EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

Using Computational Criteria to Extract Large Swadesh Lists for Lexicostatistics

Leiden, October 28, 2015

Johannes Dellert & Armin Buch

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

Table of Contents

Motivation
- Swadesh concepts
- Swadesh lists
- Swadeshness

Our Data

Measuring Swadeshness

Evaluation

Outlook

- lexical data is very valuable for studying the history and phylogeny of languages
- advantages: easy to obtain, represent, and to compare across languages
- problems: lexical substitution and borrowing
- solution: choose concepts least prone to these processes (*Swadesh lists*[1])

[1] See `http://concepticon.clld.org/contributions`

Swadesh concepts

- words likely to be expressed by cognate words in phylogenetically related languages
- traditionally selected by intuition/experience
- **item stability**: observed preservation of cognates within families
- typical Swadesh lists: little more than 200 items
- ASJP: 40 items which most reliably represent language relationships [Holman et al., 2008]
- 40 or 200 items (including non-cognates) are vastly insufficient for the *comparative method* (sound correspondences will only appear for very closely related languages)

More Swadesh concepts

- Which items to collect and test for stability?
- there are longer *basic vocabulary* lists (IDS and WOLD)
- different purpose: WOLD wants loanwords, IDS has many non-basic concepts which are difficult to extract from lexical resources (orientation towards fieldwork)
- **swadeshness**: an informal, scalar, a priori estimation of appropriateness as a Swadesh item
- many ways of measuring swadeshness are possible

Swadeshness: What makes a concept suitable?

Desired feature	Measurable correlate
basic	inclusion in wordlists / textbook glossaries
universal (in Neolithic cultures)	inclusion in wordlists (in minority languages)
high-frequency	shorter realizations
morphologically simple	shorter realizations
stable against	
• borrowing	diversity (across language families)
• metaphor	small similarity to words for related concepts
• taboo	?
no onomatopoesia	diversity across all languages
clearly delineated	well-defined (little polysemy)

Formalizing Swadeshness

- already Swadesh [1952] suggests using stability scores
- recent approaches differ in methodology:
    - WOLD [Tadmor, 2009]: borrowing and replacement (expert judgments)
    - Holman et al. [2008], Rama and Borin [2014]: preservation of cognate classes within language families using automated cognate detection / similarity scores

Our goals

- provide empirically determined concept lists for more extensive lexicostatistical databases
- measure *swadeshness* formally and reproducibly without expert knowledge/bias
- rank any number of concepts given enough data

⇒ a method for automatically generating customized Swadesh lists (any desired length, by geographical region, etc.)

Key ideas

- measure **basicness** of a concept by the average **language-specific information content** across realizations
- measure **stability** by **correlating distances** of concept realizations and overall language distances for balanced samples of language pairs

Table of Contents

Motivation

Our Data

Measuring Swadeshness

Evaluation

Outlook

Our Data

- a large dictionary database; version used in this study contains 975,953 translations into German
- substantial amounts of data for 88 languages covering 20 primary language families of Eurasia
- uniform phonetic representation (in ASJP code), based on automated conversion of orthography (for most languages) or additional pronunciation information contained in the database (e.g. for English, Danish, Japanese, Chinese, Persian)

Our Data: Concepts

- German glosses were automatically grouped into overlapping clusters based on a polysemy network over our data
- manual resolution into a master list of thousands of concepts, each expressed by a tuple of German gloss lemmas
  - ▷ “queue”: Warteschlange::N/Schlange::N
  - ▷ “snake”: Schlange::N/Schlange[Tier]::N
- automated lookup of concept realizations in the database, only use concepts for which realizations in at least 10 languages were found (5,039 concepts)

Our Data: Realization Distances

- for realization distances, we use the Needleman-Wunsch algorithm over ASJP sound classes
- global segment similarities inferred iteratively from alignments (see LexStat [List, 2012]).
- additional weighting by information content

Our Data: Language Distances

- for the overall language distance, we computed the dER measure as described in Jäger [2013]
- essentially, align all words for the language pair, and measure how far up in ranking of pairs with different meanings pairs with identical meanings would occur
- only based on the top-50 concepts from the Holman ranking (1,250 pairs are enough for a good estimate, avoids problems with massive loans in basic vocabulary)

Table of Contents

Motivation

Our Data

Measuring Swadeshness

- Information Content
- Preservation and Diversity
- Swadeshness Score

Evaluation

Outlook

Information content

- word length reflects both basicness and monomorphemicity, but needs to be defined in a cross-linguistically valid way
- unit of length: phoneme in ASJP encoding, weighted for its **perplexity** in n-gram model
  - information content of a word form: add up segment-wise perplexity scores: InfOrmEISn
- small phoneme inventory ⇒ each segment contains less information ⇒ longer words are needed to have the same cumulative information content: $|\mathtt{arbaro}| \approx |\mathtt{tqe}|$
- *inf*(*c*): average information content over all realizations of a concept *c*

Family-internal preservation and cross-family diversity

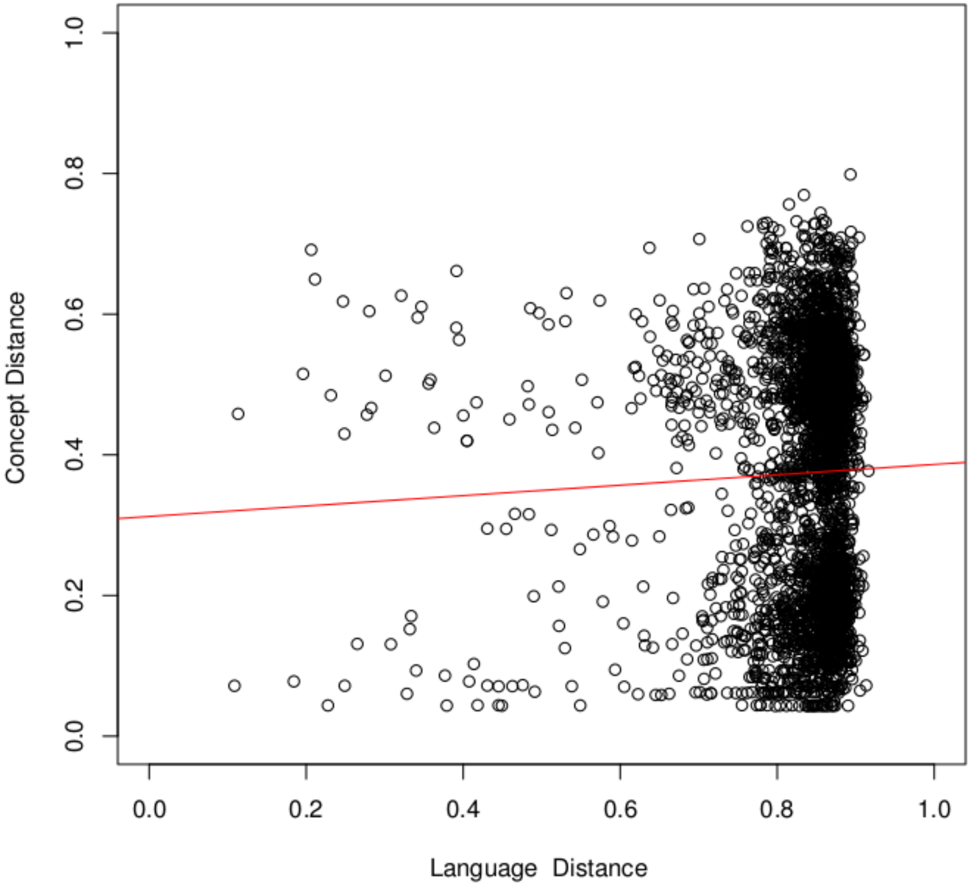
- traditional approaches: cognate preservation ratios in pairs of related languages
- our earlier approach: 1 language per family, opposite measure (words likely to be expressed by *non*-cognate words in phylogenetically *un*-related languages)
- our new synthesis: how well do the distances between concept realizations represent overall language distance?
- no need for costly expert judgments or unreliable automatic cognacy detection

Local-Global Distance Correlation

- **local-global distance correlation** $lgc(c)$ is the average Pearson correlation between concept-specific distances and global distances over 10 balanced samples of language pairs
- penalizes similar realizations in unrelated languages
  - ▷ English *door* and Japanese *doa*, a borrowing
- penalizes dissimilar realizations in closely related languages
  - ▷ Spanish *pájaro* (from Lat. *passer* “sparrow”), replaced *avis* as the most usual word for “bird”

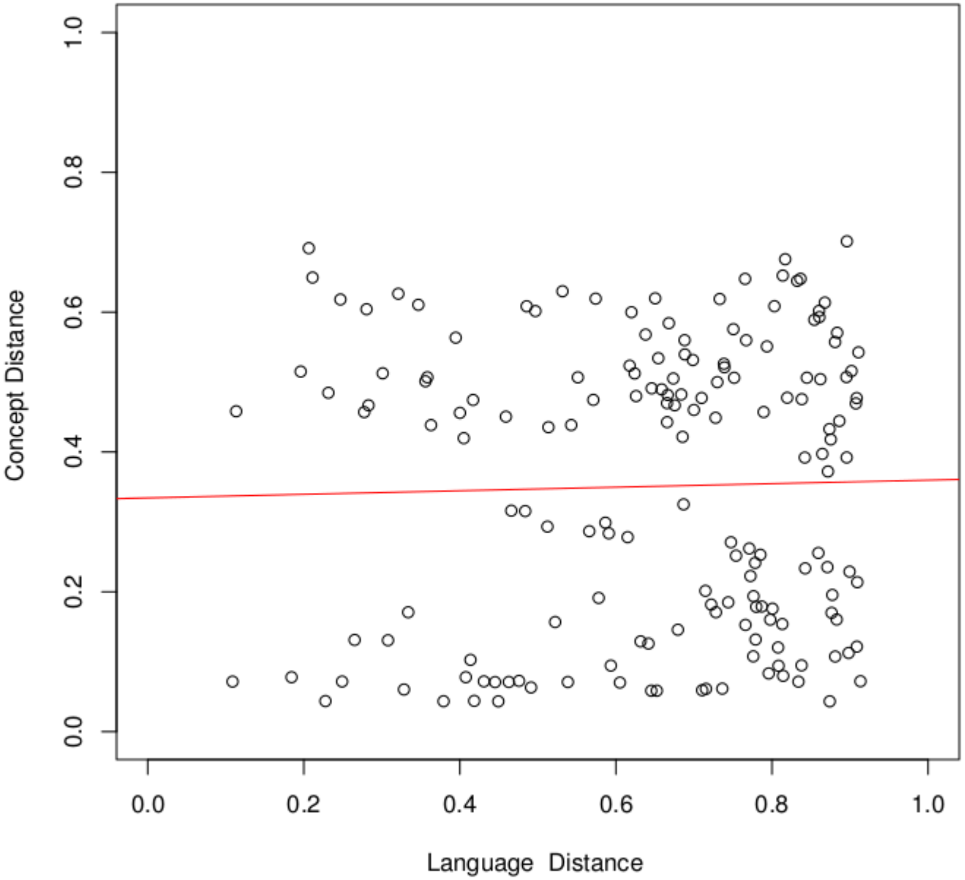
Local-Global Distance Correlation: Example 1

All realization distances for “April”:



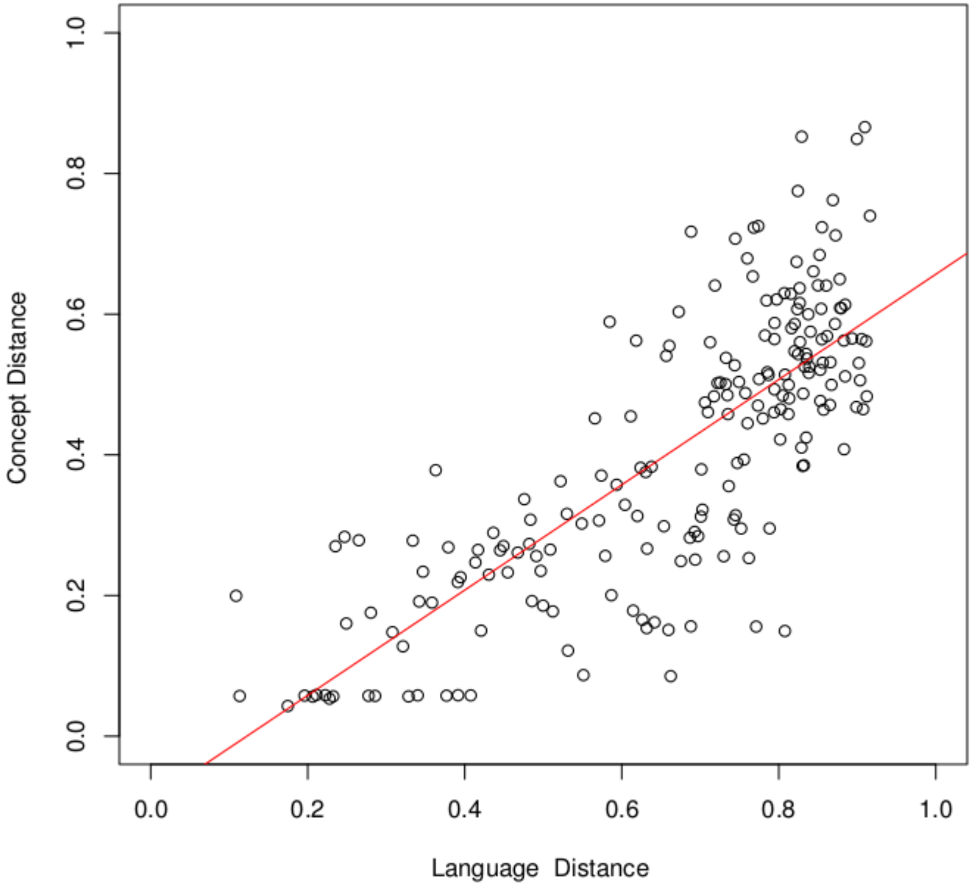
Local-Global Distance Correlation: Example 1

A sample of realization distances for “April”:



Local-Global Distance Correlation: Example 2

A sample of realization distances for “fish”:



Combining the measures

- ranking is based on the **swadeshness score**
  $sc(c) := inf(c) - 3 \cdot lgc(c)$,
  a trivially simple linear combination of our two measures
- weight ratio is optimized for coverage of the Swadesh list (i.e. a single parameter ⇒ no risk of overfitting)
- more complex combinations did not lead to significant improvements

Table of Contents

Motivation

Our Data

Measuring Swadeshness

Evaluation

- Top 300 concepts

Outlook

Evaluation in a nutshell

- Spearman rank correlation values of 0.276 with the ranking in [Holman et al., 2008] and 0.468 with [Rama and Borin, 2014]
- not extremely high, but also not our main goal
- 98 of 207 Swadesh concepts are also in our top-207 list (a good result, given that we filtered these out from over 5,000 concepts)
- internal stability: volatility of the ranking (not today)

Top 300 concepts: 1-100

- Swadesh concepts (117/207) in red
- sparse data ($<$ 20 languages) in gray

to be, what, to give, I, one, thou, water, to arrive, to put, to drink, to go, to come, him/her, they, she, to take, to walk, he, hand, you, this, in, two, and, day, if, blood, at, mouth, to carry, month, river, to say, with, side, we, ten, to burn (intransitive), to live, arm, tooth, soil, head louse, who, person, one (pronoun), to melt, skin, ice, to go away, to see, by means of, to float, grease/fat, is, foot, six, that, to sleep, hair, to listen, to become, to travel, to fetch, there, to buy, to tie, to hear, name, head, here, man, son, fingernail, to bring, house, to water, to know, since, night, where, end (temporal), good, road, ashes, us, branch, dog, not, path, to wear, to cry, stone, to stand up, like that, to sew, to fall, thee, direction, them

Top 300 concepts: 101-200

five, fish, like this, to cover, to divide, under, to get, his/her, fire, to pound, a, thread, bone, to hold, wood, to remain, throat, to blow, snow, thy, place, eye, to smelt, to call (give a name), moon, to step, to fly, new, land, to read, in order to, big, to cease, salt, to boil, via (direction), woman, to make, to open, to you (pl.), whom, to plait, nest, breast, strength, but, to mow, to cook, to flow, too, to pursue (a business), whether, neck, leg, mother, old age, to seem, beech, can, thither, word, to wash, nose, cold, towards, work, ear, old man, three, to pour, a hundred, to cut, horn, to throw, to share, to weave, father, wool, to stand, or, to be allowed, grass, door, to milk, to dig, bile, my, peel/husk, god, mind, to have, to wipe, warm, while, language, to smell (intransitive), long, to bite, to burn (transitive), me

Top 300 concepts: 201-300

to suck, to find, chest, tongue, other, to sink, cow, meadow, to grow, to roast, edge, sea, country, to kick, to finish, shaft, half, than (comparison), seat, to tread, lake, horse, wet, to die, tree, to lick, arrow, mountain, to look, in (time span), four, clay, under (direction), bad, to dry (intransitive), to hang up, to swim, to elevate, beeswax, town square, to thee, to abandon, to eat, leaf, to incline, smoke, to recognize, hither, topic, pain, to plough, stick, through, to choose, bark, hedge, time, bar, to fill, page, above, let, time (occasion), top, as (comparison), handle, green (n.), to row, price, meat, law (subject), out (direction), to feel, more, moss, to hit, to begin, wide, deep, to want, steady, winter, to rinse, amount, to move, self, wind, to pull, for, sail, belt, hole, to suit, to tear, to sell, to preserve, age, saddle, ago, to sting

Table of Contents

Motivation

Our Data

Measuring Swadeshness

Evaluation

Outlook

Summary and future directions

- this approach is reproducible, adaptive, data-driven and automatic, scalable
- captures some of the intuitions and insights behind Swadesh lists
- within EVOLAEMP, we are collecting data for 1,016 concepts in 103 languages of Northern Eurasia (NorthEuraLex)
- this list of concepts was derived earlier using previous versions of our method and database

References I

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. Explorations in automated language classification. *Folia Linguistica*, 42(3-4): 331–354, 2008.

Gerhard Jäger. Phylogenetic Inference from Word Lists Using Weighted Alignment with Empirically Determined Weights. *Language Dynamics and Change*, 3(2):245–291, 2013.

Johann-Mattis List. LexStat. Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125, Stroudsburg, 2012.

References II

Taraka Rama and Lars Borin. N-Gram Approaches to the Historical Dynamics of Basic Vocabulary. *Journal of Quantitative Linguistics*, 21(1):50–64, 2014.

Morris Swadesh. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4):pp. 452–463, 1952.

Uri Tadmor. Loanwords in the world's languages: Findings and results. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the world's languages. A comparative handbook*, pages 55–75. 2009.