CLICS 2.0
Towards an Improved Handling of Cross-Linguistic Colexification Patterns

Johann-Mattis List

Research Group “Computer-Assisted Language Comparison”
Department of Linguistic and Cultural Evolution
Max-Planck Institute for the Science of Human History
Jena, Germany

2017/07/03
A long, long time ago...
Predecessors: People and Ideas
Predecessors: People and Ideas

Predecessors: People and Ideas

Predecessors: People and Ideas

Predecessors: People and Ideas

- Steiner, Stadler, and Cysouw (2011): A pipeline for computational historical linguistics.
Predecessors: People and Ideas

- Steiner, Stadler, and Cysouw (2011): A pipeline for computational historical linguistics.
- Urban (2011): Assymetries in overt marking and directionality in semantic change.
A long, long time ago...

Predecessors: Data
Predecessors: Data

- Intercontinental Dictionary Series (IDS, Key and Comrie 2016) offers 1310 concepts translated into about 360 languages, an earlier version offered ca. 200 languages.
Predecessors: Data

- Intercontinental Dictionary Series (IDS, Key and Comrie 2016) offers 1310 concepts translated into about 360 languages, an earlier version offered ca. 200 languages.

- World Loanword Typology (WOLD, Haspelmath and Tadmor 2009) offers 1430 concepts translated into 41 languages (some overlap with IDS).
Predecessors: Techniques

- Steiner, Stadler, and Cysouw (2011) present the idea to model similarities between concepts by constructing a matrix from parts of the IDS data that shows how often individual languages colexify certain concepts.
Predecessors: Techniques

- Steiner, Stadler, and Cysouw (2011) present the idea to model similarities between concepts by constructing a matrix from parts of the IDS data that shows how often individual languages colexify certain concepts.
- Cysouw (2010) shows how to use polysemy data to draw networks.
Initial Ideas

List, Terhalle, and Urban (2013) build on ideas of Cysouw (2010) and Steiner, Stadler, and Cysouw (2011) in using IDS data for polysemy studies and in using network techniques to study the data. In contrast to earlier approaches, they use techniques for community detection (Girvan and Newman 2002) to further analyze the network and to partition the concepts into communities which seem to make intuitively sense, reminding of naturally derived semantic fields.
Initial Ideas

List, Terhalle, and Urban (2013) build on ideas of Cysouw (2010) and Steiner, Stadler and Cysouw (2011) in using IDS data for polysemy studies and in using network techniques to study the data.
List, Terhalle, and Urban (2013) build on ideas of Cysouw (2010) and Steiner, Stadler and Cysouw (2011) in using IDS data for polysemy studies and in using network techniques to study the data.

In contrast to earlier approaches, they use techniques for *community detection* (Girvan and Newman 2002) to further analyse the network, and to partition the concepts into communities which seem to make intuitively sense, reminding of naturally derived semantic fields.
Further Ideas

A long, long time ago...

Mayer, List, Terhalle, and Urban (2014) present an interactive way to visualize cross-linguistic colexification data. List, Mayer, Terhalle, and Urban (2014) publish the database and the web-application online, under the name CLICS (Database of Cross-Linguistic Colexifications).

In contrast to earlier attempts, they increased the database by merging IDS, WOLD, and additional datasets which they collected themselves, thus containing 220 languages in total. They also improved the community detection procedure by using Infomap (Rosvall and Bergstrom 2008), an advanced algorithm based on random walks in complex networks.
Further Ideas

Mayer, List, Terhalle, and Urban (2014) present an interactive way to visualize cross-linguistic colexification data.
Mayer, List, Terhalle, and Urban (2014) present an interactive way to visualize cross-linguistic colexification data.

List, Mayer, Terhalle, and Urban (2014) publish the database and the web-application online, under the name CLICS (*Database of Cross-Linguistic Colexifications*).
Further Ideas

- Mayer, List, Terhalle, and Urban (2014) present an interactive way to visualize cross-linguistic colexification data.
- List, Mayer, Terhalle, and Urban (2014) publish the database and the web-application online, under the name CLICS (*Database of Cross-Linguistic Colexifications*).
- In contrast to earlier attempts, they increased the data by merging IDS, WOLD, and additional datasets which they collected themselves, thus containing 220 languages in total.
Further Ideas

- Mayer, List, Terhalle, and Urban (2014) present an interactive way to visualize cross-linguistic colexification data.
- List, Mayer, Terhalle, and Urban (2014) publish the database and the web-application online, under the name CLICS (*Database of Cross-Linguistic Colexifications*).
- In contrast to earlier attempts, they increased the data by merging IDS, WOLD, and additional datasets which they collected themselves, thus containing 220 languages in total.
- They also improved the community detection procedure by using Infomap (Rosvall and Bergstrom 2008), an advanced algorithm based on random walks in complex networks.
CLICS 1.0
IDS (Key and Comrie 2007 version), of 233 language varieties, 178 included in CLICS.

WOLD (Haspelmath and Tadmor 2009), of 41 languages in WOLD, 33 are included in CLICS.

Logos Dictionary (Logos Group), of dictionaries for more than 60 different languages, 4 languages were manually extracted and included in CLICS.

Språkbanken project (University of Gothenburg) offers 8 wordlists for SEn languages, 6 were included in CLICS.
IDS (Key and Comrie 2007 version), of 233 language varieties, 178 included in CLICS.
IDS (Key and Comrie 2007 version), of 233 language varieties, 178 included in CLICS.

WOLD (Haspelmath and Tadmor 2009), of 41 languages in WOLD, 33 are included in CLICS.
Data

- **IDS** (Key and Comrie 2007 version), of 233 language varieties, 178 included in CLICS.
- **WOLD** (Haspelmath and Tadmor 2009), of 41 languages in WOLD, 33 are included in CLICS.
- **Logos Dictionary** (Logos Group), of dictionaries for more than 60 different languages, 4 languages were manually extracted and included in CLICS.
IDS (Key and Comrie 2007 version), of 233 language varieties, 178 included in CLICS.

WOLD (Haspelmath and Tadmor 2009), of 41 languages in WOLD, 33 are included in CLICS.

Logos Dictionary (Logos Group), of dictionaries for more than 60 different languages, 4 languages were manually extracted and included in CLICS.

Språkbanken project (University of Gothenburg) offers 8 word lists for SEA languages, 6 were included in CLICS.
Methods

Problems
Methods

Problems

(A) Data cannot be displayed fully, complexity needs to be reduced.

(B) Data is noisy and needs to be corrected.
## Problems

1. (A) Data cannot be displayed fully, complexity needs to be reduced.
2. (B) Data is noisy and needs to be corrected.

## Solutions
Methods

Problems

(A) Data cannot be displayed fully, complexity needs to be reduced.

(B) Data is noisy and needs to be corrected.

Solutions

(A) Show communities instead of showing all the data, offer a subgraph-view that cuts out the nearest neighbors of one concept to compensate for data loss in the community view.

(B) Filter by language families and weight the concept links by frequency of occurrence, following Dellert’s (2014) suggestion. This will cut most of the links resulting from homophony and leaves the links which are due to polysemy.
Interface is written in JavaScript for the visualizations and PHP for querying the data. The interactive component of the network browser was specifically designed for CLICS and builds on the D3 framework by Bostock et al. (2011). The underlying network with the inferred communities is offered for download from the website, and the whole code which was used to create the website is available for download at http://github.com/clics/clics.
Interface

- Interface is written in JavaScript for the visualizations and PHP for querying the data.

The interactive component of the network browser was specifically designed for CLICS and builds on the D3 framework by Bostock et al. (2011).

The underlying network with the inferred communities is offered for download from the website, and the whole code which was used to create the website is available for download at http://github.com/clics/clics.
Interface

- Interface is written in JavaScript for the visualizations and PHP for querying the data.
- The interactive component of the network browser was specifically designed for CLICS and builds on the D3 framework by Bostock et al. (2011).
Interface

- Interface is written in JavaScript for the visualizations and PHP for querying the data.
- The interactive component of the network browser was specifically designed for CLICS and builds on the D3 framework by Bostock et al. (2011).
- The underlying network with the inferred communities is offered for download from the website, and the whole code which was used to create the website is available for download at http://github.com/clics/clics/clics.
DEMO
Motivation

Problems in CLICS 1.0:
- Difficult to curate (error-correction, linking data, adding data)
- Difficult to collaborate (the CLICs team is separated and everybody is extremely busy with stuff other than CLICS)
- Difficult to communicate (not all users understand how we arrived at the data, and oftentimes think that it is us who messed up datasets up, etc., although we only took the data to produce something new out of it)
- Difficult to expand (new datasets cannot be added without having a true guiding principle)
- Difficult to catch up (we know much, much better now how to curate datasets, but we did not know this when preparing CLICS initially)
Motivation

Problems in CLICS 1.0

- difficult to curate (error-correction, linking data, adding data)
Motivation

Problems in CLICS 1.0

- difficult to curate (error-correction, linking data, adding data)
- difficult to collaborate (the CLICS team is separated and everybody is extremely busy with stuff other than CLICS)
Motivation

Problems in CLICS 1.0

- difficult to curate (error-correction, linking data, adding data)
- difficult to collaborate (the CLICS team is separated and everybody is extremely busy with stuff other than CLICS)
- difficult to communicate (not all users understand how we arrived at the data, and often think that it is us who messed datasets up, etc., although we only take the data to produce something new out of it)
Motivation

Problems in CLICS 1.0

- difficult to curate (error-correction, linking data, adding data)
- difficult to collaborate (the CLICS team is separated and everybody is extremely busy with stuff other than CLICS)
- difficult to communicate (not all users understand how we arrived at the data, and often think that it is us who messed datasets up, etc., although we only take the data to produce something new out of it)
- difficult to expand (new datasets cannot be added without having a true guiding principle)
Motivation

Problems in CLICS 1.0

- difficult to curate (error-correction, linking data, adding data)
- difficult to collaborate (the CLICS team is separated and everybody is extremely busy with stuff other than CLICS)
- difficult to communicate (not all users understand how we arrived at the data, and often think that it is us who messed datasets up, etc., although we only take the data to produce something new out of it)
- difficult to expand (new datasets cannot be added without having a true guiding principle)
- difficult to catch up (we know much, much better now, how to curate datasets, but we did not know this when preparing CLICS initially)
CLICS 2.0: Uses state-of-the-art available data, separates data from display (CLICS 2.0 does not host data but simply uses it). Assembles data with help of the Concepticon (List, Forkel, and Cysouw 2016). Assembles information on languages exclusively from Glottolog (Hammarström et al. 2017). Curates the code and polysemy data with help of a transparent API. Regularly releases data in release circles of about 1 per year (following the practice of Glottolog and other CLLD projects). Normalizes the data which is analyzed by CLICS.
use the state of the art of available data
Ideas

- use the state of the art of available data
- separate data from display (CLICS 2.0 does not host data, but simply uses it)
Ideas

- use the state of the art of available data
- separate data from display (CLICS 2.0 does not host data, but simply uses it)
- assemble data with help of the Concepticon (List, Forkel, and Cysouw 2016)
use the state of the art of available data
separate data from display (CLICS 2.0 does not host data, but simply uses it)
assemble data with help of the Concepticon (List, Forkel, and Cysouw 2016)
assemble information on languages exclusively from Glottolog (Hammarström et al. 2017)
Ideas

- use the state of the art of available data
- separate data from display (CLICS 2.0 does not host data, but simply uses it)
- assemble data with help of the Concepticon (List, Forkel, and Cysouw 2016)
- assemble information on languages exclusively from Glottolog (Hammarström et al. 2017)
- curate the code and the polysemy data with help of a transparent API
Ideas

- use the state of the art of available data
- separate data from display (CLICS 2.0 does not host data, but simply uses it)
- assemble data with help of the Concepticon (List, Forkel, and Cysouw 2016)
- assemble information on languages exclusively from Glottolog (Hammarström et al. 2017)
- curate the code and the polysemy data with help of a transparent API
- regularly release the data in release circles of about 1 per year (following the practice of Glottolog and other CLLD projects)
Ideas

- use the state of the art of available data
- separate data from display (CLICS 2.0 does not host data, but simply uses it)
- assemble data with help of the Concepticon (List, Forkel, and Cysouw 2016)
- assemble information on languages exclusively from Glottolog (Hammarström et al. 2017)
- curate the code and the polysemy data with help of a transparent API
- regularly release the data in release circles of about 1 per year (following the practice of Glottolog and other CLLD projects)
- normalize the data which is analysed by CLICS
# Excursus: Concepticon

<table>
<thead>
<tr>
<th>Concept List</th>
<th># Items</th>
<th>Concept Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen (2007)</td>
<td>500</td>
<td>animal oil; 动物油 (脂肪)</td>
</tr>
<tr>
<td>Gregersen (1976)</td>
<td>217</td>
<td>fat-grease*fat-grease</td>
</tr>
<tr>
<td>Heggarty (2005)</td>
<td>150</td>
<td>fat (grease); grasa</td>
</tr>
<tr>
<td>Swadesh (1955)</td>
<td>100</td>
<td>fat (grease)</td>
</tr>
<tr>
<td>Alpher and Nash (1999)</td>
<td>151</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Hale (1961)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>OGrady and Klokeid (1969)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Blust (2008)</td>
<td>210</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Matisoff (1978)</td>
<td>200</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>218</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Dunn et al. (2012)</td>
<td>207</td>
<td>fat</td>
</tr>
<tr>
<td>Swadesh (1950)</td>
<td>215</td>
<td>fat</td>
</tr>
<tr>
<td>Zgraggen (1980)</td>
<td>380</td>
<td>fat</td>
</tr>
<tr>
<td>Jachontov (1991)</td>
<td>100</td>
<td>fat n.</td>
</tr>
<tr>
<td>Wiktionary (2003)</td>
<td>207</td>
<td>fat (noun)</td>
</tr>
<tr>
<td>Starostin (1991)</td>
<td>110</td>
<td>fat n.; жир</td>
</tr>
<tr>
<td>TeilDautrey et al. (2008)</td>
<td>430</td>
<td>fat, oil</td>
</tr>
<tr>
<td>Swadesh (1952)</td>
<td>200</td>
<td>fat (organic substance)</td>
</tr>
<tr>
<td>Shiro (1973)</td>
<td>200</td>
<td>grease (fat)</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>100</td>
<td>grease; graisse; Fett; grasa</td>
</tr>
<tr>
<td>Wang (2006)</td>
<td>200</td>
<td>pig oil; 猪油</td>
</tr>
<tr>
<td>Haspelmath and Tadmor (2009)</td>
<td>1460</td>
<td>the grease or fat</td>
</tr>
</tbody>
</table>
## Excursus: Concepticon

<table>
<thead>
<tr>
<th>Concept List</th>
<th># Items</th>
<th>Concept Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen (2007)</td>
<td>500</td>
<td>animal oil; 动物油 (脂肪)</td>
</tr>
<tr>
<td>Gregersen (1976)</td>
<td>217</td>
<td>fat-grease*fat-grease</td>
</tr>
<tr>
<td>Heggarty (2005)</td>
<td>150</td>
<td>fat (grease); grasa</td>
</tr>
<tr>
<td>Swadesh (1955)</td>
<td>100</td>
<td>fat (grease)</td>
</tr>
<tr>
<td>Alpher and Nash (1999)</td>
<td>151</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Hale (1961)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>OGrady and Klokeid (1969)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Blust (2008)</td>
<td>210</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Matisoff (1978)</td>
<td>200</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>218</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Dunn et al. (2012)</td>
<td>207</td>
<td>fat</td>
</tr>
<tr>
<td>Swadesh (1950)</td>
<td>215</td>
<td>fat</td>
</tr>
<tr>
<td>Zgraggen (1980)</td>
<td>380</td>
<td>fat</td>
</tr>
<tr>
<td>Jachontov (1991)</td>
<td>100</td>
<td>fat n.</td>
</tr>
<tr>
<td>Wiktionary (2003)</td>
<td>207</td>
<td>fat (noun)</td>
</tr>
<tr>
<td>Starostin (1991)</td>
<td>110</td>
<td>fat n.; жир</td>
</tr>
<tr>
<td>TeilDautrey et al. (2008)</td>
<td>430</td>
<td>fat, oil</td>
</tr>
<tr>
<td>Swadesh (1952)</td>
<td>200</td>
<td>fat (organic substance)</td>
</tr>
<tr>
<td>Shiro (1973)</td>
<td>200</td>
<td>grease (fat)</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>100</td>
<td>grease; graisse; Fett; grasa</td>
</tr>
<tr>
<td>Wang (2006)</td>
<td>200</td>
<td>pig oil; 猪油</td>
</tr>
<tr>
<td>Haspelmath and Tadmor (2009)</td>
<td>1460</td>
<td>the grease or fat</td>
</tr>
</tbody>
</table>
## Excursus: Concepticon

<table>
<thead>
<tr>
<th>Concept List</th>
<th># Items</th>
<th>Concept Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen (2007)</td>
<td>500</td>
<td>animal oil; 动物油 (脂肪)</td>
</tr>
<tr>
<td>Gregersen (1976)</td>
<td>217</td>
<td>fat-grease*fat-grease</td>
</tr>
<tr>
<td>Heggarty (2005)</td>
<td>150</td>
<td>fat (grease); grasa</td>
</tr>
<tr>
<td>Swadesh (1955)</td>
<td>100</td>
<td>fat (grease)</td>
</tr>
<tr>
<td>Alpher and Nash (1999)</td>
<td>151</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Hale (1961)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>OGrady and Klokeid (1969)</td>
<td>100</td>
<td>fat, grease</td>
</tr>
<tr>
<td>Blust (2008)</td>
<td>210</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Matisoff (1978)</td>
<td>200</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>218</td>
<td>fat/grease</td>
</tr>
<tr>
<td>Dunn et al. (2012)</td>
<td>207</td>
<td>fat</td>
</tr>
<tr>
<td>Swadesh (1950)</td>
<td>215</td>
<td>fat</td>
</tr>
<tr>
<td>Zgrabagen (1980)</td>
<td>380</td>
<td>fat</td>
</tr>
<tr>
<td>Jachontov (1991)</td>
<td>100</td>
<td>fat n.</td>
</tr>
<tr>
<td>Wiktionary (2003)</td>
<td>207</td>
<td>fat (noun)</td>
</tr>
<tr>
<td>Starostin (1991)</td>
<td>110</td>
<td>fat n.; жир</td>
</tr>
<tr>
<td>TeilDautrey et al. (2008)</td>
<td>430</td>
<td>fat, oil</td>
</tr>
<tr>
<td>Swadesh (1952)</td>
<td>200</td>
<td>fat (organic substance)</td>
</tr>
<tr>
<td>Shiro (1973)</td>
<td>200</td>
<td>grease (fat)</td>
</tr>
<tr>
<td>Samarin (1969)</td>
<td>100</td>
<td>grease; graisse; Fett; grasa</td>
</tr>
<tr>
<td>Wang (2006)</td>
<td>200</td>
<td>pig oil; 猪油</td>
</tr>
<tr>
<td>Haspelmath and Tadmor (2009)</td>
<td>1460</td>
<td>the grease or fat</td>
</tr>
</tbody>
</table>
### Excursus: Concepticon

<table>
<thead>
<tr>
<th>Concept List</th>
<th># Items</th>
<th>Concept Label</th>
<th>Concept ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen (2007)</td>
<td>500</td>
<td>animal oil; 动物油 (脂肪)</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Gregersen (1976)</td>
<td>217</td>
<td>fat-grease*fat-grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Heggarty (2005)</td>
<td>150</td>
<td>fat (grease); grasa</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Swadesh (1955)</td>
<td>100</td>
<td>fat (grease)</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Alpher and Nash (1999)</td>
<td>151</td>
<td>fat, grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Hale (1961)</td>
<td>100</td>
<td>fat, grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>OGrady and Klokeid (1969)</td>
<td>100</td>
<td>fat, grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Matisoff (1978)</td>
<td>200</td>
<td>fat/grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Samar (1969)</td>
<td>218</td>
<td>fat/grease</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Dunn et al. (2012)</td>
<td>207</td>
<td>fat</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Swadesh (1950)</td>
<td>215</td>
<td>fat</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Zraggen (1980)</td>
<td>380</td>
<td>fat</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Jachontov (1991)</td>
<td>100</td>
<td>fat n.</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Wiktionary (2003)</td>
<td>207</td>
<td>fat (noun)</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Starostin (1991)</td>
<td>110</td>
<td>fat n.; жир</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>TeilDautrey et al. (2008)</td>
<td>430</td>
<td>fat, oil</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Swadesh (1952)</td>
<td>200</td>
<td>fat (organic substance)</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Shiro (1973)</td>
<td>200</td>
<td>grease (fat)</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Samar (1969)</td>
<td>100</td>
<td>grease; graisse; Fett; grasa</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
<tr>
<td>Haspelmath and Tadmor (2009)</td>
<td>1460</td>
<td>the grease or fat</td>
<td>GREASE (CONCEPTICON-ID:323)</td>
</tr>
</tbody>
</table>
Excursus: Concepticon

<table>
<thead>
<tr>
<th>Concepticon (List et al. 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>link concept labels in published concept lists (questionnaires) to concept sets</td>
</tr>
<tr>
<td>link concept sets to meta-data</td>
</tr>
<tr>
<td>define relations between concept sets</td>
</tr>
<tr>
<td>never link one concept in a given list to more than one concept set (guarantees consistency)</td>
</tr>
<tr>
<td>provide an API to check the consistency of the data and to query the data</td>
</tr>
<tr>
<td>provide a web-interface to browse through the data</td>
</tr>
</tbody>
</table>
Concepticon

[Diagram showing relationships between source, compiler, concept list, and concept set labels]

- Source
- Compiler
- Concept List
- Concept Labels
- Concept Sets

CLICS 2.0
Excursus
http://concepticon.clld.org
## Excursus: Data

<table>
<thead>
<tr>
<th>DATASET</th>
<th>EDITORS</th>
<th>LANGUAGES</th>
<th>CONCEPTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDS</td>
<td>Key and Comrie (2016)</td>
<td>367</td>
<td>1310</td>
</tr>
<tr>
<td>WOLD</td>
<td>Haspelmath and Tadmor (2008)</td>
<td>41</td>
<td>1430</td>
</tr>
<tr>
<td>BaiDial*</td>
<td>Allen (2007)</td>
<td>8</td>
<td>500</td>
</tr>
<tr>
<td>HuberReed</td>
<td>Huber and Reed (1992)</td>
<td>71</td>
<td>374</td>
</tr>
<tr>
<td>BantuBVD*</td>
<td>Teil-Dautrey (2008)</td>
<td>10</td>
<td>430</td>
</tr>
<tr>
<td>Tryon1983*</td>
<td>Tryon (1983)</td>
<td>111</td>
<td>324</td>
</tr>
<tr>
<td>Madang*</td>
<td>Zgraggen (1980)</td>
<td>100</td>
<td>380</td>
</tr>
<tr>
<td>Cihui*</td>
<td>Beijing Daxue (1964)</td>
<td>17</td>
<td>905</td>
</tr>
<tr>
<td>TBL*</td>
<td>Huang (1992)</td>
<td>50</td>
<td>1800</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>Dellert and Jäger (2017)</td>
<td>106</td>
<td>1000</td>
</tr>
</tbody>
</table>

Datasets with an asterisk are currently in preparation and will be most likely released already within this year.
Excursus: Data

By linking these datasets to the Concepticon (which we have already done with most of them), we can easily combine the data into a bigger dataset that we use as our basic data for CLICS 2.0. Given problems with concept overlap in the datasets, we can make different selections for the users, including datasets with many concepts but not many languages and datasets with many languages but less concepts.
By linking these datasets to the Concepticon (which we have already done with most of them), we can easily combine the data into a bigger dataset that we use as our basic data for CLICS 2.0.
Excursus: Data

- By linking these datasets to the Concepticon (which we have already done with most of them), we can easily combine the data into a bigger dataset that we use as our basic data for CLICS 2.0.
- Given problems with concept overlap in the datasets, we can make different selections for the users, including datasets with many concepts but not so many languages and datasets with many languages but less concepts.
Excursus: Data

<table>
<thead>
<tr>
<th>Subset</th>
<th>Datasets</th>
<th>Concepts</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Low</td>
<td>&gt;= 2</td>
<td>&gt;= 1000</td>
<td>&gt;= 300</td>
</tr>
<tr>
<td>Mid-Mid</td>
<td>&gt;= 5</td>
<td>&gt;= 500</td>
<td>&gt;= 600</td>
</tr>
<tr>
<td>Low-High</td>
<td>&gt;= 10</td>
<td>&gt;= 250</td>
<td>&gt;= 1000</td>
</tr>
</tbody>
</table>
Effectively this means, that with CLICS 2.0, we can immediately offer different views on the data, which allow scholars to investigate very broad patterns of semantic associations, as well as fine-grained patterns with a lower attestation.
Excursus: Software API

With the Python API that we are currently preparing for CLICS 2.0, users will be able to use their own data or run their own network analyses, since all data is shipped with CLICS. Users can also use the data we selected for CLICS 2.0. We will try to offer cookbooks accompanying the software API, to help users use it efficiently.

By shifting to the CLLD framework, scholars can also create their own CLICS websites, since the source code for the creation of interactive networks will be transparently shipped with the data.

Springs schools and further events carried out at the MPI-SH have part of my ERC project on Computer-Assisted Language Comparison will cover – among others – introductory tutorials to all the software APIs that are shipped with the different tools and datasets developed at our department.
Excursus: Software API

With the Python API that we are currently preparing for CLICS 2.0, users will be able to use their own data to run their own network analyses, since all data is shipped with CLICS, users can also use the data we selected for CLICS 2.0.
Excursus: Software API

- With the Python API that we are currently preparing for CLICS 2.0, users will be able to use their own data to run their own network analyses, since all data is shipped with CLICS, users can also use the data we selected for CLICS 2.0.

- We will try to offer cookbooks accompanying the software API, to help users to use it efficiently.
Excursus: Software API

- With the Python API that we are currently preparing for CLICS 2.0, users will be able to use their own data to run their own network analyses, since all data is shipped with CLICS, users can also use the data we selected for CLICS 2.0.
- We will try to offer cookbooks accompanying the software API, to help users to use it efficiently.
- By shifting to the CLLD framework, scholars can also create their own CLICS websites, since the source code for the creation of interactive networks will be transparently shipped with the data.
Excursus: Software API

- With the Python API that we are currently preparing for CLICS 2.0, users will be able to use their own data to run their own network analyses, since all data is shipped with CLICS, users can also use the data we selected for CLICS 2.0.

- We will try to offer cookbooks accompanying the software API, to help users to use it efficiently.

- By shifting to the CLLD framework, scholars can also create their own CLICS websites, since the source code for the creation of interactive networks will be transparently shipped with the data.

- Spring schools and further events carried out at the MPI-SHH as part of my ERC project on Computer-Assisted Language Comparison will cover – among others – introductory tutorials to all the software APIs that are shipped with the different tools and datasets developed at our department.
<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICS 2.0</td>
</tr>
<tr>
<td>Features</td>
</tr>
</tbody>
</table>

- Drastic increase in data
- Drastic increase in transparency
- Drastic increase in replicability
- Regular floating releases which feature new data
- Strict and clear-cut collaboration guidelines
- New methods (see demo on next slide)
- Rigid policy towards open data (since we heavily profit from all of our colleagues who publish their data!)
Features

- drastic increase in data
Features

- drastic increase in data
- drastic increase in transparency
Features

- drastic increase in data
- drastic increase in transparency
- drastic increase in replicability
Features

- drastic increase in data
- drastic increase in transparency
- drastic increase in replicability
- regular floating releases which feature new data
Features

- drastic increase in data
- drastic increase in transparency
- drastic increase in replicability
- regular floating releases which feature new data
- strict and clear-cut collaboration guidelines
Features

- drastic increase in data
- drastic increase in transparency
- drastic increase in replicability
- regular floating releases which feature new data
- strict and clear-cut collaboration guidelines
- new methods (see demo on next slide)
Features

- drastic increase in data
- drastic increase in transparency
- drastic increase in replicability
- regular floating releases which feature new data
- strict and clear-cut collaboration guidelines
- new methods (see demo on next slide)
- rigid policy towards open data (since we heavily profit from all of our colleagues who publish their data!)
Features: Coverage
New Methods

Following Urban (2011) we are currently testing an automated variant of partial colexifications which can help us to direct our networks and shed light on compositional aspects of semantic associations. By improving our insights into graph theory and available algorithms, we can now enhance the analysis of the networks. Articulation points, for example, show key players in a network which connect between different communities.
New Methods

Following Urban (2011) we are currently testing an automatized variant of *partial colexifications* which can help us to direct our networks and shed light on compositional aspect of semantic associations.
New Methods

- Following Urban (2011) we are currently testing an automatized variant of *partial colexifications* which can help us to direct our networks and shed light on compositional aspect of semantic associations.

- By improving our insights into graph theory and available algorithms, we can now enhance the analysis of the networks. *Articulation points*, for example, show key players in a network which connect between different communities.
New Methods
New Methods

Undirected network in which articulation points are marked.
CLICS 2.0 DEMO
We are working hard on assembling more data and building up the new API as well as the web interface, but we are currently not many who work on CLICS or in its periphery. We hope that we can publish CLICS 2.0 very late this year, and in a worst case, in early 2018. But we would argue that it is better to publish the next version a bit later rather than publishing a version that we will need to update immediately after we first publish it.

If we can learn one thing from CLICS 1.0, it is that we need to keep the code and the data in a state that we can easily curate them. We hope we will achieve this with CLICS 2.0.
Schedule

- We are working hard on assembling more data and building up the new API as well as the web-interface, but we are currently not many who work on CLICS or in its periphery.

We hope that we can publish CLICS 2.0 very late this year, and in a worst case, in early 2018. But we would argue that it is better to publish the next version a bit later rather than publishing a version that we will need to update immediately after we first publish it. If we can learn one thing from CLICS 1.0, it is that we need to keep the code and the data in a state that we can easily curate them. We hope we will achieve this with CLICS 2.0.
Schedule

- We are working hard on assembling more data and building up the new API as well as the web-interface, but we are currently not many who work on CLICS or in its periphery.

- We hope that we can publish CLICS 2.0 very late this year, and in a worst case, in early 2018.
We are working hard on assembling more data and building up the new API as well as the web-interface, but we are currently not many who work on CLICS or in its periphery.

We hope that we can publish CLICS 2.0 very late this year, and in a worst case, in early 2018.

But we would argue that it is better to publish the next version a bit later rather than publishing a version that we will need to update immediately after we first published it.
We are working hard on assembling more data and building up the new API as well as the web-interface, but we are currently not many who work on CLICS or in its periphery.

We hope that we can publish CLICS 2.0 very late this year, and in a worst case, in early 2018.

But we would argue that it is better to publish the next version a bit later rather than publishing a version that we will need to update immediately after we first published it.

If we can learn one thing from CLICS 1.0, it is that we need to keep the code and the data in a state that we can easily curate them. We hope we will achieve this with CLICS 2.0.
Outlook
It is still a rather long way from CLICS 1.0 to CLICS 2.0.
It is still a rather long way from CLICS 1.0 to CLICS 2.0.
But we hope that we are on the right track by now, and that won’t disappoint those who came to like the Cross-Linguistic Colexification Database.
It is still a rather long way from CLICS 1.0 to CLICS 2.0. But we hope that we are on the right track by now, and that won’t disappoint those who came to like the Cross-Linguistic Colexification Database.

CLICS 2.0 won’t be perfect, but it will be entertaining and hopefully very interesting for our colleagues working on historical linguistics and lexical typology.
Thanks for your attention!