EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

Correlating borrowing events across concepts to derive a data-driven source of evidence for loanword etymologies

Verena Blaschke and Johannes Dellert

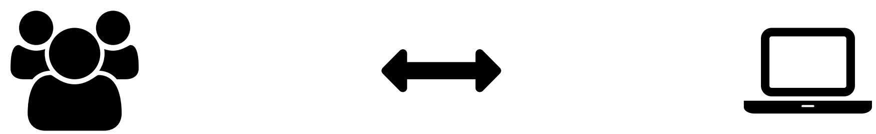
February 26, 2021

MaEiQCL workshop at the DGfS annual meeting

European Research Council
Established by the European Commission

Table of Contents

Motivation



- Historical linguistics often relies on heuristics based on the shared experience of experts
- To integrate this into theoretical frameworks that can be used in conjunction with computational methods:
  How can we formalize and statistically validate such reasoning patterns in historical linguistics?

Motivation: Automated loanword detection

- Usually based on exceptions from sound laws
- What if we don't know the sound laws already / if there aren't any useful exceptions?

LAT mɛn**t**a DEU mɪn**t͡s**ə

Motivation: Automated loanword detection

- Usually based on exceptions from sound laws
- What if we don't know the sound laws already / if there aren't any useful exceptions?

LAT mɛn**t**a DEU mɪn**t͡s**ə

- If some language has already been established as a donor language for some words, it becomes more likely as a candidate donor for other words
- Assumption that words from the same semantic field get borrowed together
- Historical/cultural knowledge

Can we model these in a *data-driven* way?

Table of Contents

Data: WOLD

World loanword database
(Haspelmath and Tadmor, 2009)

- 41 (recipient) languages
  (26 families)
- 1,500 concepts
  (24 semantic fields)
- Loanword status:
  1. clearly borrowed
  ... 5. no evidence of borrowing

Borrowability

1. Religion/belief
2. Clothing/grooming
3. The house
4. The law

...

21. Kinship
22. The body
23. Spatial relations
24. Sense perception

(Tadmor, 2009, p. 64)

Data: CLICS^2

Cross-linguistic colexifications (List et al., 2018)

- 1,200 languages
- 2,500 concepts
- Network: two concepts share an edge if 3+ unrelated languages use the same lexical unit for both concepts

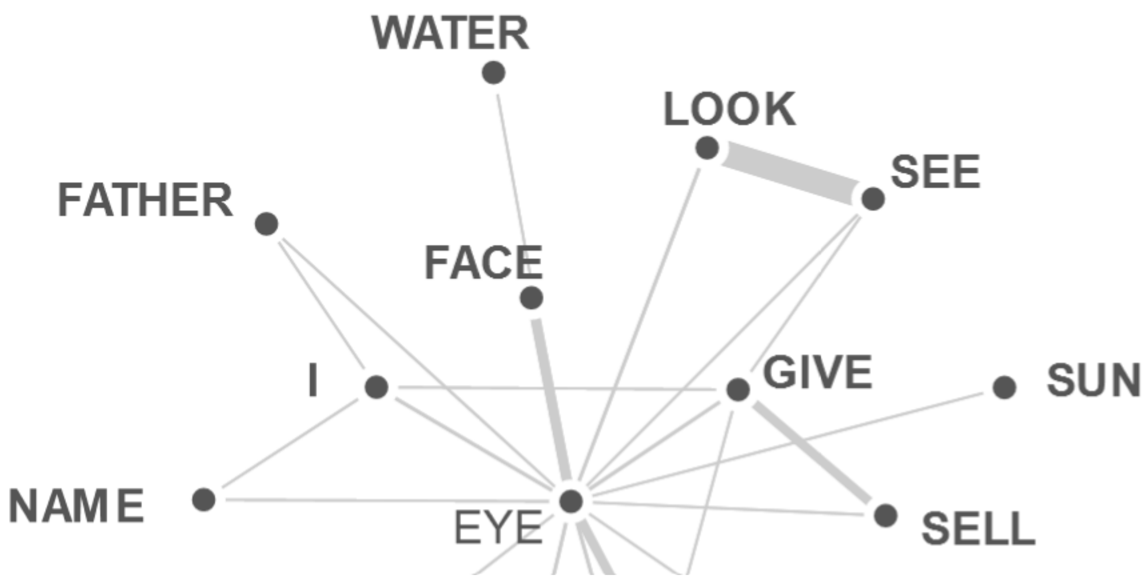


Table of Contents

Method

- 14,000 clearly/probably borrowed loanwords that do not have inherited synonyms and were borrowed at least 3x

TO BORROW	Hausa	<	Arabic
	Sakha	<	Hausa
	...		...
THE HAND	Thai	<	Sanskrit
	Malagasy	<	Malay
	...		...

Method: Implication strength

Given concepts $\textsc{x}$ and $\textsc{y}$, does $\textsc{x}$ being a loanword imply that $\textsc{y}$ was borrowed by and from the same languages?

$$\text{impl\_strength}(\textsc{x},\textsc{y}) = \frac{\text{\# of donor-target pairs that borrowed } \textsc{x} \text{ and } \textsc{y}}{\text{\# of donor-target pairs that borrowed } \textsc{x}}$$

Method: Implication strength

Given concepts X and Y, does X being a loanword imply that Y was borrowed by and from the same languages?

$$\text{impl\_strength}(\textsc{x},\textsc{y}) = \frac{\text{\# of donor-target pairs that borrowed X and Y}}{\text{\# of donor-target pairs that borrowed X}}$$

- THE EAST: borrowed 7x (**aqc < ava**, **kap < ava**, **rif < ara**, rmc < ces, sjd < gmq, **swa < ara**, **vie < cmn**)
- THE CAUSE: borrowed 15x (**aqc < ava**, eng < fra, gwd < amh, hau < ara, ind < ara, **kap < ava**, knc < ara, qvi < spa, **rif < ara**, ron < fra, sah < xgn, sjd < rus, **swa < ara**, tha < san, **vie < cmn**)

$$\text{implication\_strength}(\text{THE EAST, THE CAUSE}) = 5/7 \approx 71\%$$
$$\text{implication\_strength}(\text{THE CAUSE, THE EAST}) = 5/15 \approx 33\%$$

Method: NPMI

Co-occurrence

- What if the concepts were borrowed at very different rates?
- Normalized pointwise mutual information (+1 complete co-occurrence ... 0 independence ... -1 no co-occurrence)

$$\mathrm{NPMI}(x, y) = \frac{\ln \frac{p(x,y)}{p(x)p(y)}}{-\ln p(x, y)}$$

$$\begin{aligned} p(\textsc{the east, the cause}) &= 5/41 \\ p(\textsc{the east}) &= 7/41 \\ p(\textsc{the cause}) &= 15/41 \\ \mathrm{NPMI}(\textsc{the east, the cause}) &\approx 0.32 \end{aligned}$$

Method: Intra-pair similarity

How semantically similar are the concept pairs?

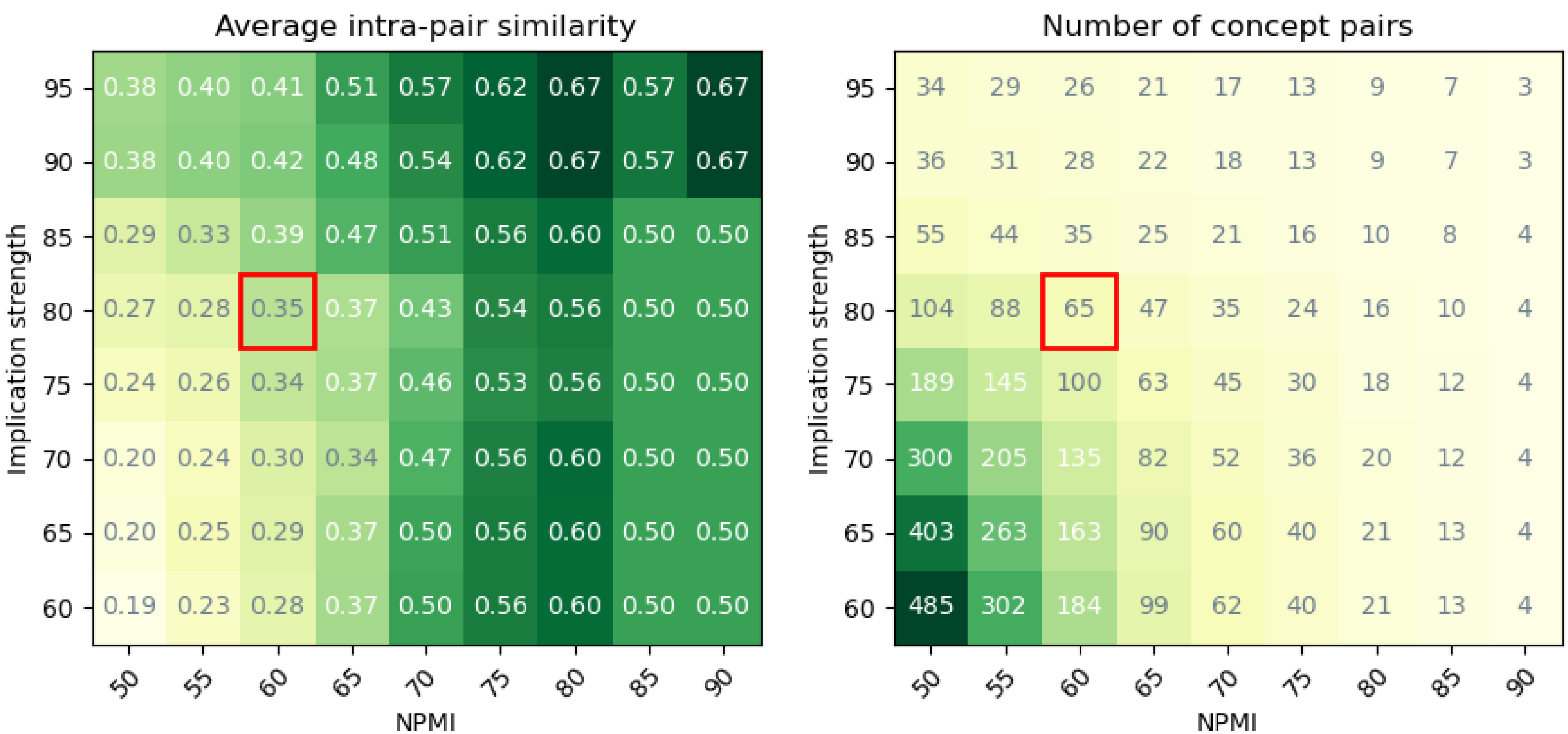
- Inverse correlation with CLICS node distance

Distance (in edges)	Similarity
1 (colexified)	1
2	0.5
3	0.33
4+	0

Method: Thresholds

Thresholds (determined a posteriori)

- High implication strength (≥ 0.8) that still shows meaningful connections (NPMI ≥ 0.6)



Method: Bootstrapping

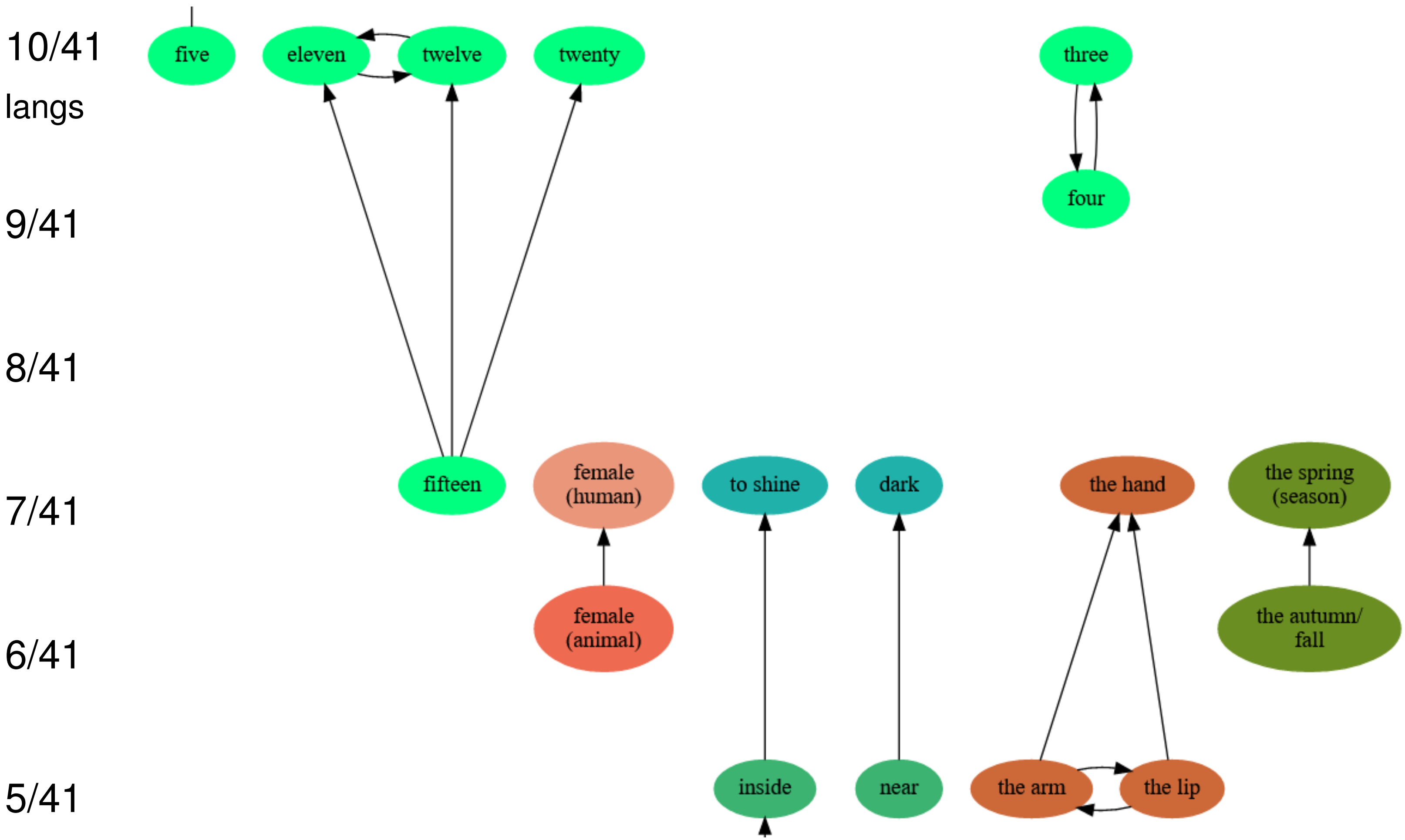
1,000 bootstrap samples of language sets (with replacement; one-sided 95 % confidence interval)

- ‘Noisy’ observations that describe only a small number of instances tend to vary more and tend to get filtered out
- Higher implication and correlation scores

	implication strength	
	before	after bootstrapping
THE PARENTS -> TO PEEL	0.5	0.42
THE ARM -> THE LIP	0.8	0.8

Table of Contents

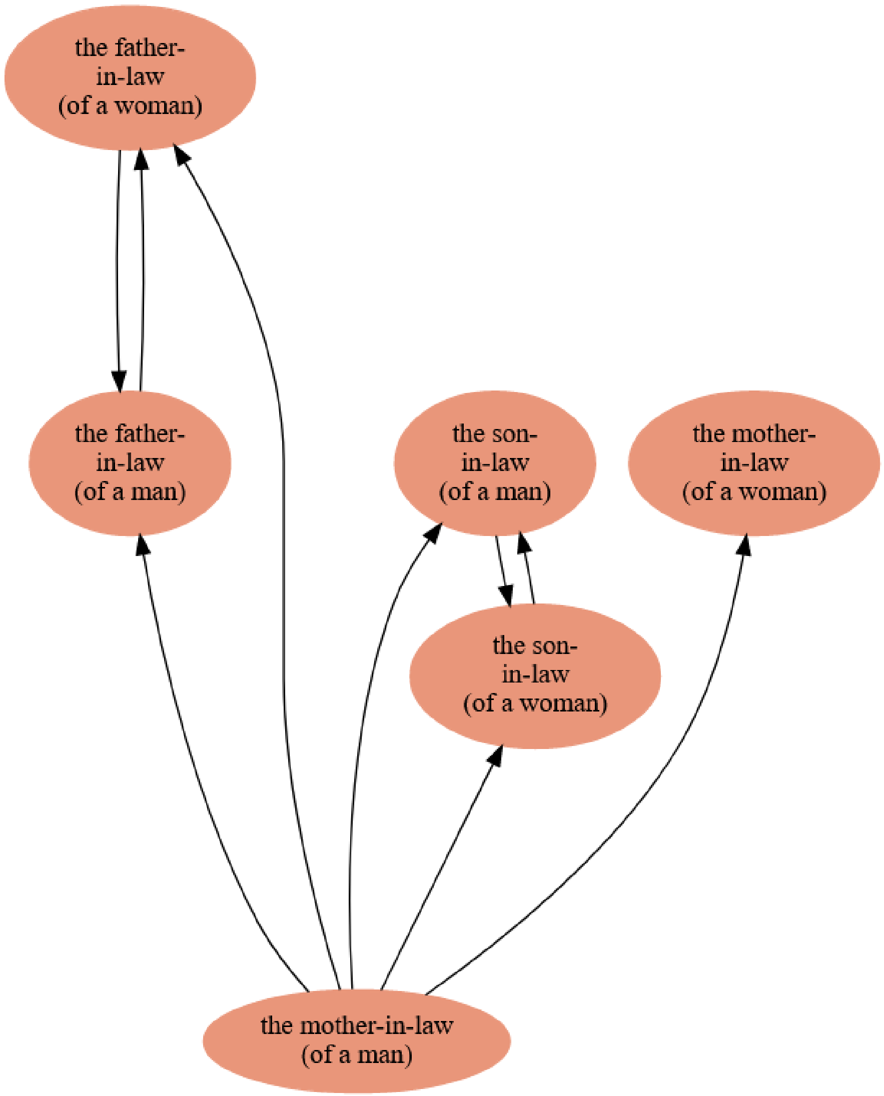
Results


10/41 langs
9/41
8/41
7/41
6/41
5/41
five
eleven
twelve
twenty
three
four
fifteen
female (human)
to shine
dark
the hand
the spring (season)
female (animal)
the autumn/ fall
inside
near
the arm
the lip

Results

65 concept pairs

- 45 of these are within-domain correlations
- Common semantic fields:
  - ▷ KINSHIP (!)
  - ▷ QUANTITY (!)
  - ▷ THE BODY
  - ▷ SPATIAL RELATIONS
- ...specifically the rarely-borrowed fields!
- Clusters ('package deals')



Results

Only 20 cross-domain pairs

- 2 largely due to colexification

female(2) (ANIMALS)	>	female(1) (KINSHIP)
the knife(1) (FOOD/DRINK)	>	the knife(2) (BASIC ACTIONS/TECHNOLOGY)

- Some are somewhat plausible, but most appear to be random

to kneel (MOTION)	>	the defeat (WARFARE/HUNTING)
the beeswax (ANIMALS)	>	the kidney (THE BODY)
...		...
to make (BASIC ACTIONS/TECHNOLOGY)	>	inside (SPATIAL RELATIONS)
sometimes (TIME)	>	to sneeze (THE BODY)

To be investigated...

- In some cases, the concept pairs are colexified concepts—how often is that?
- More languages, language families
- How much is there to the cross-domain relations?

Future plans

Incorporate correlation information into model for etymological inference to combine measures of:

- Adherence to soundlaws
- Language contact
- Borrowing frequency by concept, in general and given other borrowed concepts

Conclusion

- Even with a limited sample of languages, we can extract some meaningful borrowing patterns.
- Kinship terms and numerals are not borrowed very often, but when they are, there exist some ‘package deals.’

`https://github.com/verenablaschke/borrowing-correlations`

Acknowledgments

This work is being funded by the project CrossLingference, a grant from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 834050, awarded to Gerhard Jäger).

European Research Council
Established by the European Commission

References

Haspelmath, M. and Tadmor, U., editors (2009). *World loanword database*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at `https://wold.clld.org/`.

List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., and Forkel, R. (2018). CLICS2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A comparative handbook*, 55:75.

Appendix: Bootstrapping

- Sample the set of languages 1,000 times with replacement
- Calculate implication strength and NPMI for each sample
- The scores are the lower bounds of the corresponding (one-sided, right-open, 95%) confidence intervals:

$$\bar{a} - \frac{1.64485s}{\sqrt{1000}}$$

where

- $\bar{a}$ is the arithmetic mean of the given score in all 1,000 samples
- s is the standard deviation of the mean

Appendix: Etymological inference

$$sem\_corr(c, L_D, L_R) = \frac{\sum_{c' \in C} P\_borr(c', L_D, L_R) \cdot sem\_sim(c, c')}{|C|}$$

$$lang\_corr(c, L_D, L_R) = \frac{\sum_{c' \in C} P\_borr(c', L_D, L_R)}{|C|}$$

$$event\_corr(c, L_D, L_R) = \frac{\sum_{c' \in C} P\_borr(c', L_D, L_R) \cdot borr\_impl(c', c)}{|C|}$$

$$\begin{aligned} P\_borr(c, L_D, L_R) = f(&soundlaws(c, L_R), \\ &borr\_freq(c), \\ &sem\_corr(c, L_D, L_R), \\ &lang\_corr(c, L_D, L_R), \\ &event\_corr(c, L_D, L_R)) \end{aligned}$$