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

Johannes Dellert

Seminar für Sprachwissenschaft, Universität Tübingen

jdellert@sfs.uni-tuebingen.de

ABSTRACT

Recently, cross-linguistic polysemies have been discovered as a valuable resource for lexical semantics. Co-occurrences of gloss lexemes in multilingual wordlists can be summarized into polysemy networks, a new type of semantic resource. In this work, this relatively new paradigm is applied to an entire dictionary database, resulting in a polysemy network that spans more than 30,000 German lexemes.

Within computational historical linguistics, polysemy networks have been discussed as a possible computational model of semantic change, where short paths in such networks are expected to express possible semantic shifts. To assess the validity of the polysemy network for this purpose, I apply it to the task of finding cognate candidates, i.e. words in related languages which might have developed from the same word in a common ancestor language. On a test set of cognates shared between Finnish and Hungarian, I investigate the number of true cognate pairs whose translations are connected in the polysemy network by shortest paths of different lengths. The results are very promising, providing evidence that cross-linguistic polysemies are indeed closely connected to plausible semantic shifts.

KEYWORDS: Cognate Detection, Semantic Change, Semantic Shift, Cross-Linguistic Polysemies, Polysemy Network, Finno-Ugric.

1 Introduction

In historical linguistics, the ubiquity of semantic change is a central obstacle for establishing etymologies. The principal problem is that allowing for too much semantic latitude will make it much more likely to find parallels between words which are in fact only due to chance. In the etymological literature, informal arguments for the plausibility of claimed semantic shifts are used to address this risk. Very often, such arguments rely on parallel developments in other languages. For instance, a historical linguist will find a proposed semantic shift from "sun" to "day" much more plausible than a shift from "moon" to "night", just because the first shift is much more common cross-linguistically. If we want to automate the task of cognate detection in large wordlists, this informal knowledge of plausible shifts needs to be modeled explicitly.

In recent work from the field of lexical typology such as (Croft et al., 2009), synchronous **polysemies** (where multiple concepts are expressed by the same form) are considered as a possible source of evidence for plausible semantic shifts. First steps towards automating the process of aggregating cross-linguistic polysemy data were already made by (Croft et al., 2009) and (Perrin, 2010), who introduced the concept of a **polysemy network**, a graph over concepts expressed by glosses, where each link represents the fact that at least one language has a lexical item which covers both concepts.

(Steiner et al., 2011) present a toolchain for computational historical linguistics, where the module for modeling semantic change is based on two similarity matrices. The first aggregates form distances across languages to arrive at a crude measure of lexical relatedness, the second encodes a polysemy network extracted from a small number of wordlists. (List et al., 2013) elaborate on the extraction of such networks from wordlists, and presents a polysemy network that connects 1,286 concepts on the basis of typologically very diverse dictionary data.

In this work, I use a large multilingual dictionary to infer a much larger polysemy network over German glosses. Evaluation on a test set of 200 Finnish and Hungarian cognate pairs provides evidence that this network is indeed a good model of plausible semantic shifts.

2 The Dictionary Data

The polysemy network was extracted from my personal dictionary database of language-to-German wordlists for 80 (mainly Eurasian) languages, altogether comprising more than 400,000 entries. Most of the wordlists have been extracted from phrasebook and textbook glossaries (most typically between 1,000 and 3,000 entries), but the lists for 30 languages contain rather complete basic vocabularies of 5,000 or more entries, and the database has more than 15,000 entries for English, Dutch, Swedish, Polish, Russian, Persian, Arabic, Turkish, Hungarian, Finnish, and Chinese. Despite a strong bias in favor of Indo-European and Uralic, ten Eurasian language families are represented by a word list of more than 5,000 entries, leading to a reasonable amount of cross-linguistic diversity.

The database can be formalized as a set of dictionaries D_L where L is any ISO-639-3 language code. Each dictionary D_L is a collection of entries $\langle l, T \rangle$, where l is a lemma in language L, and T a tuple of German glosses approximating one of the senses of l. Furthermore, we use a partition $\mathscr{F}$ which groups the language codes into genetic units F. These genetic units are chosen at approximately the time depth of the primary branches of the Indo-European and Uralic language families, which is roughly equivalent to the genus level.

3 The Polysemy Network

In essence, a polysemy network can be inferred automatically as a graph over dictionary glosses, with weighted edges counting the instances where the connected glosses occur together on the same side of a dictionary equation (**colexifications**). The problem with this simple definition is that not every colexification represents polysemy or vagueness. The dictionary data contain many **chance homonymies** such as *arm* "arm; poor" from Swedish and Dutch, or *kuus(i)* "fir; six" from Finnish and Estonian. Similar to (List et al., 2013), I solve this problem by discarding any colexification edge that is only attested for one genetic unit. The reduction of the bias in favor of Indo-European and Uralic languages is worth the loss of information incurred.

Formally, we model the polysemy network as an undirected graph $G = \langle V, E \rangle$ over a set of German glosses V. An edge between two glosses $\{g_1, g_2\}$ is included in the network if and only if there are at least two genetic units where g_1 and g_2 were both used as glosses for some lemma in some language:

$$\{g_1, g_2\} \in E :\Leftrightarrow |\{F \ \in \mathscr{F} \mid \exists L \in F, \exists l : g_1, g_2 \in \bigcup_{\langle l,T\rangle \in D_L} T\}| \geq 2$$

The polysemy network G over the current dictionary database consists of 32,653 gloss nodes and 47,648 edges. 14,391 nodes form unconnected islands. The rest either belongs to a surprisingly large and rather densely connected central component of 13,073 glosses, or to one of 1,940 smaller components. The sizes of these components are distributed as follows:

size	2	3	4	5	6	7	8	9	10	11	12	14	15	35
count	1,325	325	143	67	28	19	8	9	4	4	5	1	1	1

To give an impression of the network structure, Figure 1 shows a graph visualization representing the surroundings of the gloss *Atem:N* "breath". The node size represents the frequency of each gloss in the dictionary database. Interesting colexifications with *Atem:N* include *Seele:N* "soul", *Geist:N* "mind, spirit", and *Leben:N* "life".

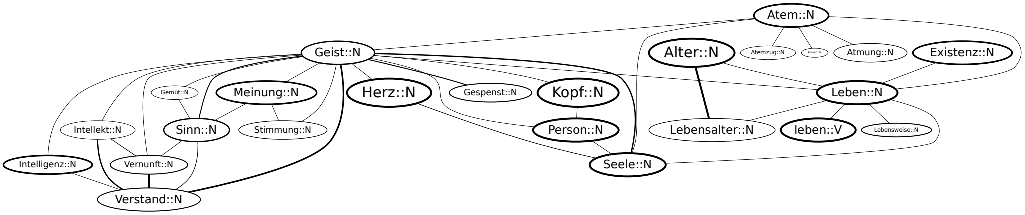

Figure 1: The depth-2 environment of *Atem:N* "breath" in the polysemy network.

4 The Test Set for Cognate Finding

For the experiment described here, I chose to use cognates shared between Finnish and Hungarian, two quite distant languages from the Uralic language family which separated about 4,000 years ago. Due to lexical replacement by borrowing or innovation, only a few hundred cognates between the two languages have survived. For the test set, I collected a quite exhaustive set of 306 cognate pairs by going through the *Uralisches Etymologisches*

Wörterbuch (UEW) "Uralic Etymological Dictionary" edited by (Rédei, 1988), which is still considered the standard source on shared Finno-Ugric vocabulary. Only about one third of the cognate pairs have not undergone any semantic change (e.g. *veri* and *vér* "blood", *kolme* and *három* "three")[1], all the others have undergone semantic shift, often in very interesting ways. For instance, the Proto-Finno-Ugric word reconstructed by Rédei as **kojwa* "birch" (Finnish *koivu*) is possibly related to Hungarian *hajó* "ship". While this etymology (like a few dozen others) in my test set is disputed, I still chose to include such candidate pairs because they showcase the types of semantic shifts which are at least deemed possible.

To investigate the network's potential for helping us to find these cognate pairs, I use two small Finnish-German (Semrau and Rump, 1996) and Hungarian-German (Maczky-Váry, 1997) dictionaries of about 15,000 entries each. While much larger dictionaries are available for these two language pairs, dictionaries of this size are a realistic test case for wider application because they are similar in size and coverage to the best lexical resources we have for many less researched languages.

5 Evaluation

For the evaluation, I computed how many true cognate pairs would actually be taken into consideration by an automated method that relies on the polysemy network to judge semantic similarity. The polysemy network gives us a principled way to guide the search for cognate candidates. Starting from a concept, we can consider all glosses which are at most k steps away in the graph. The resulting accessibility relations can be defined recursively as

$$G_0 = \{\{a,a\} \mid a \in V\}; \quad G_{k+1} := \{\{x,z\} \mid \exists y : \ \{x,y\} \in G_k, \{y,z\} \in E\}$$

The standard lexicostatistical approach based on Swadesh lists thus corresponds to only considering G_0, and the nodes in Figure 1 represent the environment of "breath" in G_2.

We assess the usefulness of the polysemy network for cognate detection by testing how many of the semantic shifts contained in the test set are covered by short paths in the network. For path depth k, we are thus interested in the question whether $\exists i,j : \{g_i, h_j\} \in G_k$ for each cognate pair $\langle l_1, \{g_1, ..., g_{i_m}\}\rangle \in D_{fin}$ and $\langle l_2, \{h_1, ..., h_{j_n}\}\rangle \in D_{hun}$. We also need to consider how large the search environments become for every path depth k, since for higher k we will soon cover too large areas in the semantic space.

For 106 out of the 306 cognate pairs in the test set, at least one of the lexical items was missing from the dictionaries. Usually, this was either because the lexicographers considered the concept in question as no longer basic to modern life (e.g. certain types of fish and birds, or hunting and fishing implements), or because the cognates can only be found in dialects, and not in the modern standard variants covered by the dictionaries. For the remaining 200 cognate pairs, the results are summarized in Figure 2.

First of all, we see that a typical lexicostatistical method (G_0) could only find at most 89 cognates, even if all the relevant concepts were contained in the Swadesh list. As expected, the search environment size grows very quickly for higher k, making chance similarities far too likely. To the author's opinion, the most interesting search depth is given by G_2, since considering about 50 additional glosses per cognate pair will keep this risk manageable even

[1] In all the cognate pairs discussed in this paper, the Finnish form is always quoted first, then the Hungarian form.

search depth k	G_0	G_1	G_2	G_3	G_4	G_5
avg. size of environment in G_k	1	9.2	55.7	229.7	718.6	1720
cognate pairs connected by G_k	89	116	130	137	144	154
not connected by G_k	111	84	70	63	56	46

Figure 2: The results for cognate finding on the Finnish and Hungarian test set.

for automated methods. For G_2 compared to G_0, we get 41 additional true cognate pairs in our candidate list, a very promising improvement of about 46%. Looking at a few examples, the connection between *ääni* "voice; sound" and *ének* "singing; song" is found at $k = 1$, *vuori* "mountain" and *orr* "nose" at $k = 2$, while *salko* "staff" and *szálko* "splinter" are only connected at $k = 4$. For disputed cognates like the aforementioned *koivu* "birch" and *hajó* "ship", there often is no connection.

6 Conclusion and Outlook

In this paper, I have used a large multilingual dictionary to construct a polysemy network which encodes cross-linguistic polysemies on an unprecedented scale. A small experiment on an almost exhaustive list of shared cognates between Finnish and Hungarian has shown that the network helps to make distant cognate pairs available for automated detection. From a wider perspective, the results provide some evidence that cross-linguistic polysemy patterns are a good model of the plausibility judgments for semantic shifts used in historical linguistics.

In future work, the performance of the polysemy network for the chosen task will be evaluated against other measures of semantic similarity which can be derived from existing lexical-semantic resources. The network as such will also be useful for investigating the value of cross-linguistic polysemies as an alternative data source in many other branches of computational linguistics where lexical similarity judgments are needed.

References

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.

List, J.-M., Terhalle, A., and Urban, M. (2013). Using Network Approaches to Enhance the Analysis of Cross-Linguistic Polysemies. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Maczky-Váry, M. (1997). *Langenscheidts Universal-Wörterbuch Ungarisch*. Langenscheidt KG, Berlin und München. 7. Auflage.

Perrin, L.-M. (2010). Polysemous Qualities and Universal Networks, Invariance and Diversity. *Linguistic Discovery*, 8:1–22.

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

Semrau, R. and Rump, R.-I. (1996). *Langenscheidts Universal-Wörterbuch Finnisch*. Langenscheidt KG, Berlin und München. 12. Auflage.

Steiner, L., Stadler, P. F., and Cysouw, M. (2011). A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change*, 1(1):89–127.