EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

Evaluating Cross-Linguistic Polysemies as a Model of Semantic Change for Cognate Finding

STRiX workshop — Gothenburg, November 24, 2014

Johannes Dellert

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Table of Contents

Motivation: Computational Historical Linguistics

CHL develops computational methods for analyzing phenomena of interests to historical linguistics

- ▷ phylogenetic relationships
- ▷ language contacts
- ▷ language change on different levels of linguistic description

Goals of our EVOLAEMP project in Tübingen:

- ▷ evaluate existing methods borrowed directly from bioinformatics
- ▷ attempt to enhance these methods by bringing linguistic knowledge back into the models (sound correspondences, semantic change)

Problem for evaluation: not enough data available

- ▷ existing wide-coverage lexicostatistical databases have at most 200 concepts per language
- ▷ more concepts ⇒ only samples from each linguistic

Motivation: The Idea

There is some informal notion of plausibility when cross-semantic etymologies are discussed in the literature:

- ▷ a semantic shift from “sun” to “day” is plausible
- ▷ a shift from “moon” to “night” is much less so
- ▷ “nose” $\rightarrow$ “mountain” is good, “nose” $\rightarrow$ “swamp” is not

How can we capture and model these constraints?

Basic Idea: If there is any language in which two concepts can be expressed by the same word, this makes a semantic shift between these concepts much more plausible.

Implementation: Observe which gloss language lemmas often **occur together as translation glosses** in dictionaries, use the counts to build a weighted **polysemy graph**.

Table of Contents

The Dictionary Data

Colexification data is based on a (German-based) multilingual dictionary database of 610,000 lemma-gloss pairs, including:

- ▷ **more than 10,000 entries** in: Arabic, Chinese, Dutch, English, Finnish, Hungarian, Japanese, Kazakh, Karelian, Lithuanian, Meadow Mari, Nganasan, Persian, Polish, Russian, Spanish, Swedish, Tundra Nenets, Turkish
- ▷ **more than 5,000 entries** in: Basque, Chukchi, Erzya, Estonian, French, Georgian, Indonesian, Italian, Kalmyk, Ket, Khanty, Kildin Sami, Komi, Livonian, Mansi, Moksha, Mongolian, Northern Sami, Norwegian, Portuguese, Selkup, Skolt Sami, Tatar, Udmurt, Uzbek, Veps
- ▷ **less than 5,000 entries** in more than 50 additional languages (all Eurasian, under continual expansion)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Table of Contents

The Polysemy Network: Basic Idea

polysemy network: weighted undirected graph over gloss language lemmas, edge weights represent degree of colexification

Basic idea resembles that of a **semantic network** (e.g. François 2008), but link structure is less meaningful:

- ▷ transitivity of edges not relevant
- ▷ no strict relationship to isolectic areas

Weighted variant first proposed by List ea. (2013), who also mention possible applications to an improved computational treatment of semantic change.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Problem of Chance Homophonies

Not every colexification represents polysemy or vagueness, there are quite a few instances of **chance homophonies**: `fa` *šir* “milk; lion”, `ru` *luk* “bow; onion”, `sv` *arm* “arm; poor”

List ea. (2013) approach this problem by discarding any edge that is only present in one language family.

On our database, the problem is less severe if **orthographic forms** are used for determining colexification. (especially helpful for `en`, `fr`, `ja`, and `zh`)

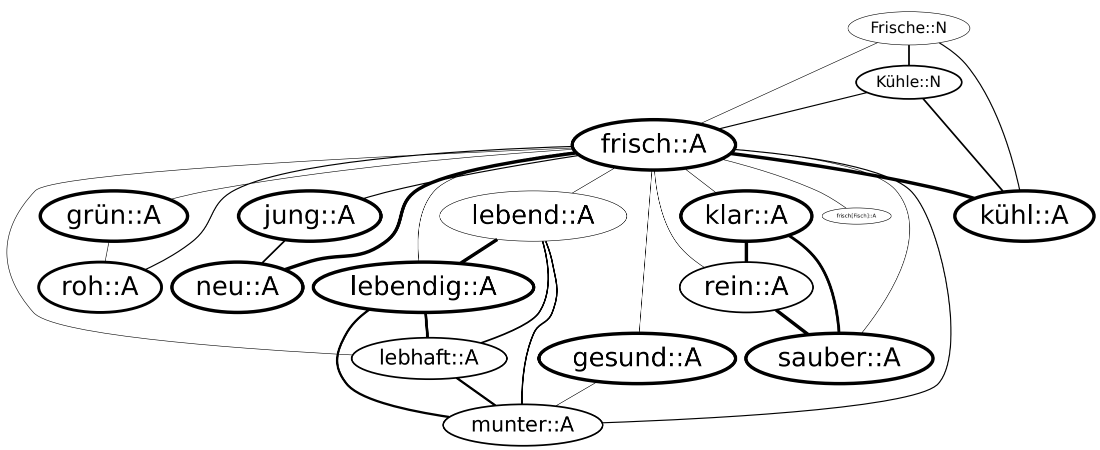
We apply the language-family criterion on the genus level:

- ▷ (milk::N,lion::N), (poor::A,arm::N), (bow::N,onion::N) only appear in one language ⇒ edges are discarded
- ▷ (six::NUM,fir::N) occurs in `fi` and `vep`, but same genus ⇒ also count 1, edge is discarded

The Polysemy Network: Metrics

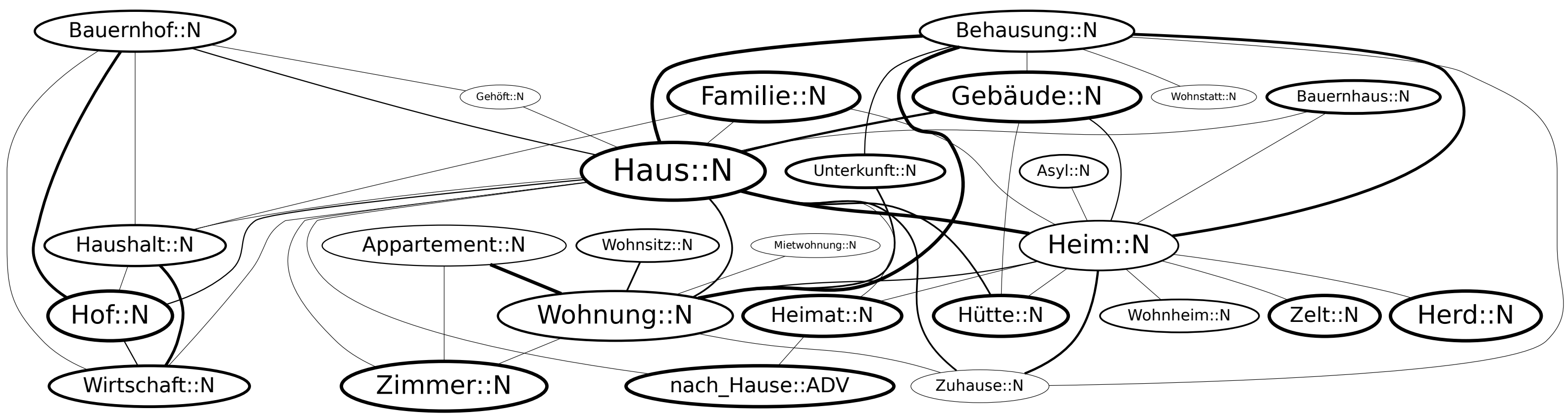
The current version of the resulting polysemy network has:

- ▷ 39,707 nodes and 80,399 edges
- ▷ a central **connected component of 19,194 lemmas** (!)
- ▷ 70 other components of 6 or more lemmas
- ▷ 519 components of sizes 3-5
- ▷ 1,672 components of size 2
- ▷ 14,854 unconnected “islands” of size 1



The Polysemy Network: Example

The following subnetwork is generated if we start with the lemma *Zuhause::N* “home” and follow all edges up to depth 2:



“house”, “family”, “household”, “farm”, “dwelling”, “homeland”, ...

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Table of Contents

Experiment: Cognate Finding

Question: Do synchronic polysemies provide a useful model of semantic change for computational historical linguistics?

> Croft ea. (2009): “The first step in a semantic change is extension of a word to a new meaning. [...] A crosslinguistic sample will allow us to quantify the likelihood of semantic change in a particular time slice.”

The network allows us to test this assumption on a large scale.

Experiment: Cognate Finding

A **pilot study** (more extensive experiments ongoing):

Take two distantly related languages (Finnish and Hungarian).

Compile a list of cognates from the etymological literature (306 cognate pairs, quite exhaustive).

Get standard bilingual dictionaries of both languages (in my case, `fi-de` and `hu-de` dictionaries of 15,000 entries).

For each cognate pair, determine the minimum number of hops through the polysemy network which is needed to link some pair of German glosses from the dictionaries.

Investigate the relationship between the percentage of cognate pairs covered, and the amount of semantic latency introduced at each search depth k.

Experiment: Cognate Finding

Results:

For the 200 cognate pairs covered by the dictionaries:

search depth k	0	1	2	3	4
env. size after k hops	1	18.6	185.2	1057.3	3324.6
connected by k hops	89	122	140	155	166
not connected by k hops	111	78	60	45	34

Only 89 cognates share a gloss
(too few for detecting regular sound correspondences).

Environment size grows quickly, $k = 2$ being the highest depth where risk of chance similarity seems manageable.

51 additional pairs (57% improvement) using G_2.

Answer: Yes, cross-linguistic polysemies do promise to be useful.

Experiment: Cognate Finding

Some **example paths** (*fi* on the left, *hu* on the right):

ääni “voice” : Stimme – Lied : *ének* “song”

kerjätä “beg” : betteln – bitten : *kér* “ask for”

vuori “mountain”: Berg – Gipfel – Spitze : *orr* “nose”

tunkea “thrust”: drängen – schieben – stecken : *dug* “stick”

kumpu “hill”: Hügel – Erdhügel – Erdwall – Welle : *hab* “wave”

valo “light”: Licht – Feuer – Funke – Blitz: *villám* “spark”

terä “blade”: Spitze – Maul – Loch – Grube – Falle : *tőr* “trap”

metsä “forest”: Wald – Baum – Stange
– Sandbank – flach – weit : *messze* “far”

Table of Contents

Polysemies as a Model of Semantic Similarity

Approaches to modeling semantic similarity in applications:

- WordNet, FrameNet etc. (**ontologies**)
- LSA, PMI etc. (**co-occurrence models**)

Cross-linguistic polysemies are largely orthogonal to both of these:

- Ontology: *house* $\rightarrow$ *building*, *hut*, *cottage*
- Co-occurrences: *house* $\rightarrow$ *build*, *live*, *destroy*, *mortgage*
- Polysemies: *house* $\rightarrow$ *home*, *family*, *tent*, *hearth*, *farm*

$\Rightarrow$ Polysemy networks could be relevant for many applications as an **additional source of similarity judgments**.

References

FRANÇOIS, A. (2008). **Semantic maps and the typology of colexification: intertwining polysemous networks across languages.** In M. Vanhove (Ed.), *From polysemy to semantic change*, pp. 163–215. Amsterdam: Benjamins.

PERRIN, L.-M. (2010). **Polysemous qualities and universal networks, invariance and diversity.** Linguistic Discovery 8(1), 259–280.

LIST J.-M., TERHALLE A., URBAN M. (2013). **Using network approaches to enhance the analysis of cross-linguistic polysemies.** *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*: 347-353.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

References

KONDRAK, G. (2001). **Identifying cognates by phonetic and semantic similarity.** *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies.*

CROFT, W., BECKNER, C., SUTTAN, L., WILKINS, J., BHATTACHARYA, T., AND HRUSCHKA, D. (2009). **Quantifying semantic shift for reconstructing language families.** *Talk, held at the 83rd Annual Meeting of the Linguistic Society of America.*

RÉDEI, K., editor (1988). **Uralisches etymologisches Wörterbuch.** *Akadémiai Kiadó, Budapest.*