The phylogeny of word meanings
Inferring the directionality of semantic change from word lists

Gerhard Jäger

joint work with Alla Münch and Johannes Dellert

Seminar für Sprachwissenschaft, Tübingen

Wissenschaftskolleg Berlin

January 26, 2016
“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. [...] We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. [...] The frequent presence of rudiments, both in languages and in species, is still more remarkable. [...] Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues.”

(Darwin, The Descent of Man)
Evolution and language change

Vater Unser im Himmel, geheiligt werde Dein Name

Onze Vader in de Hemel, laat Uw Naam geheiligd worden

Our Father in heaven, hallowed be your name

Fader Vor, du som er i himlene!
Helliget vorde dit navn
Evolution and language change

Diagram of phylogenetic relationships among Germanic languages.
Evolution and language change

**Middle High German:**
Got vater unser, då du bist in dem himelrîche gewaltic alles des dir ist, geheiliget sô werde dîn nam

**Old High German:**
Fater unser thû thâr bist in himile, si giheilagôt thîn namo

**Gothic:**
Atta unsar þu in himinam, weihnai namo þein
Genetic language relationships

- In most cases, we do not have written records of earlier stages
- Regular sound correspondences provide evidence for genetic relationship though
  - Correspondences indicate common ancestor + different sound shifts
  - The more cognates two languages share and the fewer sound shifts separate them, the closer they are related
### Erste bzw. Germanische Lautverschiebung (Indoeuropäisch → Germanisch)

**Phase**

1. */b/ → */p/  
2. */d/ → */t/  
3. */g/ → */k/  
4. */k'/ → */b'/  
5. */p'/ → */p'/  

### Zweite bzw. Hochdeutsche Lautverschiebung (Germanisch → Althochdeutsch)

**Beispiele (Neuhochdeutsch)**

1. */p/ → */f/ → */f/
   - niederdeutsch: *slapen* → *schenken*; niederdeutsch und englisch: *ship* → *Schiff*
   - niederdeutsch: *scherp* → *scharf*
   - Beispiel: *Schiff* → *scharf*  

2. */p/ → */pf/
   - niederdeutsch: *Peppe* → *Pfeffer*
   - niederdeutsch: *Plauch* → *Plugh* → *Pflug*
   - niederdeutsch: *scherp* → *sharp*  
   - Beispiel: *scherp* → *sharp*  

3. */t/ → */ss/ → */s/
   - niederdeutsch: *dat* → *that*
   - Beispiel: *dat* → *that*  

4. */k/ → */xk/ → */x/
   - niederdeutsch: *ik* → *ich*
   - niederdeutsch und englisch: *makne* → *mauchen*
   - niederdeutsch: *auk* → *auch*
   - Beispiel: *ik* → *ich*  

5. */b'/ → */b/  
   - Beispiel: *Berg* → *bairisch: Berg*

6. */d'/ → */d/  
   - Beispiel: *Dag* → *Tag*  

7. */g'/ → */k/  
   - Beispiel: *Gott* → *bairisch: Gott*

8. */t'/ → */p/  
   - Beispiel: *thorn* → *Dorn*  

**Jahrhundert**

1. 4/5  
2. 6/7  
3. 4/5  
4. 4/5  
5. 5/6  
6. 5/6  
7. 7/8  
8. 8/9  
9. 9/10  

**Dialektgebiete**

1. oberdeutsch und mitteldeutsch  
2. ober- und mitteldeutsch  
3. südbairisch, hoch- und höchstalemannisch  
4. teilweise bairisch und alemannisch  
5. teilweise bairisch und alemannisch  
6. gesamtes deutsches Dialektkontinuum
How historical linguists do it

Example: Polynesian languages

- Taken from Crowley & Bowern (2010)
How historical linguists do it

**Table 5.1: Data from Four Polynesian Languages**

<table>
<thead>
<tr>
<th></th>
<th>Tongan</th>
<th>Samoan</th>
<th>Rarotongan</th>
<th>Hawaiian</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>tapu</td>
<td>tapu</td>
<td>tapu</td>
<td>kapu</td>
</tr>
<tr>
<td>2.</td>
<td>pito</td>
<td>pute</td>
<td>pito</td>
<td>piko</td>
</tr>
<tr>
<td>3.</td>
<td>puhi</td>
<td>feula</td>
<td>pu?i</td>
<td>puhi</td>
</tr>
<tr>
<td>4.</td>
<td>tafa?aki</td>
<td>tafa</td>
<td>ta?a</td>
<td>kaha</td>
</tr>
<tr>
<td>5.</td>
<td>ta?a</td>
<td>tae</td>
<td>tae</td>
<td>kae</td>
</tr>
<tr>
<td>6.</td>
<td>tagata</td>
<td>tagata</td>
<td>tagata</td>
<td>kanaka</td>
</tr>
<tr>
<td>7.</td>
<td>tahi</td>
<td>tai</td>
<td>hai</td>
<td>'sea'</td>
</tr>
<tr>
<td>8.</td>
<td>malahi</td>
<td>malosi</td>
<td>ka?a</td>
<td>?aha</td>
</tr>
<tr>
<td>9.</td>
<td>kalo</td>
<td>?alo</td>
<td>karo</td>
<td>?alo</td>
</tr>
<tr>
<td>10.</td>
<td>aka</td>
<td>a?a</td>
<td>aka</td>
<td>a?a</td>
</tr>
<tr>
<td>11.</td>
<td>?ahu</td>
<td>au</td>
<td>au</td>
<td>au</td>
</tr>
<tr>
<td>12.</td>
<td>?ulu</td>
<td>ulu</td>
<td>unu</td>
<td>po?o</td>
</tr>
<tr>
<td>13.</td>
<td>?ufi</td>
<td>ufi</td>
<td>u?i</td>
<td>uhi</td>
</tr>
<tr>
<td>14.</td>
<td>afi</td>
<td>afi</td>
<td>a?i</td>
<td>ahi</td>
</tr>
<tr>
<td>15.</td>
<td>faa</td>
<td>faa</td>
<td>?aa</td>
<td>baa</td>
</tr>
<tr>
<td>16.</td>
<td>feke</td>
<td>fe?e</td>
<td>?eke</td>
<td>he?e</td>
</tr>
<tr>
<td>17.</td>
<td>ika</td>
<td>ika</td>
<td>i?a</td>
<td>'fish'</td>
</tr>
<tr>
<td>18.</td>
<td>ihu</td>
<td>isu</td>
<td>putangio</td>
<td>ihu</td>
</tr>
<tr>
<td>19.</td>
<td>hau</td>
<td>sau</td>
<td>?au</td>
<td>hau</td>
</tr>
<tr>
<td>20.</td>
<td>tafua?i</td>
<td>si?a</td>
<td>?ika</td>
<td>hi?a</td>
</tr>
<tr>
<td>22.</td>
<td>hake</td>
<td>a?e</td>
<td>a?e</td>
<td>'up'</td>
</tr>
<tr>
<td>23.</td>
<td>huu</td>
<td>ulu</td>
<td>uru</td>
<td>komo</td>
</tr>
<tr>
<td>24.</td>
<td>maka</td>
<td>mapa</td>
<td>mapa</td>
<td>mana</td>
</tr>
<tr>
<td>25.</td>
<td>ma?u</td>
<td>mau</td>
<td>mau</td>
<td>mau</td>
</tr>
<tr>
<td>26.</td>
<td>maa</td>
<td>maka</td>
<td>mara</td>
<td>mala</td>
</tr>
<tr>
<td>27.</td>
<td>na?a</td>
<td>fa?aga</td>
<td>maninia</td>
<td>naa</td>
</tr>
<tr>
<td>28.</td>
<td>nofo</td>
<td>nofo</td>
<td>no?o</td>
<td>noho</td>
</tr>
<tr>
<td>29.</td>
<td>galu</td>
<td>galu</td>
<td>?aru</td>
<td>nalu</td>
</tr>
<tr>
<td>30.</td>
<td>gutu</td>
<td>gutu</td>
<td>gutu</td>
<td>nuku</td>
</tr>
<tr>
<td>31.</td>
<td>vaka</td>
<td>va?a</td>
<td>vaka</td>
<td>va?a</td>
</tr>
<tr>
<td>32.</td>
<td>va?e</td>
<td>va?e</td>
<td>va?e</td>
<td>va?e</td>
</tr>
<tr>
<td>33.</td>
<td>laho</td>
<td>laho</td>
<td>r?o</td>
<td>r?o</td>
</tr>
<tr>
<td>34.</td>
<td>lobo</td>
<td>lobo</td>
<td>l?o</td>
<td>l?o</td>
</tr>
</tbody>
</table>

*Phylogeny of word meanings*
Guidelines for reconstruction

- Only establish sound correspondences if you are reasonably sure the words are cognate
- Assume sound shifts that are plausible (are known to occur frequently)
- Assume as few sound changes as possible for reconstructing a proto-language
- The reconstructed proto-language should have a typologically plausible sound system
Polynesian example

- Vowels in Proto-Polynesian are unchanged in daughter languages (otherwise we would stipulate unnecessary sound shift)
- Likewise, $p$, $m$ and $n$ are unchanged
- Majority rule:
  - pp. $^{\star}t$, $^{\star}N$, $^{\star}v \rightarrow hw. \ k$, $n$, $w$
- Lenition is more likely than fortition
- Also, Proto-Polynesian has $p$ and $t$, so it should also have a $k$, hence:
  - pp. $^{\star}k \rightarrow sm.$, hw. $7$ (rather than $^{\star}7 \rightarrow tg./rg.$ $k$)
Polynesian example

- majority rule:
  - pp. *f → rg. 7, hw. h
- not enough data to reconstruct the / and r
- majority rule:
  - pp. *h, *7 → sm., rg., hw. 0
- change s → h is known to be more common than h → s, hence (against majority rule):
  - pp. *s → tg./hw. h, rg. 7
Polynesian example

- constructing a tree

Proto-Polynesian

Tongan
s->h
k->7
h->0
7->0

Samoan

Rarotongan

Hawaiian
t->k
N->n
v->w
k->7
f->h
h->0
7->0
s->7
Polynesian example

- constructing a tree

Proto-Polynesian

Tongan
  - s->h

Rarotongan
  - f->7
  - h->0
  - 7->0
  - s->7

Samoan

Hawaiian
  - t->k
  - N->n
  - v->w
  - f->h
  - s->h

Polynesian example
● constructing a tree
Proto-Polynesian
Tongan
Samoan
Rarotongan
Hawaiian
t->k
N->n
v->w
f->h
s->7
How historical linguists do it

Polynesian example

Proto-Polynesian

Tongan
- s->h

Rarotongan
- f->7
- s->7

Samoan
- 7->0
- h->0

Hawaiian
- t->k
- N->n
- v->w
- f->h
- s->h

Gerhard Jäger (Tübingen) Phylogeny of word meanings 1/26/2016 15 / 65
Polynesian example

- reconstruction seems reasonable because
  - only one shift is assumed twice (s->7), and this type is known to occur frequently
  - reconstruction assumes (pull-) chain shifts
    - Rarotongan and Proto-Samoan/Hawaian restore the lost 7
    - Hawaiian additionally restores the lost k and h
- this procedure started from a reconstructed proto-language; usually tree construction and reconstruction of ancestral forms go hand in hand
How computational biologists do it
Transition probabilities in a two-state model

In a symmetric two-state model with rate of change \( r \) per unit time, where

\[
0 \xrightarrow{r} 1
\]

we have

\[
\text{Prob} (1 \mid 0, t, r) = \frac{1}{2} (1 - e^{-2rt})
\]

When \( rt \) is small then to very good approximation the probability of the events in the branch is \( rt \) or \( 1 - rt \). The latter is nearly 1.
### Rooted, bifurcating, labelled trees

<table>
<thead>
<tr>
<th>species</th>
<th>number of trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
</tr>
<tr>
<td>11</td>
<td>654,729,075</td>
</tr>
<tr>
<td>12</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>13</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>14</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>15</td>
<td>213,458,046,676,875</td>
</tr>
<tr>
<td>16</td>
<td>6,190,283,353,629,375</td>
</tr>
<tr>
<td>17</td>
<td>191,898,783,962,510,625</td>
</tr>
<tr>
<td>18</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>19</td>
<td>221,643,095,476,699,771,875</td>
</tr>
<tr>
<td>20</td>
<td>8,200,794,532,637,891,559,375</td>
</tr>
<tr>
<td>30</td>
<td>$4.9518 \times 10^{38}$</td>
</tr>
<tr>
<td>40</td>
<td>$1.00985 \times 10^{57}$</td>
</tr>
<tr>
<td>50</td>
<td>$2.75292 \times 10^{76}$</td>
</tr>
</tbody>
</table>
How computational biologists do it

- computer doggedly
  - looks at one damned tree after another
  - computes the likelihood of the data given that tree, and
  - moves on to the next tree

- exhaustive search is impossible, given the astronomical number of different trees

- no guarantee to find the single best tree, but there are good heuristics

- highly informative results:
  - estimates of branch lengths
  - confidence values for branches of a tree
  - probabilistic reconstructions of ancestral states
  - ...

Different cognate classes for a given meaning are treated just like different alleles of a gene.

Once the cognacy data are in this format, the full gamut of computational phylogenetics can be unleashed.
Spectacular results

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

1 Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New

Family Tree of Languages Has Roots in Anatolia, Biologists Say

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel, Quentin D. Atkinson, Andrea S. Calude, and Andrew Maudsley

Author Affiliations

PNAS
Skeptically received by traditional historical linguists

Asya Pereltsvaig (historical linguist), blog post  
The Malformed Language Tree of Bouckaert and His Colleagues

The study also groups Frisian with Flemish and Dutch, rather than English, which many Germanic scholars find objectionable.

It should also be pointed out that the authors of the Science article make a critical mistake in that they do not distinguish shared innovation from shared retention.

Another unsupportable configuration of the Science model concerns the internal grouping of the Slavic languages.

Yet again, we see that accepting the model proposed by Bouckaert and his colleagues requires one to believe not just three but actually dozens of impossible things before breakfast (with apologies to Lewis Carroll).

Martin Lewis (geographer), blog post  
Mismodeling Indo-European Origin and Expansion: Bouckaert, Atkinson, Wade and the Assault on Historical Linguistics

In the Science piece, the painstaking work of generations of historical linguists who have rigorously examined Indo-European origins and expansion is shrugged off as if it were of no account, even though the study itself rests entirely on the taken-for-granted work of linguists in establishing relations among languages based on words of common descent (cognates).

While purporting to offer a truly scientific approach, Bouckaert et al. actually forward an example of scientism, or the inappropriate and overweening application of specific scientific techniques to problems that lie beyond their own purview.
A closer look at cognate classes

- based on expert judgments
- usually cover ca. 200 concepts (Swadesh list)
- publicly available for very few language families (mainly Indo-European and Austronesian)

http://ielex.mpi.nl/
Phylogenetic inference with cognate classes

- defines a classification of languages into discrete states for each Swadesh concept
- example: Indo-European, concept foot
cognate classes are mathematically treated like alleles of a gene, or different DNA letters

- evolution as continuous time Markov chain:
  - evolution along continuous time line (no discrete generations)
  - a language spontaneously, at each point in time
    - split into two daughter languages, or
    - change it’s state for any concept
  - state transition probability density described by rate matrix

- for a given tree structure (with branch lengths) and rate matrix, each distribution of observed states at the leaves has a well-defined likelihood
Phylogenetic inference with cognate classes

- **Inference methods:**
  - **Maximum Likelihood:**
    - find the tree structure/rate matrix combination maximizing the likelihood of observed distribution of cognacy classes over languages
  - **Bayesian inference:**
    - assume prior distributions over tree structures and rate matrices
    - may incorporate historical, archaeological, linguistic etc. information
    - generate posterior distribution over trees
Practical issues

- for $n$ cognate classes, $n^2 - n$ transition rates have to be fitted $\Rightarrow$ often insufficient data
- therefore *binary recoding* of cognate classification:
  - each cognate class is a separate feature
  - values are 1 (present in a language) and 0 (absent)
- only two rates need to be fitted per concept
## Practical issues

<table>
<thead>
<tr>
<th>language</th>
<th>cognate class</th>
<th>orthographic form</th>
<th>A</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>K</th>
<th>Q</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>A</td>
<td>cane</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spanish</td>
<td>-</td>
<td>perro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>French</td>
<td>A</td>
<td>chien</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Portuguese</td>
<td>A</td>
<td>cão</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Romanian</td>
<td>A</td>
<td>căine</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Greek</td>
<td>A</td>
<td>κων, skilos</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>English</td>
<td>-</td>
<td>dog</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>German</td>
<td>A</td>
<td>Hund</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Danish</td>
<td>A</td>
<td>hund</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Icelandic</td>
<td>A</td>
<td>hundur, rakki</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Norwegian</td>
<td>A</td>
<td>hund, rakke, bikkje</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>E, K</td>
<td>PES, kuce</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Serbocroatian</td>
<td>E</td>
<td>pas</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Polish</td>
<td>E</td>
<td>pies</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Russian</td>
<td>A</td>
<td>собака</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Irish</td>
<td>A, I</td>
<td>cú, madra</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Marathi</td>
<td>H</td>
<td>kutrā</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hindi</td>
<td>H</td>
<td>kutta</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Latin</td>
<td>A</td>
<td>cansis</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Classical Greek</td>
<td>A</td>
<td>χών</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gothic</td>
<td>A</td>
<td>hunds</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Old English</td>
<td>A</td>
<td>hund</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Parallel evolution

- by definition, each cognate class can emerge only once
- transition rate $0 \rightarrow 1$ should thus be exceedingly small
- no parallel evolution $0 \rightarrow 1$ to be expected
- only expected exception: undetected borrowings
- However (see next slide...)
Parallel evolution
Parallel evolution

- apparently independent emergence of *leg:*Q in Greek, Icelandic, Romanian and Proto-Indo-Iranian
- traditional scholarship:
  - members of class *leq:*Q are cognate to members of *foot:*B
  - both cognate classes descend from Proto-Indo-European *ped-
  - parallel semantic change *foot* → *leg* in several branches of Indo-European
Testing for correlated evolution

Following Pagel and Meade (2006) (see also Dunn et al. 2011), using software BayesTraits (http://www.evolution.rdg.ac.uk/BayesTraits.html)

- model comparison

- log Bayes Factor $\approx 243$ in favor of second model $\Rightarrow$ strong evidence
- downsides:
  - strong evidence in favor of dependence, but weak evidence in favor of directionality
  - information about Swadesh entries of a single family does not afford scaling up
A polysemy network (cf. Dellert and Münch 2015, see also List et al. 2013) is a graph over concepts (expressed by glosses) where each link represents the fact that at least one language has a lexical item which can denote both concepts (or rather: can be translated by both glosses). If we count the number of instances in different language families, link strength measures cross-linguistic colexification between concepts.

Polysemy networks are a potential source of evidence for plausibility of semantic shifts:

- Intuition: every instance of semantic shift needs to pass an intermediate stage where the word in question is polysemous.
- A massively cross-linguistic polysemy network provides us with a snapshot of semantic evolution in action.

⇒ Polysemy networks might be a good model for modeling semantic change in computational historical linguistics!
Polysemy networks

**Question:** How well does a polysemy network model attested semantic shifts?

**Data:**
- a massively cross-linguistic polysemy network (TUE) developed in Tübingen within EVOLAEMP (Dellert, 2014)
- the Catalogue of Semantic Shifts (ZAL) maintained by the Russian academy of sciences (Zalizniak, 2008; Zalizniak et al., 2012)

**Method:** for each attested change, determine the length of the shortest paths between the relevant concepts in the network
Tübingen Polysemy Network (TUE)

- is derived from a large German-based dictionary database containing more than 750,000 entries in 114 languages
- has near-complete coverage of a 1,000-item list of basic concepts in 88 languages
- is built on more than 10,000 translations each for 31 languages from 10 primary language families of Eurasia
- is available to other researchers upon request
Tübingen Polysemy Network (TUE): Some data

- 32,653 concepts and 47,647 links
- a central connected component of 13,073 nodes (!)
- 1,640 smaller components and 14,391 unconnected islands
- an average of 2.4 neighbors for each node
- a typical shortest path length of 8
Small world characteristics

Graphs showing the relationship between size of graphs and the length of shortest paths, and the frequency of degrees.

Gerhard Jäger (Tübingen)
The Catalogue of Semantic Shifts

crosslinguistically recurring shifts collected manually by experts
no restrictions or preferences in the range of meanings or languages (de facto: focus on Eurasian)
contained 3,650 semantic shifts at the time of publication
classified into six types, each supported by up to 40 realizations
publicly available via a web interface: http://semshifts.iling-ran.ru/
Procedure

- English version of the catalogue was extracted from the website
- considerable semi-automated cleanup work
- the total number of shift pairs in the result: 6,174

Preperatory steps

- metalanguages of the catalogue: English and Russian
- primary language of the Tübingen Polysemy Network: German
- electronic English-German dictionary needed to be used as an intermediary for comparing the links in both resources

Experiment

- evaluate pairs of English glosses in the catalogue against the network by determining the shortest paths in the network between any pair of German translation equivalents
- do this separately for each type, and compare recall
Types of Semantic Shifts in the Catalogue

- Polysemy
- Diachronic Semantic Evolution
- Syncretism
- Cognates
- Borrowings
- Morphological Derivation

Only Diachronic Semantic Evolution, Cognates, and Borrowings contain instances of semantic change in the stricter sense.
Polysemy

- **synchronic polysemy**: A and B are meanings of a polysemous word
- ZAL: tree — forest / TUE: Baum ‘tree’ — Holz ‘wood’ — Wald ‘forest’
- Example: Tuvan *yjash* ‘tree; forest’
Diachronic semantic evolution

- Diachronic semantic evolution of a word in one language or from an ancestor language to a descendant language
- ZAL: board, plank — table, desk / TUE: Brett ‘board’ — Tisch ‘table’
- Example: Latin *tabula* ‘plate, tablet’ → Italian *tavola* ‘table’
Diachronic semantic evolution

- Diachronic semantic evolution of a word in one language or from an ancestor language to a descendant language
- ZAL: catch — hunt / TUE: jagen ‘hunt’ — fangen ‘catch’
- Example: Latin capto ‘various meanings related to catching’ → Italian cacciare ‘to hunt’
Diachronic semantic evolution

- Diachronic semantic evolution of a word in one language or from an ancestor language to a descendant language
- Example: Hungarian orr ‘nose’ ← ‘top, mountain’
Cognates

- meanings A and B belong to words of two sister languages which have developed from the same root in their common ancestor
- ZAL: hear — understand / TUE: hören ‘hear’ - verstehen ‘understand’
- Example: French entendre ‘to hear’ — Spanish entender ‘to understand, to be knowledgeable’
Borrowings

- B is the meaning of a borrowed word, while A is the meaning of its source in the donor language.
- ZAL: bell — watch/clock / TUE: Glocke ‘bell’ - Uhr ‘watch/clock’
- Example: Medieval Latin *clocca* ‘bell’ → English *clock* ‘clock, watch’
Table: Shifts in the catalogue covered by the polysemy network.

<table>
<thead>
<tr>
<th>Semantic shift type</th>
<th>Number of instances</th>
<th>No path</th>
<th>Path length 1 or 2</th>
<th>Path length 3 and more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polysemy</td>
<td>2315</td>
<td>20.6 %</td>
<td>35.0 %</td>
<td>44.4 %</td>
</tr>
<tr>
<td>Semantic Evolution</td>
<td>107</td>
<td>26.2 %</td>
<td>33.6 %</td>
<td>40.2 %</td>
</tr>
<tr>
<td>Morphological Derivation</td>
<td>597</td>
<td>28.5 %</td>
<td>29.0 %</td>
<td>42.5 %</td>
</tr>
<tr>
<td>Syncretism</td>
<td>43</td>
<td>25.6 %</td>
<td>55.8 %</td>
<td>18.6 %</td>
</tr>
<tr>
<td>Borrowing</td>
<td>58</td>
<td>31.0 %</td>
<td>41.4 %</td>
<td>27.6 %</td>
</tr>
<tr>
<td>Cognates</td>
<td>393</td>
<td>23.7 %</td>
<td>37.9 %</td>
<td>38.4 %</td>
</tr>
</tbody>
</table>
Towards detecting directional biases in meaning change

Assumptions
- meaning change proceeds along the edges of the polysemy graph
- for each possible transition, there is a fixed transition rate, across language families and times → ultimately founded in cognition

Research questions
- Can we estimate those rates from data?
- Can we detect directional biases?

Requirement
- Many data — across language families and beyond the Swadesh list
- Massive data collection effort of the Tübingen EVOLAEMP project
- (currently) translations of 1,017 concepts into 103 (mostly) Northern Eurasian languages (cf. Dellert, 2015)
- everything transcribed in IPA
- (so far) no manual cognate coding
Automatic cognate classification

- clustering of the words for a given concept
- heuristic technique based on phonetic similarity
- so far only rough approximation to expert classification

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>Cognate class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dra.Southern Dravidian.Kannada</td>
<td>paada</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Baltic.Lithuanian</td>
<td>peeda</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Greek.Greek</td>
<td>po8i</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Indic.Hindi</td>
<td>pEEr</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Indic.Hindi</td>
<td>paaw</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Iranian.Ossetic</td>
<td>fad</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Iranian.Pashto</td>
<td>psa</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Iranian.Persian</td>
<td>paa</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Romance.French</td>
<td>pye</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Romance.Italian</td>
<td>piede</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Romance.Portuguese</td>
<td>pE</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>IE.Romance.Spanish</td>
<td>pye</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>Kor.Korean.Korean</td>
<td>pal</td>
<td>Fuß:12</td>
</tr>
<tr>
<td>Alt.Tungusic.Manchu</td>
<td>pothxo</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>Brs.Burushaski.Burushaski</td>
<td>ut</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>Brs.Burushaski.Burushaski</td>
<td>utis</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Armenian.Armenian</td>
<td>votkh</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.Danish</td>
<td>foo78</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.Dutch</td>
<td>vut</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.English</td>
<td>fut</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.German</td>
<td>fuus</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.Icelandic</td>
<td>foutir</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.Norwegian</td>
<td>fuut</td>
<td>Fuß:6</td>
</tr>
<tr>
<td>IE.Germanic.Swedish</td>
<td>fuut</td>
<td>Fuß:6</td>
</tr>
</tbody>
</table>
Getting a handle on meaning change

- two cognate classes \( A \) and \( B \), for concepts \( c_A \) and \( c_B \), are merged if
  - \( c_A \) and \( c_B \) are neighbors in the polysemy graph
  - in at least one language, \( A \) and \( B \) are instantiated by the same phonetic string

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>Cognate class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dra.Southern Dravidian.Kannada</td>
<td>paada</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Baltic.Lithuanian</td>
<td>peeda</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Greek.Greek</td>
<td>po8i</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Indic.Hindi</td>
<td>pEEr</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Indic.Hindi</td>
<td>paaw</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Ossetic</td>
<td>fad</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Pashto</td>
<td>psa</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Persian</td>
<td>paa</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.French</td>
<td>pye</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Italian</td>
<td>piede</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Portuguese</td>
<td>pE</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Spanish</td>
<td>pye</td>
<td>foot</td>
</tr>
<tr>
<td>Kor.Korean.Korean</td>
<td>pal</td>
<td>foot</td>
</tr>
<tr>
<td>Ura.Permic.Udmurt</td>
<td>p3d</td>
<td>foot</td>
</tr>
<tr>
<td>Alt.Tungusic.Nanai</td>
<td>b3gdi</td>
<td>Bein:9</td>
</tr>
<tr>
<td>IE.Greek.Greek</td>
<td>po8i</td>
<td>Bein:9</td>
</tr>
<tr>
<td>IE.Indic.Bengali</td>
<td>pa</td>
<td>Bein:9</td>
</tr>
<tr>
<td>IE.Iranian.Pashto</td>
<td>psa</td>
<td>Bein:9</td>
</tr>
<tr>
<td>IE.Iranian.Persian</td>
<td>paa</td>
<td>Bein:9</td>
</tr>
</tbody>
</table>

\(^0\) see also (Dellert, 2016) for a different approach
Getting a handle on meaning change

- treat each merged cognate class as feature and meaning as value
- fit a model with two parameters:
  - transition rate \( \text{foot} \rightarrow \text{leg} \)
  - transition rate \( \text{leg} \rightarrow \text{foot} \)

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt. Tungusic.Nanai</td>
<td>b3gdi</td>
<td>leg</td>
</tr>
<tr>
<td>Dra.Southern Dravidian.Kannada</td>
<td>paada</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Baltic.Lithuanian</td>
<td>peeda</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Greek.Greek</td>
<td>po8i</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Indic.Bengali</td>
<td>pa</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Indic.Hindi</td>
<td>pEEr/paaw</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Ossetic</td>
<td>fad</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Pashto</td>
<td>psa</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Iranian.Persian</td>
<td>paa</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.French</td>
<td>pye</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Italian</td>
<td>piede</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Portuguese</td>
<td>pE</td>
<td>foot</td>
</tr>
<tr>
<td>IE.Romance.Spanish</td>
<td>pye</td>
<td>foot</td>
</tr>
<tr>
<td>Kor.Korean.Korean</td>
<td>pal</td>
<td>foot</td>
</tr>
<tr>
<td>Krt.Kartvelian.Georgian</td>
<td>phExi</td>
<td>leg</td>
</tr>
<tr>
<td>Ura.Permic.Udmurt</td>
<td>p3d</td>
<td>foot</td>
</tr>
</tbody>
</table>
Results

- no evidence of directional bias for
  - foot ↔ leg
  - moon ↔ months

\[ \text{foot/leg bias} \]
\[ \text{mean} = -0.021 \]
\[ 95\% \text{ HDI} = [-0.674, 0.808] \]

\[ \text{moon/month bias} \]
\[ \text{mean} = 0.0238 \]
\[ 95\% \text{ HDI} = [-0.964, 0.871] \]

Model fitted using BayesTraits (Pagel and Meade, 2014); tree sample generated with BayesPhylogenies (Pagel and Meade, 2015)
Results

- clear evidence of directional bias for
  - \textit{sun} \leftrightarrow \textit{day}
  - \textit{tongue} \leftrightarrow \textit{language}
Conclusion

- method suffers from poor quality of cognate classification
- results therefore to be taken with a pound of salt
- still, proof of concept that crosslinguistic lexical data provide a handle on studying laws of meaning evolution

Great potential for combining cognitive with phylogenetic methods!


Pagel, M. and A. Meade (2014). BayesTraits 2.0. software distributed by the authors.
Pagel, M. and A. Meade (2015). BayesPhylogenies 2.0. software distributed by the authors.

