Contacts of Finnic and Saami languages in computational trees and networks

Jyri Lehtinen (BEDLAN/UraLex / University of Helsinki)

Language Contact: the State of the Art, Helsinki, 28–30 Aug 2014
Computational linguistic phylogeny

Using computational methods for studying evolutionary phylogeny

**Phylogeny** = representation of divergence history represented by the data, e.g. in evolutionary biology

<table>
<thead>
<tr>
<th>Concept</th>
<th>Biological phylogeny</th>
<th>Linguistic phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon:</td>
<td>Species, population, etc.</td>
<td>Language, dialect, etc.</td>
</tr>
<tr>
<td>Character:</td>
<td>DNA locus, anatomical characteristic, etc.</td>
<td>Lexical cognate set, structural trait, etc.</td>
</tr>
</tbody>
</table>

Allows basing interpretation on large amounts of data

- Compare with **qualitative** examination of few important innovations in historical linguistics
- Computational phylogeny is a **quantitative** approach, but **depends** on traditional hist. ling. study for good data
Historical linguistics in evolutionary framework

Why use methods for evolutionary relationships between biological organisms?

Parallels between the processes of biological evolution and language change

<table>
<thead>
<tr>
<th>Concept</th>
<th>Biological evolution</th>
<th>Language change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmitted material</td>
<td>DNA (biochemical structure)</td>
<td>Phonemes, morphemes, etc. (culturally distributed representations)</td>
</tr>
<tr>
<td>Vertical transmission</td>
<td>Inheritance of DNA from parent organism(s)</td>
<td>Learning of linguistic forms and constructions by speaker    (usually from multiple sources)</td>
</tr>
<tr>
<td>Horizontal transmission</td>
<td>Lateral gene flow (esp. in single-celled organisms)</td>
<td>Borrowing, structural contact influence</td>
</tr>
<tr>
<td>Adaptive evolution</td>
<td>Natural selection</td>
<td>Spread of innovated linguistic norms</td>
</tr>
</tbody>
</table>

→ Justification from parallel properties of biological and linguistic processes
Phylogenetic methods in historical linguistics

Typically according to the family tree model

- **Algorithms**: UPGMA, Neighbor-Joining, Maximum parsimony, “Perfect phylogeny” (Ringe et al. 2002), Bayesian MCMC maximum likelihood...
- Language families: Austronesian, Indo-European, Bantu, Arawak, Semitic, etc.

Also different methods for representing reticulation (conflicting connections)

- **Algorithms**: Network, Split Decomposition, NeighborNet...

Both family tree-like and network-like models represent different realistic mechanisms

Tree-like pattern: discrete divergence typically associated with expansions

Network-like pattern: convergence (borrowing), wave-like divergence in dialect continua
On trees

In many cases trees are too simplistic, as different data point to different divergence histories

- Different models for conflicting criteria: Wave model; Glottometry
- Computational trees with vocabulary data represent:
  - Clearest treelike signal of border formation between lexical diffusion areas,
  - not, for instance:
    - ”The” divergence history of ”languages” (various conceptualizations of ”language”, no clear hierarchical separation of language communities in linkages)
    - Divergence history that would correspond with other domains like phonological innovations

Does the data actually support a tree model?

Alternative: networks

- Show conflicting splits in a reticulated pattern
Character-based trees in comparison with distance networks

a) **Bayesian MCMC tree**
   - Model of 5 languages (simulated data)
   - Algorithm goes through data each character (cognate/correlate set) at a time to find phylog. signal
   - Discrete branchings with support values, low values indicate conflict in data

b) **NeighborNet** with same data
   - Networks with distances between whole datasets of different languages
   - Splits display conflicting connections simultaneously
   - Can include bootstrap values indicating statistical robustness
Lexical data of Uralic languages

Concentrating here on **Saami** and **Finnic** languages:

- South Saami
- Ume Saami
- Pite Saami
- North Saami
- Inari Saami
- Skolt Saami
- Kildin Saami
- Finnish (Std)
- Karelian Proper
- Ingrian
- Veps
- Votic
- Estonian (Std)
- South Estonian (Võro)
- Livonian
Lexical data in linguistic phylogeny

Lexical data: cognate-based coding vs. correlate-based
- Cognate-based: Directly descended from the same ancestral form (every internal borrowing creates a new cognate set!)
- Correlate-based: Any kind of same origin, through descent, borrowing etc.

Characters: cognates or correlates having absence/presence/unknown values (0/1/?)

Meaning-based data collection to detect innovations in form and function

Alternative data types
- Phonological/morphological innovations (e.g. Ringe et al. 2002)
- Absence/presence of typological features (Dunn et al. 2005, Dunn et al. 2008)
### Example from Uralic data

<table>
<thead>
<tr>
<th>Proto-Uralic (outgroup)</th>
<th>FISH</th>
<th>WATER</th>
<th>EAR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>South Saami</strong></td>
<td>kala</td>
<td>weti</td>
<td>[Not rec’able]</td>
</tr>
<tr>
<td><strong>North Saami</strong></td>
<td>guolli</td>
<td>čáhci</td>
<td>beallji</td>
</tr>
<tr>
<td><strong>Inari Saami</strong></td>
<td>kyeli</td>
<td>čāći</td>
<td>pelji</td>
</tr>
<tr>
<td><strong>Kildin Saami</strong></td>
<td>kūll’</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Standard Finnish</strong></td>
<td>kala</td>
<td>–</td>
<td>korva</td>
</tr>
<tr>
<td><strong>Ingrian</strong></td>
<td>kala</td>
<td>–</td>
<td>korva</td>
</tr>
<tr>
<td><strong>Western Votic</strong></td>
<td>kaa</td>
<td>–</td>
<td>kër̄v</td>
</tr>
<tr>
<td><strong>Standard Estonian</strong></td>
<td>kala</td>
<td>–</td>
<td>körv</td>
</tr>
<tr>
<td><strong>Võro South Estonian</strong></td>
<td>kala</td>
<td>–</td>
<td>körv</td>
</tr>
<tr>
<td><strong>Courland Livonian</strong></td>
<td>kalà</td>
<td>–</td>
<td>kúora</td>
</tr>
<tr>
<td><strong>Erzya</strong></td>
<td>kal</td>
<td>ved’</td>
<td>–</td>
</tr>
<tr>
<td><strong>Meadow Mari</strong></td>
<td>kol</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Komi-Zyrian</strong></td>
<td>–</td>
<td>va</td>
<td>pel’</td>
</tr>
<tr>
<td><strong>Udmurt</strong></td>
<td>–</td>
<td>vu</td>
<td>pel’</td>
</tr>
<tr>
<td><strong>Hungarian</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Vakh-Vasyugan Khanty</strong></td>
<td>kul</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FISH</th>
<th>WATER</th>
<th>EAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 0 1 0 1 0</td>
<td>1 0 0 1 0 1 0 0</td>
<td>1 0 0 0 1 0 0 0</td>
</tr>
<tr>
<td>1 0 0 0 1 0 1 0 0</td>
<td>1 0 0 1 0 0 1 0</td>
<td>1 0 0 0 1 0 0 0</td>
</tr>
<tr>
<td>1 0 0 1 0 0 0 1 0</td>
<td>1 0 1 0 0 0 0 1 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
<tr>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
<td>1 0 1 0 0 0 1 0 0</td>
</tr>
</tbody>
</table>

**Proto-Uralic (outgroup)**
- *guelie* etc.
- "ćeri" etc.
- *vesi* etc.
- "tjaetsie" etc.
- "bieljie" etc.
- *korva* etc.
- "xa" etc.
- *kala* – *weti* – [Not rec’able] [Not rec’able] [Not rec’able]
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
- *kala* – *vesi* – *korva* –
Uralic meaning lists

All: 25 languages
- 7 Saami, 8 Finnic, 1 Mordvin, 1 Mari, 3 Permic, 3 Ugric, 2 Samoyed

**Full 313** meanings dataset
- Both “basic vocabulary” and less stable vocabulary
- 313 meaning items, altogether 3514 correlate sets

**Ura149** stable meanings list
- Historically stable vocabulary in Uralic (little borrowing, little cross-branch variation)
- Meanings for which words have: a) at most 3 loan etymologies, b) at most 3 absences of word and c) at most 14 cognate sets that the words belong to in different languages (determined from set of 24 Uralic languages)
- 149 meaning items, 1125 correlate sets

**WOLD401-500** less stable vocabulary list
- Meanings that based on WOLD data (Haspelmath and Tadmor 2009) are more easily borrowed, replaced etc. than crosslinguistically more stable meanings (ranked 401-500 from over 1400)
- 101 meaning items, 1288 correlate sets
Network representation of Uralic vocabulary connections

\( \delta \) (delta) is a statistical measure of "treelikeness" of the data, \( \delta \) close to 0 means very treelike.
Bayesian MCMC tree based on Uralic specific basic vocabulary meaning list (149 meaning items)
Alternative visualization of splits as "historical isoglosses"

\[ w > 0.02 \]
\[ 0.02 > w > 0.005 \]
\[ 0.005 > w > 0.001 \]
\[ w < 0.001 \]

(\( w \): weight of split; only robust splits with bootstrap > 0.75 included)

(neighborhood visualization of splits represented as line length)
Finnic and Saami in close inspection

Finnic
- Livonian
- Votic
- Standard Estonian
- South Estonian (Võro)
- Standard Finnish
- Karelian Proper
- Ingrian
- Veps

Saami
- South Saami
- Ume Saami
- Pite Saami
- North Saami
- Inari Saami
- Kildin Saami
- Skolt Saami

Full 313 data
Connections based on most stable vocabulary

Ura149 stable vocabulary
Comparison: stable vs. unstable vocabulary (impact of borrowing and more recent innovation)
Comparison: stable vs. unstable vocabulary (impact of borrowing and more recent innovation)
Comparison with historical-comparative study

Kallio 2007
Phonological innovations;
Qualitative study
Conclusions from vocabulary

- Trees provide accurate histories of divergence
- Networks show factors behind unclear branching
- Networks display conflicting connections of Finnic and Saami languages
- Discrete Finno-Saami, Finnic and Saami branches
- Contact between Finnic and Saami evident
Typological data in phylogenetic analysis

- Trees and clustering of the languages of Sahul region (Melanesia, Australia) based on typological data
- 160 typological features in Reesink et al. 2009 (mostly binary)

- Typological data on Uralic languages
  - Features from Dunn et al. (some not varying within Uralic)
  - Features from WALS that are useful for Uralic
  - Other Uralic-specific features
  - Phonology, lexicon, gramm. category marking and morphosyntax
### Examples of typological features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JustUncle</td>
<td>Lexicon</td>
<td>Same word used for mother’s brother and father’s brother</td>
</tr>
<tr>
<td>RelAgeDistRelat</td>
<td>Lexicon</td>
<td>Different words for younger and older relative for other relations than siblings</td>
</tr>
<tr>
<td>MorphIdioscrGiveVerb</td>
<td>Lexicon</td>
<td>Is the verb for ‘give’ morphologically idiosyncratic, e.g. irregularly inflected, compared to most other verbs</td>
</tr>
<tr>
<td>Cases</td>
<td>NPCategories</td>
<td>No. of productive inflectional cases</td>
</tr>
<tr>
<td>DualInN</td>
<td>NPCategories</td>
<td>Dual marker in nouns</td>
</tr>
<tr>
<td>MarkedPossd</td>
<td>NPCategories</td>
<td>Marker of possession can be on nominal possessed entity</td>
</tr>
<tr>
<td>3PPronGend</td>
<td>NPCategories</td>
<td>Gender or animacy distinction in 3rd person pronouns (and not elsewhere in the grammar)</td>
</tr>
<tr>
<td>DualInPron</td>
<td>NPCategories</td>
<td>Dual category in pronouns</td>
</tr>
<tr>
<td>U3DistVisInDem</td>
<td>NPCategories</td>
<td>Opposition between three or more distance/visibility terms in the demonstrative system</td>
</tr>
<tr>
<td>MainVerbNegMark</td>
<td>VerbalCategories</td>
<td>Can negation be marked on the main verb</td>
</tr>
<tr>
<td>NegAux</td>
<td>VerbalCategories</td>
<td>Is there an auxiliary verb for forming negative constructions (negative auxiliary)</td>
</tr>
<tr>
<td>TAMMarkdOnNegAux</td>
<td>VerbalCategories</td>
<td>Can tense, aspect and/or mood be marked on negative auxiliary</td>
</tr>
<tr>
<td>PersNumTAMFusion</td>
<td>VerbalCategories</td>
<td>Can person, number and TAM be marked with same portmanteau morph on verbs</td>
</tr>
<tr>
<td>ImpersWPersMorph</td>
<td>VerbalCategories</td>
<td>Expression of impersonality with same forms as person/number marking</td>
</tr>
</tbody>
</table>
Goals for Uralic typological data

End result: Quantitative typological mini-profiles of Finnic, Saami and many other Uralic languages

Uralic typological phylogenetics
Trees with typological features? Or better stick with networks?
Comparison between typological and lexical distances
Comparison with other language families
Thank you!

Get:
- These slides
- New preprint version of accepted article (Lehtinen et al. 2014, *Language Dynamics and Change*)
- Other material from research by BEDLAN

“in the coming days” on my page:

http://www.ling.helsinki.fi/~jdlehtin/