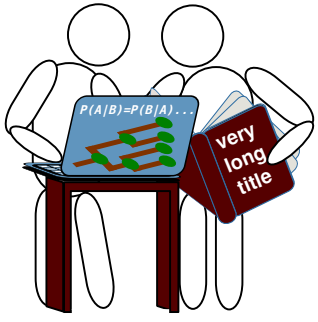
CLICS²

A computer-assisted framework for the investigation of lexical motivation 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

2018-06-27



MAX-PLANCK-GESELLSCHAFT

European
Research
Council

From Semantic Maps...

... to Cross-Linguistic Polysemies

Early Accounts: People and Ideas

Early Accounts: People and Ideas

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

Early Accounts: People and Ideas

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

Early Accounts: 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.

Early Accounts: 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.

Early Accounts: 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.

Early Accounts: Data

Early Accounts: Data

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

Early Accounts: 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).

Early Accounts: 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.

Early Accounts: 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 increase the data by merging IDS, WOLD, and additional datasets, 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 increase the data by merging IDS, WOLD, and additional datasets, thus containing 220 languages in total.
- They also improve 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`.

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`.
- The full wordlists underlying the original CLICS database are now also available from Zenodo (published in List 2018, `https://zenodo.org/record/1194088`).

CLICS DEMO

CLICS^2

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 things 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 things 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 up datasets, 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 things 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 up datasets, 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 things 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 up datasets, 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 wordlist data

Ideas

- use the state of the art of available wordlist data
- separate data from display (CLICS² does not host data, but simply uses it)

Ideas

- use the state of the art of available wordlist data
- separate data from display (CLICS² does not host data, but simply uses it)
- curate data following the recommendations developed for the Cross-Linguistic Data Formats (CLDF, `http://cldf.clld.org`) initiative (Forkel et al. 2017)

Ideas

- use the state of the art of available wordlist data
- separate data from display (CLICS² does not host data, but simply uses it)
- curate data following the recommendations developed for the Cross-Linguistic Data Formats (CLDF, `http://cldf.clld.org`) initiative (Forkel et al. 2017)
- curate the code and the data with help of a transparent API

Ideas

- use the state of the art of available wordlist data
- separate data from display (CLICS² does not host data, but simply uses it)
- curate data following the recommendations developed for the Cross-Linguistic Data Formats (CLDF, `http://cldf.clld.org`) initiative (Forkel et al. 2017)
- curate the code and the 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 wordlist data
- separate data from display (CLICS² does not host data, but simply uses it)
- curate data following the recommendations developed for the Cross-Linguistic Data Formats (CLDF, `http://cldf.clld.org`) initiative (Forkel et al. 2017)
- curate the code and the 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)

Excursus: The Cross-Linguistic Data Initiative

Cross-Linguistic Data Formats (Forkel et al. 2017)

- aims at increasing the comparability of cross-linguistic data and analyses
- supports methods for standardization via reference catalogues like Glottolog (Hammarström et al. 2018) and Concepticon (List et al. 2017)
- provides software APIs which help to test whether data conforms to standards
- offers working examples for best practice
- supported by different software frameworks (LingPy, BEASTling, EDICTOR)

CLDFDEMO

Excursus: Reference Catalogues

- The advantages of linking one's data to reference catalogs like Glottolog (Hammarström et al. 2018, `http://glottolog.org`) are obvious: Since Glottolog harvests various types of additional information regarding language varieties all over the world that can be used effortlessly, once linked.

Excursus: Reference Catalogues

- The advantages of linking one's data to reference catalogs like Glottolog (Hammarström et al. 2018, `http://glottolog.org`) are obvious: Since Glottolog harvests various types of additional information regarding language varieties all over the world that can be used effortlessly, once linked.
- The Concepticon project (`http://concepticon.clld.org`, List et al. 2016, List et al. 2018) is much less well known among scholars, but it offers the same advantages when dealing with wordlist data that was built by means of a questionnaire of "elicitation glosses".

Excursus: Concepticon

Concepticon (List et al. 2016)

- link **concept labels** (“elicitation glosses”) 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

Concept Set FAT (ORGANIC SUBSTANCE)

Esters of three fatty acid chains and the alcohol glycerol which form a semi-solid substance in room temperature and occur in animals and plants.

Related concept sets

FAT (ORGANIC SUBSTANCE)	narrower	FAT (FOR NOURISHMENT)
ORGANIC FAT OR OIL	narrower	FAT (ORGANIC SUBSTANCE)

ID	Concept in Source	English Gloss	Conceptlist
Alpher-1999-151-27	fat, grease [english]		Alpher 1999 151
He-2010-207-145	脂肪 [chinese]	fat	He 2010 207
Janhunan-2008-235-96	fat / grease [english]		Janhunan 2008 235
Gudschinsky-1956-200-42	fat-grease [english]		Gudschinsky 1956 200
Swadesh-1952-200-43	fat (organic substance) [english]		Swadesh 1952 200
Swadesh-1955-100-26	fat (grease) [english]		Swadesh 1955 100
...	...	...	...

Concepticon

Selected language: en (●) English ( ) German ( ) Chinese ( ) Russian ( ) French ( ) Portuguese ( ) Spanish

fece|

MATCH	ID	GLOSS	DEFINITION	SIMILARITY
face	1560	FACE	The front part of the head, featuring the eyes, nose, and mouth and the surrounding area.	3
feces	675	FAECES (EXCREMENT)	Substance that human and animal bodies release from time to time as a little pile of waste remaining from digestion, after it has been collected in the colon.	3
fence	1690	FENCE	Delimitation for an area.	3

CONCEPTICON DEMO

Excursus: Data in CLDF

#	Dataset	Source	Range	Glosses	Concepticon	Varieties	Glottolog	Families
1	allenbai	Allen (2007)	Bai (ST)	500	499	9	9	1
2	bantubvd	Greenhill & Gray (2015)	Bantu	430	415	10	9	1
3	beidasinitic	Běijīng Dàxué (1964)	Sinitic (ST)	905	700	18	18	1
4	bowernpny	Bowern & Atkinson (2011)	Pama-Nyungan	348	342	171	164	2
5	hubercolumbian	Huber & Reed (1992)	Colombian	374	343	69	65	16
6	ids	Key & Comrie (2016)	World-wide	1305	1305	324	234	61
7	kraft	Kraft (1981)	Chadic	434	428	67	60	3
8	northeuralex	Dellert & Jäger (2017)	North-Eurasian	1016	940	107	105	21
9	robinsonap	Robinson & Holton (2012)	Alor-Pantar	398	386	13	11	1
10	satterthwaitetb	Satterthwaite-Phillips (2011)	Sino-Tibetan	423	418	18	15	1
11	sunztb	Sūn (1991)	Sino-Tibetan	1005	906	50	44	1
12	tls	Nurse and Phillipson (1975)	Tanzanian	1533	808	131	97	1
13	tryonsolomon	Tryon and Hackman (1983)	Solomon Islands	324	311	111	96	5
14	wold	Haspelmath & Tadmor (2009)	World-wide	1460	1457	41	40	25
15	zgraggenmadang	Z'graggen (1980abcd)	Madang	336	306	100	98	1
	TOTAL / OVERLAP				2482	1266	1036	91

Datasets are all released under https://zenodo.org/communities/clics.

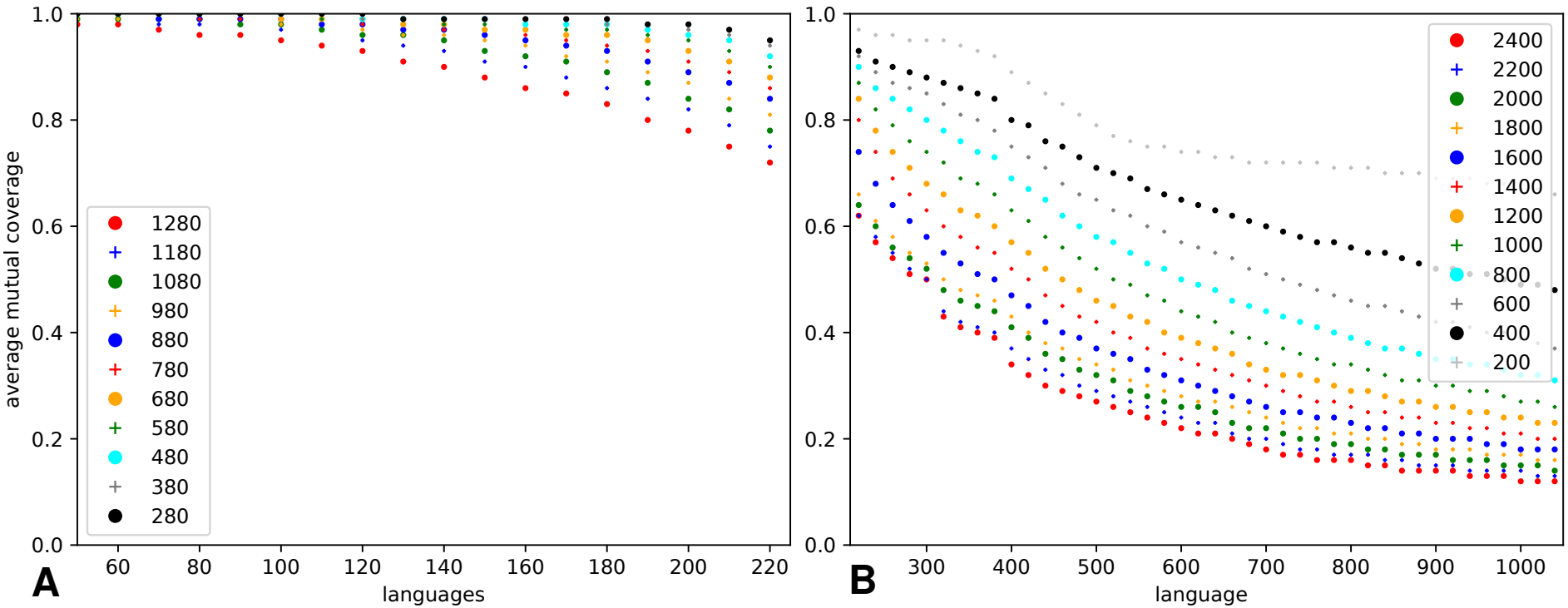
Excursus: Data in CLDF

- Since our datasets are all available in CLDF format, we can easily aggregate them for our new version of CLICS².

Excursus: Data in CLDF

- Since our datasets are all available in CLDF format, we can easily aggregate them for our new version of CLICS².
- Given problems with concept overlap in the datasets, we offer code examples that can be used to compute mutual coverage statists allowing users to select subsets of the data optimal for a given analysis.

Excursus: Data in CLDF



Excursus: Software API

Excursus: Software API

- With the Python API that we have prepared for CLICS² (`https://github.com/clics/clics2`), users are able to use their own data to run their own network analyses. Since all data for CLICS² is independently shared and curated, users can also use the data we selected for CLICS² but test different parameters of our API.

Excursus: Software API

- With the Python API that we have prepared for CLICS² (`https://github.com/clics/clics2`), users are able to use their own data to run their own network analyses. Since all data for CLICS² is independently shared and curated, users can also use the data we selected for CLICS² but test different parameters of our API.
- We offer examples of how the data we use for CLICS² can be computed with help of the API and plan to make them available in form of code cookbooks.

Excursus: Software API

- With the Python API that we have prepared for CLICS² (`https://github.com/clics/clics2`), users are able to use their own data to run their own network analyses. Since all data for CLICS² is independently shared and curated, users can also use the data we selected for CLICS² but test different parameters of our API.
- We offer examples of how the data we use for CLICS² can be computed with help of the API and plan to make them available in form of code cookbooks.
- By shifting to the CLLD framework, scholars can also create their own CLICS websites, since the source code for the creation of interactive networks is transparently shipped with the data.

Features: Summary

Features: Summary

- drastic increase in data

Features: Summary

- drastic increase in data
- drastic increase in transparency

Features: Summary

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

Features: Summary

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

Features: Summary

- 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

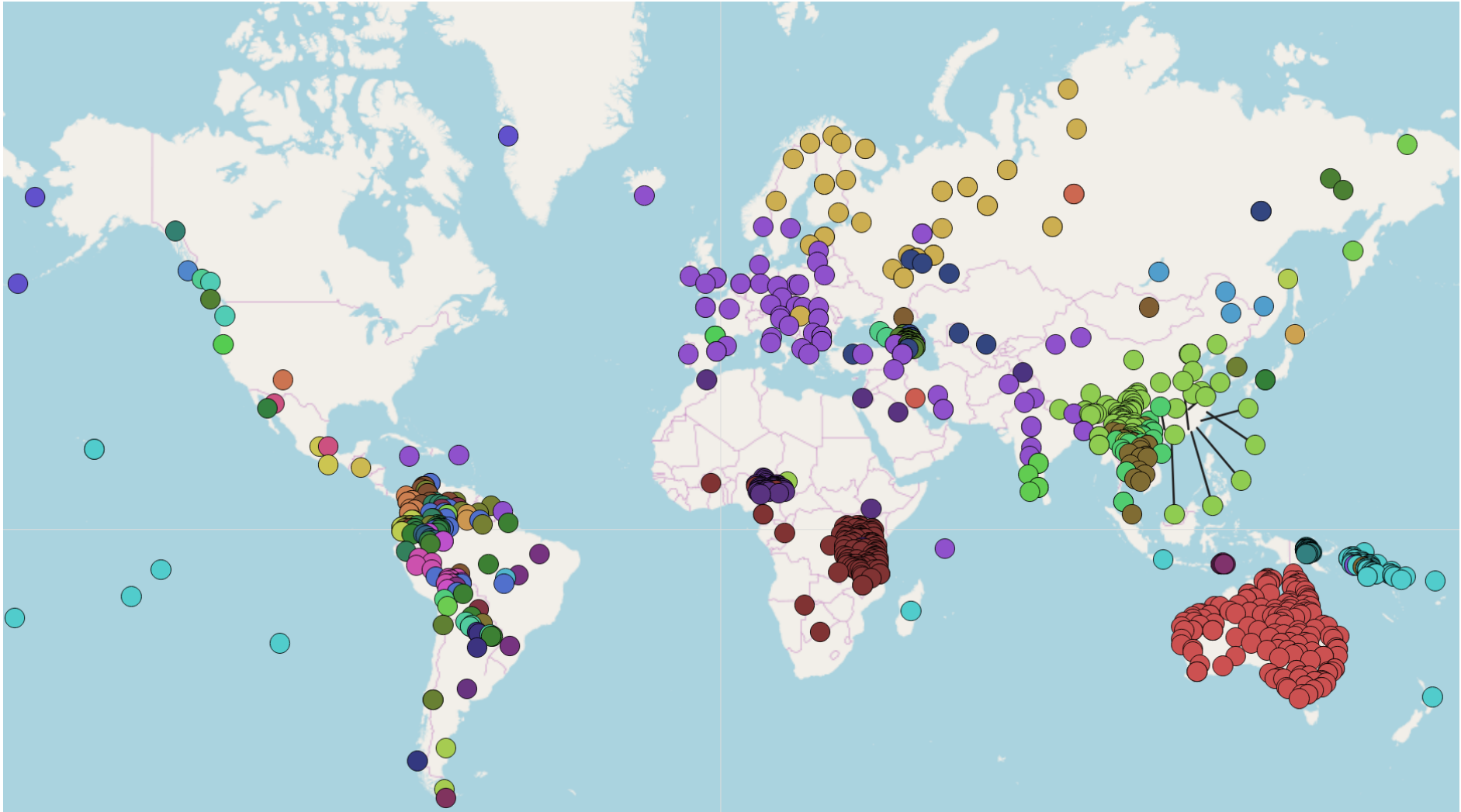
Features: Summary

- 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: Summary

- 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



Features: Enhanced Browsing

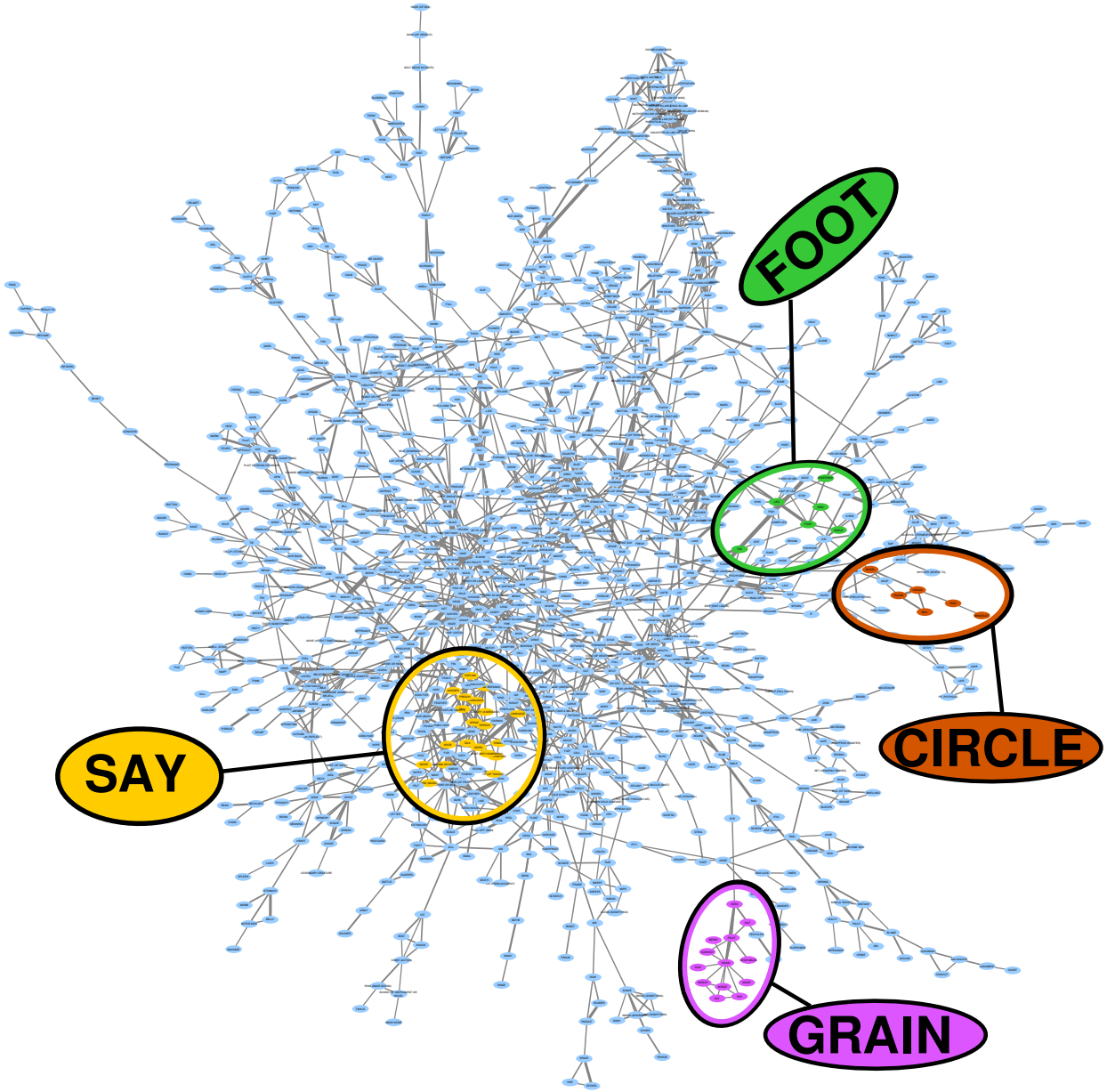
Features: Enhanced Browsing

- Thanks to the CLLD framework, the data is now much easier to browse, and all data is clearly linked to the original datasets.

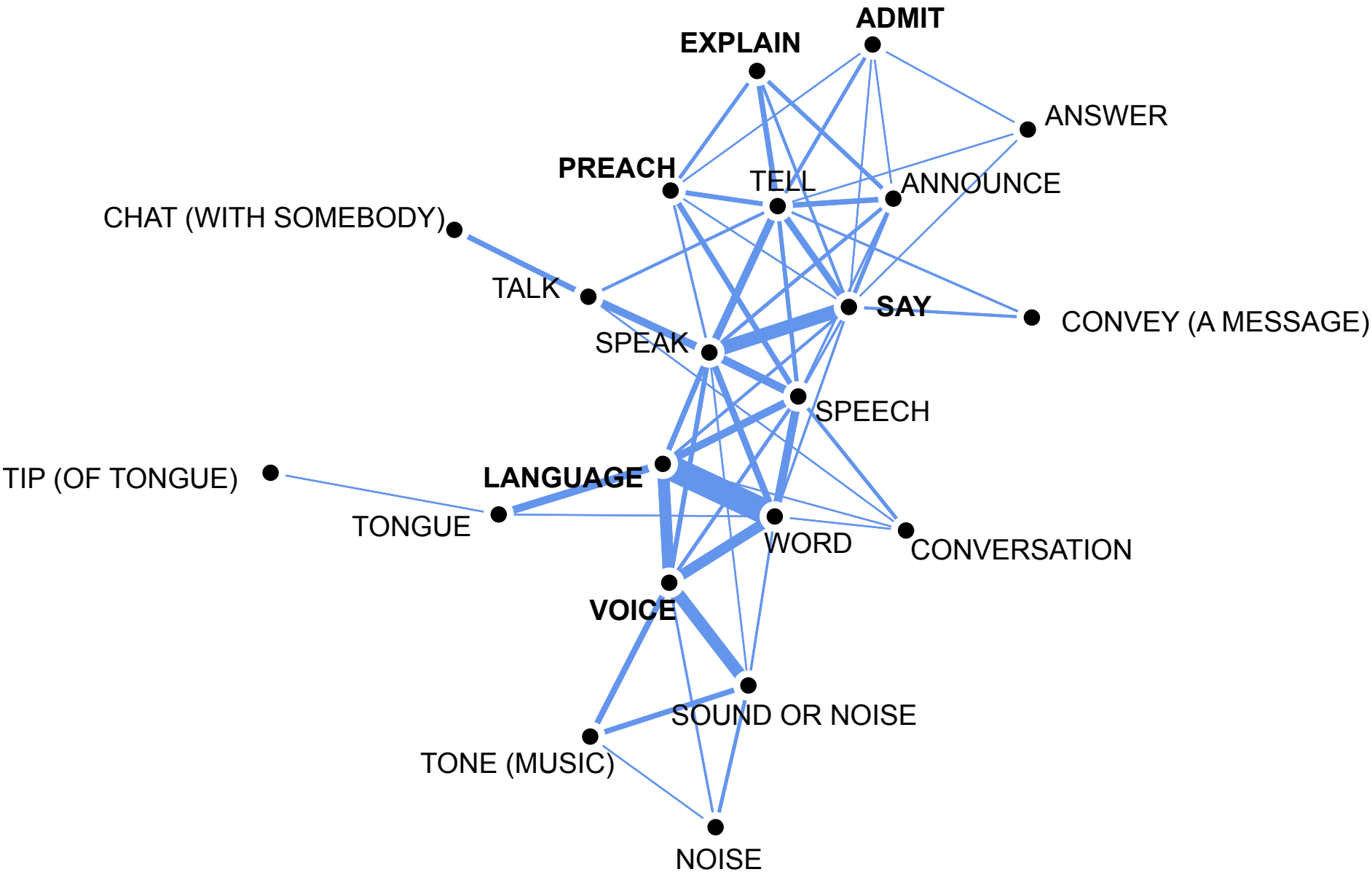
Features: Enhanced Browsing

- Thanks to the CLLD framework, the data is now much easier to browse, and all data is clearly linked to the original datasets.
- Thanks to a standalone app that can be created from our data in pure HTML format, users can still browse CLICS² data with the old look-and-feel, and even use the standalone application to deploy their own data in form of CLICS networks.
- In addition, we are currently experimenting with a new visualization that allows users to inspect the CLICS² network in all its complexity, following visualization methods developed for the inspection of Galaxies (contributed by Thomas Mayer).

Features: Examples

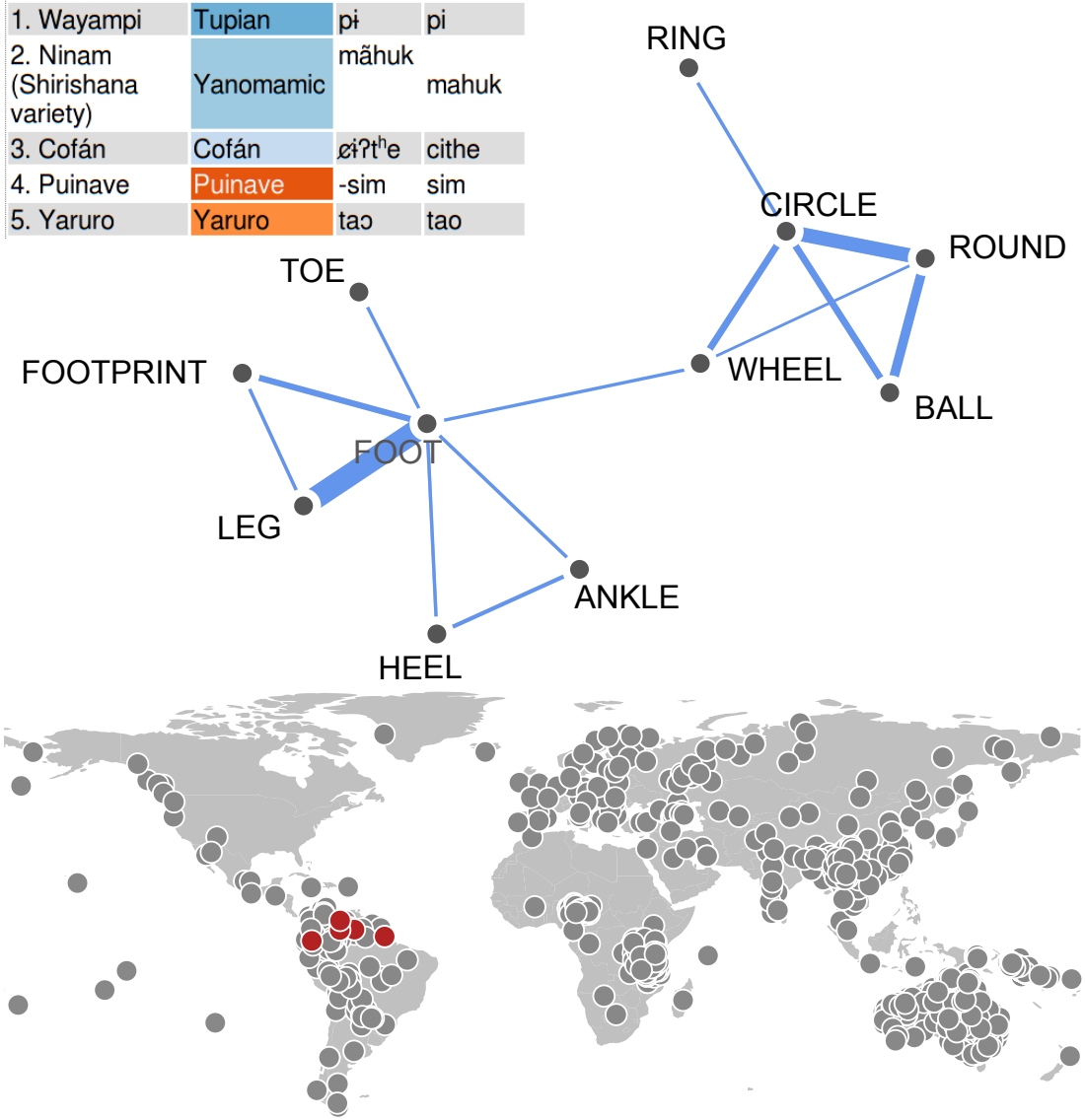


Features: Examples

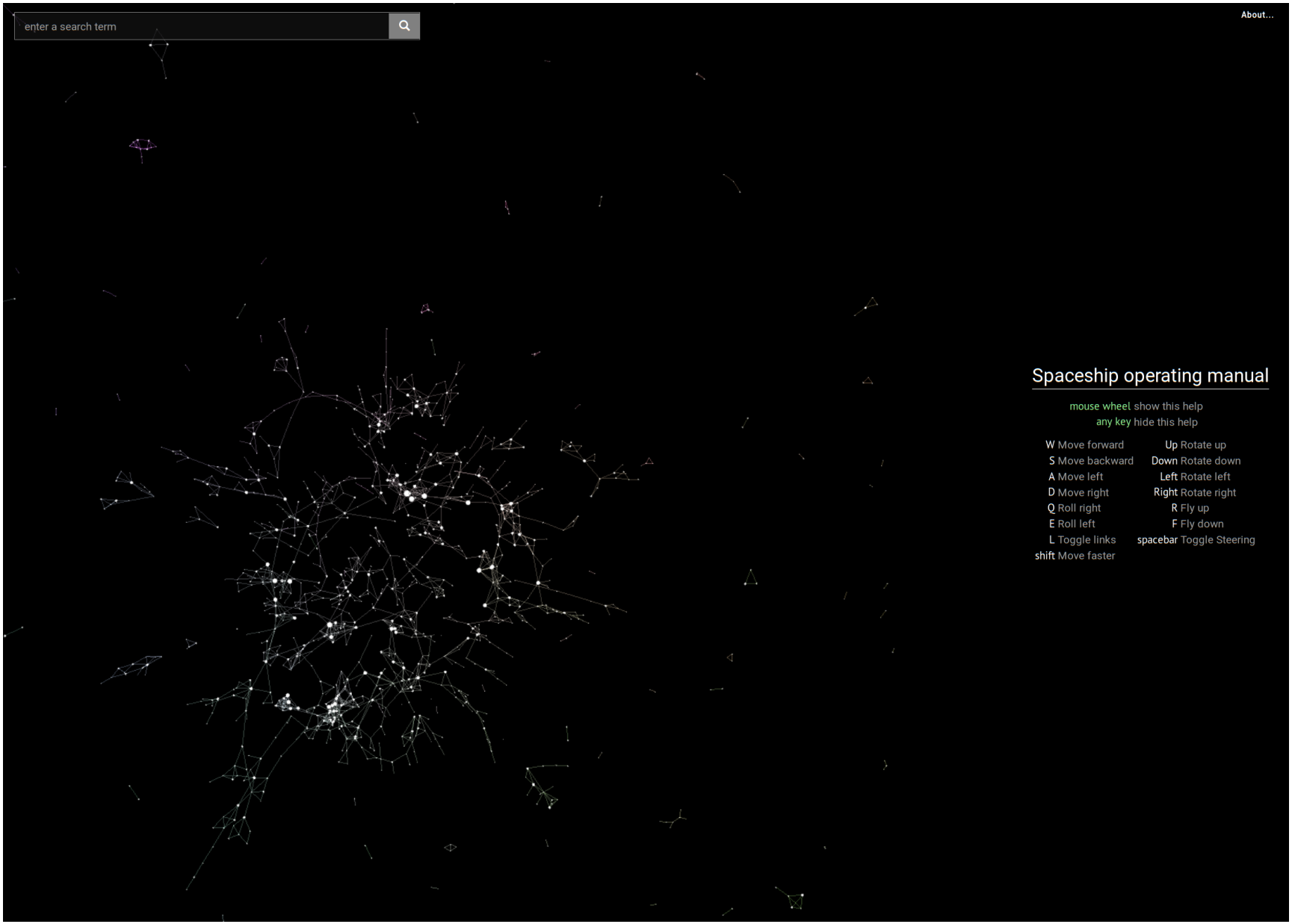


Features: Examples

Language	Family	Value	Norm. Value
1. Wayampi	Tupian	pɨ	pi
2. Ninam (Shirishana variety)	Yanomamic	mãhuk	mahuk
3. Cofán	Cofán	ɕiʔtʰe	cithe
4. Puinave	Puinave	-sim	sim
5. Yaruro	Yaruro	taɔ	tao



Features: Examples



CLICS^2 DEMO

Schedule

Schedule

- CLICS data is currently being released, see `https://zenodo.org/communities/clics`.

Schedule

- CLICS data is currently being released, see `https://zenodo.org/communities/clics`.
- CLICS² is deployed online in a beta-version (0.1) at `http://clics.clld.org` and published by List, Greenhill, Anderson, Mayer, Tresoldi and Forkel (2018).

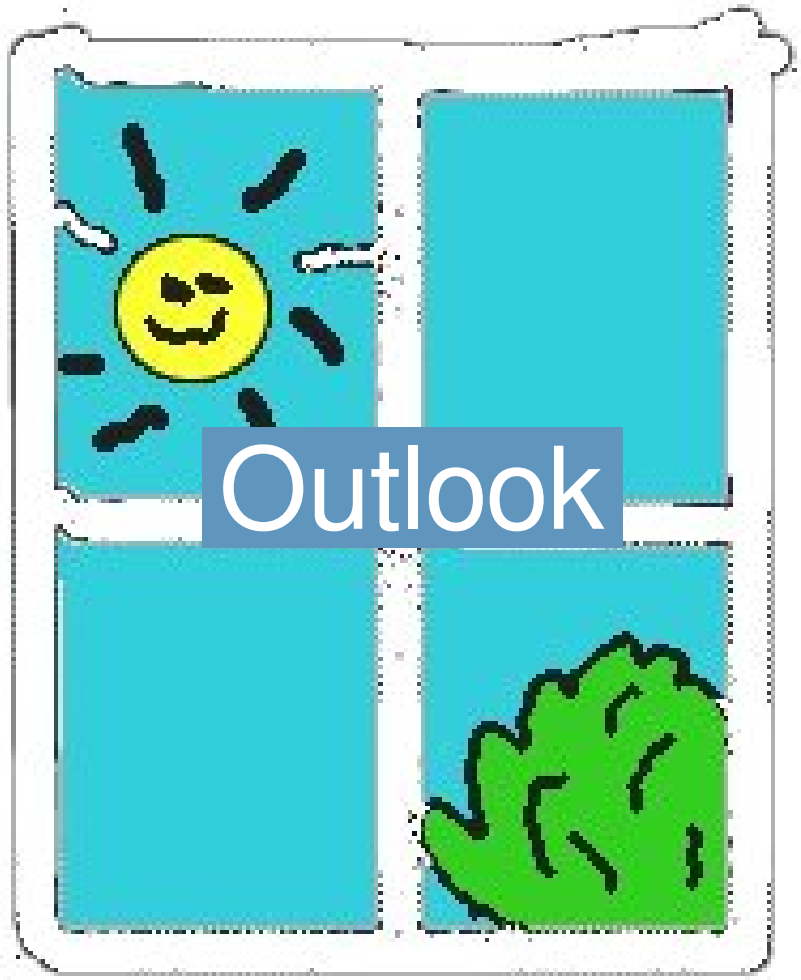
Schedule

- CLICS data is currently being released, see `https://zenodo.org/communities/clics`.
- CLICS² is deployed online in a beta-version (0.1) at `http://clics.clld.org` and published by List, Greenhill, Anderson, Mayer, Tresoldi and Forkel (2018).
- The official version will be published along with our paper on CLICS² (List et al. forthcoming, Linguistic Typology), approximately by the end of July.

Schedule

- CLICS data is currently being released, see `https://zenodo.org/communities/clics`.
- CLICS² is deployed online in a beta-version (0.1) at `http://clics.clld.org` and published by List, Greenhill, Anderson, Mayer, Tresoldi and Forkel (2018).
- The official version will be published along with our paper on CLICS² (List et al. forthcoming, Linguistic Typology), approximately by the end of July.
- The space-ship visualization will be deployed online later this year.

Outlook



- With CLICS^2, we provide a new framework for the collection and curation of data for the purpose of studying cross-linguistic colexification patterns.

- With CLICS^2, we provide a new framework for the collection and curation of data for the purpose of studying cross-linguistic colexification patterns.
- Future updates are planned, and we assume that we will be able to increase the data further by at least five more larger datasets.

- With CLICS^2, we provide a new framework for the collection and curation of data for the purpose of studying cross-linguistic colexification patterns.
- Future updates are planned, and we assume that we will be able to increase the data further by at least five more larger datasets.
- CLICS^2 is not perfect, and it does not come with any warranty. However, we hope that the improvements in terms of data transparency will make it much easier for scholars to work with the new cross-linguistic colexification database than its predecessor.

Thanks to our CLICS^2 team:

Simon Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel

Thanks to our CLICS^2 team:

Simon Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel

Thank You for your attention!