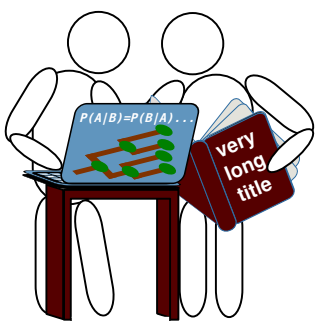
CLICS 2.0

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



MAX-PLANCK-GESELLSCHAFT

European Research Council

A long, long time ago...

Predecessors: People and Ideas

Predecessors: People and Ideas

- Haspelmath (2003): The geometry of grammatical meaning.

Predecessors: People and Ideas

- Haspelmath (2003): The geometry of grammatical meaning.
- François (2008): Semantic maps and the typology of colexification.

Predecessors: People and Ideas

- Haspelmath (2003): The geometry of grammatical meaning.
- François (2008): Semantic maps and the typology of colexification.
- Cysouw (2010): Drawing networks from recurrent polysemies.

Predecessors: People and Ideas

- Haspelmath (2003): The geometry of grammatical meaning.
- François (2008): Semantic maps and the typology of colexification.
- Cysouw (2010): Drawing networks from recurrent polysemies.
- Steiner, Stadler, and Cysouw (2011): A pipeline for computational historical linguistics.

Predecessors: People and Ideas

- Haspelmath (2003): The geometry of grammatical meaning.
- François (2008): Semantic maps and the typology of colexification.
- Cysouw (2010): Drawing networks from recurrent polysemies.
- Steiner, Stadler, and Cysouw (2011): A pipeline for computational historical linguistics.
- Urban (2011): Assymetries in overt marking and directionality in semantic change.

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

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.

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 analyse the network, and to partition the concepts into communities which seem to make intuitively sense, reminding of naturally derived semantic fields.

Further Ideas

Further Ideas

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

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*).

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

Data

Data

- IDS (Key and Comrie 2007 version), of 233 language varieties, 178 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.

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.

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.
- 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.

Methods

Problems

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

Interface

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

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`.

DEMO

CLICS 2.0

Motivation

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)

Ideas

Ideas

- 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)

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)

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

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

Excursus: Concepticon

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

Excursus: Concepticon

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

Excursus: Concepticon

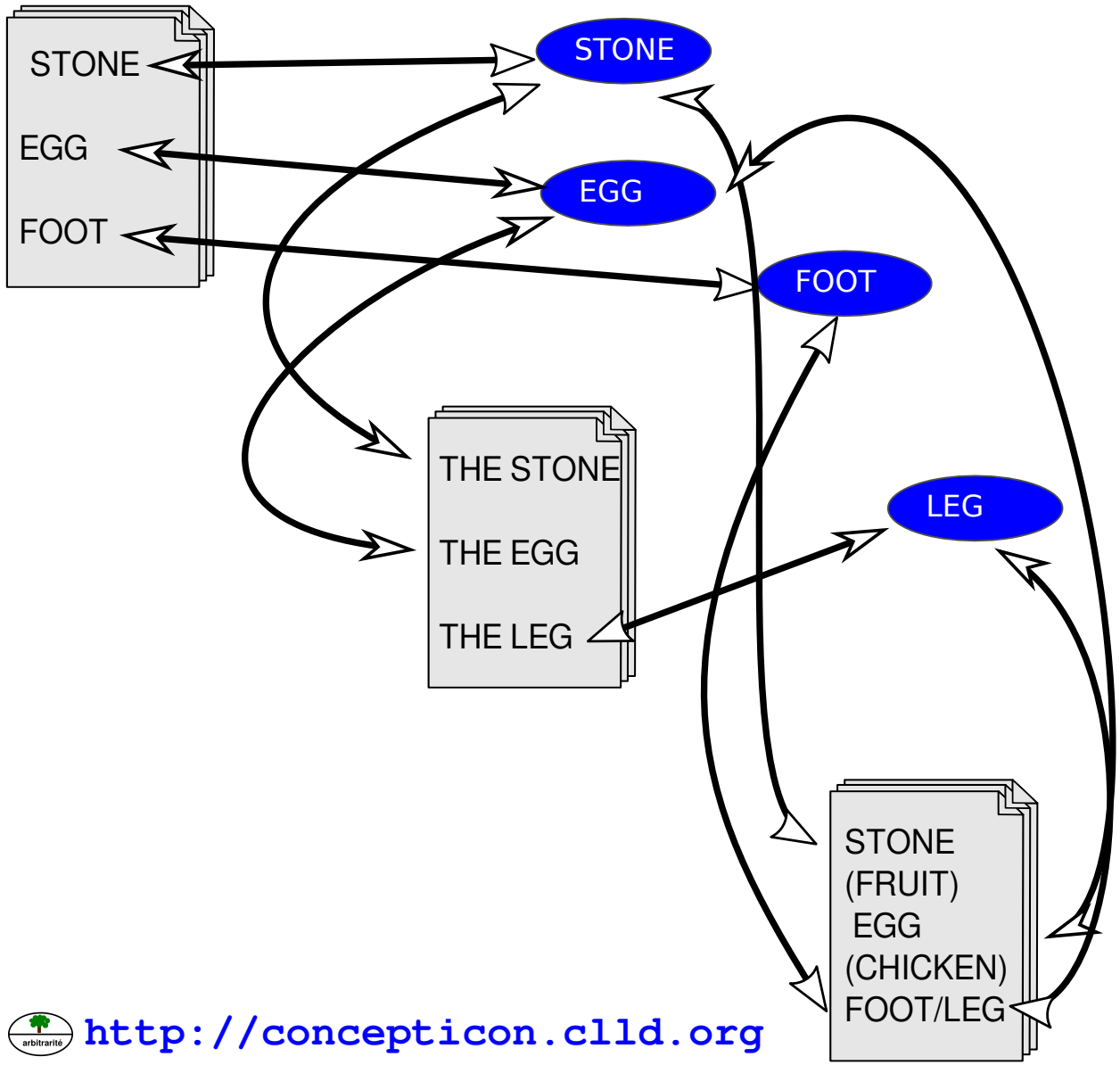
Concept List	# Items	Concept Label	Concept ID
Allen (2007)	500	animal oil; 动物油（脂肪）	GREASE (CONCEPTICON-ID:323)
Gregersen (1976)	217	fat-grease*fat-grease	GREASE (CONCEPTICON-ID:323)
Heggarty (2005)	150	fat (grease); grasa	GREASE (CONCEPTICON-ID:323)
Swadesh (1955)	100	fat (grease)	GREASE (CONCEPTICON-ID:323)
Alpher and Nash (1999)	151	fat, grease	GREASE (CONCEPTICON-ID:323)
Hale (1961)	100	fat, grease	GREASE (CONCEPTICON-ID:323)
OGrady and Klokeid (1969)	100	fat, grease	GREASE (CONCEPTICON-ID:323)
Blust (2008)	210	fat/grease	GREASE (CONCEPTICON-ID:323)
Matisoff (1978)	200	fat/grease	GREASE (CONCEPTICON-ID:323)
Samarin (1969)	218	fat/grease	GREASE (CONCEPTICON-ID:323)
Dunn et al. (2012)	207	fat	GREASE (CONCEPTICON-ID:323)
Swadesh (1950)	215	fat	GREASE (CONCEPTICON-ID:323)
Zgraggen (1980)	380	fat	GREASE (CONCEPTICON-ID:323)
Jachontov (1991)	100	fat n.	GREASE (CONCEPTICON-ID:323)
Wiktionary (2003)	207	fat (noun)	GREASE (CONCEPTICON-ID:323)
Starostin (1991)	110	fat n.; жир	GREASE (CONCEPTICON-ID:323)
TeilDautrey et al. (2008)	430	fat, oil	GREASE (CONCEPTICON-ID:323)
Swadesh (1952)	200	fat (organic substance)	GREASE (CONCEPTICON-ID:323)
Shiro (1973)	200	grease (fat)	GREASE (CONCEPTICON-ID:323)
Samarin (1969)	100	grease; graisse; Fett; grasa	GREASE (CONCEPTICON-ID:323)
Wang (2006)	200	pig oil; 猪油	GREASE (CONCEPTICON-ID:323)
Haspelmath and Tadmor (2009)	1460	the grease or fat	GREASE (CONCEPTICON-ID:323)

Excursus: Concepticon

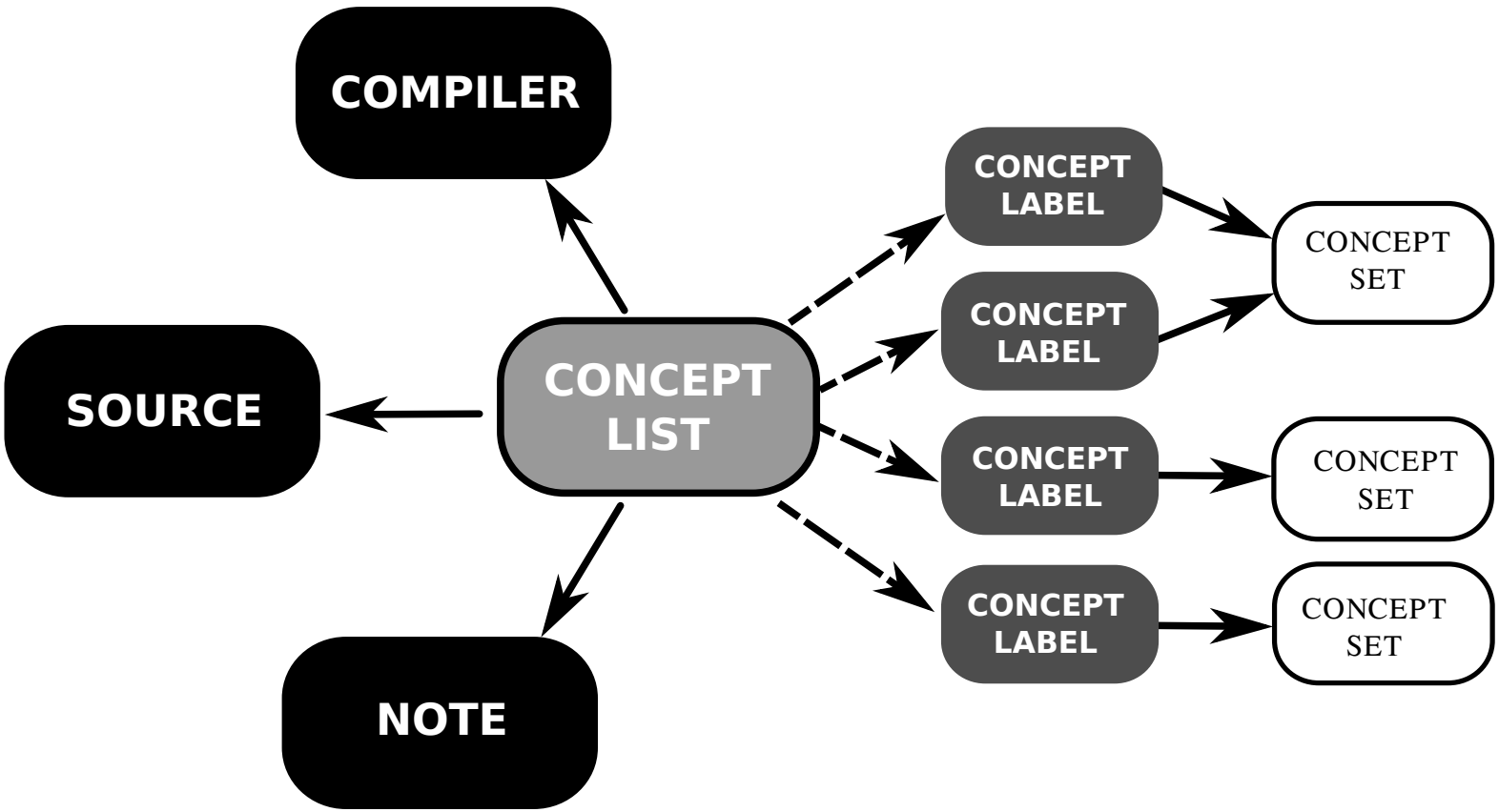
Concepticon (List et al. 2016)

- link **concept labels** in published **concept lists** (questionnaires) to **concept sets**
- link **concept sets** to meta-data
- define relations between concept sets
- never link one concept in a given list to more than one concept set (guarantees consistency)
- provide an API to check the consistency of the data and to query the data
- provide a web-interface to browse through the data

Concepticon



Concepticon



http://concepticon.clld.org

Excursus: Data

DATASET	EDITORS	LANGUAGES	CONCEPTS
IDS	Key and Comrie (2016)	367	1310
WOLD	Haspelmath and Tadmor (2008)	41	1430
BaiDial*	Allen (2007)	8	500
HuberReed	Huber and Reed (1992)	71	374
Kraft1981	Kraft (1981)	68	434
BantuBVD*	Teil-Dautrey (2008)	10	430
Tryon1983*	Tryon (1983)	111	324
Madang*	Zgraggen (1980)	100	380
Cihui*	Beijing Daxue (1964)	17	905
TBL*	Huang (1992)	50	1800
NorthEuraLex	Dellert and Jäger (2017)	106	1000

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

Excursus: Data

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.

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

Subset	Datasets	Concepts	Languages
High-Low	>= 2	>= 1000	>= 300
Mid-Mid	>= 5	>= 500	>= 600
Low-High	>= 10	>= 250	>= 1000

Excursus: Data

Subset	Datasets	Concepts	Languages
High-Low	>= 2	>= 1000	>= 300
Mid-Mid	>= 5	>= 500	>= 600
Low-High	>= 10	>= 250	>= 1000

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

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.

Features

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

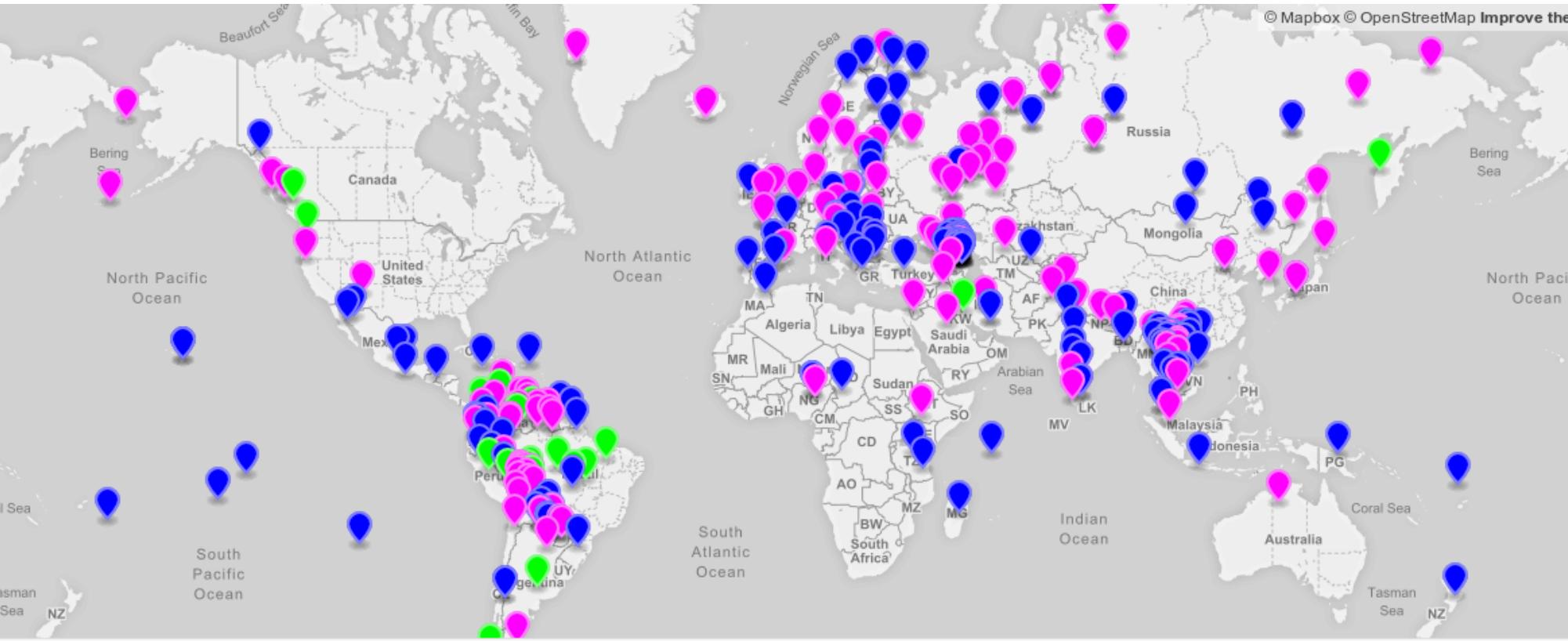
Features

- 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

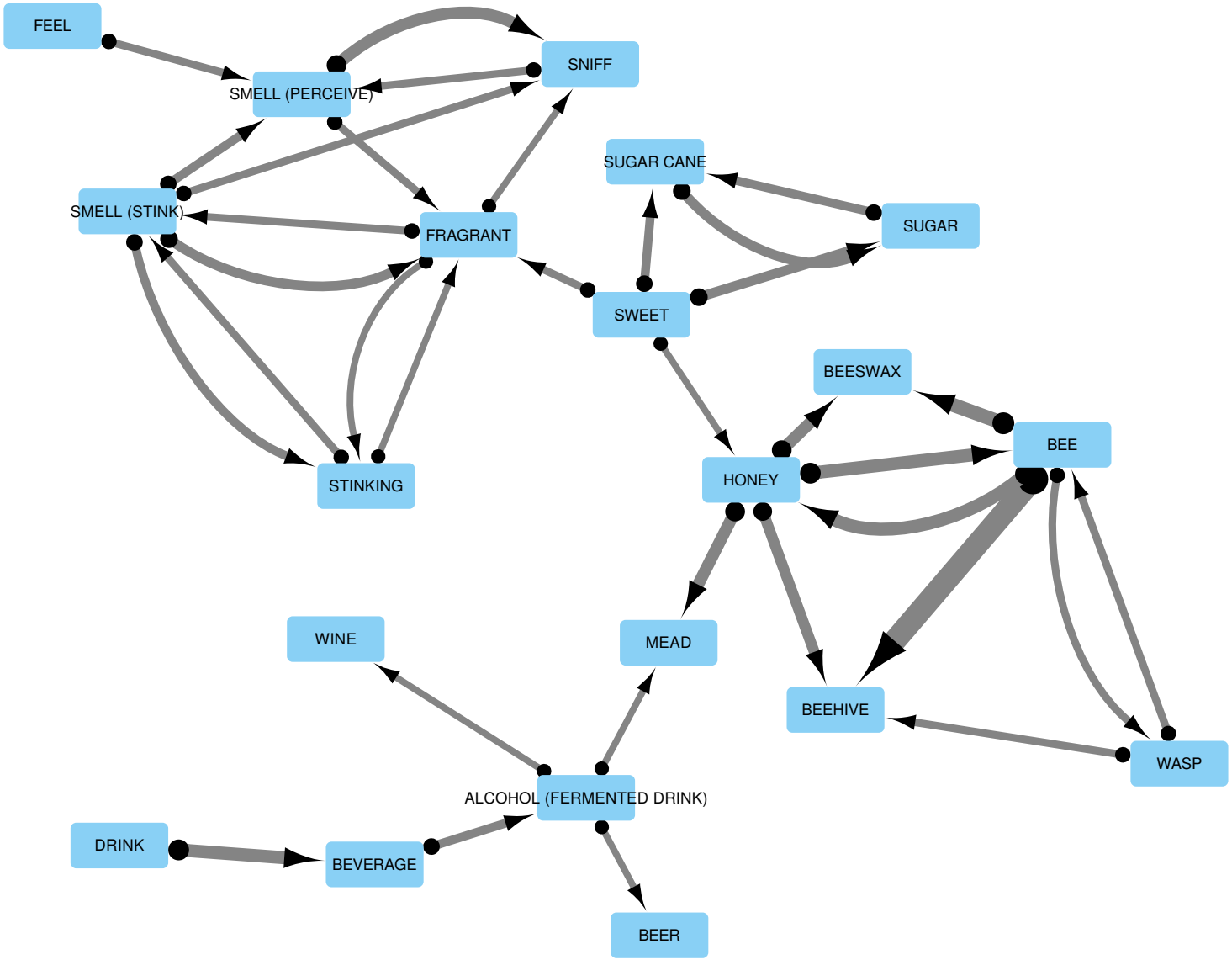
New Methods

- 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.

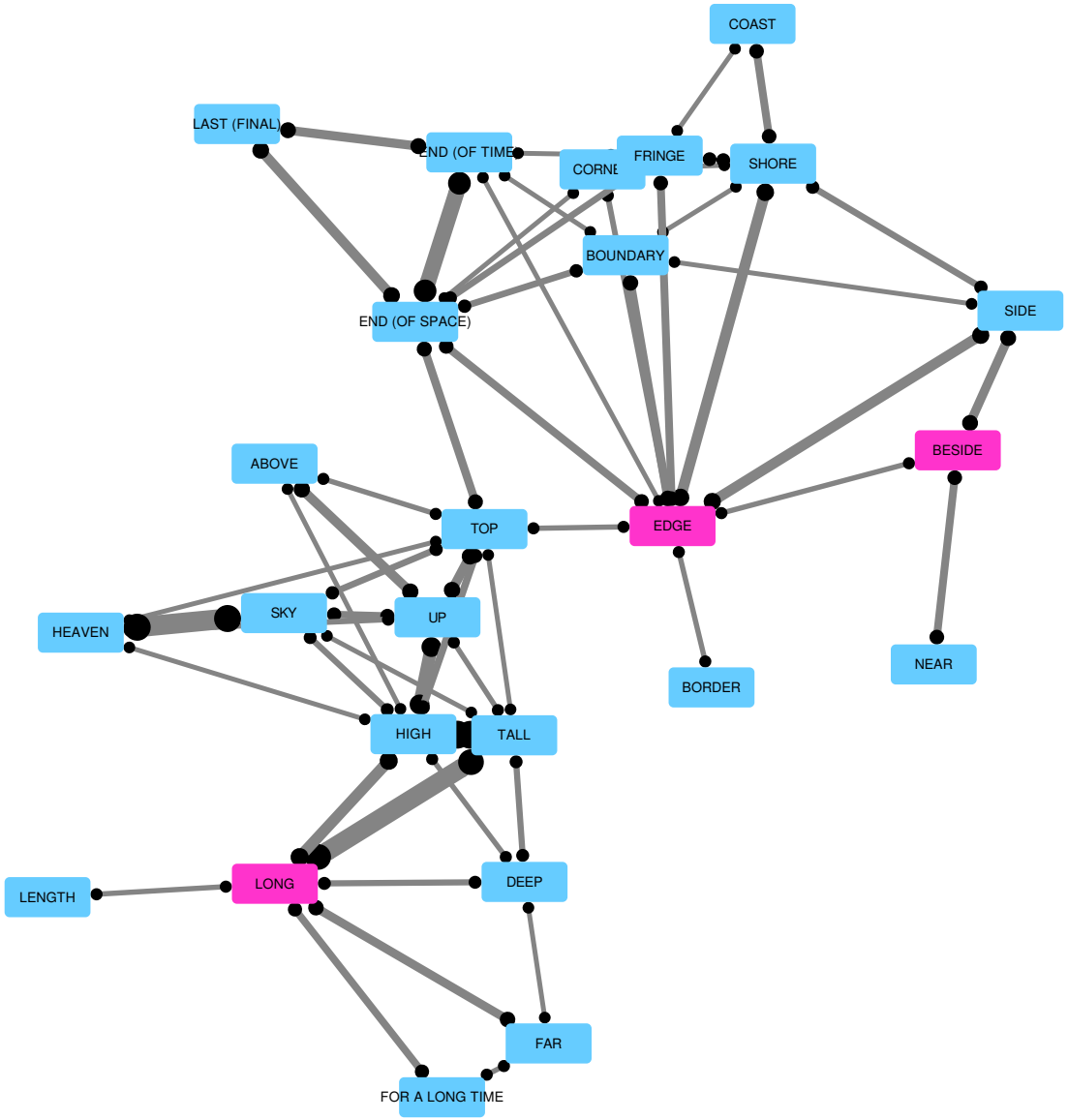
New Methods

- 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


FEEL
SNIFF
SMELL (PERCEIVE)
SUGAR CANE
SMELL (STINK)
FRAGRANT
SUGAR
SWEET
BEESWAX
BEE
HONEY
STINKING
WINE
MEAD
BEEHIVE
WASP
ALCOHOL (FERMENTED DRINK)
DRINK
BEVERAGE
BEER

New Methods


COAST
LAST (FINAL)
END (OF TIME
CORNE
FRINGE
SHORE
BOUNDARY
SIDE
END (OF SPACE)
BESIDE
ABOVE
TOP
EDGE
HEAVEN
SKY
UP
NEAR
BORDER
HIGH
TALL
LENGTH
LONG
DEEP
FAR
FOR A LONG TIME

CLICS 2.0 DEMO

Schedule

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.

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.

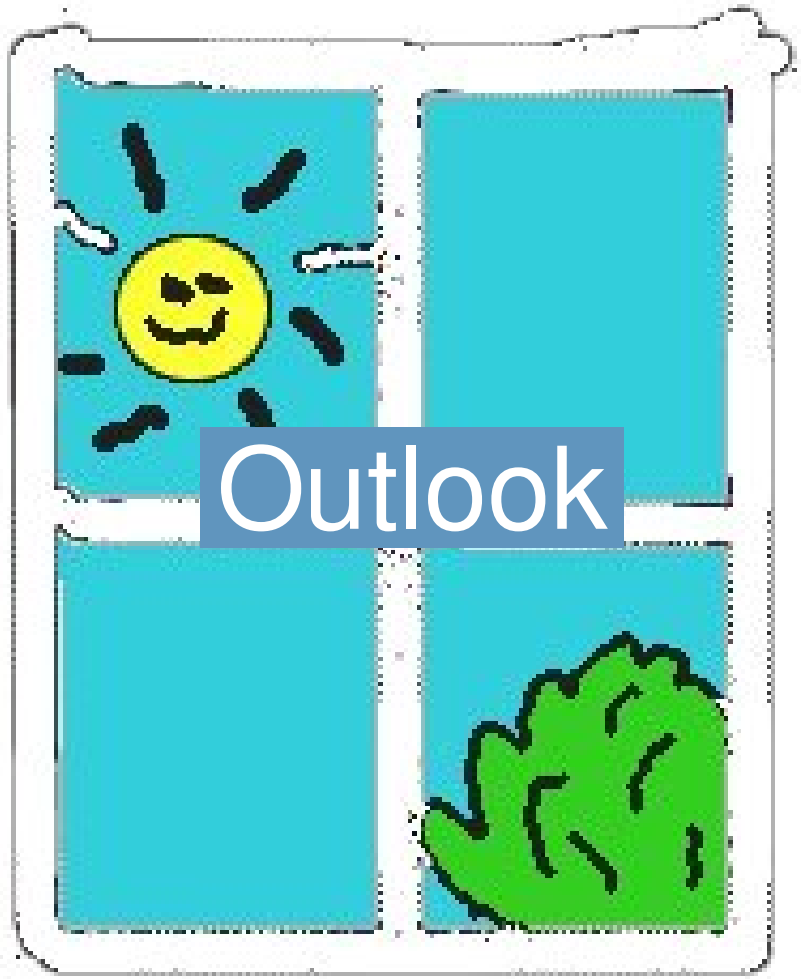
Schedule

- 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 published it.

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 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!