Using Network Approaches to Enhance the Analysis of Cross-Linguistic Polysemies

Johann-Mattis List
Philipps-University Marburg
mattis.list@uni-marburg.de

Anselm Terhalle
Heinrich Heine University Düsseldorf
terhalle@phil.hhu.de

Matthias Urban
Philipps-University Marburg
matthias.urban@uni-marburg.de

Abstract

Since long it has been noted that cross-linguistically recurring polysemies can serve as an indicator of conceptual relations, and quite a few approaches to model and analyze such data have been proposed in the recent past. Although – given the nature of the data – it seems natural to model and analyze it with the help of network techniques, there are only a few approaches which make explicit use of them. In this paper, we show how the strict application of weighted network models helps to get more out of cross-linguistic polysemies than would be possible using approaches that are only based on item-to-item comparison. For our study we use a large dataset consisting of 1252 semantic items translated into 195 different languages covering 44 different language families. By analyzing the community structure of the network reconstructed from the data, we find that a majority of the concepts (68%) can be separated into 104 large communities consisting of five and more nodes. These large communities almost exclusively constitute meaningful groupings of concepts into conceptual fields. They provide a valid starting point for deeper analyses of various topics in historical semantics, such as cognate detection, etymological analysis, and semantic reconstruction.

1 Introduction

What do "milk" and "udders" have to do with each other? Conceptually, they are closely related, since the former is the product and the content of the latter. Linguistically, they may even look the same, being referred to by identical word forms in many different languages, such as, e.g., by [nax] in Judeo-Tat (an Indo-European language), by [ukun] in Oroqen (an Altaic language), or by [mis] in Miao (a Hmong-Mien language, all data taken from Key and Comrie 2007). Historically, the conceptual relation between "milk" and "udders" may show up in the form of semantic shifts where a word which was formerly used to express one of the concepts in a given language is henceforth used to express the other one. Thus, in Standard Chinese, the word [niou^{35}nai^{214}] "milk" is a compound of [niou^{35}] "cow" and [nai^{214}] "milk" which originally meant "udder" (as well as "breast").[1]

The situation in which a set of conceptually related meanings is expressed by the same form in a given language is known as *polysemy*. In this paper, we show how an analysis of a weighted network reconstructed from a large database of cross-linguistic polysemies can help to shed light on common conceptual associations in the world's languages. We find that the reconstructed network has a strong *community structure*: it can be easily separated into communities, i.e. groups of nodes that share more connections with each other than with nodes outside the group. We detected 104 large communities consisting of five and more nodes. These communities cover 879 out of 1286 concepts (68%) and almost exclusively constitute meaningful groupings of concepts into conceptual fields, thus providing a valid starting point for further analyses in historical linguistics.

[1] Superscript numbers indicate tones.

2 Polysemy and Conceptual Relations in Synchrony and Diachrony

The term *polysemy* was first used by Bréal (1897, 154), who explicitly introduced the notion as a direct consequence of semantic change. Indeed, recent approaches to semantic change (e.g. Traugott and Dasher 2002) emphasize that the development of an initially secondary polysemous reading is the first stage of a complete semantic change to take place. Given that for semantic change a certain 'association [...] between the old meaning and the new [...] might be regarded as a necessary condition' (Ullmann, 1972, 211), the same kind of "association" can also be assumed to hold for polysemy. There are recent approaches in historical semantics that explicitly start from polysemous lexical items to extract information about what kinds of concepts are associated and what kinds of semantic change seem to be plausible (Croft et al., 2009). This procedure has the great advantage of providing 'an important antidote to the unbridled imagination in postulating etymologies' (Evans and Wilkins, 2000, 550), where intuitive assessments regarding plausibility as far as semantics is concerned are still frequent. Our approach operates in a similar vein.

3 Modeling Cross-Linguistic Polysemies as Weighted Networks

The idea to model polysemies as networks itself is not new. It was already underlying Haspelmath's (2003) *semantic map* approach which is used as a heuristic tool to analyze grammatical categories in linguistic typology. François (2008) applied this approach to the lexical domain, followed by further work by Croft et al. (2009), Perrin (2010), Cysouw (2010a, 2010b), and Steiner et al. (2011), who also introduced a simplified procedure to retrieve putative polysemies from semantically aligned word lists. What is new in our approach, however, is the strict modeling of cross-linguistic polysemies as *weighted* networks that shall be briefly introduced in the following.

3.1 Reconstruction

Networks (or graphs) constitute a system representation tool which is used by many different disciplines (Newman, 2004). We make use of a weighted network model to display and analyze conceptual relations reconstructed on the basis of cross-linguistic polysemies. Our network has the following structure:

> Let C be a set of n concepts $c_1, \ldots, c_n$ whose linguistic representation we want to analyze on the basis of a set L of m different languages belonging to $o \leq m$ language families. Our network is an undirected weighted graph $G = (V, E, f)$, with $V = C$ and $E \subseteq \{e_{ij} := \{c_i, c_j\} \mid c_i, c_j \in C \text{ and } i \neq j\}$. f is a mapping from E into $\mathbb{N}$ with $f(e_{ij}) \leq o$ being the number of language families which use one word for both c_i and c_j, and $f(E) = \{f(e_{ij}) \mid e_{ij} \in E\}$.[2]

In less formal terms, we reconstruct a weighted network by representing all concepts in a given multilingual word list as nodes (vertices), and draw edges between all nodes that show up as polysemies in the data. The edge weights reflect the number of language families in which these polysemies are attested.

3.2 Analysis

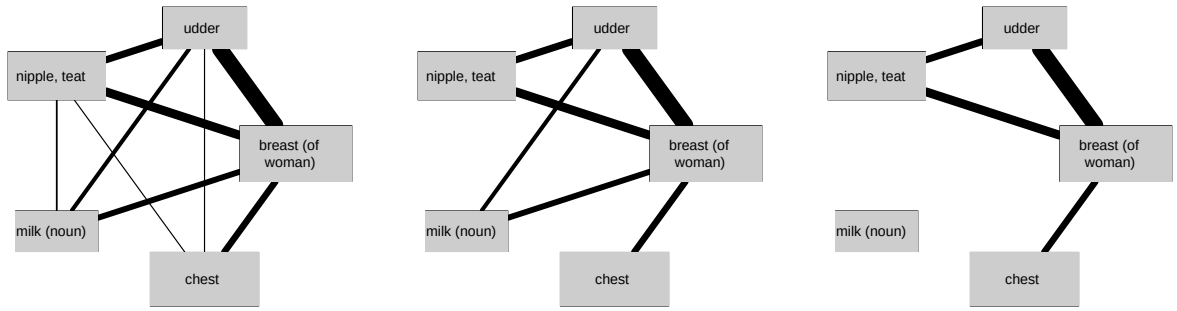
Statistical accounts on cross-linguistic polysemies retrieved from semantically aligned word lists make it possible to define the similarity between concepts on an item-to-item basis. Here, problems may arise from the fact that our approach cannot make an a priori distinction between polysemy and semantic vagueness on the one hand (Geeraerts, 1993), and polysemy and homophony on the other (compare François' 2008 notion of *colexification*). While, as Haspelmath (2003, 231) and François (2008, 169f) argue, the distinction between (true) polysemy and semantic vagueness does not matter from a cross-linguistic perspective (and we will make no attempt to distinguish the two), the failure to single out homophones can lead to wrong assessments regarding concept relations. In German, Dutch, and Yiddish, for example, the stems expressing the concepts "arm" and "poor" are identical in form. If one naively counted the number

[2]Many thanks to Daniel Schulzek for helpful comments on the maths.

of languages where the concepts are expressed by the same word, the link would appear to find much more support in the languages of the world than that between "udder" and "chest", which is only reflected in two languages in our data, namely in Aymara and Sirionó (see Supplemental Material), although in this latter case, the connection between the two concepts seems much more meaningful than in the former one.

A first way to solve this problem is to count occurrences of links between concepts not on the basis of languages but of language families, thus avoiding that they result from genetic inheritance. This would resolve the problem in favor of "udder" and "chest", since the link between "arm" and "poor" only occurs in the closely related Germanic languages, whereas Aymara and Sirionó belong to different language families. The problem can further be addressed by setting up a threshold of occurrences and to ignore links that exhibit less occurrences. Such an analysis is illustrated in Figure 1, where, starting from a given concept network, the threshold is successively increased and more and more edges are successively removed. While such an analysis surely yields the most reliable links, it has the drawback of ignoring many links that might reflect true – although less frequently attested – conceptual relations.

In order to solve the problem of separating the wheat from the chaff without losing too much wheat, the network perspective as opposed to the item-to-item perspective can be of great help. For the kind of networks we are dealing with in this study, an analysis of *community structure* seems to be specifically useful. Community structure refers to the property of many networks that do not consist of random collections of nodes with random connections between them but of 'distinct "communities" – groups of vertices within which the connections are dense but between which they are sparser' (Newman, 2004, 4). It is straightforward to assume that a network reflecting relations between concepts should exhibit some kind of community structure, given that it is often assumed that concepts can be grouped into specific conceptual fields. Analyzing the community structure of cross-linguistic polysemy networks should therefore give us some interesting insights into general concept relations.



(a) no cutoff (b) cutoff < 5 (c) cutoff < 10

Figure 1: Setting up a Threshold of Occurrences.

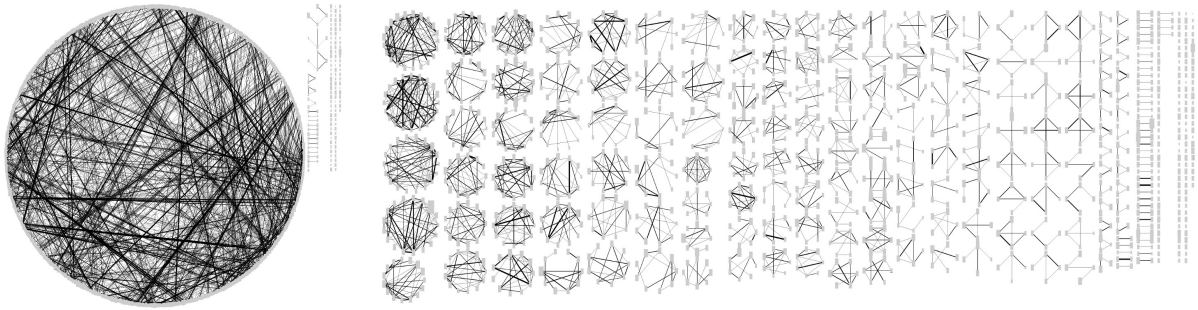
4 Application

4.1 Data

Our analysis is based on a large multilingual word list consisting of 1252 glosses ("concepts") translated into 195 different languages, covering 44 different language families (see Supplemental Material). The data was taken from three different sources, namely the Intercontinental Dictionary Series (IDS, Key and Comrie 2007, 133 languages), the World Loanword Database (WOLD, Haspelmath and Tadmor 2009, 30 languages), and a multilingual dictionary provided by the Logos Group (Logos, Logos Group 2008, 32 languages). The data from each of these sources was automatically cleaned and normalized with help of Python scripts. While the original sources of IDS and WOLD have a total of 1310 glosses, we selected only those glosses which were at least reflected in 100 languages. The structure of the input data is illustrated in Figure 2, with two instances of polysemies in Russian and German marked in bold font.

Key	Concept	Russian	German	...
1.1	world	mir, svet	Welt	...
1.21	earth, land	zemlja	**Erde**, Land	...
1.212	ground, soil	počva	**Erde**, Boden	...
1.420	tree	**derevo**	Baum	...
1.430	wood	**derevo**	Wald	...
...	...	...	...	...

Figure 2: Structure of the Data



(a) Complete Network (b) Community Structure

Figure 3: Comparing the Network and its Community Structure.

4.2 Analysis

We created a weighted network from the data, using the number of language families in which a particular polysemy was reflected as edge weights. We further analyzed the community structure of the data with help of a weighted version (Newman, 2004) of the Girvan-Newman algorithm for community detection (Girvan and Newman, 2002).[3] This algorithm successively removes the edges with the highest *betweenness* from a given network. Edge betweenness is defined as 'the number of shortest paths between pairs of vertices that run along it' (Girvan and Newman, 2002, 7822). We followed Newman (2004) in using *modularity*, i.e. the 'fraction of edges that fall within communities minus the expected value of the same quantity if edges are assigned at random' (Newman, 2004, 6), as a criterion to find the best split for the analysis.

4.3 Results

In Figure 3 the original network (a) and the network's community structure (b) are contrasted. The original network consists of one very large connected component of 1141 nodes and only a spurious number of unconnected nodes. The analysis of the network with help of the Girvan-Newman algorithm yielded a total of 337 communities of which 104 are rather large, consisting of 5 and more nodes, and covering a majority of the concepts (879 out of 1289, 68%). Most of these large communities constitute meaningful groupings of concepts into conceptual fields. Community 5, for example, groups concepts that deal with the cover of bodies ("feather", "hair", "bark", etc.). Community 28 deals with learning ("study", "count", "try, attempt", etc.). And community 70 contains concepts related to transport vehicles ("canoe", "boat", and "carriage, wagon, cart", etc., see Supplemental Material).

Apart from the general question of which items are grouped together in one community, it is also interesting to investigate the internal structure of the communities more closely. Community 3, for example, consists of 18 items which all center around the concepts "tree" and "wood" (see the network representation using force-directed layout in Figure 4). Cross-linguistically, the conflation of "tree" and

[3] We are well aware of the fact that there are many other, supposedly better, algorithms for community detection. However, given that this is a pilot study, we decided to take a rather simple algorithm whose basic ideas are nicely described and easy to understand, thus limiting our presuppositions about conceptual structures to a minimum.

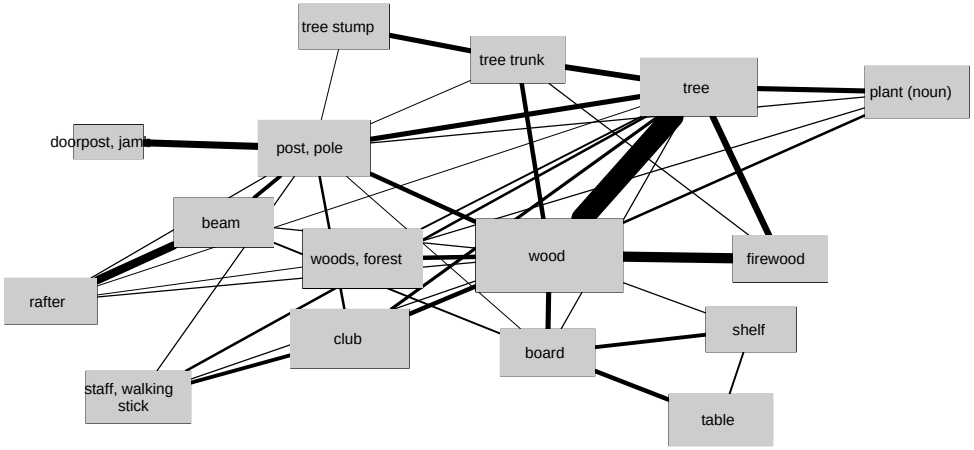


Figure 4: Force-Directed Representation of Cluster 3.

"wood" into one lexical form occurs very frequently, and it has been identified previously by several scholars (Hjelmslev, 1963; Witkowski et al., 1981). Our data is congruent with the traditional literature in this respect. However, in our network, these central concepts are connected with a substantial number of contiguously related ones. Thus, "wood" has something to do with "clubs", "staffs", "walking sticks", and the like, in that these all are artifacts "made from wood", and "trees" (and "wood") have also a contiguous relation with "forests" in so far as forests are, simply spoken, "agglomerations of trees". Unlike the well-documented association between "wood" and "tree", to our knowledge most of these further semantic connections have not yet been documented explicitly from a cross-linguistic point of view. Hence, as even this single example shows, our analysis contributes to enhancing our knowledge regarding cross-linguistically common conceptual associations. From the perspective of historical semantics, as Witkowski et al. (1981) argue, polysemies between "tree" and "wood" can be interpreted in diachronic terms, with terms for "wood" frequently giving rise to terms for "tree". But there is a diachronic correlate also for other concepts in this cluster. For example, the Proto-Indo-European root **dóru-* has a direct descendant in Russian *dérevo* "tree, wood". In other languages, however, reflexes are used to denote all kinds of weapons that were originally made from wood, such as Avestan *dāuru* "club" (but also "tree trunk"), or Modern Greek δόρυ "spear", which regularly meant "wood, tree" (but also "beam, pole") in Old Greek (Nussbaum 1968, 147, footnote).

5 Conclusion

Polysemies offer powerful evidence to study conceptual relations and semantic change. In this paper, we tried to show how the analysis of such data can be enhanced with the help of weighted network approaches. By applying them to a large dataset, we illustrated the potential of cross-linguistic polysemy data for semantic analyses. The analysis of the community structure of our network not only reproduces findings from the literature, but also reveals additional cross-linguistic regularities in the conceptual structures underlying semantic change.

Supplemental Material

The supplemental material accompanying this study contains the word lists of all 195 languages, upon which the reconstruction of the weighted network was based, and all essential results in form of text files, including the network, a detailed list and description of all inferred communities, and additional statistics regarding the languages.

References

Bréal, M. (1897). *Essai de sémantique*. Paris: Hachette.

Croft, W., C. Beckner, L. Sutton, T. Wilkins, J.and Bhattacharya, and D. Hruschka (2009). *Quantifying semantic shift for reconstructing language families*. Talk, held at the 83rd Annual Meeting of the Linguistic Society of America. PDF: `http://www.unm.edu/~wcroft/Papers/Polysemy-LSA-HO.pdf`.

Cysouw, M. (2010a). Drawing networks from recurrent polysemies. *Linguistic Discovery 8*(1), 281–285.

Cysouw, M. (2010b). Semantic maps as metrics on meaning. *Linguistic Discovery 8*(1), 70–95.

Evans, N. and D. Wilkins (2000). In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language 76*(3), 546–592.

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.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics 4*(3), 223–272.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 99*(12), 7821–7826.

Haspelmath, M. (2003). The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language*, pp. 211–242. Mahwah, NJ: Lawrence Erlbaum.

Haspelmath, M. and U. Tadmor (2009). *World Loanword Database*. Munich: Max Planck Digital Library.

Hjelmslev, L. (1963). *Prolegomena to a theory of language*. Madison: University of Wisconsin Press.

Key, M. R. and B. Comrie (2007). *IDS – The Intercontinental Dictionary Series*. URL: `http://lingweb.eva.mpg.de/ids/`.

Logos Group (2008). *Logos Dictionary*. URL: `http://www.logosdictionary.org/index.php`.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E 70*(5), 056131.

Nussbaum, A. J. (1968). *Head and horn in Indo-European. The words for "horn," "head," and "hornet"*. Berlin and New York: de Gruyter.

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

Steiner, L., P. F. Stadler, and M. Cysouw (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change 1*(1), 89–127.

Traugott, E. C. and R. B. Dasher (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Ullmann, S. (1972). *Semantics*. Blackwell.

Witkowski, S. R., C. H. Brown, and P. K. Chase (1981). Where do tree terms come from? *Man 16*(1), 1–14.