The position and internal connections of Finno-Saami: A computational perspective

JYRI LEHTINEN

MAPPING METHODS, TARTU 8-10 MAY 2014
Background

- **The BEDLAN project, 2009-2013**
  (Funded by Kone Foundation; three Finnish universities and the Institute for the Languages of Finland, Kotus)
  - Publications: Honkola et al. 2013 (*Journal of Evolutionary Biology*);
    Syrjänen et al. 2013 (*Diachronica*);
    Lehtinen et al. Forthcoming (*Language Dynamics and Change*)

- Two successor projects from beginning of 2014
  - **SUMURA-syyyni** (led by Outi Vesakoski, Univ. Turku, MTT):
    Microevolution of **Finnish dialects**;
    Computational analysis of **Finnic and Saami** connections
  - **UraLex** (Antti Leino, Univ. Tampere)
    Database of **Uralic vocabulary** (semasiological and etymological);
    Computational analysis of **typological variation** of Uralic languages

> [http://kielievantuutio.uta.fi](http://kielievantuutio.uta.fi)
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 of languages separated after expansion
  - Network-like pattern: convergence (borrowing), wave-like divergence in dialect continua
Linguistic data in phylogenetic methods

- 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)
Computational family trees of Uralic languages

- Syrjänen et al. 2013: Basic vocabulary data with Bayesian MCMC methodology
- Represents widely accepted subgroups clearly with minimal confusing effects of loanwords
- Supports Finno-Ugric, Ugric and Finno-Saami branches
- Limited support for Finno-Volgaic and Finno-Mordvin
- No clear Finno-Permic lexical layer
Basic vocabulary optimization for Uralic

- Earlier: Ura100 (no borrowings in vocabulary sources)
- New: Ura149 (larger dataset, based on more criteria of basic vocabulary)

Branch order essentially the same; illustrates the low impact of less stable vocabulary on phylogenetic analysis

226 item basic vocabulary

Ura149 optimized basic vocabulary
Beyond trees

Trees only show splits, no convergence

Why weak support for a split?
  ◦ Which alternative connections make splitting ambiguous?

Tree algorithms always produce trees
  ◦ Does the data actually support a tree model?

Alternative: networks
  ◦ Show conflicting splits in a network-like pattern
Comparison:
Character-based trees vs. distance-based networks

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

- **NeighborNet** with same data
  - Algorithm operates by distances between whole datasets of different languages
  - Splits display conflicting connections simultaneously
  - Can include bootstrap values indicating statistical robustness
The distance network of Uralic basic vocabulary

- Network shows generally low reticulation: good fit for a tree
- Weak connections between subgroups: supports rapid diversification
- More complexity in Finnic and Saami
Finnic and Saami in distance networks

- Tree with Bayesian maximum likelihood finds mostly bifurcating pattern in Finnic but not Saami
- However, network shows much conflict within Finnic as well
- Many crisscrossing splits in Saami show the complexity of lexical connections
- Robust splits separating expected groups (South-Ume; South-Ume-North; Skolt-Kildin)
- Conflicting splits connecting neighbouring varieties

- Finnic north-south separation reflected in two robust splits (99.1, 97.1)
- Shows that Izhor-Finn-Karelian branch corresponds with a weak but fairly robust split (81.9)
- No Est-Võro-Votic split, but Livonian is separated from these along with Veps as conservative within Finnic (80.1)
Comparison: Optimized basic voc. vs. less stable voc.
Saami

Kildin Saami
Skolt Saami
North Saami
Inari Saami
Ume Saami
South Saami

Veps
Karelian Proper
Izhorian
Finnish

Ura149 optimized basic vocabulary

WOLD401-500 less stable vocabulary

Finnic

Livonian
Võro
Votic
Estonian
Izhorian
Karelian
Finnish
Võro
Estonian
Votic

Veps
Karelian
Izhorian
Finnish
Population clustering methods and variation in language

- **STRUCTURE**: algorithm for estimating the share of ancestral populations in studied individuals/populations
  - A prominent example: *The genetic structure of human populations* (Rosenberg et al. 2002)
- **STRUCTURE** with linguistic data
  - Clustering of the languages of Sahul region (Melanesia, Australia) based on typological data (Reesink et al. 2009)
  - Classifying extinct Tasmanian languages based on word lists (Bowern 2012)
- In BEDLAN, clustering of Finnish dialects with STRUCTURE
Uralic vocabulary in STRUCTURE (preliminary)

- Uralic vocabulary: estimated best value of K: 6
- However, with K=6 (as with other values), uninformative clustering, especially for individual STRUCTURE runs

Inference:
Method for population analysis not suited for studying variation between languages?
(at least with vocabulary data)
Next:
Computational analysis of Uralic typological variation

  - 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
  - End result: Quantitative typological mini-profiles of Finnic, Saami and many other Uralic languages
- Uralic typological phylogenetics
  - Trees with typological features?
  - Comparison between typological and lexical distances
  - Comparison with other language families
Conclusions

- 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
- STRUCTURE shows no useful picture of “ancestral linguistic populations” as contributing to Uralic lexical variation
References