Shared structural features in Transeurasian languages: borrowed or inherited?

Nataliia Neshcheret

Eurasia3angle,
Max Planck Institute for the Science of Human History

29 Aug 2018, SLE Tallinn
Language sample
Introduction

How are these methods and this kind of data useful?

Q1 How can my research contribute to the debate on the internal structure of the Transeurasian family?

Q2 Can we define a “Transeurasian” area, based on structural features, which stands out among other language families in the area?

Q3 What methods do other disciplines offer for investigation of questions from macro-typology and historical linguistics?

Q4 What can structural features tell us about the relationships between languages in question?

Q5 What is the impact of language contact on the structural change of the Transeurasian languages?

Q6 Are there differences in structural features regarding the amount of genealogical signal?

Q7 How does the topology change, if structural features with the lowest phylogenetic signal are excluded from the analysis?
The challenge: Structural features

- “where the lexical signal has been lost, a faint structural signal might still be discernible” (Dunn et al. 2005)
- “the most stable structural features of languages could be useful for deep historical reconstruction just like the most conservative portion of the vocabulary” (Dediu and Levinson 2012)
- “Structural features necessarily have a more attenuated historical signal than lexical features, since shared structural features may originate from borrowing and convergent evolution (homoplasy) as well as from inheritance.” (Reesink et al. 2009)
- “[...] on average, most grammatical features actually change faster than items of basic vocabulary” (Greenhill et al. 2017)
Material and methods

What and how?

- **data:**
  - 38 Transeurasian languages (9,576 data points)
  - 4 non-Transeurasian languages
  - 228 structural features (189 Grambank features, 39 features on phonology and formal representation)

- **sources:**
  - language descriptions,
  - dictionaries,
  - native speakers,
  - language specialists

- **methods:**
  - Bayesian tree-sampling,
  - neighbour-joining
  - phylogenetic comparative methods
Feature set

- **morphosyntactic features:**
  - person, number, possession, interrogation, negation, derivation patterns, valency operations, numeral systems, comparison, argument marking, deixis)

- **phonological features**
  - voicing distinction in plosives/fricatives, l/r distinction, constraints on initial consonants, availability of initial consonant clusters, vowel harmony, vowel length
Coding example

(1) Udehe (Tungusic; Nikolaeva and Tolskaya 2001: 840)

\[
mamasa\ ule:-we\ olokto-\ ini
\]
old.woman meat-ACC cook-3SG

‘The old woman is cooking meat.’

(2) Khalkha (Mongolic; Janhunen 2012: 246)

\[
oxai\ mo:\r-i:g\ bari-eb
\]
dog cat-ACC catch-TERM

‘The dog caught the cat.’

- Is pragmatically unmarked word order verb-final for transitive clauses? → yes, 1
- Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause? → yes for Udehe, 1, → no for Khalkha, 0
### Raw data

<table>
<thead>
<tr>
<th>language</th>
<th>verb-final word order</th>
<th>A argument marked on the verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Udehe</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Khalkha</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Q1, Q4: The TEA (=Transeurasian) forest

- lexicostatistics tree (vocabulary) (Starostin et al. 2003)
  - Turkic
  - Mongolic
  - Tungusic
  - Koreanic
  - Japonic

- classical comparative tree (Robbeets 2015)

- lexicostatistics tree (case suffixes) (Blážek and Schwarz 2014)
  - Turkic
  - Mongolic
  - Tungusic
  - Koreanic
  - Japonic

- Bayesian tree (Robbeets and Bouckaert 2018)
Q1, Q4: Transeurasian topology based on structural features

- Japono-Koreanic vs. Altaic branches
- Tungusic splits off first from the Altaic ancestor
- this structure is stable across all tested models, if neighbouring languages are excluded
Q1, Q4: Transeurasian topology based on structural features

weight threshold = 0.00568

Splitstree4, Huson and Bryant (2010)
Q2, Q5: How do non-Transeurasian languages relate to the Transeurasian languages?

weight threshold = 0.00568
Splitstree4, Huson and Bryant (2010)
Q3: What are the languages with the highest conflicting signal?
Q1-Q5: Transeurasian vs. neighbours: the best-fitting model

BEAST2, Bouckaert et al. (2014)
Q6: How reliable are structural features in deriving phylogenies?

- is there a phylogenetic signal in structural features?
  - calculate Fritz and Purvis’ D (Fritz and Purvis 2010):
    - lower D value indicate a higher phylogenetic signal, higher values are a sign of overdispersion
    - a feature with a high signal will have the same state in sister languages
- how does a phylogeny based on features with a high phylogenetic signal differ from the one based on all the coded features?
  - compare the results of distance-based methods
  - compare the maximum clade credibility trees
- what are the differences in the phylogenetic signal across features? [future research]
  - compare D values across features
  - compare D values across features on different language domains
Q6: GB028 Is there an inclusive/exclusive distinction?  
Estimated $D = 0.08$
Q6: Phylogenetic signal across structural features: Fritz & Purvis’ D

blue = ideal topology, red = the best-fitting model

function *phylo.d*, package CAPER, Orme et al. (2013)
Q6: Are high D values due to feature uniformity?

- if a feature has 41 times a “0” value and 1 time a “1” value, most of the sister branches will have the same value.
Q6: Estimated $D < 0.5$: core Transeurasian

weight threshold = 0.00568
Splitstree4, Huson and Bryant (2010)
Q7: Estimated D < 0.5: Neighbours

weight threshold = 0.00568

Splitstree4, Huson and Bryant (2010)
Q7: Estimated $D < 0.5$: Neighbours

Phylogenetic signal

BEAST 2 Bouckaert et al. (2014)
Conclusions

- topology based on structural features = topology based on basic vocabulary, if neighbours excluded
- “Transeurasian” area not clearly definable due to typological similarity of the Uralic
- Bayesian and neighbour-joining tree-building methods useful for macro-typology
- phylogenetic comparative methods need to be applied with caution
- the phylogenetic signal can be veiled by contact, if extensive
- exclusion of “unstable” features did not provide a topology similar to the one based on basic vocabulary
Acknowledgements

I am grateful to Simon Greenhill, Annemarie Verkerk and Ron Hübler for their methodological support.

Thanks to Aleksandr Savelyev and Sofia Oskolskaya for support in data collection.

The research leading to these results has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 646612) granted to Martine Robbeets.


Reesink, Ger, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity

Robbeets, Martine. 2015. *Diachrony of verb morphology: Japanese and the
Transeurasian languages*, volume 291. Walter de Gruyter GmbH & Co KG.

Robbeets, Martine, and Remco Bouckaert. 2018. Bayesian phylolinguistics reveals the
internal structure of the Transeurasian family. *Journal of Language Evolution*
3:145–162.

dictionary of the Altaic languages*. Brill Leiden.