EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
LANGUAGE EVOLUTION:
THE EMPIRICAL TURN

Synchronic Lexical Associations and the Directionality of Semantic Change

XPrag Meets Historical Pragmatics

Cologne, November 14, 2016

Johannes Dellert

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

Simulation Experiment

Conclusions

Examples from Dellert (2016)

Introduction: General Idea

General ideas behind my talk:

- cross-linguistic polysemies provide a snapshot of semantic change in progress (polysemy is an intermediate stage)
- we can treat **concepts as information-theoretic variables**, and their realizations in each language as samples
- polysemies define information geometry for concept space
- vanishing conditional mutual information can be used to test for **conditional independence between concepts**
- principles of causal inference sometimes allow us to infer that one concept “causes” another
- directionality of causal signal can be interpreted as the **dominant direction of semantic change**

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

Simulation Experiment

Conclusions

Examples from Dellert (2016)

Causal Inference: Basic Idea

- algorithmic techniques to infer causal relationships between variables from observational data alone (Pearl, 2009)
- not possible for two variables: “correlation is not causation”
- but: interaction between more than two variables often provides hints about underlying causal scenario
- underlying theory (Reichenbach’s **Common Cause Principle**) states that whenever two variables are correlated, there must be either a directed causal path in exactly one direction, or a common cause (“no correlation without causation”)
- the Common Cause Principle is problematic in our application, implying we can only partially apply causal inference

Conditional Independence and Causal Graphs

- core building block: a **conditional independence** relation
- $(X \perp\!\!\!\perp Y \mid Z)$ intuitively means:
  "any dependence between the variables X and Y can be explained by the influence of Z"
- PC algorithm: sequence of conditional independence tests reduces a complete graph to a **causal skeleton**, where no link can be explained away by conditioning on other variables
- removal of link $X - Y$ relies on finding a **separating set**, i.e. a set of variables $\{Z_1, \ldots, Z_n\}$ such that $(X \perp\!\!\!\perp Y \mid Z_1, \ldots, Z_n)$

Unshielded Collider Criterion

- directionality inference on the causal skeleton
- for each pattern of the form $X - Z - Y$ (**unshielded triple**), ask whether the central variable was part of the separating set that was used for explaining away the link $X - Y$
- underlying idea: if Z was not necessary to explain away $X - Y$, this excludes all patterns except $X \rightarrow Z \leftarrow Y$ (a **v-structure**)
- reason: we would expect some information flow in all three scenarios $X \leftarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$, and $X \rightarrow Z \rightarrow Y$
- this relies on a causal **faithfulness** assumption: we can measure $(X \perp\!\!\!\perp Y \mid Z)$ iff this is implied by the true causal graph

Propagating Directionality Information

- faithfulness implies we can be sure to have detected exactly the true v-structures
- this implies an inference rule $X \rightarrow Z - Y \Rightarrow X \rightarrow Z \rightarrow Y$
- BUT: this is only true if we can assume **causal sufficiency** (all possible common causes are observed)
- the PC algorithm would use this to propagate directionality information through the graph, in many case assigning a direction to each node in the causal skeleton

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

- Conditional Independence between Concepts
- Deriving Directionality

Simulation Experiment

Conclusions

Examples from Dellert (2016)

Applying Causal Inference to Polysemy Data

- treat concepts as information-theoretic variables
- define concepts to have non-zero mutual information
  if they can be expressed by the same word in some language
- ⇒ it becomes possible to derive causal conclusions from massively cross-linguistic polysemy data!
- formally, such data can be expressed in terms of **isolectic sets** (sets of concepts covered by some lexeme in some language):
  {Foot,FootOfStairs,FootOfTable,Foot[Measure],...}
- *iso*(c): the set of all isolectic sets in the data which include the concept c; the above set is an element of *iso*(Foot)

Conditional Independence between Concepts

- joint information measure for sets of concepts $c_1, \ldots, c_n$:

$$R(c_1, \ldots, c_n) := \left| \bigcup_{i=1}^{n} iso(c_i) \right|$$

- from this we get **conditional mutual information between concepts** given a set of concepts $S := \{s_1, \ldots, s_n\}$:

$$\begin{aligned} I(c_i, c_j; S) := {} & R(c_i, s_1, \ldots, s_n) + R(c_j, s_1, \ldots, s_n) \\ & - R(c_i, c_j, s_1, \ldots, s_n) - R(s_1, \ldots, s_n) \end{aligned}$$

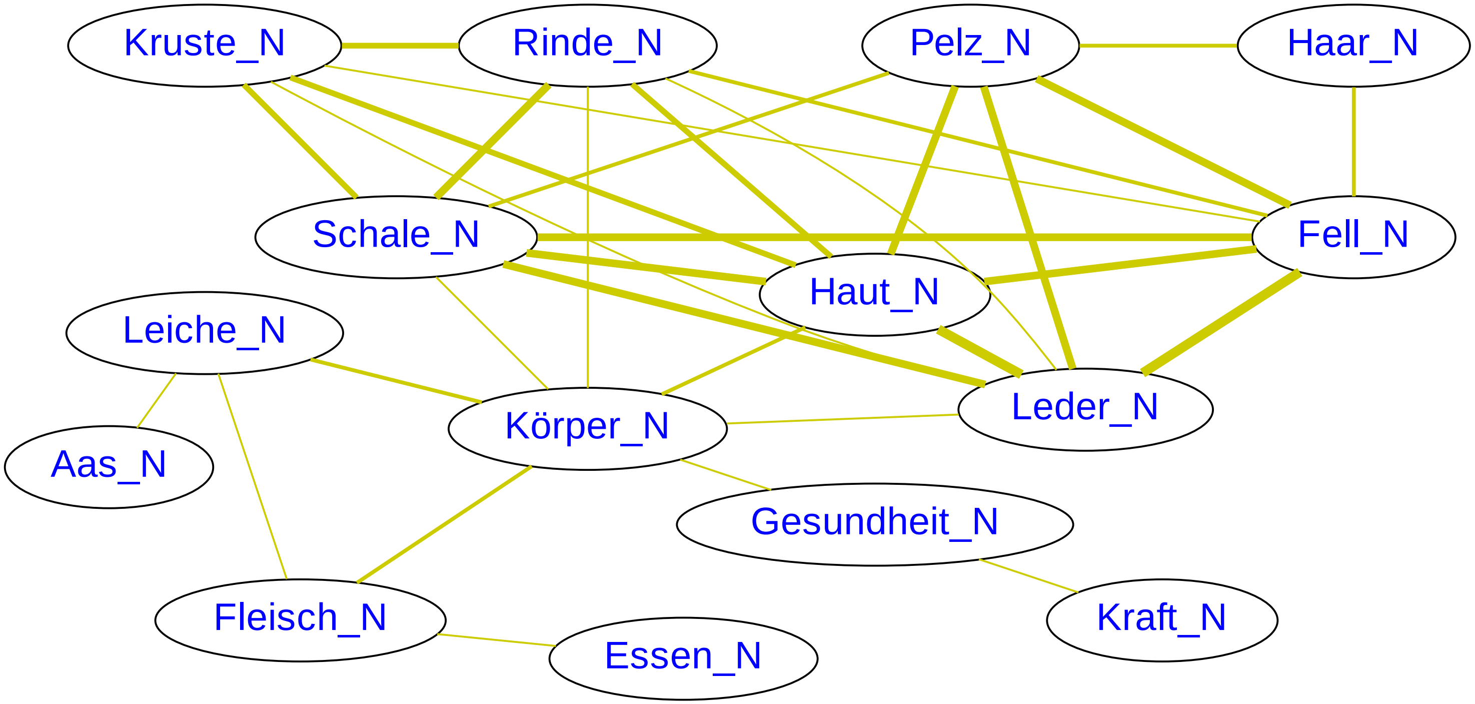
- R is **submodular**; Steudel et al. (2010) show that checking for non-zero I gives us a consistent conditional independence test
- intuitively: are there colexifications between c_i and c_j which cannot be explained away by colexification with any of the concepts in S?

Causal Skeleton and Semantic Map

If we add a constraint on conditioning sets, the **causal skeleton will be a semantic map** as defined by Haspelmath (2003)

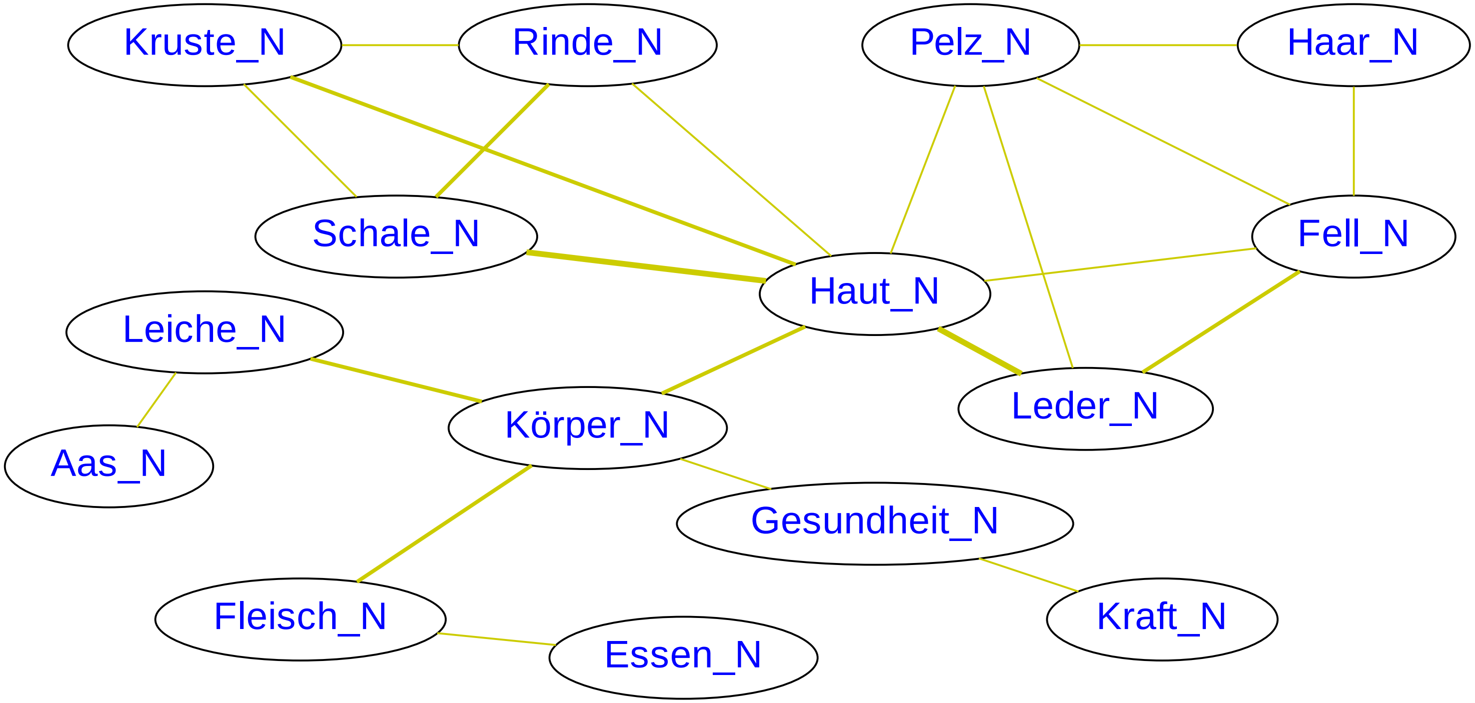
- when deciding whether to delete a link between concepts A and B, the vanilla PC algorithm checks subsets of all neighbors of A and B in the current skeleton
- instead, check for all separating sets composed of neighbors on paths between A and B in the current skeleton
- a link will then be deleted iff every isolectic set it represents also contains other concepts which still connect A and B
- this is equivalent to the condition for a **semantic map**: every isolectic set must cover a connected component in the map

Example: Polysemy Network around SKIN:N



(based on 1.654 isolectic sets from the NorthEuraLex database)

Example: Semantic Map = Derived Causal Skeleton



(inferred on data covering 124 languages from 22 families)

Directionality from the PC algorithm

- PC: v-structure $X \rightarrow Z \leftarrow Y$ iff Z not needed to separate X, Y
- testing for conditional independence using our measure only satisfies **monotone faithfulness**: $(X \perp\!\!\!\perp Y \mid Z)$ iff this is implied by the true causal graph, and Z is minimal with this property
- this means we will never have $X \perp\!\!\!\perp Y$, but $(X \not\perp\!\!\!\perp Y \mid Z)$
- problem: this (in addition to the data sparseness problem) makes the PC algorithm unstable, order in which links are considered can change the result considerably
- workaround in Dellert (2016): aggregate evidence from different unshielded triples into a directionality score
- this causes some errors to cancel out, arrows with high aggregate scores are much more reliable

A specialized v-structure test

New alternative:

- a simple test based on the hypergeometric distribution
- in a v-structure $X \rightarrow Z \leftarrow Y$, we would expect the number k of isolectic sets covering X, Y, Z to be low
- we want to model the distribution of k under null hypothesis that it is not a v-structure
- we get probability of getting k sets covering all three variables if we randomly draw sets for covering Z and Y from all sets covering Z, some of which also cover X
- $k \sim Hypergeo(N, K, n)$ with N (red balls) the number of sets covering X and Z, K (black balls) the number of sets covering Z, but not X, and n (sample size) the sets covering Y and Z
- we can simply check whether $chyper(k, N, K, n) < p$ for a p-value of our choice (in the experiments: p = 0.1)

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

Simulation Experiment

- Simulating Isolectic Sets
- Evaluate Recovery of Directionality

Conclusions

Examples from Dellert (2016)

Simulation Experiment

- to validate the method, we would need hundreds of test cases with enough data (hundreds of languages)
- getting enough real data to do this would require years
- as a preliminary solution, synthesize data based on existing colexification data, and evaluate on them:

1. randomly add directionality to edges in a semantic map (which ensures that the structure is realistic)
2. simulate the development of isolectic sets in the network
3. perform the algorithm on large amounts of data to see whether the directed edges are successfully identified

Simulating Isolectic Sets

Some **properties of isolectic sets** which must be covered:

- there is a maximal size, i.e. sets must both grow and shrink
- on a directed link, growth should only take place in one direction
- shrinking only if result still covers connected component
- resulting size distribution should be similar to real data

My preliminary **algorithm** to achieve this:

- randomly select a starting concept
- evolve the set for a duration of 20 steps, where the area can shrink, expand, or stay stable
- decision whether to shrink or expand is based on Markov chain with true distribution of set sizes as equilibrium
- shrink: remove random peripheral concept from set
- expand: choose neighbor of current set in semantic map, with probability proportional to summed-up connection strengths

Simulation Model: Example Evolution

```
{Zelt::N}
{Nomadenzelt::N, Zelt::N}
{Nomadenzelt::N, Wohnzelt::N, Zelt::N}
{Wohnzelt::N, Zelt::N}
{Nomadenzelt::N, Wohnzelt::N, Zelt::N}
{Nomadenzelt::N, Wohnzelt::N}
{Nomadenzelt::N, Wohnzelt::N, Zelt::N}
{Wohnzelt::N, Zelt::N}
{Haus::N, Wohnzelt::N, Zelt::N}
{Haus::N, Hütte::N, Wohnzelt::N, Zelt::N}
{Haus::N, Hütte::N, Zelt::N}
{Haus::N, Hütte::N}
```

Simulation Model: Examples of Isolectic Sets

```
{Aufschüttung::N, Damm::N, Staudamm::N}
{verträglich::A}
{von_Neuem::ADV, wieder::ADV, wiederum::ADV}
{Transitverkehr::N}
{wer_auch_immer::PRN}
{Gesamtbetrag::N}
{Schweine-::A}
{lieb::ADV}
{irgendwo::ADV, irgendwohin::ADV}
{ranzig::A}
{Muster::N,Patrone::N,Schablone::N,Vorlage::N}
{hierdurch::ADV}
{Büro::N, Geschäftsstelle::N, Kontor::N}
{Autoreifen::N, Rad::N}
{losgehen::V, schleudern::V, schwingen::V, sich_trennen::V, werfen::V}
```

Evaluating Lexical Change Direction Inference

Evaluating directionality inference on the simulated data:

- take a small number of concepts *n*
- consider all unshielded triples in the semantic map
- randomly turn one of the triples into a v-structure
- simulate many isolectic sets on the partially directed map
- use the test on the simulated data to detect v-structures
- compute precision and recall for arrows generated in this way

Evaluating Lexical Change Direction Inference

- two categories: —, →
- based on 2.000 seeds per concept (≈ 2.000 languages)
- average precision and recall on 10 runs on test map:

Category	Precision	Recall
—	66.23%	45.31%
→	15.47%	99.42%

- arrow precision is bad because we find too many v-structures
- changed p-value for hypergeometric test does not help
- much more extensive evaluation pending

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

Simulation Experiment

Conclusions

Examples from Dellert (2016)

Conclusions

- causal inference can be applied to an information geometry defined by cross-linguistic polysemies in order to measure causal influences between concepts
- interpretation of resulting causal pattern:
  indicates probable vectors of semantic expansion
- heuristic: to decide whether $X \rightarrow Y$ or $Y \rightarrow X$, find a third concept Z we can use to test for a v-structure $X \rightarrow Y \leftarrow Z$
- method does not aim to provide objective proofs of historical events, but an unbiased summary of large amounts of easily available data which are too varied and extensive to be processed by a human expert
- $\Rightarrow$ a tool for quickly deriving initial hypotheses about possible directional patterns of semantic evolution

Problems

- we can detect **only v-structures**, everything else is uncertain
- **data sparseness**: statistically significant vanishing correlations are rare! (we would need hundreds of languages)
- operating on dictionary data amounts to very large **sampling errors** (unattested meanings, different choices for glosses)
- cross-linguistically valid concepts often difficult to look up in dictionaries (not enough disambiguating information)

Possible Solutions and Future Work

- data sparseness problems can feasibly be overcome if we are interested in one specific area of conceptual space
- sampling errors could be diminished by corpus analysis & checking back with native speakers
- possible source for more data:
  loose colexification (e.g. derivation, compounds)
- moving from German glosses to concepts by means of automated inference of concepts, and automated lookup of concepts in the dictionary database (work in progress)
- taking account of the fact that some **polysemies are inherited** (a hidden common cause we did not control for!)

Acknowledgments

Thanks are due to:

- Gerhard Jäger (supervision)
- Igor Yanovich (important suggestions and helpful objections)
- student assistants and colleagues who contributed lexical data: Mohamed Balabel, Zalina Baysarova, Isabella Boga, Armin Buch, Natalie Clarius, Thora Daneyko, Ilja Grigorjew, Alina Ladygina, Roland Mühlenbernd, Alla Münch

References

Dellert, J. (2016). Using causal inference to detect directional tendencies in semantic evolution. In Roberts, S., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Fehér, O., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*. Online at `http://evolang.org/neworleans/papers/139.html`.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Tomasello, M., editor, *The New Psychology of Language*, volume 2, pages 211–242. Mahwah, NJ: Lawrence Erlbaum.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Steudel, B., Janzing, D., and Schölkopf, B. (2010). Causal markov condition for submodular information measures. In Kalai, A. and Mohri, M., editors, *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 464–476, Madison, WI, USA. OmniPress.

Table of Contents

Introduction

Causal Inference

Applying Causal Inference to Polysemy Data

Simulation Experiment

Conclusions

Examples from Dellert (2016)

Aggregating Evidence from Unshielded Triples

- decision procedure for deciding whether $c_1 \rightarrow c_2$ or $c_2 \rightarrow c_1$
- consider each unshielded triple $c_1 \rightarrow c_2 \leftarrow c_3$
- define $w(c_1 \rightarrow c_2; c_3) := \frac{|iso(c_1) \cap iso(c_2)| \cdot |iso(c_2) \cap iso(c_3)|}{|iso(c_2)|}$,
  i.e. the number of colexifications between c_1 and c_3 we would have expected if the true pattern had been $c_1 \leftarrow c_2 \rightarrow c_3$ or $c_1 \leftarrow c_2 \leftarrow c_3$
- aggregate the scores from all unshielded triples into a **counterevidence score** $sc(c_1 \rightarrow c_2) := \sum_{c_3} w(c_1 \rightarrow c_2; c_3)$
- if $\frac{sc(c_1 \rightarrow c_2)}{sc(c_2 \rightarrow c_1)} \leq \theta$, infer $c_1 \rightarrow c_2$ (current implementation: $\theta = 0.8$)

Data for Examples

The isolectic sets for the examples were extracted from bilingual dictionaries for about 100 languages.

- primary language German, semi-automated translations for languages where no German dictionary was available
- some data from 20 language families of Northern Eurasia
- languages which are represented best:
  Hindi, Persian, Irish, Russian, Polish, English, Swedish, Dutch, French, Spanish, Italian, Portuguese, Finnish, Hungarian, Khanty, Nenets, Mongolian, Turkish, Georgian, Arabic, Mandarin, Japanese, Thai, Indonesian
- full data for all the examples (in German) available in a machine-readable format as supplementary materials

Example 1: The Eye of a Needle

Language	Lemma	Isolectic Set
Basque	*begi*	{EYE:N, KNAG:N, EYEOFNEEDLE:N, STITCH:N, DROPOFGREASE:N, CHEESEHOLE:N}
Dutch	*oog*	{EYE:N, LOOP:N, EYEOFNEEDLE:N}
Korean	*gwi*	{EAR:N, SPOUT:N, CORNER:N, EYEOFNEEDLE:N}
Livonian	*sīlma*	{EYE:N, LOOP:N, SHACKLE:N, EYEOFNEEDLE:N}
Nenets	*xa*	{EAR:N, HANDLE:N, EYEOFNEEDLE:N}
Polish	*ucho*	{EAR:N, HANDLE:N, EYEOFNEEDLE:N}

- a classic example of metaphorical extension
- *sc*(EYEOFN → EYE)/*sc*(EYE → EYEOFN) = 1.269, i.e. evidence favors EYE → EYEOFN (130 sets, 77 langs, 19 fams)
- *sc*(EYEOFN → EAR)/*sc*(EAR → EYEOFN) = 2.765, i.e. evidence favors EAR → EYEOFN (112 sets, 76 langs, 20 fams)

Example 2: Counting and Calculating

Language	Lemma	Isolectic Set
Coptic	*op*	{COUNT:V, CALCULATE:V, ESTIMATE:V}
Czech	*počítat*	{COUNT:V, CALCULATE:V}
Indonesian	*membilang*	{COUNT:V, CALCULATE:V, NARRATE:V}
Udmurt	*lydjany*	{COUNT:V, CALCULATE:V}
Spanish	*contar*	{COUNT:V, CALCULATE:V, NARRATE::V}

- arguably, COUNT came earlier than CALCULATE
- *sc*(CALCULATE → COUNT)/*sc*(COUNT → CALCULATE) = 1.162, i.e. too little evidence, but in favor of COUNT → CALCULATE (134 isolectic sets, 68 languages, 21 families)
- reason: imperfect mapping of concepts to German glosses used for data retrieval, unresolved polysemy of gloss *zählen* “to count”

Example 3: Hoping and Expecting

Language	Lemma	Isolectic Set
Chinese	*xīwàng*	{HOPE:V, EXPECT:V, WISH:V}
Hebrew	*jixel*	{EXPECT:V, HOPE:V}
Japanese	*nozomu*	{EXPECT:V, HOPE:V, WISH:V}
Portuguese	*esperar*	{HOPE:V, EXPECT:V, WAIT:V}
Turkish	*ummak*	{HOPE:V, EXPECT:V, WAIT:V}

- HOPE:V and EXPECT:V frequently colexified;
  theory makes no prediction, historical evidence gives some hints (Latin *spērāre* “to hope” into Spanish *esperar* “to wait; to hope”)
- $sc(\text{EXPECT} \rightarrow \text{HOPE})/sc(\text{HOPE} \rightarrow \text{EXPECT}) = 2.813$,
  i.e. strong evidence in favor of HOPE:V → EXPECT:V
  (203 isolectic sets, 70 languages, 22 families)