1. Classification – G & A claim statistical improvements on analyzing lexicostatistics based on PIE and daughter languages. – better absolute chronologies. (a) How do lexicostat analyses compare with phono/morph analysis? Most would say lexicostat and phonology different
2. How well does it work for Slavic?
3. Look at data: (a) Dyen + G&A recognize tree structures for Slavic are not well supported. (b) Therefore Dyen claims 2-dimensional pseudomaps may improve situation.
4. Redd + Green: (a) quantify similarities or differences b/w different sets of data (Dyen vs. Manczak); (b) quantify similarities or differences b/w lexical vs phono/morphological; and (c) to quantify the correlation between geography and the lexical and phon/morphological data sets.

Family tree and or map-like approaches to Slavic languages?

Abstract
Lexicostatistics is decades old, but newer techniques for computational approaches to historical linguistics have gained attention with the rise of more sophisticated methods of data handling. Thus, for example, Gray and Atkinson (2003, Figure 1) claim to have established, using cognates and a Bayesian tree analysis, an authoritative Stammbaum for the Indo-European (IE) language family, including absolute chronologies of its branching.

The present paper examines a smaller subset of IE languages—Slavic—using Bayesian methods and map-like methods in attempt to compare the computational results and model assumptions with received analyses that are closer to the present. We assume that examining a group of languages closer in time to the present, where the splits are more easily verifiable, allows a more fine-grained comparison of different analysis methods. If a close fit can be found between Bayesian trees and maps and traditional analysis in Slavic, it should allow extension to greater time depths and larger families such as Indo-European.

The present paper applies Bayesian trees and map methods to two corpora: the Slavic subset of Indo-European in Gray and Atkinson (2003); and the Slavic text-token set in Mańczak (2004).

Gray and Atkinson 2003 have claimed that new models of analysis may be applied to glottochronology that answer previous criticism of the method and overcome the shortcomings. The outcome of their glottochronological experiment demonstrated impressive results in establishing absolute chronologies for Indo-European which correlate with archaeological (Renfrew’s out-of-Anatolia and Gimbutas’ Kurgan expansion) and genetic evidence (Near-Eastern contribution to the IE gene-pool during the Neolithic) (438). This establishes a root of IE at 8700 BP (Hittite), with Tocharian splitting off at 7900, Greek and Armenian at 7300, Indo-Aryan at 6900, Celto-Germano-Romance at 6100, and Balto-Slavic at 3400.

Slide 1: Slavic languages map and Gray & Atkinson Slavic results
Need Dyen et al Quote about inadequacy of family-tree model for Slavic & Celtic b/c of continued contact. This correlates with low posterior probabilities in Slavic splits vs. higher posterior probabilities in other branches. However, G & A find that Slavic has the lowest PP
whereas Celtic and other branches have high PPs among the well-accepted daughter families. (There are other weak points at deeper time depths, e.g., Indo-Iranian + Albanian.)

In G & A Slavic is rooted at 1300 BP, assuming a date of 700 AD for a terminus post- quem for the dissolution of Proto-Slavic, thus roughly corresponding to the traditional date of 500 AD for the beginning of Slavic migrations from Ukraine. Both the low PP & apparent incorrect clustering of Polish with ESl mean that the tree model does not allow absolute dating for Slavic splits. As Dyen suggests, Slavic requires the use of 2-dimensional maps.

**Figure 1: Balto-Slavic Detail (Gray & Atkinson 2003)**

![Balto-Slavic Detail](image)

**SLIDE: SCAN OF DYEN’s MDS plot**

Dyen et al had run the data but claimed that because of contact after the languages had split, Slavic is better represented as a “psuedomap” (add in page).
SLIDE: REDD plot of Dyen

Dyen’s data, which is also used by G & A, is a Swadesh-style list (200 semantics items for all IE) with 2449 realizations in form (i.e., tokens possible to match) among 842 languages. Dyen’s distance matrix is the lexicostatistical percentage of shared cognates. There is some support for classical groups: E, W, S. Polish again approaches East. Slovene is an outlier. Find commentary in Dyen why they think this is the case.

Mańczak 2004 – distances expressed as raw N of correspondences between pairs

To look at another sample of lexical correspondence Slavic data we looked at Mańczak 2004, which is not a Swadesh list. Rather, it is a set of correspondences in parallel translations of a Gospel text. Each match between pairs is registered for each time that same form (root, where applicable) is used for the same meaning, thus, POL w = UKR v, but POL w ≠ UKR do. Mańczak expressed these as raw numbers of correspondences between pairs with 1816 total realizations.
We converted Mańczak’s raw numbers to a distance matrix and created an MDS plot. We found a better fit for the traditional three groups than Dyen et al. had found. The groups could be oriented geographically, as shown, but while the branches were oriented correctly, their situation within the geography was less straightforward. Slovene was no longer an outlier. Polish was found to be near equidistant from all branches.

In order to compare w/ Manczak’s data we threw out Macedonian and E-Cz. It still supports clustering and doesn’t significantly change the big picture. Also puts Polish to ESl and closest to Ukrainian.

Alan: what is the difference between the Dyen slides you made that are currently in positions 6 and 9 in the slide order?

This moves Slovene closer to South Slavic (in contrast to its outlier status in the Dyen MDS). And W Slavic has moved from the center to a more westerly orientation. I.e., closer fit to geography. Polish is again intermediate b/w W & E, but now closer to Russian rather than Ukrainian.

Mańczak data showing differences in lexical matching.

POL tended to match RUS more often in this corpus than POL matched UKR and BEL (yellow highlights), though this was not always the case.
SLIDE: Birnbaum. Traditional schematic isogloss map for phonological isoglosses.

SLIDE: BIRNBAUM PHONOLOGY MDS PLOT
Converted into 0s (archaisms) and 1s (shared innovations), the MDS plot yielded a similar pseudomap to previous, though with three distinct branches. Again, Polish is an outlier with higher number of innovations distinct from others.

SLIDE: CORRELATION W GEOGRAPHY & 3 data sets
Shows best fit overall with geography with G & A data, least good with Dyen. Mańczak and Birnbaum were also close fits with geography.

Conclusions

References
Family tree and or map-like approaches to Slavic languages?

Alan J. Redd (Anthropology) &
Marc L. Greenberg (Slavic)
University of Kansas

“Slavic Languages: Time and Contingency”, UC Berkeley 12–13 Feb. 2010
Slavic language evolution: tree model or exchange model?
Slavic language map: West, South, and East.
Tree model: Figure 1 Atkinson and Gray (2003)
2,449 lexical items, 87 languages
Tree model: Bayesian analysis
418 lexical items, 12 languages

LAV

POL
RUS
DSB
HSB
CES
SLK
UKR
BEL
SVN
BUL
BCS

South
West
East
Tree model: Bayesian analysis
314 lexical items, 11 languages
Tree model: Bayesian analysis
314 lexical items, 11 languages; linearized tree

1155

1166

1166

1400 1200 1000 800 600 400 200 0

years before present

South

West

East

Polish
Czech
Slovak
Sloven
Serbian
Bulgarian
Latvian
Lithuanian
Belarusian
Ukrainian
Russian
Summary slide of Tree model:

Bayesian analysis; lexical items

G&A-2003 (87 languages)

This study (12 languages)
MDS plot: Figure 2 Dyen, Kruskal & Black (1992)
200 cognates; 13 languages; % of shared cognates for Swadesh list

Figure 2. Pseudomap of Slavic Languages.
MDS plot: after Figure 2 Dyen, Kruskal & Black (1992)
200 cognates; 13 languages; % of shared cognates for Swadesh list

[Diagram showing MDS plot with labels for languages like POL, E-CES, SLK, CES, DSB, HSB, BEL, UKR, RUS, MAK, BUL, SVN, BCS.]
Mańczak 2004 – distances expressed as raw N of correspondences between pairs

<table>
<thead>
<tr>
<th>Polski</th>
<th>Czeski</th>
<th>Słowacki</th>
<th>Górnolужыcki</th>
<th>Dolnolужыcki</th>
<th>Rosyjski</th>
<th>Białoruski</th>
<th>Ukrainski</th>
<th>Słoweński</th>
<th>Serbochorwacki</th>
<th>Bułgarski</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cz</td>
<td>1158</td>
<td>1345</td>
<td>1345</td>
<td>1155</td>
<td>1155</td>
<td>1103</td>
<td>1160</td>
<td>1160</td>
<td>1041</td>
<td>1041</td>
</tr>
<tr>
<td>Sla</td>
<td>1118</td>
<td>1118</td>
<td>1118</td>
<td>1011</td>
<td>970</td>
<td>968</td>
<td>900</td>
<td>1007</td>
<td>956</td>
<td>887</td>
</tr>
<tr>
<td>Ste</td>
<td>956</td>
<td>1002</td>
<td>1011</td>
<td>1002</td>
<td>950</td>
<td>902</td>
<td>902</td>
<td>984</td>
<td>894</td>
<td>860</td>
</tr>
<tr>
<td>Br</td>
<td>951</td>
<td>950</td>
<td>950</td>
<td>950</td>
<td>835</td>
<td>900</td>
<td>856</td>
<td>887</td>
<td>860</td>
<td>813</td>
</tr>
<tr>
<td>Ros</td>
<td>902</td>
<td>835</td>
<td>799</td>
<td>815</td>
<td>763</td>
<td>894</td>
<td>719</td>
<td>835</td>
<td>835</td>
<td>836</td>
</tr>
<tr>
<td>Ukra</td>
<td>900</td>
<td>738</td>
<td>731</td>
<td>629</td>
<td>599</td>
<td>777</td>
<td>698</td>
<td>815</td>
<td>724</td>
<td>687</td>
</tr>
<tr>
<td>Sch</td>
<td>887</td>
<td>717</td>
<td>724</td>
<td>612</td>
<td>597</td>
<td>707</td>
<td>694</td>
<td>815</td>
<td>724</td>
<td>687</td>
</tr>
<tr>
<td>Gł</td>
<td>835</td>
<td>698</td>
<td>721</td>
<td>612</td>
<td>596</td>
<td>721</td>
<td>693</td>
<td>777</td>
<td>719</td>
<td>589</td>
</tr>
<tr>
<td>Dl</td>
<td>786</td>
<td>693</td>
<td>707</td>
<td>601</td>
<td>592</td>
<td>601</td>
<td>612</td>
<td>717</td>
<td>719</td>
<td>530</td>
</tr>
<tr>
<td>Bg</td>
<td>735</td>
<td>572</td>
<td>589</td>
<td>530</td>
<td>506</td>
<td>599</td>
<td>592</td>
<td>597</td>
<td>597</td>
<td>306</td>
</tr>
</tbody>
</table>

average: 50.81 48.06 47.88 42.96 41.38 45.86 45.44 42.76 46.44 44.75 39.57
MDS-ML plot: 11 languages lexical items; this study data from: Mańczak (2004), 1816 tokens from Gospel texts; % shared
MDS-ML plot: 11 languages; this study
data from: Dyen, Kruskal & Black (1992), 200 cognates
MDS-ML plot: 11 Slavic languages; this study
Data from: Atkinson & Gray (2003); 315 cognates, Jaccard distance
Slide of lexical patterns with POL towards RUS (Mańczak data);
POL = RUS ≠ UKR
Birnbaum 1966: Phono- and morphological isoglosses

- A = East Slavic
- B = Lekhitic
- C = Sorbian
- D = Czecho-Slovak
- E = Slovene/BCS
- D = Macedo-Bulg.
MDS plot 11: Slavic phonological innovations; this study data from: Birnbaum (1966); 40 isoglosses; Jaccard distance
Summary of MDS plots; this study

Birnbaum-1966

G&A-2003

Mańczak-2004
Correlations with geography and MDS plots

<table>
<thead>
<tr>
<th>Data set</th>
<th>Geography correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyen-1992</td>
<td>0.381</td>
<td>ns</td>
</tr>
<tr>
<td>G&amp;A-2003</td>
<td>0.587</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Manczak-2004</td>
<td>0.531</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Birnbaum-1966</td>
<td>0.516</td>
<td>p &lt; 0.05</td>
</tr>
</tbody>
</table>

\(^1\)Mantel Test
**Correlations among MDS plots—data sets**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G&amp;A-2003</td>
<td>0.758</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manczak-2004</td>
<td>0.319</td>
<td>0.728</td>
<td></td>
</tr>
<tr>
<td>Birnbaum-1966</td>
<td>0.501</td>
<td>0.698</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Mantel test; all comparisons $p < 0.05$