An Interactive Visualization of Cross-Linguistic Polysemies

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

Abstract

In this paper, we present an interactive web-based visualization for the CLiCS database, an online resource for synchronic lexical associations (*polysemies*, or more precisely, *colexifications*) in over 200 language varieties. The associations cover 1285 concepts and represent the tendency for concepts to be expressed by the same words in the same language varieties of the world. The complexity of the network structure in the CLiCS database calls for a visualization component that makes it easier for researchers to explore the patterns of cross-linguistic colexifications. The network is represented as a force-directed graph and features a number of interactive components that allow the user to get an overview of the overall structure while at the same time providing the opportunity to look into the data in more detail. An integral part of the visualization is an interactive listing of all languages that contribute to the strength of the cross-linguistic colexifications. Each language in the list is thereby attributed a different color depending on its genealogical or areal information. In this way, given associations can be inspected for genealogical or areal bias.

Keywords: Interactive visualization, polysemies, cross-linguistic database

1. Introduction

2. CLiCS

CLiCs (***Cross-Linguistic Colexifications***, `http://lingulist.de/clics/`) is an online database of synchronic lexical associations ("polysemies", but more precisely: *colexifications*, see below) for 1285 concepts translated into currently 215 language varieties of the world. Large databases offering lexical information on the world's languages are already readily available for research in different online sources. However, the information on tendencies of meaning associations available in these databases is not easily extractable from the sources themselves. This is why CLiCs was created. It is designed to serve as a data source for work in lexical typology, diachronic semantics, and research in cognitive science that focuses on natural language semantics from the viewpoint of cross-linguistic diversity. Furthermore, CLiCs can be used to assess the plausibility of semantic connections between possible cognates in the establishment of genetic relations between languages. Table 1 gives an example on the basic structure of the data in CLiCs.

2.1. Homonymy, polysemy, and colexification

From the analytical perspective, polysemy has to be distinguished from homonymy and semantic vagueness. *Homonymy* refers to the "accidental" verbalization of at least two meanings by the same sound chain, without a conceptual relation between n and o that is more than coincidental. *Contextual variation* designates the adaption of a lexicalized meaning to contextual factors in an utterance. Although historical and synchronic criteria have been proposed to distinguish polysemy from homonymy, and contextual variation can be tested by resorting to categorization (Blank, 1997), the differentiation depends on the individual analysis of every single word and is not entirely objective. Therefore, it is difficult for quantitative investigations to provide this differentiation in advance.

In the context of CliCs, we use the term *colexification* (coined to our knowledge by François 2008) to refer to the situation when two or more of the meanings in our lexical sources are covered in a language by the same lexical item. For instance, we would say that Russian colexifies "hand" and "arm", that is, concepts that are semantically related to each other. Roughly speaking, colexification can correspond either to polysemy or contextual variation in lexical semantic analyses. Since CLiCs is not based on such analyses that would allow us to further discriminate between the two, we chose colexification as a label that deliberately does not make a commitment with regard to this distinction. However, as we will show below, quantitative approaches are available to rule out effects of accidental homonymy.

2.2. Data and sources of CLiCs

CLiCs offers information on colexification in 215 different language varieties covering 50 different language families. All language varieties in our sample comprise a total of 290,760 words covering 1,288 different concepts. Using a strictly automatic procedure, we identified 45,282 cases of colexification that correspond to 16,043 different links between the 1,288 concepts covered by our data.

Currently CLiCs utilizes three different sources, all of which are freely available online themselves. (1) The *Intercontinental Dictionary Series* (IDS, Key and Comrie 2007) features lexical data for 233 world languages. IDS data were provided mostly by experts on the respective languages, although in some cases published written sources have been used. There are 1,310 entries to be filled for each language, though, of course, there are gaps in coverage for individual languages. The list of concepts is inspired by Buck (1949). Of all 233 languages in the IDS, 178 were automatically cleaned and included in CLiCs.[1] (2) The IDS list, in turn, provides the basis for the choice of mean-

[1] In all cases, we ignored proto-languages and archaic languages (like Latin and Old Greek), and those languages which did not have enough coverage in terms of lexical items.

ings in the *World Loanword Database* (WOLD, Haspelmath and Tadmor 2009). The principal aim of this source is to provide a basis for generalizations on the borrowability of items in different parts of the lexicon. The WOLD data consist of vocabularies of between 1,000-2,000 items for 41 languages, with annotations about the borrowing history of particular items where applicable. WOLD data was coded by experts on the respective languages, in some cases also with the aid of extant sources. Of all 41 languages in WOLD, 33 languages, which were not yet present in the IDS, are included in CLiCs. (3) The *Logos Dictionary* (`http://www.logosdictionary.org`) is a freely accessible multilingual online dictionary that is regularly updated online by a network of professional translators. It offers lexical data for more than 60 different languages. We manually extracted lexical data for 4 languages that were neither present in IDS nor in WOLD.

2.3. Network modeling of CLiCs

As mentioned above, there is no guarantee that lexical associations within CLiCs are due to historical reasons or due to chance. For example, there are three attested links between the concepts "arm" and "poor" in the current version of CLiCs, which are due to homonymy in some Germanic languages (German, Dutch, and Yiddish).

In order to distinguish strong association tendencies from spuriously occurring associations and to rule out cases of accidental homonymy, List et al. (2013) model cross-linguistic colexification data as a weighted network in which nodes represent concepts and weighted edges between the nodes represent the number of attested colexifications in the data. With the help of *community detection analyses*, strongly interconnected regions in the colexification network can be identified. List et al. (2013) apply a weighted version of the community detection algorithm by Girvan and Newman (2002) to a cross-linguistic colexification network consisting of 1252 concepts translated into 195 languages covering 44 language families. Their analysis yielded a total of 337 communities, with 104 communities consisting of 5 and more nodes and covering 68% of all concepts. A qualitative survey of the largest communities showed that most of them constitute meaningful units, and accidental homologies were successfully excluded.

Concept	IDS-Key	Families	Languages
money	11.43	15	33
coin	11.44	9	13
iron	9.67	3	3
gold	9.64	2	2
tin, tinplate	9.69	2	2
white	15.64	2	2
blunt, dull	15.79	1	1
bright	15.57	1	1
chest	4.4	1	1
clock, timepiece	14.53	1	1
copper, bronze	9.66	1	1
earring	6.77	1	1
hammer	9.49	1	1
helmet	20.33	1	1
jewel	6.72	1	1
lead (noun)	9.68	1	1
price	11.87	1	1
razor	6.93	1	1
saw	9.48	1	1

Table 1: Common colexifications involving the concept "silver" in CLiCs. Concepts which are expressed by the same word form in more than two different language families are shaded gray.

2.4. Limitations and Caveats

The structure of the data in CLiCs is a direct image of the structure of the data in IDS and WOLD and does not involve a reanalysis of any sort on our behalf. However, it must be emphasized that the meaning associations reported in CLiCs are recovered from sheer identity of form in different cells in the sources we have used, and do not necessarily rest on language-internal semantic analysis. Furthermore, we have no control over artifacts that (a) may have arisen in the process of data gathering themselves, (b) were created by mapping the predefined concepts onto the actual languages, and (c) were introduced when cleaning parts of the data automatically in which the textual coding was not provided in a consistent way.

A further problem that may arise when using CLiCs is that the coverage of the world's languages in both IDS and WOLD is biased towards certain regions of the world. In the case of IDS, South American languages and languages of the Caucasus are overrepresented. In the case of WOLD, languages of Europe figure particularly prominently. Since it is possible and even expectable that certain polysemies in the lexicon are frequent or even restricted to certain areas of the world, it is important to take appropriate measures to rule out unwarranted generalizations due to areal effects.

3. Visualization

The CLiCs database is available online at `http://lingulist.de/clics/` and offers its users a search interface to all concepts and cross-linguistic polysemies between concepts. The wealth of information in the database and the various possibilities of exploring the colexifications in the network call for an additional component that makes potentially interesting observations more easily accessible to the researcher. The idea was to equip the database with a visualization component that provides various interactive functionalities and enables users to navigate through the networks of colexifications while at the same time providing more detailed information on the actual language data. A prototype of the visualization is online accessible.[2]

3.1. Web-based visualization

We opted for a web-based implementation of the CLiCs visualization in JavaScript using the D3 library (Bostock et al., 2011). The main benefits of a web-based visualization are its platform independence and the fact that users can access it from any device with a browser supporting JavaScript. There is no need for the installation of additional software or for maintenance of the system on the part

[2] `http://tinyurl.com/clicsvis`

of the user (Murray, 2010). In addition, links to the descriptions of the external resources can easily be included to allow users to explore the CLiCs data in more detail on demand.

3.2. Data Preparation

In its current form, the data in CLiCs yields a *small world network* in which all nodes are densely connected. Browsing such a full network is very confusing and provides little insights for the user (see Figure 1). In order to break down the complexity inherent in CLiCs, we decided to split the data into communities first. Starting from 1285 concepts in CLiCs which were connected to at least one other concept, we applied the *Infomap* algorithm by Rosvall and Bergstrom (2008) to cluster all concepts into communities, using the number of attested language families per colexification as edge weights. The Infomap algorithm was chosen because of its remarkable performance on the community detection task, both in terms of computation time and quality of results (Lancichinetti and Fortunato, 2009). With help of this analysis, the 1285 concepts could be subdivided into 160 communities of which 156 contain more than one node. Of all communities, 106 are *large*, containing more than five nodes. The large communities cover 88% (1125) of all nodes in the original network (1285). In order to enable the user to quickly identify communities of specific interest, we labelled all communities by taking the concept with the highest degree as a representative. Apart from one very large community of 102 concepts ("do, make"), the rest of the communities does not differ much in size, ranging from 2 ("meeting house") to 26 concepts ("get, obtain") with an average of 8 concepts per community.

3.3. Interactive functionalities

The visualization features various interactive functionalities that are designed to enhance the exploration of the CLiCs data on the level of communities. The main component is a flexible force-directed graph layout that displays the concepts as nodes and the cross-linguistic polysemies as edges (see Figure 2). The strength of the force in the edges of the graph is dependent on the number of cases that can be attested in the languages for the respective concepts that are linked through the edge. We decided to have separate graphs for all communities, which the user can select from a drop-down menu. As described above, the communities have been automatically generated from the whole network of concepts and links with the help of the Infomap algorithm for community detection (Rosvall and Bergstrom, 2008).

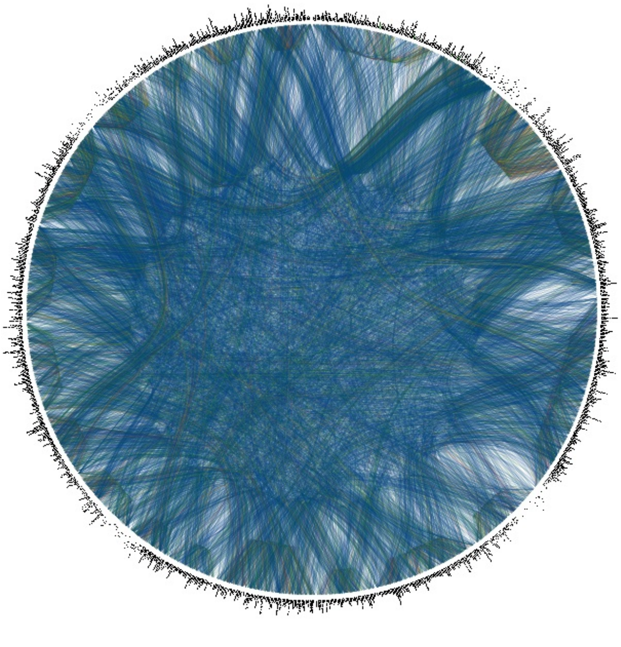


Figure 1: Full network of all connected components in CLiCs

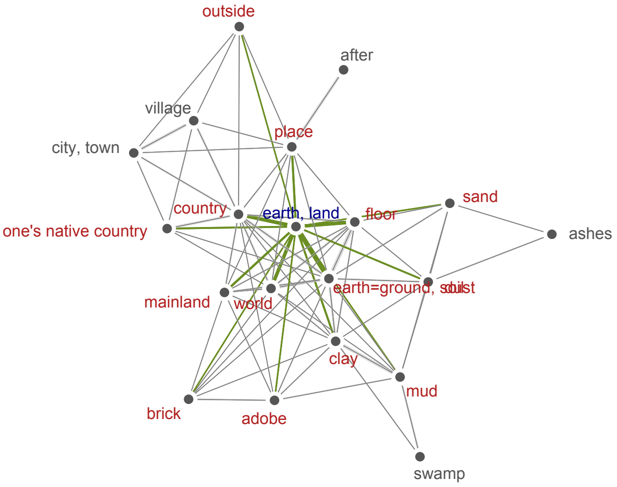
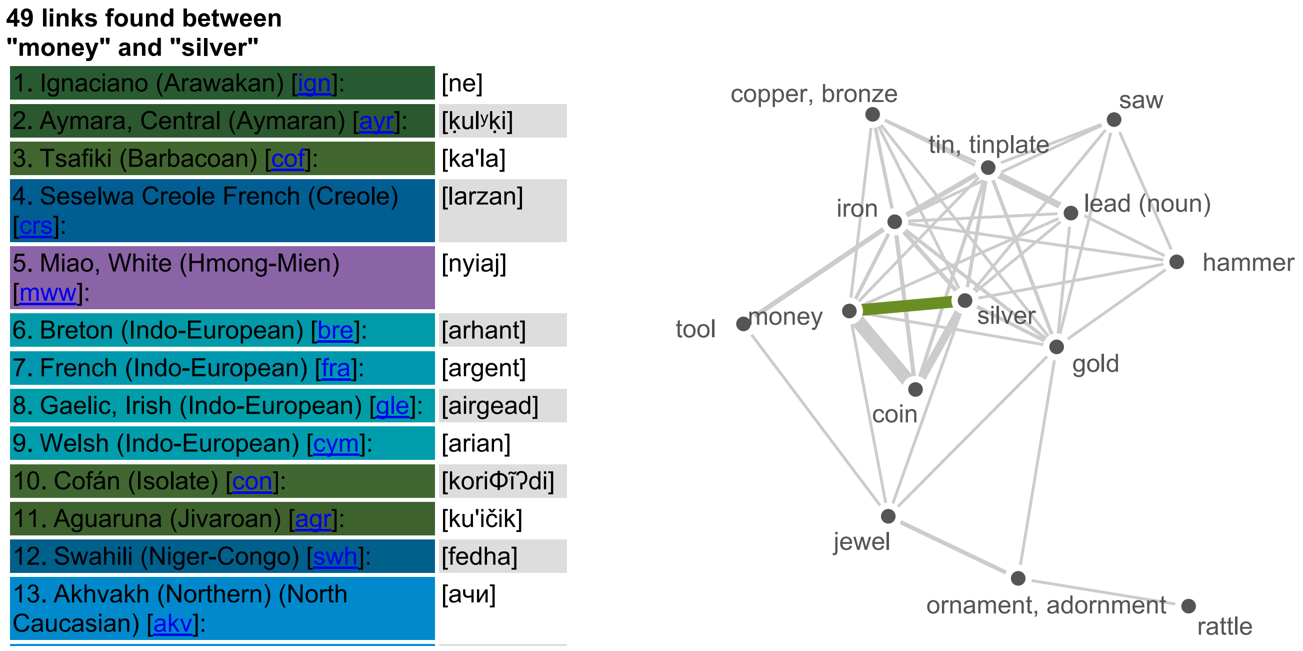


Figure 2: Force-directed graph with mouse-over functionalities highlighting all connected concepts

The force-directed graph layout ensures that all concepts are neatly arranged according to their similarity as defined by the number of cross-linguistic colexifications. As a result, concepts that are highly connected are located close to each other. To make it easier for users to explore the network that is depicted in the graph, concepts can be dragged to different positions where there is less overlap. The dragging behavior of a concept is activated when mousing over the respective node in the graph (when the cursor symbol turns into a crosshair).

As mentioned above, the edges of the graph represent the number of cases of cross-linguistic colexifications for the linked concepts. For a more detailed view on which languages contribute to the strength of the connections, the user can mouse over the links in the graph to see a list of languages featuring polysemous words for the respective link (Figure 3). The list includes additional information on the languages such as their ISO 639-3 language code and family. Furthermore, each entry in the list provides a hyperlink to the original source from where the information is taken.

Each language in the list is attributed a different background color depending on its language family or location in order to allow for an at-a-glance overview for all languages in the list. The user can choose from a drop-down menu whether to include the genealogical or areal informa-



49 links found between "money" and "silver"

1. Ignaciano (Arawakan) [ign]:	[ne]
2. Aymara, Central (Aymaran) [ayr]:	[ḳulʸḳi]
3. Tsafiki (Barbacoan) [cof]:	[ka'la]
4. Seselwa Creole French (Creole) [crs]:	[larzan]
5. Miao, White (Hmong-Mien) [mww]:	[nyiaj]
6. Breton (Indo-European) [bre]:	[arhant]
7. French (Indo-European) [fra]:	[argent]
8. Gaelic, Irish (Indo-European) [gle]:	[airgead]
9. Welsh (Indo-European) [cym]:	[arian]
10. Cofán (Isolate) [con]:	[koriΦĩʔdi]
11. Aguaruna (Jivaroan) [agr]:	[ku'ičik]
12. Swahili (Niger-Congo) [swh]:	[fedha]
13. Akhvakh (Northern) (North Caucasian) [akv]:	[ачи]

Figure 3: Force-directed graph with mouse-over functionalities showing a subset of the list of words contributing to the cross-linguistic polysemies. The entries have different background colors depending on their location in the world map (cf. Figure 4).

tion as the background color. For the genealogical information, all language families are attributed a different color value. Languages belonging to the same language families are therefore given the same background color. Moreover, the list is sorted according to language families. In this way, the user can immediately see how many languages of a given family contribute to the overall strength for the connection at hand.

As to the areal information, the world map is provided with a color gradient as shown in Figure 4. To this end, each position in the world map is attributed a color value using the L*a*b* color space. The color hue thereby indicates the position on the map in terms of the longitude (East-West) whereas the lightness of the color represents the position in terms of the latitude information (North-South).[3] The mapping from geolocation to color values allows for an easier evaluation of areal patterns in the selected connection. In this regard, users can directly detect whether a certain cross-linguistic polysemy is restricted to a certain region of the world or constitutes a more widespread colexification pattern (see the case study in Section 3.5. below).

In addition to the interactive functionalities described above, the visualization also features a variety of further components that allow for an easier exploration of the database. The graph layout is equipped with panning and zooming functionality that enables the user to navigate through the network graph. Panning is enabled when the cursor changes into a hand symbol when mousing over a link of the graph. The whole graph can then be dragged to a new position. The zooming behavior is activated with the scroll wheel. When mousing over a concept (node) in the graph all connected links and concepts are highlighted in order to provide a better overview of the connectivity of certain concepts (see Figure 2). The control panel of the visualization also includes a slider button that allows the user to show only those edges in the graph with a minimum number of cross-linguistic colexifications.

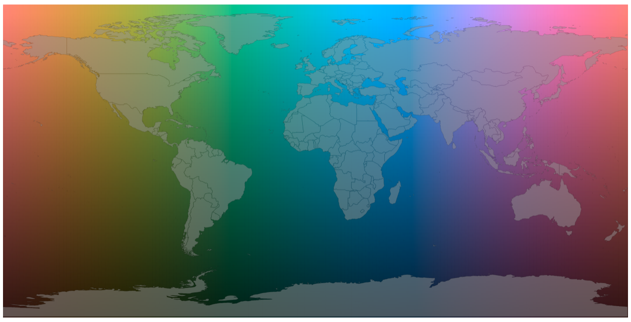


Figure 4: World map with color gradient

3.4. Implementation

The visualization is implemented in JavaScript using the D3 library (Bostock et al., 2011).[4] The force-directed graph is generated with the `force()` function from the `d3.layout` module. The layout implementation uses position Verlet integration for simple constraints (Dwyer, 2009).[5] In order to ensure that the concept labels are located close to the concept nodes, a second force layout

[3] See (Mayer et al., 2014) for a different approach of a linguistically informed color gradient of the world map.

[4] `http://d3js.org`

[5] See `https://github.com/mbostock/d3/wiki/Force-Layout` for a description of the implementation.

(with a static weight of 1) for each concept link to the node is set up.

The color values for the world map gradient scale are computed from the two-dimensional geographical coordinates that are given as an input. The latitude [-90;90] and longitude [-180;180] values are thereby normalized between [0;1] and serve as the input for the function `cl2pix`.[6]

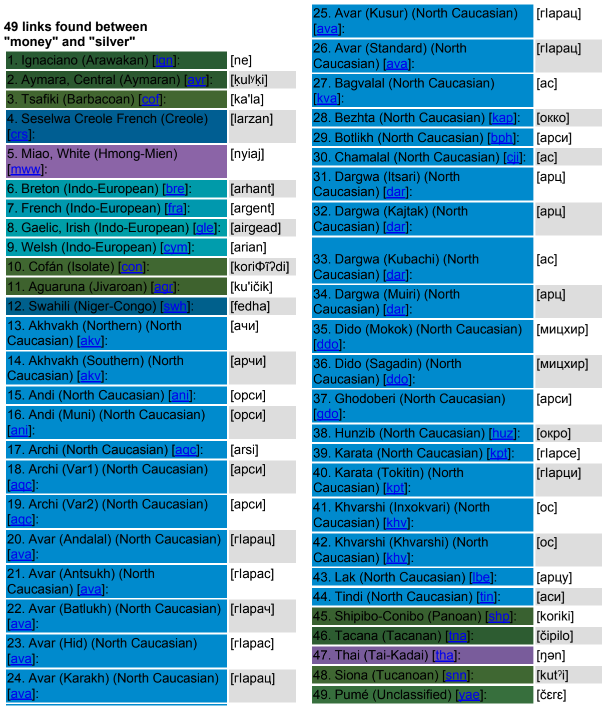
```
function cl2pix(c,l){
    var TAU = 6.2831853
    var L = l*0.61 + 0.09;
    var angle = TAU/6.0 - c*TAU;
    var r = l*0.311 + 0.125
    var a = Math.sin(angle)*r;
    var b = Math.cos(angle)*r;
    return [L,a,b];
 };
```

The actual HTML color code is generated with the function `d3.lab` from the D3 library, which takes as input the three values for `[L,a,b]`. The main reason for choosing the L*a*b* color space is a smoother transition between different color hues without any visible boundaries.[7] For the coloring of the language families, the background colors are generated with the categorical scale functions of the `d3.scale` module.

The dragging and panning functionalities of the graph are implemented with the `drag()` function from the `d3.behavior` module and the SVG `transform` and `translate` attributes.

3.5. Case study

In order to illustrate the usefulness of the visualization for the purposes of exploring the database, consider the graph in Figure 3. Among other things, it contains the connection between the concepts "money" and "silver". A subset of the languages and words contributing to this connection are shown on the left where the background color represents the language families. For instance, French contributes to the cross-linguistic colexification because both concepts are realized by the same word (viz. *argent*) in that language. When looking at the areal distribution of the languages, a clear pattern emerges at a glance (see Figure 5). Most of the languages contributing to the colexification are from two major regions: Caucasus (marked in blue) and South America (marked in green). However, as mentioned in Section 2.4., this distribution might be an artifact of the general bias for languages of the Caucasus and South America in the underlying databases. In any case, the visualization directly points the attention to this pattern. As the aim of the visualization component is not to replace linguistic research but to guide it, such patterns have to be looked at in more detail by checking the actual data.



49 links found between "money" and "silver"

1. Ignaciano (Arawakan) [ign]:	[ne]
2. Aymara, Central (Aymaran) [ayr]:	[ḳulʸḳi]
3. Tsafiki (Barbacoan) [cof]:	[ka'la]
4. Seselwa Creole French (Creole) [crs]:	[larzan]
5. Miao, White (Hmong-Mien) [mww]:	[nyiaj]
6. Breton (Indo-European) [bre]:	[arhant]
7. French (Indo-European) [fra]:	[argent]
8. Gaelic, Irish (Indo-European) [gle]:	[airgead]
9. Welsh (Indo-European) [cym]:	[arian]
10. Cofán (Isolate) [con]:	[koriΦĩʔdi]
11. Aguaruna (Jivaroan) [agr]:	[ku'ičik]
12. Swahili (Niger-Congo) [swh]:	[fedha]
13. Akhvakh (Northern) (North Caucasian) [akv]:	[ачи]
14. Akhvakh (Southern) (North Caucasian) [akv]:	[арчи]
15. Andi (North Caucasian) [ani]:	[орси]
16. Andi (Muni) (North Caucasian) [ani]:	[орси]
17. Archi (North Caucasian) [aqc]:	[arsi]
18. Archi (Var1) (North Caucasian) [aqc]:	[арси]
19. Archi (Var2) (North Caucasian) [aqc]:	[арси]
20. Avar (Andalal) (North Caucasian) [ava]:	[гӀарац]
21. Avar (Antsukh) (North Caucasian) [ava]:	[гӀарас]
22. Avar (Batlukh) (North Caucasian) [ava]:	[гӀарач]
23. Avar (Hid) (North Caucasian) [ava]:	[гӀарас]
24. Avar (Karakh) (North Caucasian) [ava]:	[гӀарац]
25. Avar (Kusur) (North Caucasian) [ava]:	[гӀарац]
26. Avar (Standard) (North Caucasian) [ava]:	[гӀарац]
27. Bagvalal (North Caucasian) [kva]:	[ас]
28. Bezhta (North Caucasian) [kap]:	[окко]
29. Botlikh (North Caucasian) [bph]:	[арси]
30. Chamalal (North Caucasian) [cji]:	[ас]
31. Dargwa (Itsari) (North Caucasian) [dar]:	[арц]
32. Dargwa (Kajtak) (North Caucasian) [dar]:	[арц]
33. Dargwa (Kubachi) (North Caucasian) [dar]:	[ас]
34. Dargwa (Muiri) (North Caucasian) [dar]:	[арц]
35. Dido (Mokok) (North Caucasian) [ddo]:	[мицхир]
36. Dido (Sagadin) (North Caucasian) [ddo]:	[мицхир]
37. Ghodoberi (North Caucasian) [gdo]:	[арси]
38. Hunzib (North Caucasian) [huz]:	[окро]
39. Karata (North Caucasian) [kpt]:	[гӀарсе]
40. Karata (Tokitin) (North Caucasian) [kpt]:	[гӀарци]
41. Khvarshi (Inxokvari) (North Caucasian) [khv]:	[ос]
42. Khvarshi (Khvarshi) (North Caucasian) [khv]:	[ос]
43. Lak (North Caucasian) [lbe]:	[арцу]
44. Tindi (North Caucasian) [tin]:	[аси]
45. Shipibo-Conibo (Panoan) [shp]:	[koriki]
46. Tacana (Tacanan) [tna]:	[čipilo]
47. Thai (Tai-Kadai) [tha]:	[ŋən]
48. Siona (Tucanoan) [snn]:	[kutˀi]
49. Pumé (Unclassified) [yae]:	[čɛrɛ]

Figure 5: Languages and words contributing to the connections of polysemies for the concepts "money" and "silver"

4. Conclusions

5. References

Andreas Blank. 1997. *Prinzipien des Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-driven documents. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)*, 17(12):2301–2309.

Carl Darling Buck. 1949. *A dictionary of selected synonyms in the principal Indo-European languages. A contribution to the history of ideas*. University of Chicago Press, Chicago and Illinois.

Tim Dwyer. 2009. Scalable, versatile and simple constrained graph layout. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'09, pages 991–1006, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Alexandre François. 2008. Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In M. Vanhove, editor, *From polysemy to semantic change*, pages 163–215. Benjamins, Amsterdam.

M. Girvan 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.

Martin Haspelmath and Uri Tadmor. 2009. *World Loanword Database*. Max Planck Digital Library, Munich.

Mary Ritchie Key and Bernard Comrie. 2007. *IDS –*

[6] The code was adapted from the GNU C code by David Dalrymple (`http://davidad.net/colorviz/`) and translated into JavaScript.

[7] See `http://davidad.net/colorviz/` for the difference between using the L*a*b* and HSV color space in terms of transitions between different color hues.

The Intercontinental Dictionary Series. URL: `http://lingweb.eva.mpg.de/ids/`.

A. Lancichinetti and S. Fortunato. 2009. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5 Pt 2):056117.

Johann-Mattis List, Anselm Terhalle, and Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, pages 347–353, Stroudsburg. Association for Computational Linguistics.

Thomas Mayer, Bernhard Wälchli, Christian Rohrdantz, and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. In *Language Processing and Grammars. The role of functionally oriented computational models*, pages 13–38. John Benjamins.

Scott Murray. 2010. *Interactive Data Visualization for the Web*. O'Reilly Media, Inc.

M. Rosvall and C. T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.*, 105(4):1118–1123, Jan.