hierarchical clustering of the phonetic strings for each concept based on their pairwise distances, and cuts the tree at a given threshold value to partition the strings into clusters of similar forms.

For any set of languages $L_1, \ldots, L_k$, I will write the number of correlates shared between all of them according to this partition as $c(L_1, \ldots, L_n)$. For a single language L, $c(L)$ then denotes the number of correlate sets covered. This number will later be used for normalization in order to compensate for the effect of the slightly uneven coverage of the concept list for the different languages.

4 Modeling Lexical Flow

The application of the PC algorithm to this dataset presupposes a useful definition of conditional independence between sets of languages. The idea of this paper is to use a **lexical flow model** to define such a conditional independence relation. Building on sets of correlates $cor(L_1, ..., L_k)$ shared by a the languages $L_1, \ldots, L_k$, the independence test can be based on a measure of conditional overlap, which I will call $I(L_1, L_2; Z)$ because of parallels with conditional mutual information:

$$I(L_1, L_2; Z) := \frac{|cor(L_1, L_2) \backslash \{c \mid \exists \{Z_1, \ldots, Z_k\} \subseteq Z \colon c \in cor(Z_1, \ldots, Z_k)\}|}{\min\{|cor(L_1)|, |cor(L_2)|\}} \tag{1}$$

Informally, $I(L_1, L_2; Z)$ quantifies the ratio of correlates between L_1 and L_2 which cannot be explained away by having been borrowed through a subset of the languages in Z. To use this measure of dependence as a conditional independence test, we simply check whether $I(L_1, L_2; Z) \leq \theta_{L_1,L_2}$ for a threshold θ_{L_1,L_2}, which could be derived from the number of false correlates between L_1 and L_2 which we expect due to automated correlate detection. In practice, I am setting $\theta_{L_1,L_2} := 0.02$ for all language pairs because the distribution of false correlates is difficult to estimate, and language-specific thresholds did not lead to better results in initial experiments on a smaller language set. On the NorthEuraLex data, this means that languages which share 20 correlates or less will be unconditionally independent, and every link the algorithm establishes will explain an overlap of at least 20 correlates.

Based on this conditional independence test, the first stage of the PC algorithm derives a causal skeleton which represents a scenario of contacts between pairs of input languages that is only as complex as necessary to explain the lexical overlaps. The model thus assumes that all similarities are primarily due to mutual influence, and never infers the existence of hidden common causes (i.e. proto-languages), although the links without any clear unidirectional signal can be interpreted in this way.

The PC algorithm is tractable because it tests separating set candidates in order of cardinality, and builds on the assumption that any separating set must be a subset of immediate neighbors of L_1 and L_2 in the current skeleton. For our model, we cannot make that assumption, because removal of a link between two languages should not rely on shared correlates with possibly unconnected neighbors. Instead, we need to explicitly model the lexical flow.

To explain away a correlate that is shared between two languages L_1 and L_2, it must have been possible for the lexeme in question to have travelled between the two languages on some other path. Therefore, any minimal separating set must form a union of acyclic paths between L_1 and L_2. My implementation uses a depth-first search of the current graph to get all such paths which contain four nodes or less, and generates all combinations of these paths which lead to separating set candidates of a given cardinality. Longer paths would need to be considered in theory, but did not lead to different results on my data, at a much higher computational cost.

5 Deciding Directionality

In the second stage of the standard PC algorithm, directionality inference on the causal skeleton is performed by asking whether the central variable in each pattern of the form $X - Z - Y$ was part of the separating set that was used for explaining away the link $X - Y$. The idea is that if Z was not necessary to explain away $X - Y$, this

excludes all causal patterns except $X \rightarrow Z \leftarrow Y$. Since there will often be many separating sets of the same size, the result of this decision procedure can be highly dependent on the order in which separating set candidates are tried out. In practice, this means that many possible orders have to be tested, often giving rise to conflicting evidence which needs to be reconciled. Moreover, this type of inference relies on the very strong assumption that every scenario in which X has an influence on Y and Y on Z, this would become visible as a dependence between X and Z. While this assumption may be unproblematic for continuous statistical variables, it is certainly not true for our notion of independence, since it is easily conceivable that if a language L_1 borrows from a language L_2 which in turn borrows from a language L_3, none of the lexical material from L_3 will appear in L_1.

Still, the essential idea behind this reasoning can also be applied to our case. For each triple of languages (L_1, L_2, L_3) and a given causal scenario, we can measure the difference between the expected number of correlates shared between all three languages, and the observed number of such correlates. More precisely, if the number of observed correlates $c(L_1, L_2, L_3)$ is significantly lower than the number we expect under any causal assumption which includes $L_1 \leftarrow L_2$, this gives us evidence in favor of the arrow $L_1 \rightarrow L_2$. So what is the expected number of shared correlates between all three languages under the assumption $L_1 \leftarrow L_2$? Assuming independent instances of language contact, both scenarios $L_1 \leftarrow L_2 \leftarrow L_3$ and $L_1 \leftarrow L_2 \rightarrow L_3$ allow us to multiply the ratios $r(L_1, L_2) := c(L_1, L_2)/\min\{c(L_1), c(L_2)\}$ and $r(L_2, L_3) := c(L_2, L_3)/\min\{c(L_2), c(L_3)\}$ to arrive at the percentage of $c(L_1, L_3)$ that we expect to also be shared with L_2.

In a triangle $L_1 - L_2 - L_3$, the amount of the information we can derive about the directionality of $L_1 - L_2$ in this way becomes higher the more correlate overlap there is between L_2 and L_3. This gives rise to a definition of the **counterevidence score** $sc(L_1 \rightarrow L_2)$ for the arrow $sc(L_1 \rightarrow L_2)$ based on a weighted sum over all triples:

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

Based on these scores, directionality decisions for each language pair $L_1 - L_2$ can be made by comparing the strength of counterevidence for $sc(L_1 \rightarrow L_2)$ and $sc(L_2 \rightarrow L_1)$. For the experiment, my implementation assumes that the evidence favors one direction if the ratio of counterevidence scores is lower than 0.9. As in the standard interpretation of causal graphs returned by the PC algorithm, counterevidence score ratios near 1.0 can be interpreted as being a consequence of either bidirectional influence (mutual borrowing) or a hidden common cause (ancestral relationship). Algorithm 1 gives an overview of the entire resulting inference procedure in pseudocode.

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 $P_1, ..., P_k$ 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

6 Results

Using my Java implementation, applying the method to the NorthEuraLex correlate sets takes less than five minutes on a single core with 2.2 GHz. Figure 1 shows the resulting network. For the visualization, languages (symbolized by their ISO 639-3 codes) are placed roughly at their geographical positions. Contact arrows are colored green, and links for which directionality evidence did not exceed the threshold are in black. The thickness of each edge symbolizes the amount of unexplained flow, i.e. the amount of lexical flow which the model assumes must have gone through the link in question.

Note that Indo-European (red nodes) and Uralic (blue nodes) form two clusters of black edges which are only connected by contact edges, i.e. the model correctly separates these two language families, and can explain all lexical similarity between them by contact alone. The same is not true for the separation of Turkic (purple nodes) and Uralic, though. The complex interaction between Chuvash (chv) and the

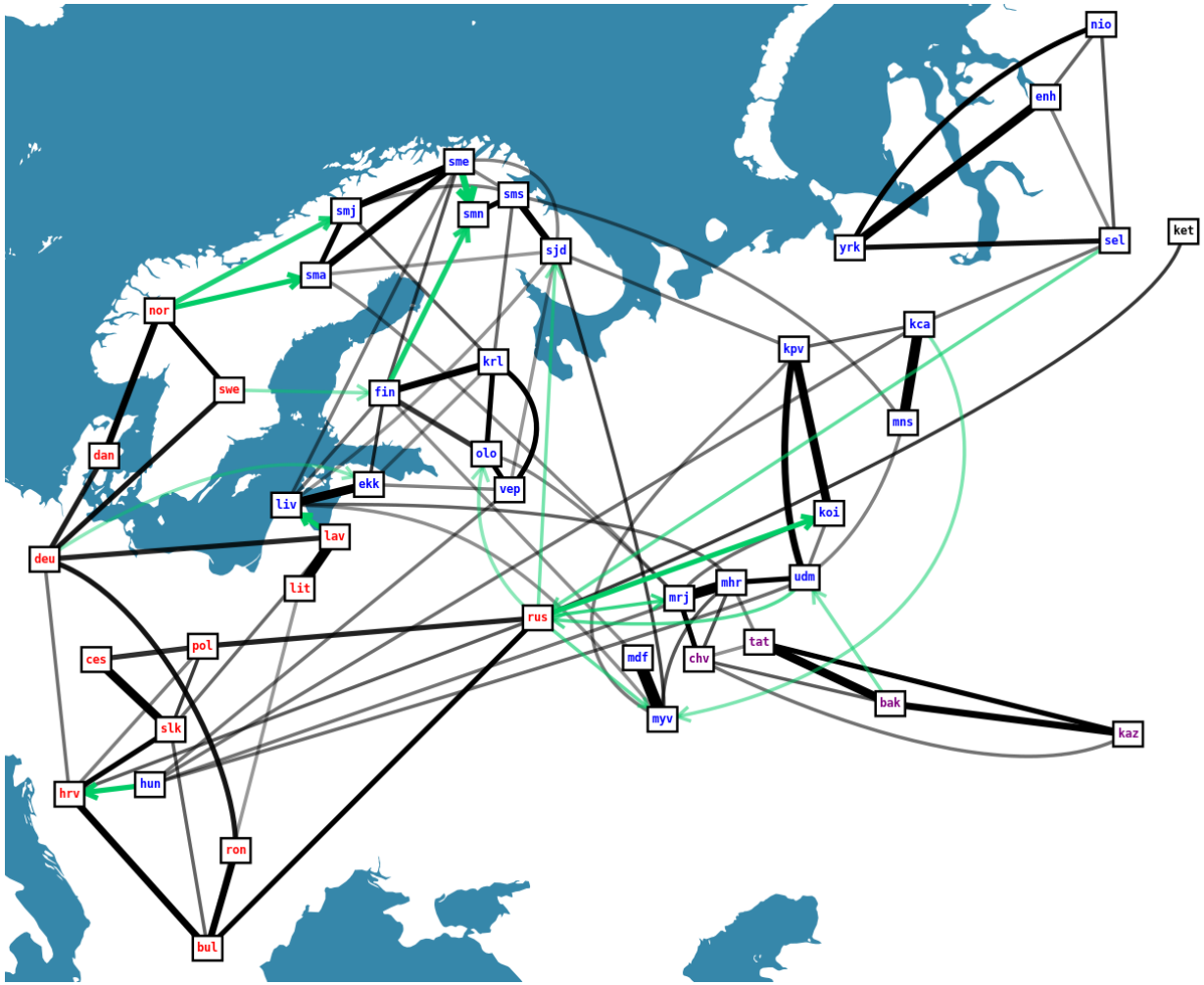

Figure 1: The lexical flow network derived from the data.

two variants of Mari prevents the method from deriving a directionality. One problem may be that Chuvash actually exerted most of its considerable influence on an earlier historical stage of Mari, and not separately on Hill Mari (mrj) and Meadow Mari (mhr), as the model was forced to infer. Otherwise, all the contacts with Turkic languages inferred by the model are correct, though of course far from exhaustive [17].

The internal structure of the language families becomes visible rather nicely in the different intensities of inferred lexical flow among members of the same branch and across branches. For instance, the connection between Latvian (lav) and Lithuanian (lit) is much stronger than the one between Latvian and German (deu). Also, the dialect chain structure of Saami [18] and Finnic [19] becomes quite clearly visible, as geographical neighbors share more lexical material than do more distant pairs of languages from these branches.

Considering the contacts around the Baltic sea, the model correctly detects strong influence of North Germanic [20] (represented by Norwegian) on Lule Saami (smj) and Southern Saami (sma). Interestingly, the remaining lexical overlap between Northern Saami (sme) and Norwegian is not enough to infer direct contact, since all the shared correlates may have entered Northern Saami through either Lule or Southern Saami. The strong influence of Swedish (swe) on Finnish (fin) is detected just as well as the influence of German on Estonian (ekk) [21]. Since this arrow can be interpreted to show the influence of the Teutonic Order, the link between German and Latvian could have displayed the same direction, but the flow representing the ancestral relationship of the two languages is stronger than the layer of German loans in Latvian. The contact between Livonian (liv) and Latvian is seen as monodirectional, which is justifiable for the lexicon because there are many Baltic loanwords in Finnic, and an additional stratum of later Latvian loans into Livonian [22]. This link is further strengthened by material from German in Livonian, all of which can be explained as either going through Estonian or Latvian. Finally, the lexical overlap between Russian (rus) and Livvi-Karelian (olo) as well as Kildin Saami (sjd) is correctly recognized as being due to heavy Russian influence on the other two languages [23].

Russian is also correctly detected as exerting considerable influence on many of the Uralic minority languages [23]. In all cases except Udmurt (udm) and Selkup (sel), Russian is correctly recognized as the donor language. Interestingly, any lexical material shared between Russian and the Northern Samoyed languages (yrk, enh, nio) is also shared with Selkup, causing the model to assume that all Russian material in Northern Samoyed was transmitted via Selkup. This is unexpected, because Russian influence on Tundra Nenets (yrk) was actually much stronger than on Selkup [23], but the inferred pattern may be true for the more basic lexicon covered by the database.

The erroneous black edge between Ket (ket) and Russian illustrates that the method runs into problems when faced with an isolate. A possible way towards resolving this

issue would be to check whether the shared correlates belong to the most stable basic vocabulary or to later strata, which would indicate contact as opposed to a genealogical relationship.

7 Conclusion and Outlook

We have seen that causal inference built on a conditional independence relation defined by vanishing lexical flow is a powerful tool for inferring network models of language contact. This new type of network has the advantage of also expressing hypotheses about the dominant direction of lexical borrowing.

The evaluation on Uralic and its contact languages has shown that the major cross-family contact events in the history of the current Uralic languages are correctly detected. Moreover, the method was always right when it inferred the existence of a link, although it could not detect the directionality for all instances of language contact, especially when an isolate was involved. Altogether, the method promises to be a worthwhile tool for providing initial hypotheses about language relationships in less well-researched linguistic areas.

The major problem of the current version is that it does not model the existence of proto-languages, and will therefore always model all instances of contact as occurring between observed languages. In reality, many of the detected contacts will actually have occurred between proto-languages. In future work, the lexical flow model will therefore be combined with ancestral state reconstruction [24] to provide a hypothesis about the correlate sets present at proto-languages, which should make it possible to also infer the existence and directionality of contacts between proto-languages, if they explain the data more parsimoniously than the assumption that only the observed languages influenced each other.

Acknowledgments

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP, which is gratefully acknowledged. My thanks go to Alina Ladygina for contributing the data for Tundra Enets, Hill Mari, and Komi-Permyak. Furthermore, thanks are due to Alla Münch, Ilja Grigorjew, Thora Daneyko, Natalie Clarius, and Roland Mühlenbernd for their help in collecting the data for the Turkic languages and Ket. Finally, I would like to thank Pavel Sofroniev for implementing the Sanavirta visualization component which was used to create the map.

References

[1] Clare Janaki Holden. 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, 2002.

[2] Michael Dunn, Stephen C Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia. *Language*, 84(4):710–759, 2008.

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

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

[5] David Bryant and Vincent Moulton. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265, 2004.

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

[7] Jyri Lehtinen, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg, and Outi Vesakoski. Behind Family Trees - Secondary Connections in Uralic Language Networks. *Language Dynamics and Change*, 4(2):189–221, 2014.

[8] Judea Pearl. *Causality.* Cambridge University Press, 2009.

[9] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT Press, 2nd edition, 2000.

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

[11] Cecil H. Brown, Eric W. Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals*, 61(4):285–308, 2008.

[12] Johann-Mattis List. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France, April 2012. Association for Computational Linguistics.

[13] Bradley Hauer and Grzegorz Kondrak. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 865–873, 2011.

[14] Taraka Rama. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 31 – June 5, 2015 Denver, Colorado, USA*, pages 1227–1231, 2015.

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

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

[17] András Róna-Tas. Turkic influence on the Uralic languages. In Sinor [25], pages 742–780.

[18] Pekka Sammallahti. Saamic. In Abondolo [26], pages 43–95.

[19] Tiit-Rein Viitso. Fennic. In Abondolo [26], pages 96–114.

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

[21] Sándor Rot. Germanic influences on the Uralic languages. In Sinor [25], pages 682–705.

[22] Seppo Suhonen. Die baltischen Lehnwörter der finnisch-ugrischen Sprachen. In Sinor [25], pages 596–615.

[23] Gyula Décsy. Slawischer Einfluss auf die uralischen Sprachen. In Sinor [25], pages 616–637.

[24] Gerhard Jäger and Johann-Mattis List. Investigating the potential of ancestral state reconstruction algorithms in historical linguistics. Workshop "Capturing Phylogenetic Algorithms in Linguistics", Lorentz Center, Leiden, 2015.

[25] Denis Sinor, editor. *The Uralic Languages. Description, History and Foreign Influences.* Handbuch der Orientalistik 8. Brill, Leiden, 1988.

[26] Daniel M. Abondolo, editor. *The Uralic Languages.* Language Family Descriptions Series. Routledge, 1998.