# An Interactive Visualization of Cross-Linguistic Colexification Patterns

**Thomas Mayer[1], Johann-Mattis List[1], Anselm Terhalle[2], Matthias Urban[3]**

[1]Philipps University of Marburg, [2]Heinrich-Heine University of Düsseldorf, [3]Universiteit Leiden
thomas.mayer@uni-marburg.de, mattis.list@uni-marburg.de,
terhalle@phil.uni-duesseldorf.de, m.urban@hum.leidenuniv.nl

**Abstract**

In this paper, we present an interactive web-based visualization for the CLiCS database, an online resource for synchronic lexical associations (*colexification patterns*) in over 200 language varieties. The associations cover 1,288 concepts and represent the tendency for concepts to be expressed by the same words in the same languages and 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 a given pattern of colexification. Each language in the list is thereby attributed a different color depending on its genealogical or areal affiliation. In this way, given associations can be inspected for genealogical or areal bias.

**Keywords:** interactive visualization, colexification, cross-linguistic database

## 1. Introduction

What does "good" have to do with "beautiful"? Logically, not everything that is good is beautiful, and not everything that is beautiful is good. However, people seem to associate these concepts quite strongly, as linguistic data suggest: "good" and "beautiful" are expressed by identical word forms in 27 languages from 8 different language families. To assess the cognitive, linguistic, and cultural implications of this fact correctly, additional information would be useful: Where are these languages located on the globe? How are they distributed amoung the 8 families? Which other concepts are verbalized by the same form as "good" and "beautiful"?

Synchronic lexical association or *colexification*, i.e. the verbalization of two or more concepts by means of the same form in a given language, is an important source of information for investigations in cognitive linguistics, linguistic typology, and historical semantics. In this paper, we present an interactive web-based visualization for CLiCs, an online database that contains a large crosslinguistic data set on colexifications worldwide. We provide a brief outline of CLiCs, describe the functionalities of the visualization, and show that the visualization is an indispensible tool that enables researches to get an overview of the data and concisely plan further quantitative analyses according to their needs.

| 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 one language family are shaded gray. In order to browse the table on the CLiCs website, use the following URL: `http://clics.lingpy.org/all.php?gloss=silver`.

## 2. CLiCS

CLiCs (***Cr**oss-**Li**nguistic **C**olexification**s***, `http://clics.lingpy.org`) is an online database of synchronic lexico-semantic associations in 215 languages and language varieties of the world. CLiCs exploits already existing large online lexical databases, but has the advantage that it makes visible the relationships between meanings and forms in the object languages, something which is not easily possible using the interfaces of its sources themselves. Table 1 gives an example on the basic structure of the data in CLiCs.

### 2.1. Homonymy, polysemy, and colexification

A well-known concept from lexical semantic analysis is that of *polysemy*. It refers to the situation in which a lexical item possesses more than one identifiable sense between which there is a conceptual relation. A number of tests are available to distinguish polysemy from *semantic vagueness*, in which a division into distinct senses is not warranted. From an analytical perspective polysemy has to be further distinguished from homonymy and contextual variation. *Homonymy* refers to the "accidental" verbalization of at least two meanings by the same sound chain, without any conceptual relation 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. Hence, it is difficult for quantitative investigations to provide this differentiation in advance. Here, we use the term *colexification* (originally from François (2008)) to refer to the situation in which two or more of the meanings in our sources correspond to the same lexical item in one of the languages. For instance, we would say that Wayuu [guc] colexifies "good" and "beautiful" by means of the word form *anasi*.[1] Colexification is thus a deliberately ambiguous label that allows us to avoid making a commitment in each case as to the adequate lexical semantic analysis. 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 (Version 1.0) offers information on colexification in 221 different language varieties covering 64 different language families.[2] All language varieties in our sample comprise a total of 301 498 words covering 1 280 different concepts.[3] Using a strictly automatic procedure, we identified 45 667 cases of colexification that correspond to 16 239 different links between the 1 280 concepts covered by our data.

At present, four sources feed into CLiCs: (1) The *Intercontinental Dictionary Series* (IDS, Key and Comrie 2007), offering lexical data for 233 languages and language varieties of the world. Ideally, datasets for each language contain 1,310 entries, though coverage differs in completeness for individual languages. Of all 233 languages in IDS, 178 were automatically cleaned and included in CLiCs. (2) The *World Loanword Database* (WOLD, Haspelmath and Tadmor 2009), the main goal of which has to do with identifying lexical borrowings, but which nevertheless also provides general lexical data for 41 languages. The vocabularies for the individual languages differ somewhat in their size, ranging somewhere between 1,000 and 2,000 items. 33 of the 41 vocabularies are included in CLiCs. (3) Data for four languages neither represented in IDS nor WOLD were added from the *LOGOS* dictionary (`http://www.logosdictionary.org`), a multilingual online dictionary. (4) Additional data for six Himalayan languages was taken from the *Sprkbanken* project (University of Gothenburg, `http://spraakbanken.gu.se`).[4]

### 2.3. Network modeling of CLiCs

As mentioned above, there is no guarantee that lexical associations within CLiCs reflect conceptual associations. 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. Communities are groups of nodes in a network 'within which the connections are dense but between which they are sparser' (Newman, 2004). 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 1,252 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.

### 2.4. Limitations and Caveats

Data structure in CLiCs directly mirrors that in the sources we used. We did not manipulate or reanalyze the data in any way, to the effect that the reliability of CLiCs largely is a function of that of its sources. However, we would like to point out that we cannot rule out entirely the possibility of artifacts arising from automatic data cleaning in cases where textual coding of the data was inconsistent. As for its actual application, it also must be borne in mind that CLiCs reflects a certain bias regarding the geographical locations of the languages included in its sources: IDS features many languages of South America and the Caucasus, while WOLD includes a disproportionate percentage of languages of Europe. Hence, the sheer frequency of instances of a particular colexification pattern in CLiCs may be misleading insofar as a pattern is very robust cross-linguistically, but actually is so only in certain regions of

[1] See also the example of "money" and "silver" in the case study in Section 3.5. below.

[2] This count includes 12 language isolates, and 3 unclassified languages, according to the classification schema of Ethnologue.

[3] Since some concepts are expressed by more than one word in the respective languages, the number of words is higher than the expected one (282 880) if multiple synonyms per concept were not allowed.

[4] 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.

the world. We have not implemented any computational method in CLiCs to balance out the picture *a posteriori*. Since we nevertheless want to present potential users of CLiCs with the possibility to assess possible areal patterns in the data, we include a powerful visualization that enables them to detect areal imbalances in colexification patterns in individual cases themselves.

## 3. Visualization

The CLiCs database is available online at `http://clics.lingpy.org` and offers its users a search interface to all concepts and cross-linguistic colexifications 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.

### 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 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 dense 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 employed two different strategies to present the data from two different perspectives. According to our first strategy, we decided to split the data into *communities* first. Starting from 1 280 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. The *Infomap* algorithm requires that weights are defined for the edges of the network. Here we used the number of attested language families per colexification as edge weights. Following a suggestion by Dellert (2014) we further normalized the number of attested language families with help of a formula

$$W = \frac{C^2}{O_A + O_B - C}, \tag{1}$$

where $C$ is the number of attested language families for the colexification of concept $A$ and concept $B$, $O_A$ is the number of language families in which concept $A$ occurs, and $O_B$ is the number of language families in which concept $B$ occurs. 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 the help of this analysis, the 1 280 concepts could be subdivided into 271 communities. Of all communities, 118 are *large*, containing more than five nodes. The large communities cover 65% (828) of all nodes in the original network (1 280). 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. The communities do not differ much in size, ranging from 2 ('men's house') to 16 concepts ('fur') with an average of 4.72 concepts per community.

The advantage of the community perspective on CLiCs is that it provides an independent automatic clustering of the data. The disadvantage is that this pre-selection deprives the user of finding alternative possibly interesting connections. Community-detection methods are not error-proof, and their performance varies depending on the algorithms being used and the data being analysed. Even more importantly, most community-detection methods are based on restricted decisions when clustering the nodes of a given network into groups. Every node is assigned to one community only. No transitions between communities are possible. In order to offer a less biased perspective on the network, we decided to use the following strategy. According to our second strategy, we decided to cut the network in parts...

### 3.3. Interactive functionalities

The visualization features various interactive functionalities that are designed to enhance the exploration of the

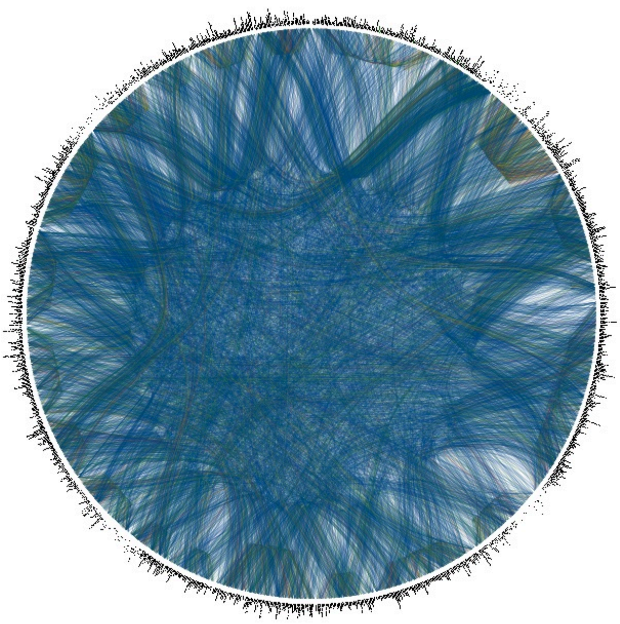


Figure 1: Full network of all 1,288 concepts in CLiCs (outer circle) together with their connections. The strength of the connections is marked in different colors, with very strong links represented in red

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 language families that can be attested to have lexical associations 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).

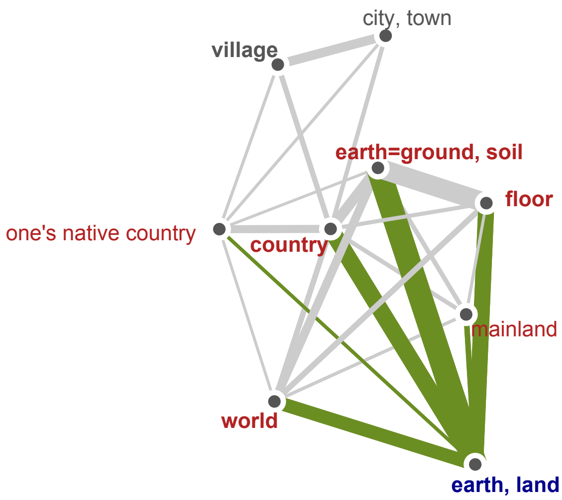



Figure 2: Force-directed graph with mouse-over functionalities highlighting all connected concepts

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 of all languages in the list. The user can choose from a drop-down menu whether to include the genealogical or areal information 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 (specifying the east-west position) whereas the lightness of the color represents the position in terms of the latitude information (specifying the north-south position).[5] 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).

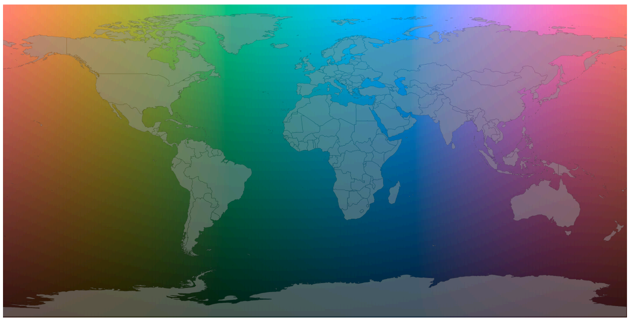


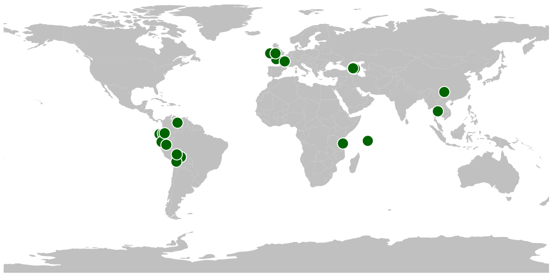
Figure 4: World map with color gradient

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.

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

**49 links for "silver" and "money":**

| Language | Family | Form |
|---|---|---|
| 1. Ignaciano | Arawakan | ne |
| 2. Aymara, Central | Aymaran | ḳulʸḳi |
| 3. Tsafiki | Barbacoan | ka'la |
| 4. Seselwa Creole French | Creole | larzan |
| 5. Miao, White | Hmong-Mien | nyiaj |
| 6. Breton | Indo-European | arhant |
| 7. French | Indo-European | argent |
| 8. Gaelic, Irish | Indo-European | airgead |
| 9. Welsh | Indo-European | arian |
| 10. Cofán | Isolate | koriΦĩʔdi |



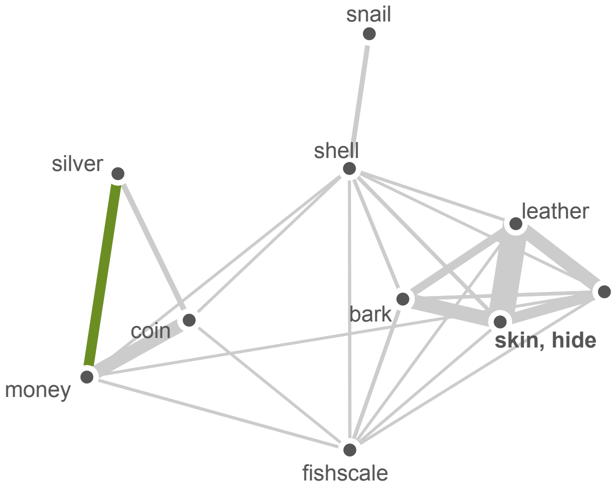



Figure 3: Force-directed graph with mouse-over functionalities showing one community in CLiCs (community 27: gold) together with a subset of the instances of a particular colexification pattern (left). The entries have different background colors depending on their location in the world map (cf. Figure 4)

### 3.4. Implementation

The visualization is implemented in JavaScript using the D3 library (Bostock et al., 2011).[6] 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).[7] In order to ensure that the concept labels are located close to the concept nodes, a second force layout (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`.[8]

```
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];
 };
```

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

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

[8] The code was adapted from the GNU C code by David Dalrymple (`http://davidad.net/colorviz/`, accessed on January 25th, 2014) and translated into JavaScript.

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. As can be seen in Figure 5, the color gradient in the L*a*b* color space exhibits a much smoother perceptual transition between the color hues on the x-axis. 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 studies

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 location of the languages. 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 6 for the full list of languages showing this colexification pattern). Most of the languages contributing to the colexification are from two major regions: Caucasus (marked in

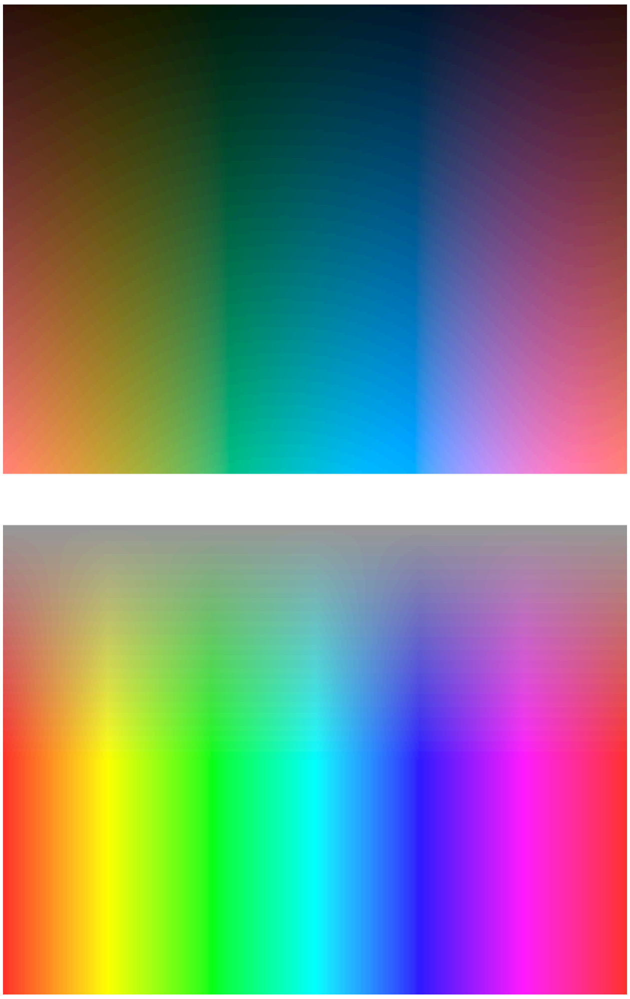


Figure 5: Comparison between two-dimensional color gradients in the L*a*b* (top) and HSV (bottom) color space

49 links for "money" and "silver".

| Language | Family | Form |
|---|---|---|
| 1. Ignaciano | Arawakan | ne |
| 2. Aymara, Central | Aymaran | ḳulʸḳi |
| 3. Tsafiki | Barbacoan | kaˈla |
| 4. Seselwa Creole French | Creole | larzan |
| 5. Miao, White | Hmong-Mien | nyiaj |
| 6. Breton | Indo-European | arhant |
| 7. French | Indo-European | argent |
| 8. Gaelic, Irish | Indo-European | airgead |
| 9. Welsh | Indo-European | arian |
| 10. Cofán | Isolate | koriΦĩʔdi |
| 11. Aguaruna | Jivaroan | kuˈičik |
| 12. Swahili | Niger-Congo | fedha |
| 13. Akhvakh (Northern) | North Caucasian | ачи |
| 14. Akhvakh (Southern) | North Caucasian | арчи |
| 15. Andi | North Caucasian | орси |
| 16. Andi (Muni) | North Caucasian | орси |
| 17. Archi | North Caucasian | arsi |
| 18. Archi (Var1) | North Caucasian | арси |
| 19. Archi (Var2) | North Caucasian | арси |
| 20. Avar (Andalal) | North Caucasian | гӀарац |
| 21. Avar (Antsukh) | North Caucasian | гӀарас |
| 22. Avar (Batlukh) | North Caucasian | гӀарач |
| 23. Avar (Hid) | North Caucasian | гӀарас |
| 24. Avar (Karakh) | North Caucasian | гӀарац |
| 25. Avar (Kusur) | North Caucasian | гӀарац |
| 26. Avar (Standard) | North Caucasian | гӀарац |
| 27. Bagvalal | North Caucasian | ac |
| 28. Bezhta | North Caucasian | окко |
| 29. Botlikh | North Caucasian | арси |
| 30. Chamalal | North Caucasian | ac |
| 31. Dargwa (Itsari) | North Caucasian | арц |
| 32. Dargwa (Kajtak) | North Caucasian | арц |
| 33. Dargwa (Kubachi) | North Caucasian | ac |
| 34. Dargwa (Muiri) | North Caucasian | арц |
| 35. Dido (Mokok) | North Caucasian | мицхир |
| 36. Dido (Sagadin) | North Caucasian | мицхир |
| 37. Ghodoberi | North Caucasian | арси |
| 38. Hunzib | North Caucasian | окро |
| 39. Karata | North Caucasian | гӀарсе |
| 40. Karata (Tokitin) | North Caucasian | гӀарци |
| 41. Khvarshi (Inxokvari) | North Caucasian | ос |
| 42. Khvarshi (Khvarshi) | North Caucasian | ос |
| 43. Lak | North Caucasian | арцу |
| 44. Tindi | North Caucasian | аси |
| 45. Shipibo-Conibo | Panoan | koriki |
| 46. Tacana | Tacanan | čipilo |
| 47. Thai | Tai-Kadai | ŋən |
| 48. Siona | Tucanoan | kutˀi |
| 49. Pumé | Unclassified | čɛrɛ |

Figure 6: Languages and words contributing to the connections of lexical associations for the concepts "money" and "silver"

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.

Another example deals with the colexification of the concepts "wheel" and "foot". In contrast to the case of "money" and "silver" above, these concepts at first glance may not immediately suggest a close association. Yet such cases do exits as the link in Figure 7 reveals. The connection links two bigger communities of nodes, including spherical objects on the one hand and parts of the lower body on the other. The list of languages for the connection "wheel" and "foot" in Figure 7 clearly shows that the association is restricted to languages of South America. This geographical restriction may reflect semantic borrowing among South American languages, but since the distribution within South America is rather erratic, independent innovation is also a possibility. At any rate, the color coding in the visualization immediately draws the researcher's attention to the potentially interesting geographical patterning.

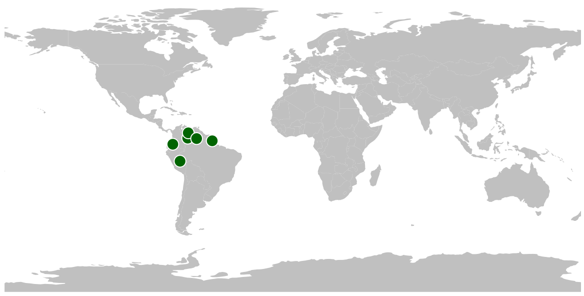
## 4. Conclusions and future work

The size and complexity of today's language resources call for a data preparation pipeline that enables researchers to find meaningful patterns among the multitude of different factors that can be taken into consideration. In our view, such a data preparation pipeline necessarily consists of two major parts, both of which are illustrated in the present paper. On the one hand, methods and techniques from data mining or computational linguistics help to detect basic trends or groups of similar objects in the search space. On the other hand, the resulting groups or trends are mapped to visual variables in order to make interesting observations readily accessible to human perception.

The CLiCs database contains a wealth of information about colexification patterns in the languages of the world. Manually inspecting the large amount of connections in the database, however, is a laborious and time-consuming task that allows for a detailed exploration of individual links but does not capture overall trends in the data. This paper presents an attempt to combine the advantages of human inspection with the strength of a computational approach (Keim et al., 2008).

**6 links for "foot" and "wheel":**

| Language | Family | Form |
|---|---|---|
| 1. Cofán | Isolate | ȼɨʔtʰe |
| 2. Puinave | Isolate | sim |
| 3. Yaminahua | Panoan | taɨ |
| 4. Wayampi | Tupi | pɨ |
| 5. Pumé | Unclassified | taɔ |
| 6. Ninam | Yanomam | mãhuk |



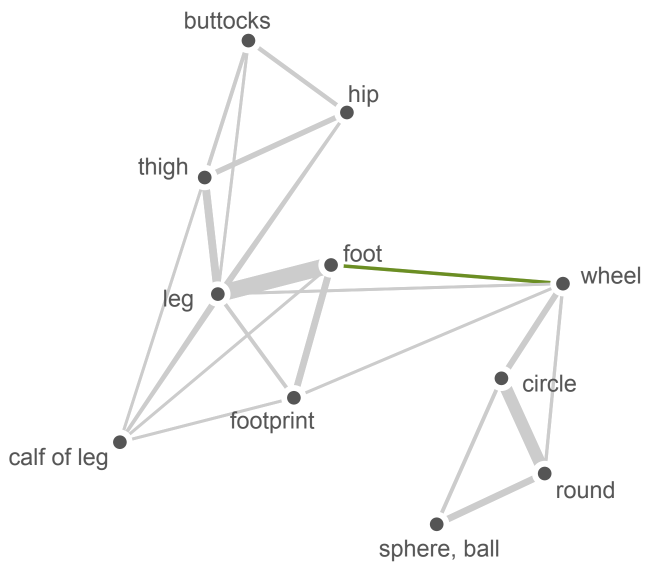



Figure 7: Force-directed graph with areal distribution for the concepts “wheel” and “foot”

The CLiCs visualization features an automatic preprocessing of the colexification links into so-called communities of the graph, groups of highly connected nodes that reveal a meaningful overall trend in the worldwide patterns of lexical associations. The communities are then graphically represented in a force-directed graph that shows all connections within the various concepts that are included. Interactive components in the visualization allow for a more detailed view of associations on the level of the languages that contribute to the colexification. Mapping the genealogical and areal information on individual languages to colors enables an at-a-glance evaluation of potentially interesting trends in individual colexifications (see the case studies in Section 3.5.). In this way, users can get an overview of the general trends in the data and at the same time have the opportunity to directly inspect the lexical associations.

In future work, we plan to enhance the visualization tool with further interactive components that allow for a better overview of the complete network of colexifications (shown in Figure 1) and facilitate the detection of genealogical or areal trends in the database. The idea is to integrate a sunburst visualization (Stasko and Zhang, 2000) for the genealogical information in order to enable a better overview of the language families that are involved in a given colexification pattern.[9] In addition, we intend to equip the user interface with further interactive components that allow users to explore the database from different perspectives (e.g., compare individual languages in terms of shared lexical associations). All components will be made publicly available online for the (linguistic) research community.

[9] See Mayer et al. (2014) for an example of using sunburst displays to represent the hierarchical structure of language families.

## 5. References

Blank, A. (1997). *Prinzipien des Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.

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

Dellert, J. (2014). Lifting a large multilingual dictionary to the level of concepts. Talk held at the “Workshop on historical and empirical evolutionary Linguistics” (February 15-16, 2014, Eberhard-Karls-Universität Tübingen).

Dwyer, T. (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.

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

Girvan, M. and Newman, M. E. (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. and Tadmor, U. (2009). *World Loanword Database*. Max Planck Digital Library, Munich.

Keim, D. A., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: definition, process, and challenges. In Kerren, Andreas, Stasko, John T., Fekete, Jean-Daniel, and North, Chris, editors, *Information Visualization*, pages 154–175. Berlin: Springer Verlag.

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

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

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 – Short Papers*, pages 347–353, Stroudsburg. Association for Computational Linguistics.

Mayer, T., Wälchli, B., Rohrdantz, C., and Hund, M. (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.

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

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

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

Stasko, J. and Zhang, E. (2000). Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 57–, Washington, DC, USA. IEEE Computer Society.