Evaluating the Potential of a Large-Scale Polysemy Network as a Model of Plausible Semantic Shifts

Alla Münch
Seminar für Sprachwissenschaft
University of Tübingen

Johannes Dellert
Seminar für Sprachwissenschaft
University of Tübingen

***Abstract*—We present a very large network of cross-linguistic polysemies, and compare the notion of semantic relatedness it encodes to the catalogue of semantic shifts maintained by the Russian Academy of Sciences. We separately evaluate all types of semantic shifts featured in the catalogue, including shifts occurring during semantic evolution, during borrowing, and during morphological derivation. The comparison shows that over one third of the attested semantic shifts take place between close neighbors in the network. This can be considered strong evidence for the usefulness of polysemy networks in modelling most types of lexical change, making them a valuable resource e.g. for semantic reconstruction or future automatization of cognate detection. We also show that the semantic shifts which occur during morphological derivation form a divergent class, and might need to be modelled separately.**

I. Introduction

In etymological research, cognacy judgements rely heavily on expert knowledge about the plausibility of semantic shifts. Such judgements are often based on evidence in the form of parallel developments in other languages, including established cases of diachronic semantic shifts. Collecting large amounts of data on attested shifts has therefore become an important prerequisite for computational historical linguistics.

Existing collections of semantic shifts [1]–[4] have the format of a dictionary or a computer database covering only a small part of conceptual space. At the same time, crosslinguistic studies of synchronous polysemy patterns have become an important focus of attention in lexical semantics [5]–[11].

Building on this research, recent work in computational historical linguistics [12] and lexical typology [13] has suggested to make use of polysemy patterns extracted from co-occurrences of gloss lexemes in multilingual wordlists. A **polysemy network** is a graph over concepts (expressed by lemmas), where each link represents the fact that at least one language has a lexical item which can be translated by both lemmas. In a slight abuse of the terminology introduced by [13], we will call such pairs of lemmas **colexified**, although the original definition of the term refers to language-independent concepts instead of lemmas in a given gloss language.

The synchronic colexification patterns encoded in polysemy networks have been proposed as potential sources of evidence for plausible semantic shifts [14], based on the intuition that every instance of semantic shift needs to pass an intermediate stage where the word in question is polysemous. To assess the validity of this idea, the predictive potential of such polysemy networks remains to be tested by comparing them to other lexical resources. Building on promising preliminary results from a pilot study [15] of cognates for a single language pair (Finnish and Hungarian), we evaluate the very large polysemy network developed in Tübingen within the framework of the EVOLAEMP project [15] against the catalogue of cross-linguistic semantic shifts developed by the Russian Academy of Sciences [16], [17].

II. Resources

The EVOLAEMP project maintains a large German-based dictionary database which contains more than 750,000 entries in 114 languages. The database features near-complete coverage of a 1,000-item list of basic concepts for about 80 languages of Northern Eurasia, and contains more than 10,000 translation pairs for 31 languages from 10 primary language families, including Indo-European, Uralic, Turkic, Afroasiatic, Dravidian, Sino-Tibetan, Japonic, Ainu, Tungusic, and Yukaghir. We formalize the database as a set of dictionaries, and define the dictionary D_L for a language L as set 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.

Our polysemy network is an undirected graph $G = \langle V, E \rangle$ over a set of German glosses V. In order to avoid the problem that chance homophones would be interpreted as polysemies and create spurious links (e.g. the famous Persian *šir* "milk; lion"), we partition the languages into a set of genetic units $\mathcal{F}$, where one unit $F \in \mathcal{F}$ roughly corresponds to a genus or subfamily such as the Romance, Semitic, or Fennic languages. An edge $\{g_1, g_2\}$ between two German glosses $g_1, g_2 \in V$ is then 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:

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

The resulting very large polysemy network contains 32,653 glosses connected by 47,647 links. Despite a strong bias in favour of Indo-European and Uralic languages, seventeen other language families (Turkic, Mongolic, Afroasiatic, Ainu, Sino-Tibetan, Northeast Caucasian, Japonic, Vasconic, Austronesian, Dravidian, Chukotko-Kamchatkan, Tungusic, Kartvelian, Tai-Kadai, Nivkh, Yeniseian, and Yukaghir) are represented by more than 5,000 links, leading to a reasonable amount of cross-linguistic diversity.

The network shows some interesting and perhaps surprising structural properties. Whereas 14,391 nodes form unconnected islands without any attested polysemy, there is a central connected component consisting of 13,073 nodes, which means that it is possible to find paths between any pair of concepts for about 40% of the concepts in the network. If we interpret the network as a model of semantic change, this implies that over time, words can change their meaning freely within almost half the conceptual space. This structural property of the network mirrors the common impression that the meanings of words can change almost arbitrarily.

The remaining nodes are distributed over 1,940 smaller components, none of which is remotely as large as the central cluster:

size	2	3	4	5	6-10	11-15	35
count	1,325	325	143	67	68	11	1

Taking a look at the number of attestations for each link, we see that almost half of them are only attested in the minimal number of two genetic units, but more than a fifth of them are attested in five or more different genetic units, making them typologically relevant:

gen. units	2	3	4	5-7	8-12	>12
%	48.3	18.9	10.0	13.2	6.5	3.1

Concerning density, each node in our network has 2.4 neighbors on average, with the following distribution:

neigh.	0	1	2	3	4-5	6-10	>10
%	21.2	38.2	13.8	7.7	8.1	7.1	3.8

Finally, we take a look at the distribution of minimal path lengths between pairs of connected concepts. The typical shortest path is of length 8, and both very long paths and very short paths are rare. Only about 20% of the node pairs which are connected at all can be reached by paths shorter than seven steps.

length	1-3	4-6	7-9	10-12	> 12
%	0.87	19.14	54.01	23.15	2.82

The second resource we are using, the Catalogue of Semantic Shifts, is an established resource which is described in much detail elsewhere [17]. All the shifts included in the catalogue are cross-linguistically recurring and were collected manually by experts, making the catalogue a reliable resource for semantic change. However, the authors' definition of semantic shifts is very liberal, as it also encompasses synchronic polysemies (these form the bulk of the catalogue) and semantic change during derivation in addition to historically attested semantic shifts. The catalogue has no restrictions in the range of meanings involved. At the time of publication, it contained 3,650 attested semantic shifts of six types, each of them supported by up to 40 realizations. The catalogue has continued to grow as additional instances were collected and made available on the web. In total, our version of the database contains 6,174 realizations from 319 languages, again mainly belonging to language families of Eurasia. Especially well-represented are Afroasiatic, Altaic, Austro-Asiatic, Northwest Caucasian, Northeast Caucasian, Indo-European, Sino-Tibetan, and Uralic.

III. Results and discussion

To make the two resources comparable, some preprocessing was necessary. The English version of the catalogue was extracted from its website and normalized by considerable semi-automated cleanup work.

The metalanguages of the catalogue are English and Russian, whereas the primary language of the network is German. The implicit assumption that each German lemma corresponds to exactly one concept is obviously false, but to some extent unavoidable if we build on translations from dictionaries and wordlists. The network thus exhibits some German-specific structure, especially idiosyncratic connections caused by polysemy or homophony of German lemmas. For instance, the network cannot reliably differentiate the concepts of "train" and "move (in a game)", because the German noun *Zug* is used for both concepts. As any reader with knowledge of German will be able to assess based on the examples, these problems are not extremely disturbing in practice, because they usually only falsely contract concepts into a single node which would otherwise still be connected at least due to polysemy in German. Still, the identification of German lemmas and concepts should be kept in mind when considering other possible use cases of our polysemy network.

The different languages of the resources made it necessary to use an electronic English-German dictionary as an intermediary for comparing the links in both resources. We decided to keep our approach simple and reproducible by allowing any attested English-German translation pair to be used as a bridge between the two resources, staying in line with our pragmatic approach to the concept-lemma relation. Formally, we define the set of German glosses equivalent to an English gloss l as $glo(l) := \bigcup_{\langle l,T\rangle \in D_{eng}} T$.

Given this simple mapping between English and German glosses, the shift pairs were then evaluated against the polysemy network by determining the length of the shortest paths connecting any pair of equivalent German glosses. Formally, the method can be described most easily by defining k-hop accessibility relations G_k:

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

For each shift pair $\langle l_1, l_2\rangle$ in the English version of the catalogue, we then compute $\min_k\{k : \exists\{g_1, g_2\} \in G_k : g_1 \in glo(l_1) \wedge g_2 \in glo(l_2)\}$.

In 43.1% of instances in the catalogue, no German translation could be found for the source and/or target concept ($glo(l_1) = \emptyset \vee glo(l_2) = \emptyset$). The relatively high number of such cases can be attributed primarily to the use of highly specialized or culture-specific concepts, the presence of hypernyms, and multi-word metalinguistic descriptions. For instance, the hypernym *[foreigner]* in the catalogue is supposed to stand for any nationality, while the hypernym itself does not necessarily stand in any relation to the target. As the polysemy network only includes individual glosses, there is no way to infer all the possible hyponyms and establish their relation to the target. For the same reason, most multi-word descriptions such as *fantastic monster* could not be matched to any node in the network either. By and large, the high percentage of such cases seems to be unavoidable due to the different approaches, scopes and aims of the datasets we are comparing.

In what follows, we only consider the results obtained for the 56.9% of cases where both the source and the target concept could be translated to glosses found in the polysemy network, which was the case for 3,513 shift pairs in total.

These results are presented in Table I in terms of recall, i.e. as percentages expressing the ratio of shifts in the catalogue covered by paths in the network. We separately consider the cases where no path between the concepts was found, the cases where there was a minimal path of length 1 or 2, and the cases where the minimal path was longer. This division is based on our previous study on the same network [15], which experimentally established that a path length of 1 or 2 roughly corresponds to the cases where the number of neighbors in the network is still low enough to make spuriously similar forms unlikely. For comparing the performance on different types of semantic shifts, we largely follow the terminology adopted in the catalogue. The total number of instances of each particular shift is provided in the first column for reference.

TABLE I. SHIFTS IN THE CATALOGUE COVERED BY THE POLYSEMY NETWORK.

Semantic shift type	Number of instances	No path	Path length 1 or 2	Path length 3 and more
Polysemy	2315	20.6 %	35.0 %	44.4 %
Semantic Evolution	107	26.2 %	33.6 %	40.2 %
Morphological Derivation	597	28.5 %	29.0 %	42.5 %
Syncretism	43	25.6 %	55.8 %	18.6 %
Borrowing	58	31.0 %	41.4 %	27.6 %
Cognates	393	23.7 %	37.9 %	38.4 %

In general, across all types of semantic shifts present in the catalogue, over one third of cases (34.6%) corresponds to short paths in the polysemy network, which we consider a surprisingly high recall given the notorious unpredictability of semantic change.

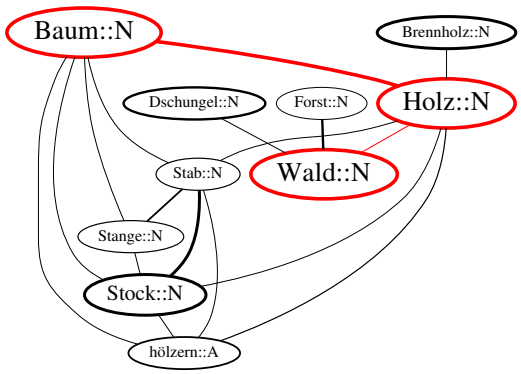

Fig. 1. Surroundings of the path from "tree" to "forest".

Going through the types of semantic shifts in the catalogue, we start by considering the results obtained for **Polysemy**, i.e. synchronic polysemies of the same nature as the colexifications the network was built on. A typical example of such a concept pair is given in Example 1. In this and all following examples, ZAL (Zalizniak) marks the concept pair in the catalogue, and TUE (Tübingen) one of the corresponding shortest paths in the polysemy network, with rough English equivalents given.

Example 1. (Polysemy)
ZAL: *tree – forest*
TUE: *Baum* "tree" – *Holz* "wood" – *Wald* "forest"

The path found for this example is also visualized in red in Figure 1, a graphical representation of the region around the path in the polysemy network. All larger nodes which are directly connected to one of the nodes on the path are included. In all the visualizations in this paper, line thickness denotes strength of colexification, and node size represents the amount of data we have for the German gloss in question. The thickness of node borders represents how often the gloss in question occurred as the only gloss in a dictionary definition, giving some indication of how basic and well-delineated the concept represented by that German gloss is.

Of the concept pairs in this category, 35.0% are connected by paths of length 1 or 2 in the network. In view of the fact that the two databases were developed independently, and are based on different sources and slightly different language samples, the overlap is surprisingly large, suggesting the existence of a large core of common polysemies which occur with high frequency across language families.

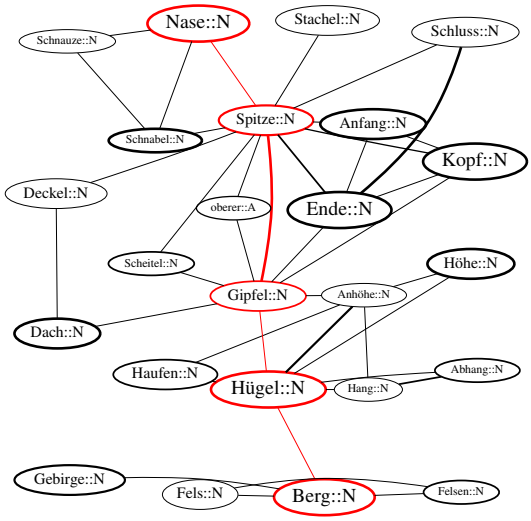

Fig. 2. Surroundings of the path from "mountain" to "nose".

Given the recall for synchronic polysemies, the main question is whether the polysemy network models other types of semantic shifts just as well. The most interesting of the other shift types in the catalogue is **Diachronic Semantic Evolution**. This category only contains attested instances of semantic shifts in the stricter sense, i.e. changes in the meaning of a word occurring between historical stages of a single language. Examples 2 and 3 are typical instances of attested shifts from the catalogue. The very interesting shortest path for Example 3 is visualized in Figure 2.

Example 2. (Diachronic Semantic Evolution)
ZAL: *property/possessions* → *cattle*
TUE: *Besitz* "property" – *Vieh* "cattle"

Example 3. (Diachronic Semantic Evolution)
ZAL: *mountain* → *nose*
TUE: *Berg* "mountain" – *Hügel* "hill" – *Gipfel* "summit" – *Spitze* "tip" – *Nase* "nose"

Remarkably, recall for semantic shifts in this stricter sense (33.6%) is very close to that of synchronic polysemy. This allows us to conclude that polysemy networks do indeed predict possible pathways of diachronic evolution just as well as synchronic polysemies in unseen data, and that the semantic processes underlying synchronic polysemies and diachronic shifts are not measurably different.

The very small **Syncretism** category turned out to contain the largest percentage (55.8%) of cases where

the concepts are immediate neighbors in the polysemy network. The reason seems to be that many of the few instances of syncretism in the catalogue are rather frequent cross-linguistically, and also appear in the polysemy network because dictionary sources do not systematically distinguish true polysemy from semantic generality.

Example 4. (Syncretism)
ZAL: *caterpillar – snake*
TUE: *Raupe* "caterpillar" – *Wurm* "worm" – *Schlange* "snake"

In Example 4, dictionaries for languages which show this instance of syncretism will normally just mention "caterpillar" and "worm" as possible translations, instead of mentioning that the word in question actually denotes a larger class of animals which contains caterpillars and worms, and cannot readily be described by any English gloss. The high recall thus indicates that the pragmatic approach usually taken by lexicographers causes instances of syncretism to be modeled very well by polysemy networks derived from dictionary data.

The shift types **Cognates** and **Borrowing** represent instances of semantic shifts that occurred during long-term separate development, or in contact situations. The shift in Example 5 is contained in the catalogue both as an instance of Cognates (for Old English and German) and an instance of Borrowing (for Old East Slavic and Komi):

Example 5. (Borrowing, Cognates)
ZAL: *boy – servant*
TUE: *Junge* "boy" – *Kellner* "waiter" – *Diener* "servant"

Our network covers 41.4% of Borrowing instances and 37.9% of Cognates instances, so in both cases recall is even slightly higher than for synchronic polysemy or semantic evolution. Thus, there seems to be no relevant difference between the semantic shifts occurring within a single language and during transfer between languages. Semantic shifts caused by internal language mechanisms and by external factors are captured equally well.

The category of **Morphological Derivation** is of special interest because in the terminology of polysemy networks, it contains prototypical instances of *loose colexification* [18], where the notion of polysemy is extended to also connecting the meanings of non-trivial derivates from the same stem. Since the polysemy network is built on strict colexifications only, we would not expect such shifts to be modeled well by the network.

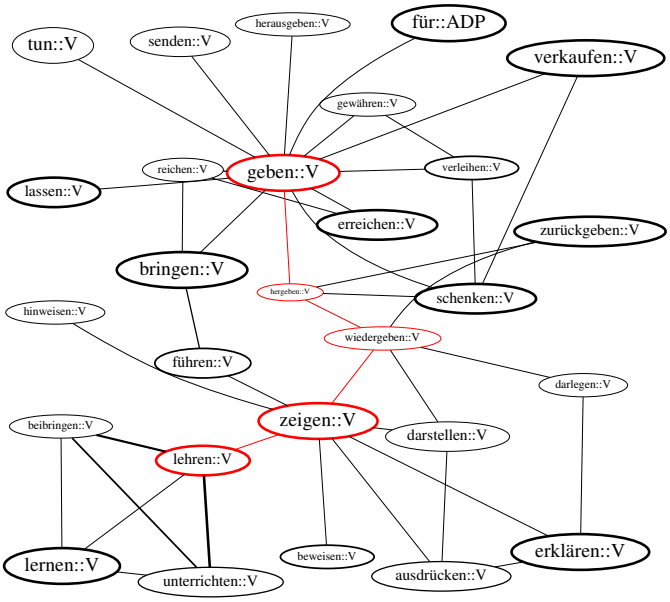

Fig. 3. Surroundings of the path from "give" to "teach".

Indeed, Morphological Derivation has the worst recall among all classes of semantic shifts, with only 29.0% of instances connected by paths of length 1 or 2 in the network. Example 6 demonstrates this for the concept "teaching", which is expressed by a word derived from a word for "to give" in some languages. The shortest paths connecting the German equivalents are of length 4, and one of them is visualized in Figure 3.

Example 6. (Morphological Derivation)
ZAL: *to give – to teach*
TUE: *geben* "give" – *hergeben* "hand over" – *wiedergeben* "render" – *zeigen* "show" – *lehren* "teach"

This result supports the intuition that the shifts attainable by derivation differ in a substantial way from the shifts attainable by plausible sequences of shifts along paths defined by strict colexification. An interesting avenue for further research is to see whether a polysemy network built on loose colexification will provide a better model of this type of semantic change.

To make our results as transparent as possible, and to ensure reproducibility, a file containing the shortest paths we found for all the shift pairs in the catalogue is publicly available on the second author's webpage. Moreover, we make the entire polysemy network available on request, allowing other researchers to use it for quantitative studies in the areas of lexical semantics and typology.

IV. Conclusions

Our evaluation of the largest available polysemy network against the largest available database of attested semantic shifts has yielded promising results. For about one third of concept pairs covered by both resources, the polysemy network contained a path of length of 1 or 2 between equivalent lemmas, making these pairs accessible e.g. for automated cognate finding methods. For about a quarter of pairs, no path between equivalent lemmas existed. For the remaining concept pairs, the shortest path found in the polysemy network was of length 3 or more, arguably indicating shifts which are too erratic to be modeled by the neighborhood relation of a polysemy network.

Overall, our results confirm the expected high potential of polysemy networks as resources for modelling the plausibility of semantic shifts.

While investigating the network's coverage of different classes of semantic shifts, we found that both diachronic processes within a language (semantic evolution) and across languages (cognates and borrowings) are modeled just as well as synchronic polysemies. In contrast, shifts occurring during morphological derivation are less frequently connected by short paths, providing evidence of a different underlying change process which is not modeled equally well by synchronic polysemy.

Moreover, our comparison has yielded an overlap of 35% between two independently developed databases of cross-linguistically recurring polysemies, suggesting that there is a common core of frequent polysemies which can in fact be used as a solid foundation in computational models of semantic evolution for historical linguistics.

V. Acknowledgements

This research was supported by the European Research Council Advanced Grant 324246 (EVOLAEMP), which is gratefully acknowledged.

References

[1] B. Heine and T. Kuteva, *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press, 2002.

[2] S. Sakhno, *Dictionnaire Russe-Français d'éthymologie comparée: Correspondances lexicales historiques*. Paris: Editions L'Harmattan, 2001.

[3] P. Gévaudan, P. Koch, and A. Neu, "Hundert Jahre nach Zauner: Die romanischen Namen der Körperteile im DECOLAR," *Romanische Forschungen*, vol. 115, no. 1, pp. 1–27, 2003.

[4] P. Gévaudan, *Typologie des lexikalischen Wandels: Bedeutungswandel, Wortbildung und Entlehnung am Beispiel der romanischen Sprachen ; mit einer Zusammenfassung in französischer Sprache*, ser. Stauffenburg Linguistik. Tübingen: Stauffenburg Verlag, 2007.

[5] E. Sweetser, *From etymology to pragmatics*, ser. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press, 1990, vol. 54.

[6] A. Blank, "Kognitive Linguistik und Bedeutungswandel," in *Interdisziplinarität und Methodenpluralismus in der Semantikforschung*, I. Pohl, Ed. Frankfurt: Peter Lang, 1999, pp. 125–147.

[7] A. Blank and P. Koch, *Historical Semantics and Cognition*. Mouton De Gruyter, 1999.

[8] E. C. Traugott and R. B. Dasher, *Regularities in semantic change*. Cambridge: Cambridge University Press, 2002.

[9] P. Koch, *Lexical Typology from a Cognitive and Linguistic Point of View*. Walter de Gruyter, 2001.

[10] ——, "Diachronic onomasiology and semantic reconstruction," in *Lexical data and universals of semantic change*, W. Mihatsch and R. Steinberg, Eds. Tübingen: Stauffenburg Verlag, 2004, pp. 79–106.

[11] ——, "Cognitive onomasiology and lexical change," in *From polysemy to semantic change: Towards a typology of lexical semantic associations*, M. Vanhove, Ed. Amsterdam and Philadelphia: John Benjamins, 2008, pp. 107–137.

[12] W. Croft, C. Beckner, L. Suttan, J. Wilkins, T. Bhattacharya, and D. Hruschka, "Quantifying semantic shift for reconstructing language families," 2009, talk, held at the 83rd Annual Meeting of the Linguistic Society of America. [Online]. Available: http://www.unm.edu/w̃croft/Papers/Polysemy-LSA-HO.pdf

[13] L.-M. Perrin, "Polysemous Qualities and Universal Networks, Invariance and Diversity," *Linguistic Discovery*, vol. 8, pp. 1–22, 2010.

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

[15] J. Dellert, "Evaluating cross-linguistic polysemies as a model of semantic change for cognate finding," 2014, proceedings of the Workshop on semantic technologies for research in the humanities and social sciences (STRiX). November 24-25, Gothenburg, Sweden.

[16] A. A. Zalizniak, "A catalogue of semantic shifts: Towards a typology of semantic derivation," in *From polysemy to semantic change. Towards a typology of lexical semantic associations*, M. Vanhove, Ed. Amsterdam and Philadelphia: John Benjamins, 2008, pp. 217–232.

[17] A. A. Zalizniak, M. Bulakh, D. Ganenkov, I. Gruntov, T. Maisak, and M. Russo, "The catalogue of semantic shifts as a database for lexical semantic typology," *Linguistics*, vol. 50, no. 3, pp. 633Ł–Ł669, 2012.

[18] A. François, "Semantic maps and the typology of colexification: Intertwining polysemous networks across languages," in *From polysemy to semantic change. Towards a typology of lexical semantic associations*, M. Vanhove, Ed. Amsterdam and Philadelphia: John Benjamins, 2008, pp. 163–216.