Lexical Typology at Scale — How Databases Transform the Study of Colexifications

Annika Tjuka

Max Planck Institute for Evolutionary Anthropology

03/06/2025

Agenda

I Introduction

II Lexical Databases

III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

Agenda

I Introduction

II Lexical Databases

III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

Introduction

About 6,500 languages are spoken worldwide.

Languages vary in how they divide the world into words.

Comparing vocabularies across languages reveals insights into human cognition and cultural variation.

Goal

Finding regularities in word meanings and causes for language variation.

Research Question

Why do words have multiple meanings?

Method

Computer-**Assisted** Language Comparison

https://calclab.org/

Agenda

I Introduction

II Lexical Databases

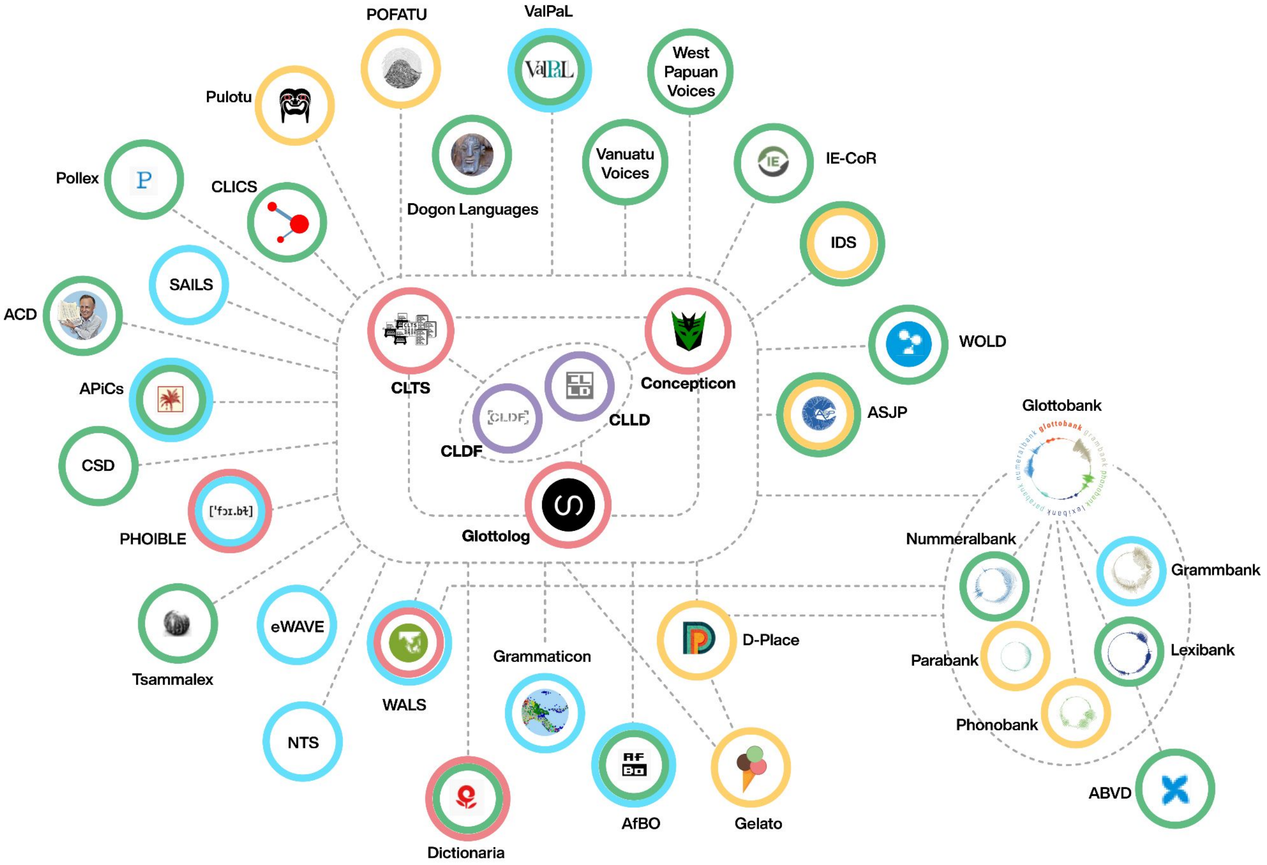
III Study I — Colexifications of Human Body Parts

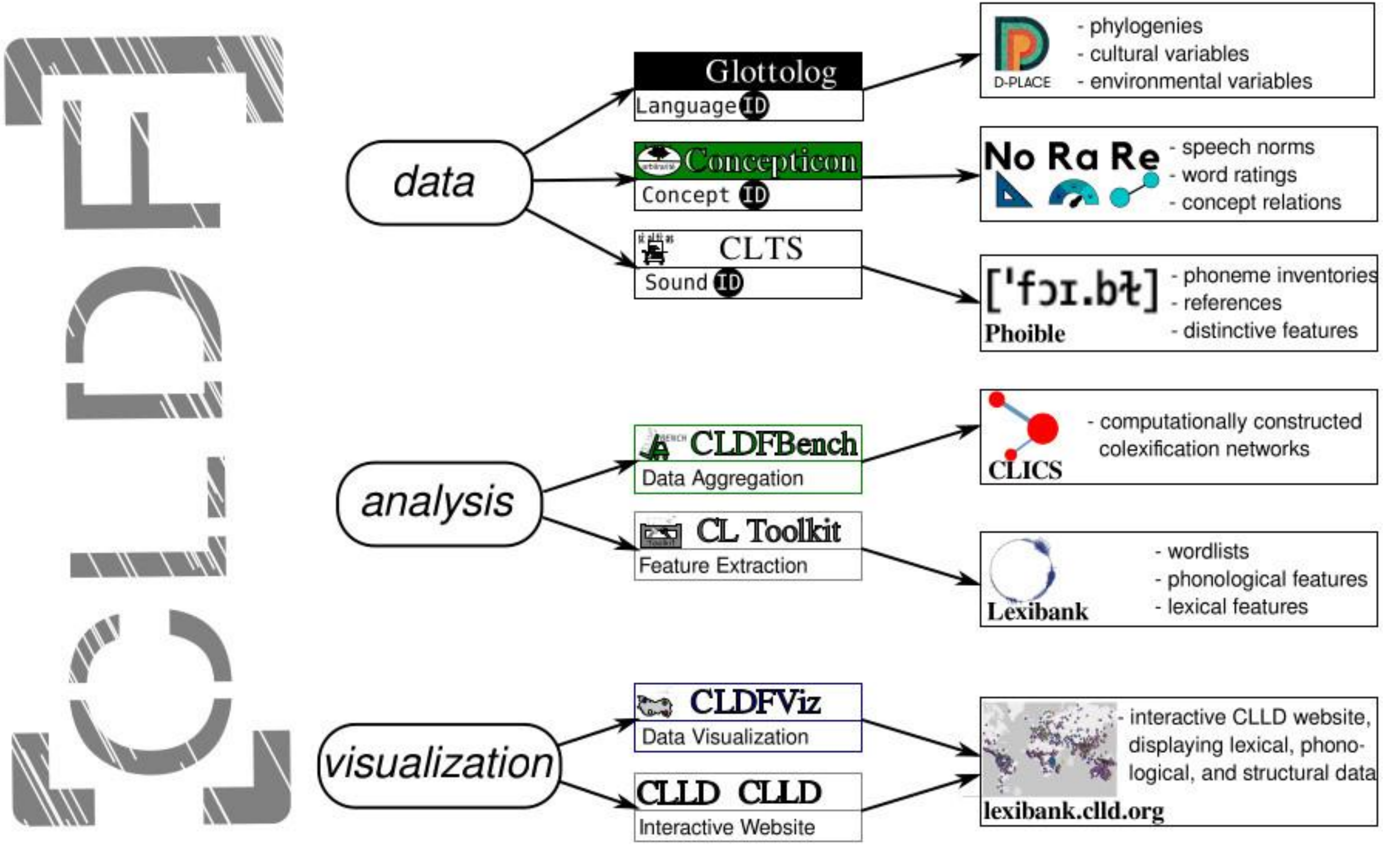
IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

Database Ecosystem at MPI-EVA





[CLDF]

data

Glottolog
Language ID

D-PLACE
- phylogenies
- cultural variables
- environmental variables

Concepticon
Concept ID

NoRaRe
- speech norms
- word ratings
- concept relations

CLTS
Sound ID

['fɔɪ.bɫ]
Phoible
- phoneme inventories
- references
- distinctive features

analysis

CLDFBench
Data Aggregation

CLICS
- computationally constructed colexification networks

CL Toolkit
Feature Extraction

Lexibank
- wordlists
- phonological features
- lexical features

visualization

CLDFViz
Data Visualization

CLLD CLLD
Interactive Website

- interactive CLLD website, displaying lexical, phonological, and structural data
lexibank.clld.org

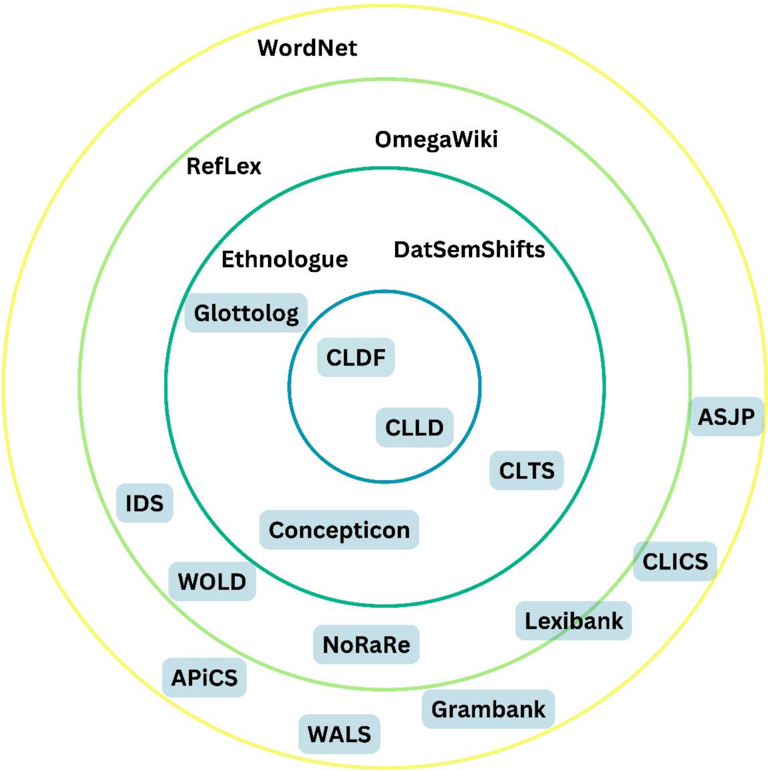
Lexical Databases

Progress: more linguistic data

Challenge: FAIR data (Wilkinson et al. 2016)

Solution: Cross-Linguistic Data Formats

(CLDF, Forkel et al. 2018)



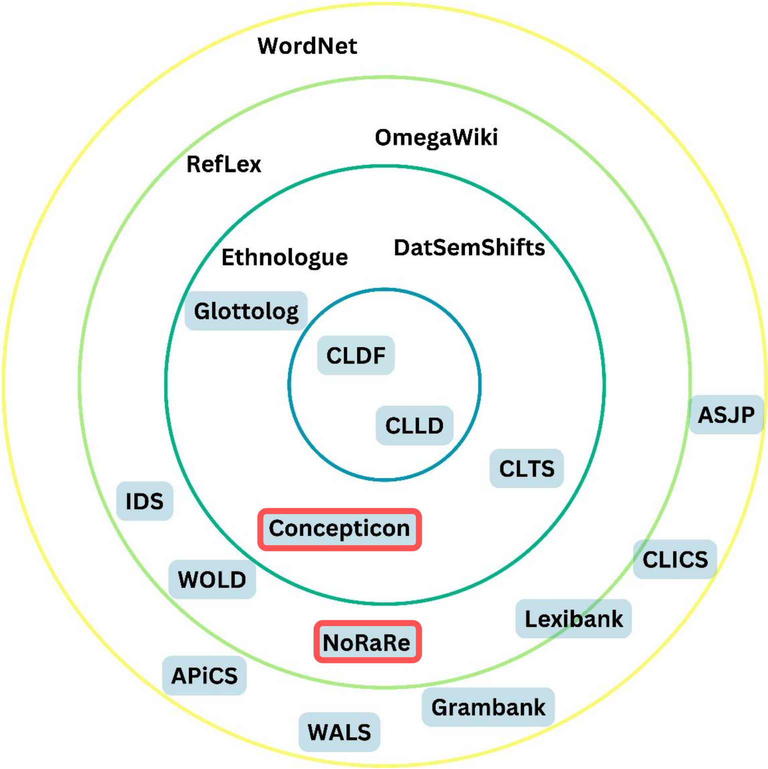
Lexical Databases

Progress: more linguistic data

Challenge: FAIR data (Wilkinson et al. 2016)

Solution: Cross-Linguistic Data Formats

(CLDF, Forkel et al. 2018)



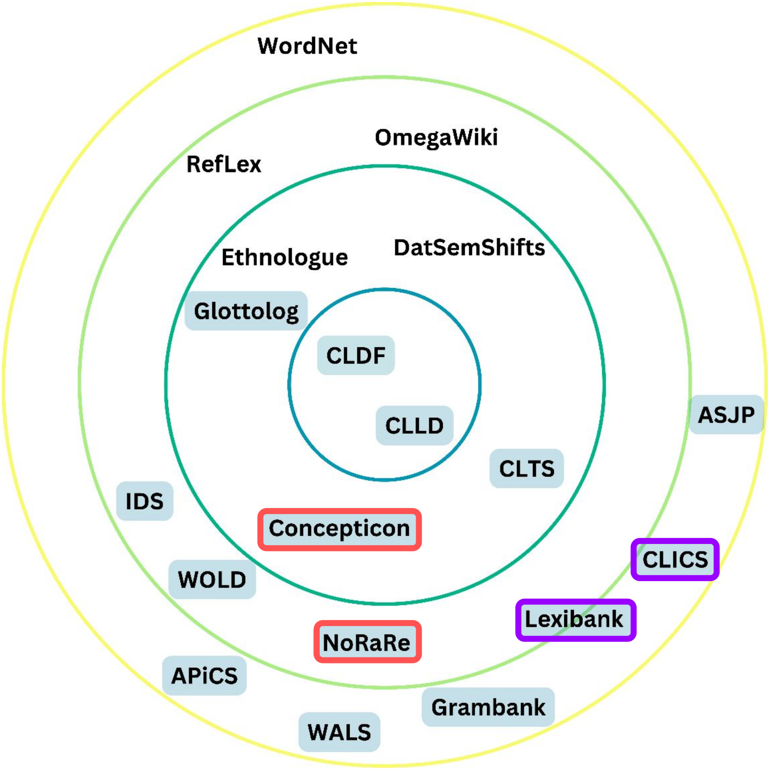
Lexical Databases

Progress: more linguistic data

Challenge: FAIR data (Wilkinson et al. 2016)

Solution: Cross-Linguistic Data Formats

(CLDF, Forkel et al. 2018)



arbitrarité

Concepticon

A resource of concept and word lists that offers standardized concept sets and links to glosses.

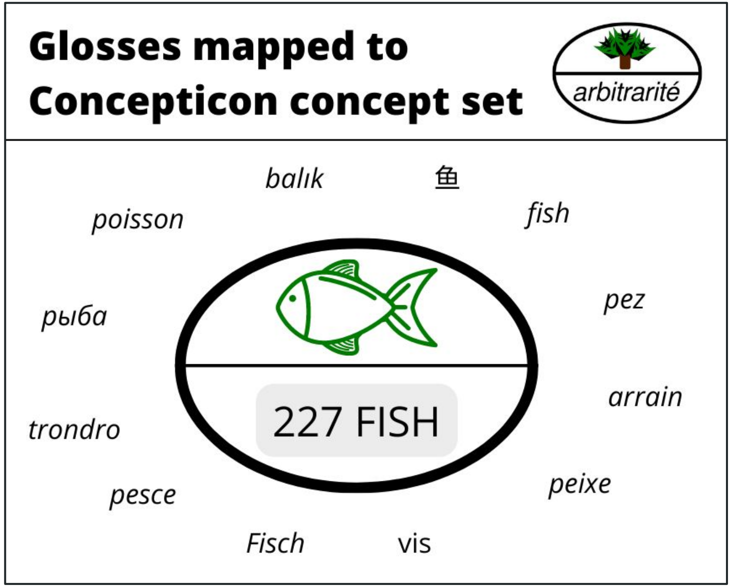
It serves as a reference catalog for historical and typological language comparison.

Concepticon Concept Sets

They consist of a unique identifier, a label, a definition, a semantic field, and an ontological category.

They reflect concepts that are deemed interesting for comparison by linguists and occur frequently in concept lists (List et al. 2016).

Elicitation glosses are established by linguists and are often based on already existing concept lists.



Data Curation

- Automatic and manual mapping to Concepticon concept sets
- Information on data types in metadata.json
- Test-driven data curation
- Python package: `pyconcepticon` (Forkel, Rzymski & List 2019)
- Accessed via command line
- Regular releases

https://concepticon.clld.org/

Tutorial: Tjuka (2020)

Workflow

Step 1: Prepare word list

ID	ENGLISH	CHINESE
Allen-2007-500-1	sky	天
Allen-2007-500-2	sun	太阳
Allen-2007-500-3	moon	月亮
Allen-2007-500-4	star	星星
Allen-2007-500-5	cloud	云
Allen-2007-500-6	wind	风
Allen-2007-500-7	rain	雨

Workflow

Step 2: Map to Concepticon

```
$ concepticon map_concepts PATH/TO/YOURLIST.tsv
```

Workflow

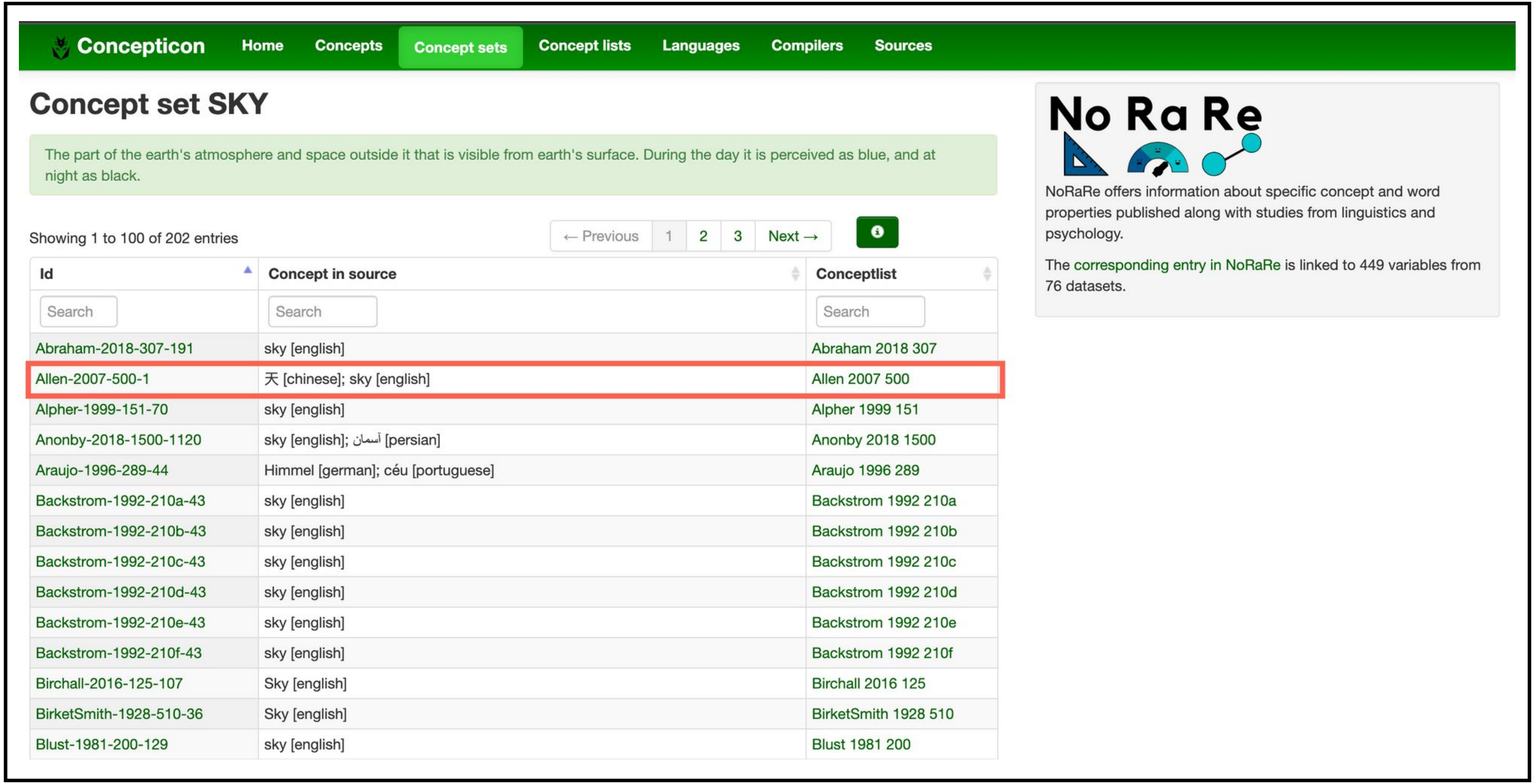
Step 2: Map to Concepticon

ID	ENGLISH	CHINESE	CONCEPTICON_ID	CONCEPTICON_GLOSS
Allen-2007-500-1	sky	天	1732	SKY
Allen-2007-500-2	sun	太阳	1343	SUN
Allen-2007-500-3	moon	月亮	1313	MOON
Allen-2007-500-4	star	星星	1430	STAR
Allen-2007-500-5	cloud	云	1489	CLOUD
Allen-2007-500-6	wind	风	960	WIND
Allen-2007-500-7	rain	雨	658	RAIN (PRECIPITATION)

Workflow

Step 2: Map to Concepticon

ID	ENGLISH	CHINESE	CONCEPTICON_ID	CONCEPTICON_GLOSS
Allen-2007-500-1	sky	天	1732	SKY
Allen-2007-500-2	sun	太阳	1343	SUN
Allen-2007-500-3	moon	月亮	1313	MOON
Allen-2007-500-4	star	星星	1430	STAR
Allen-2007-500-5	cloud	云	1489	CLOUD
Allen-2007-500-6	wind	风	960	WIND
Allen-2007-500-7	rain	雨	658	RAIN (PRECIPITATION)



Concept set SKY

The part of the earth's atmosphere and space outside it that is visible from earth's surface. During the day it is perceived as blue, and at night as black.

Showing 1 to 100 of 202 entries

Id	Concept in source	Conceptlist
Search	Search	Search
Abraham-2018-307-191	sky [english]	Abraham 2018 307
Allen-2007-500-1	天 [chinese]; sky [english]	Allen 2007 500
Alpher-1999-151-70	sky [english]	Alpher 1999 151
Anonby-2018-1500-1120	sky [english]; آسمان [persian]	Anonby 2018 1500
Araujo-1996-289-44	Himmel [german]; céu [portuguese]	Araujo 1996 289
Backstrom-1992-210a-43	sky [english]	Backstrom 1992 210a
Backstrom-1992-210b-43	sky [english]	Backstrom 1992 210b
Backstrom-1992-210c-43	sky [english]	Backstrom 1992 210c
Backstrom-1992-210d-43	sky [english]	Backstrom 1992 210d
Backstrom-1992-210e-43	sky [english]	Backstrom 1992 210e
Backstrom-1992-210f-43	sky [english]	Backstrom 1992 210f
Birchall-2016-125-107	Sky [english]	Birchall 2016 125
BirketSmith-1928-510-36	Sky [english]	BirketSmith 1928 510
Blust-1981-200-129	sky [english]	Blust 1981 200

No Ra Re

NoRaRe offers information about specific concept and word properties published along with studies from linguistics and psychology.

The corresponding entry in NoRaRe is linked to 449 variables from 76 datasets.

https://concepticon.clld.org/parameters/1732

NoRaRe

A cross-linguistic database of norms, ratings, and relations for words and concepts.

Building on Concepticon, it integrates data from psychology and linguistics.

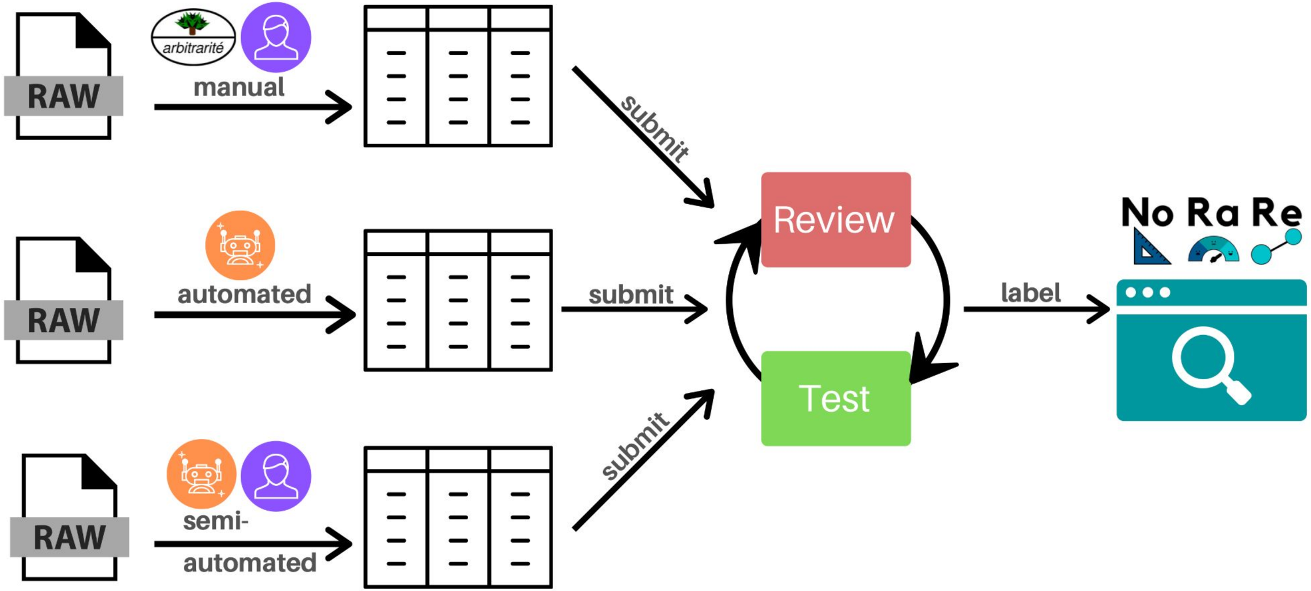
Data Curation

- Manual, automated, and semi-automated mapping
- Information on data types in metadata.json
- Test-driven data curation
- Python package: `pynorare` (List & Forkel 2020)
- Accessed via command line
- Regular releases

https://norare.clld.org/

Tutorial: Tjuka (2021a; 2021b)

Workflow



Tjuka et al. (2022): *Behavior Research Methods*

Tjuka et al. (2023): *Open Science Europe*

Agenda

I Introduction

II Lexical Databases

III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

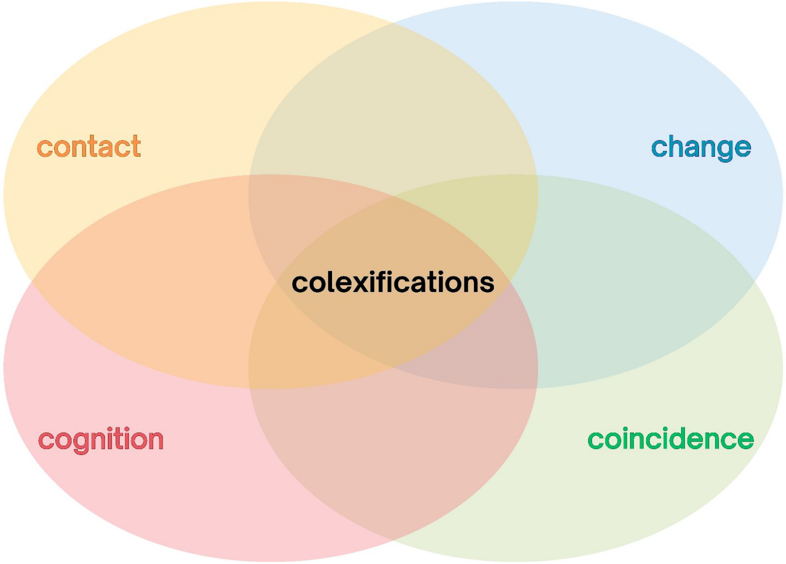
V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

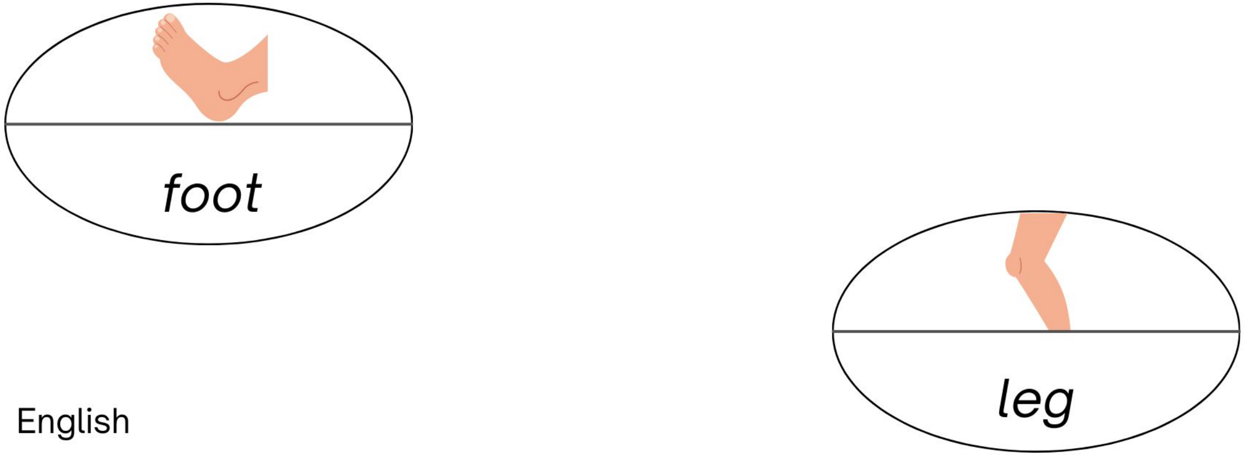
Colexifications

The same lexical form is used for two different concepts in at least two genealogically unrelated languages (François 2008).

The analysis is based on cross-linguistic data.



Colexifications

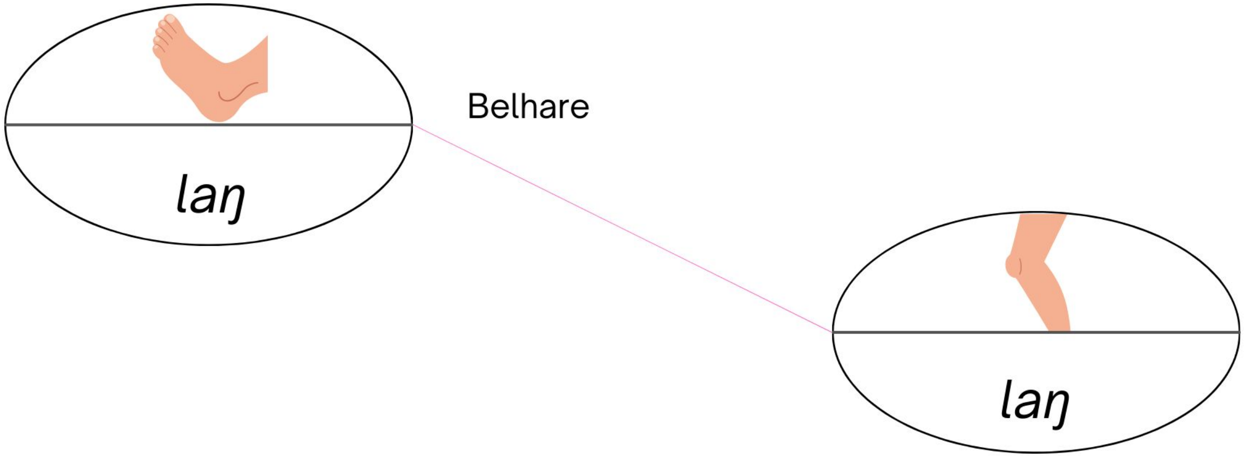


foot

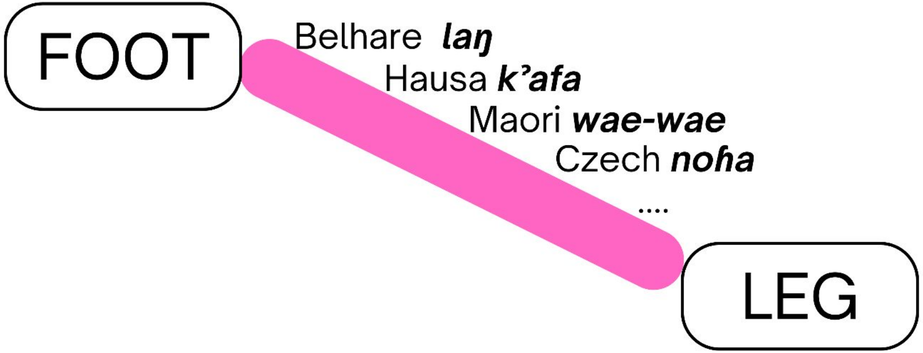
leg

English

Colexifications



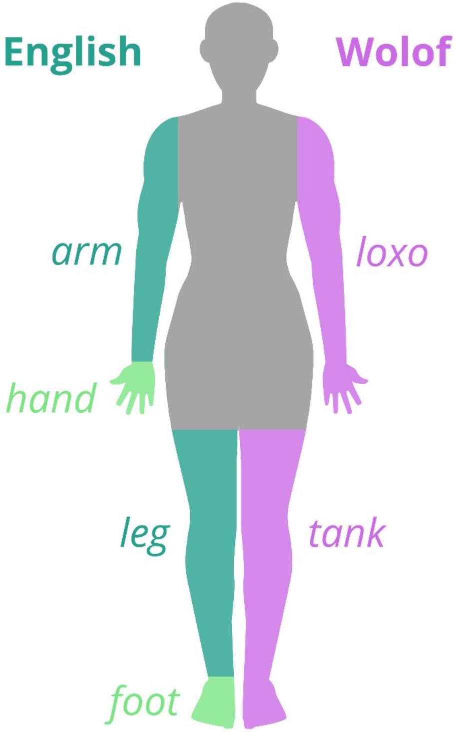
Colexifications



Body Part Vocabularies

Why do two body parts receive the same name?

Analysis of perceptual features:
contiguity, function, shape

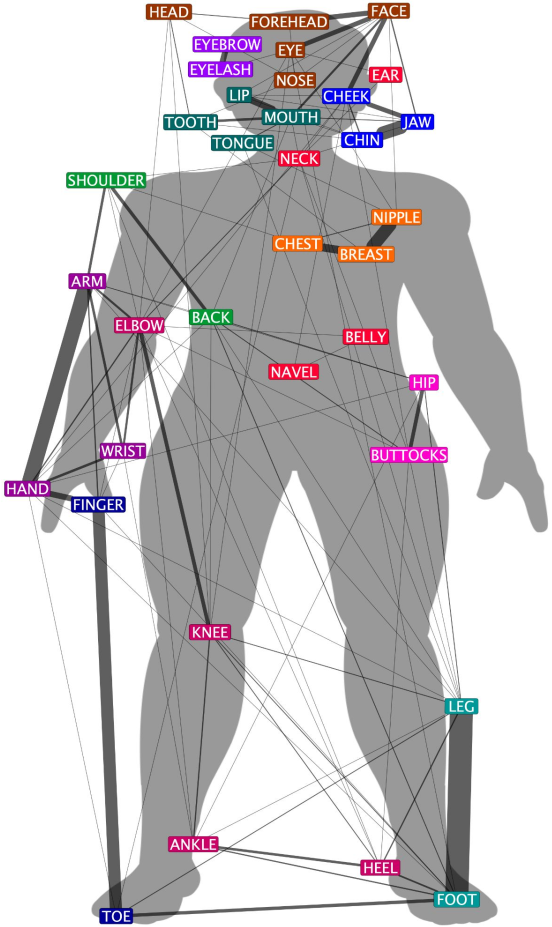


Tjuka et al. (2024): *Scientific Reports*

Materials & Methods

- 51 data sets from Lexibank (List et al. 2022) including phonetic transcriptions
- 36 human body part concepts from Concepticon v2.5
- Automated identification of full colexifications
- New, transparent workflow including cognate detection
- 110 body part colexifications across 1,028 language varieties

Tjuka (2021b; 2022b): Concept list description in *Computer-Assisted Language Comparison in Practice*



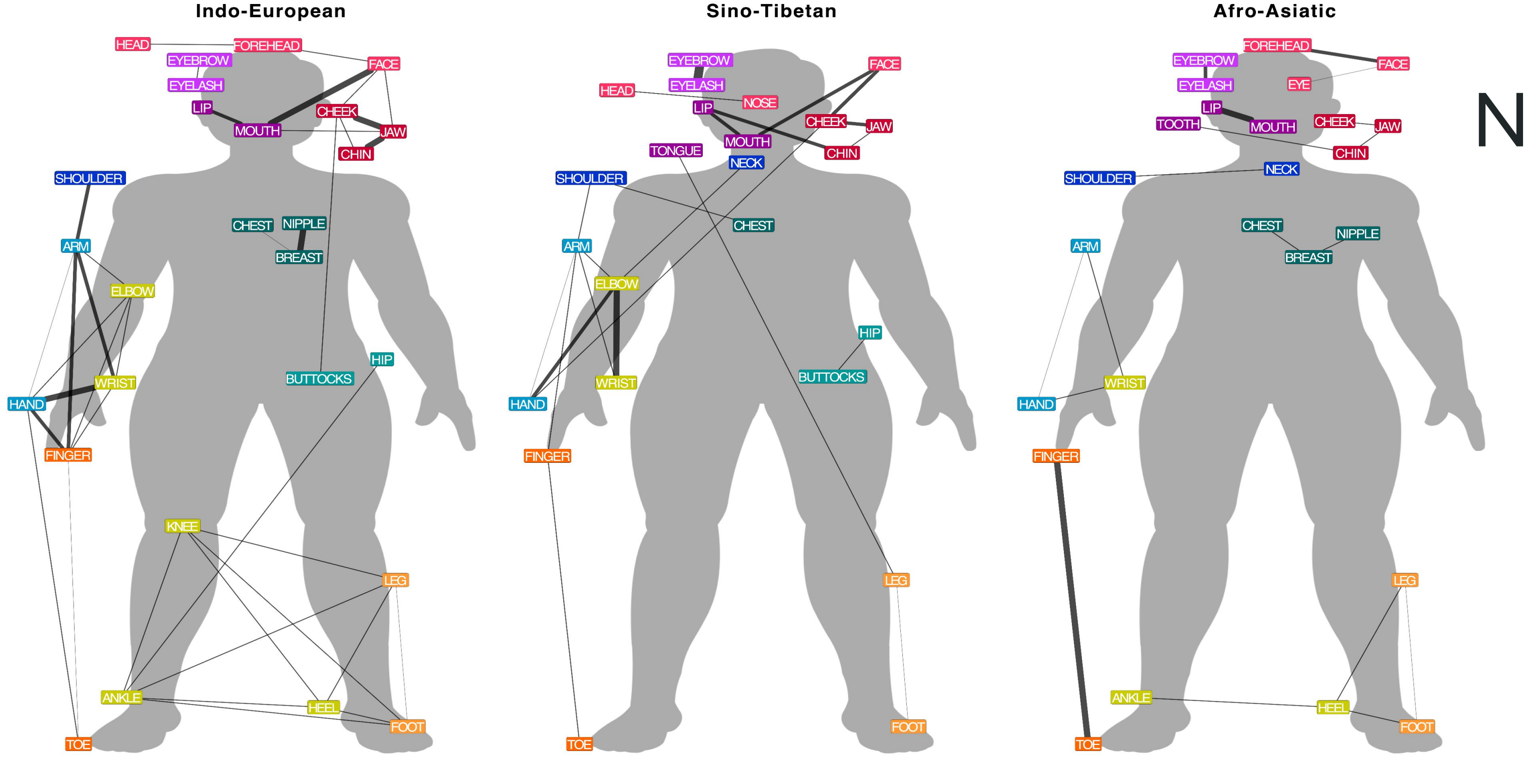
Body Part Network

Few widespread,

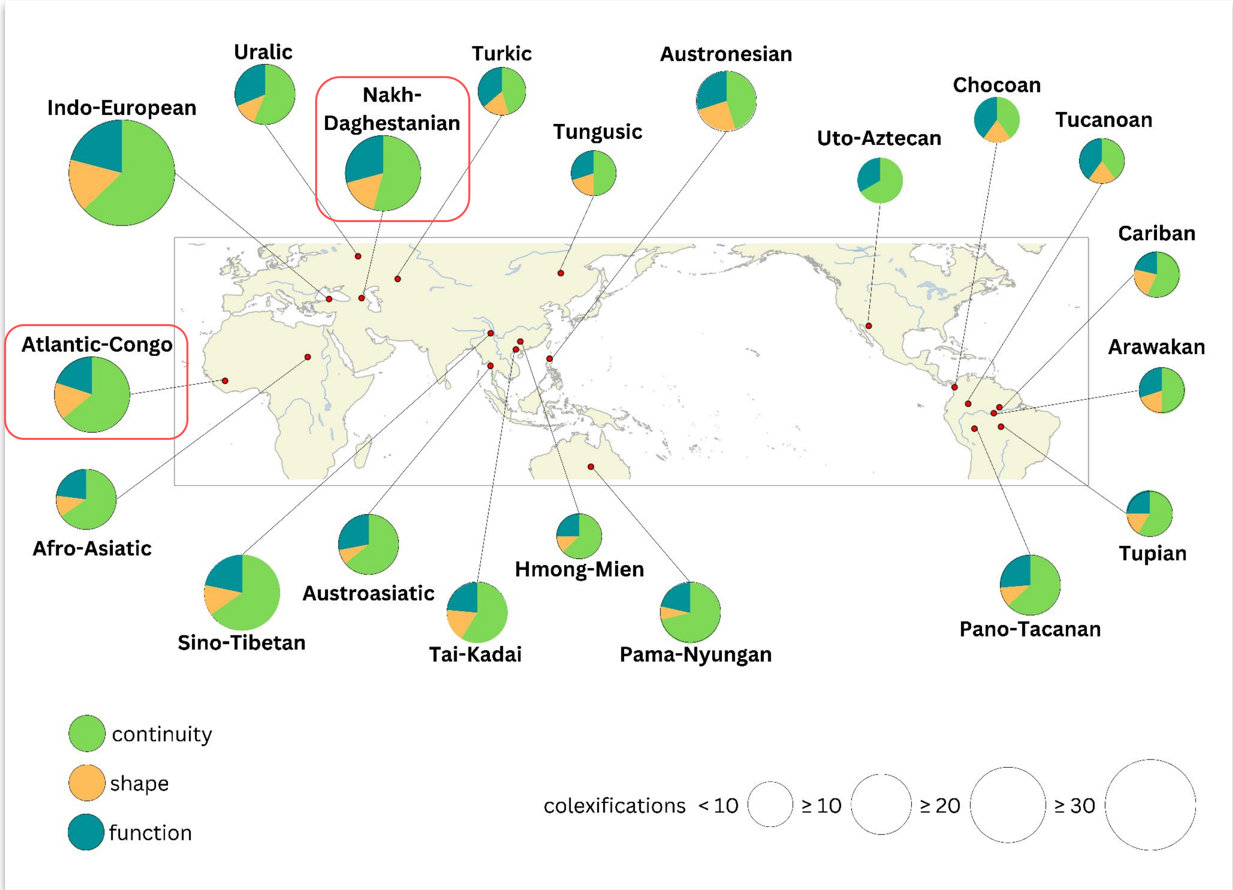
many language-specific colexifications.

Tjuka et al. (2024): *Scientific Reports*

Family Networks



Tjuka et al. (2024): *Scientific Reports*



Tjuka et al. (2024): *Scientific Reports*

Conclusions

Contiguity drives most colexifications between body parts.

Preferences for perceptual features differ across languages.

Agenda

I Introduction

II Lexical Databases

III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

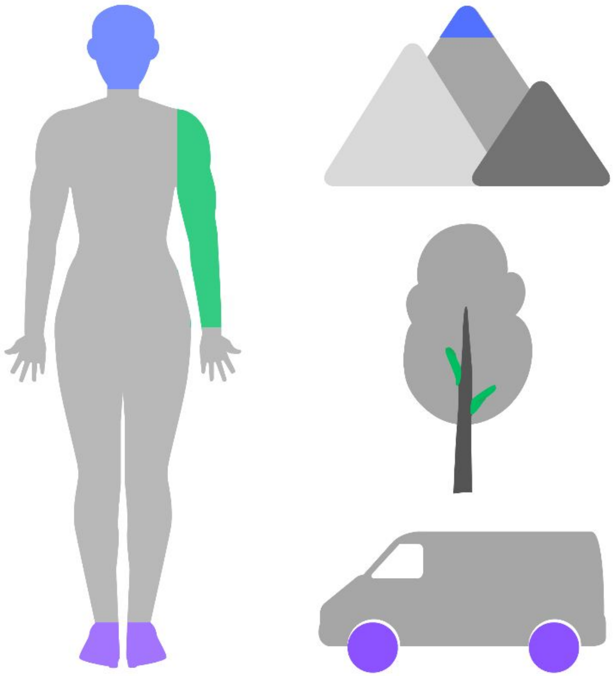
VI Summary

Aim

Exploration of the relation between the human body and objects across languages

Analysis of full colexification

Quantitative study on perceptual features (vision and touch)

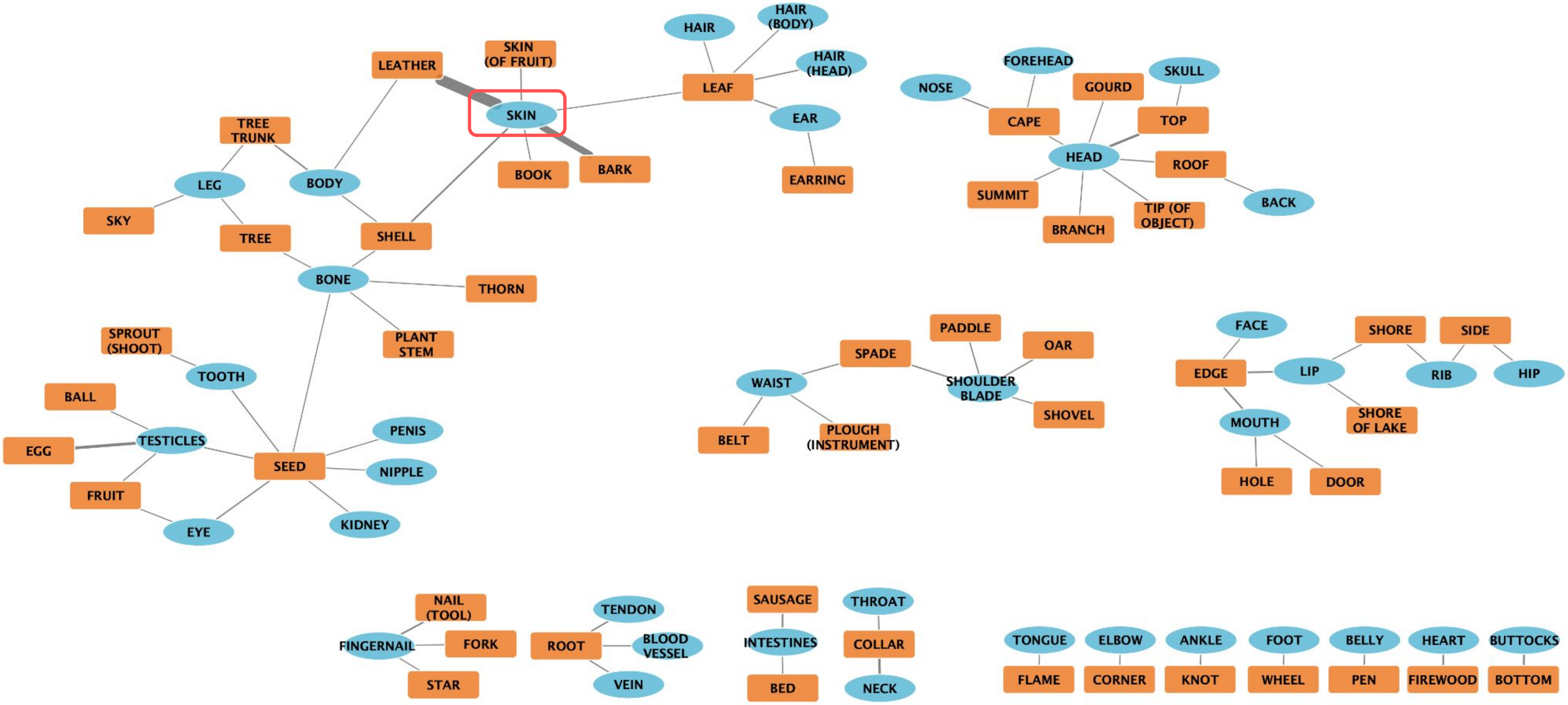


Tjuka (2024): *Linguistic Typology*

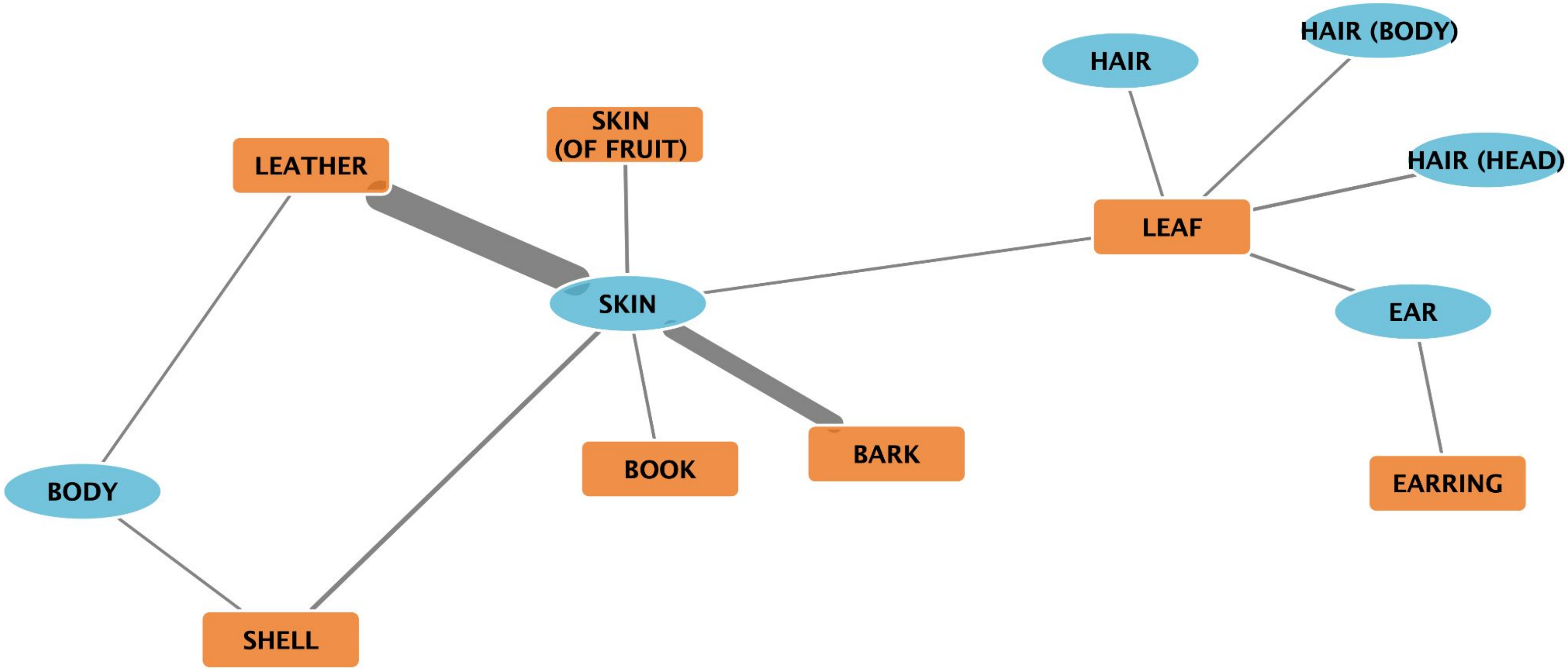
Materials & Methods

- 36 data sets from Lexibank (List et al. 2022)
- 134 human body part and 650 object concepts from Concepticon v2.5
- Automated identification of full colexifications
- 78 body-object colexifications occurring across 396 language varieties
- Analyses of frequency, distribution, cognitive relations, and coincidental cases

Tjuka (2020a; 2020b; 2022a): Concept list description in *Computer-Assisted Language Comparison in Practice*



Tjuka (2024): *Linguistic Typology*



Tjuka (2024): *Linguistic Typology*

Perceptual Features: Vision & Touch

Material: English sensory modality ratings for visual and haptic perception (Lynott et al. 2020) available in NoRaRe for 72 body-object colexifications.

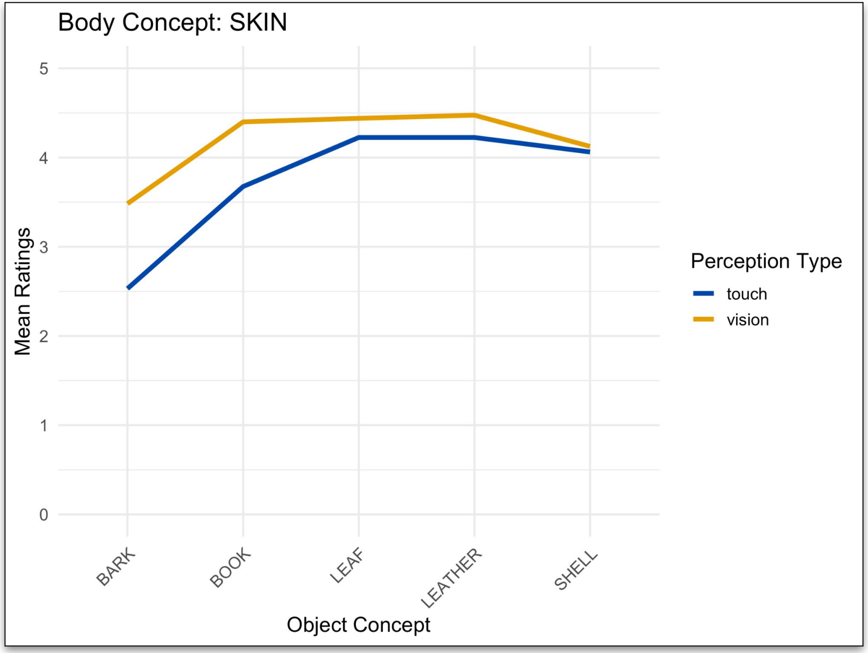
Method: Bayesian linear regression model with perception type as varying residuals.

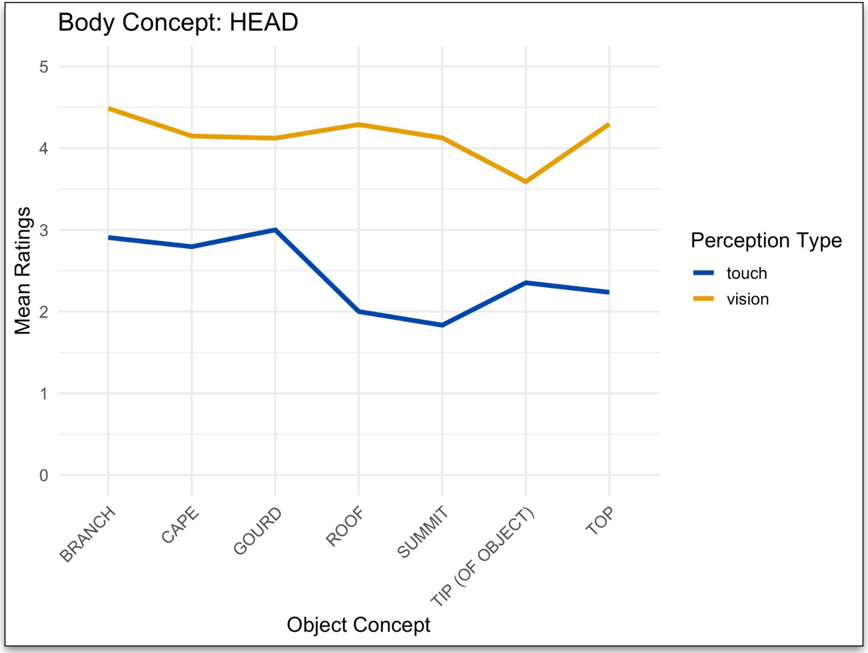
Question: Are body and object concepts perceived similarly across speakers?

Result: Body and object concepts align more closely in their visual perception (*sd* = 1.81) compared to their haptic perception (*sd* = 2.06).

Tjuka (2024): *Linguistic Typology*

Perceptual Features: Vision & Touch





Tjuka (2024): *Linguistic Typology*

Conclusions

Some widespread body-object colexifications such as SKIN-BARK or TESTICLES-EGG exist.

However, most body-object colexifications occur in a small number of languages.

Alignment of ratings on vision and touch is related to literal similarity, while divergence is related to figurative similarity and low frequency.

Agenda

I Introduction

II Lexical Databases

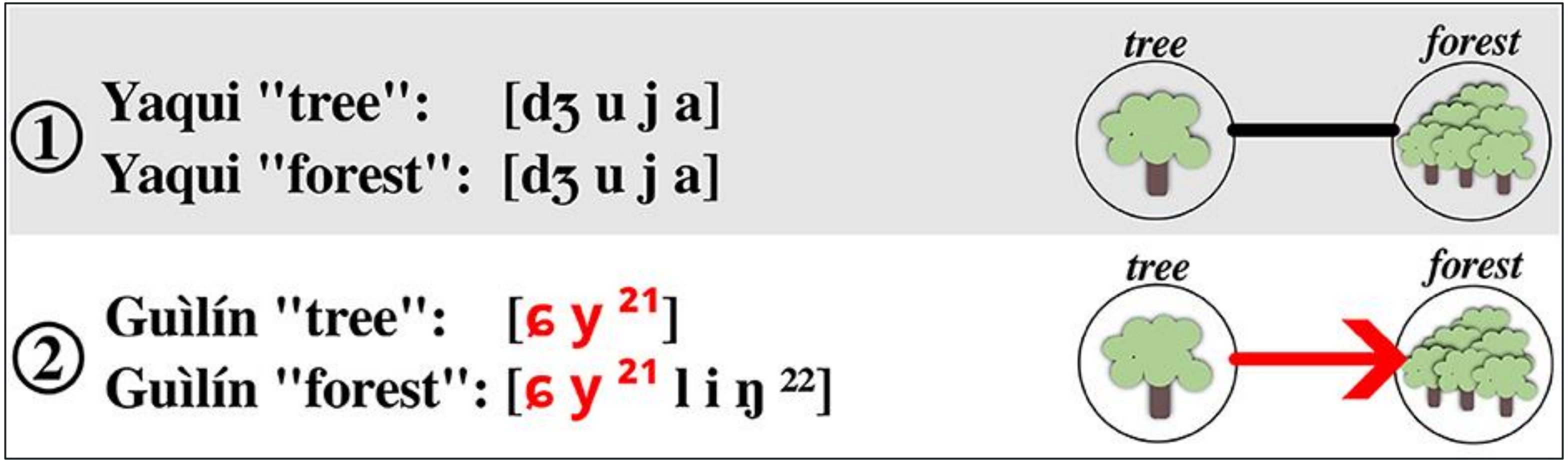
III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

Partial Colexification



① Yaqui "tree": [dʒ u j a]
Yaqui "forest": [dʒ u j a]

② Guìlín "tree": [ɕ y ^{21}]
Guìlín "forest": [ɕ y ^{21} l i ŋ ^{22}]

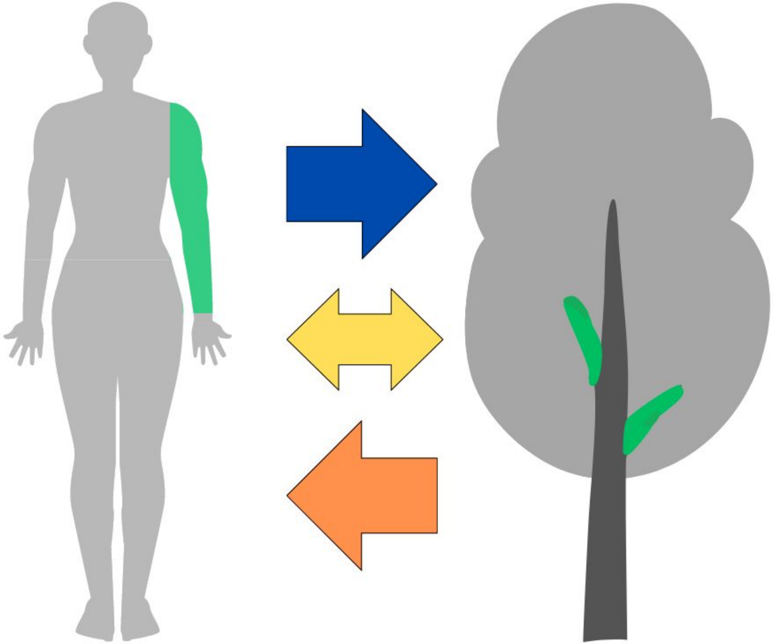
List (2023): Frontiers in Psychology

Question

In which direction does the development of body-object colexifications go?

Tjuka & List (2024): Yearbook German Cog. Ling. Assoc.

?

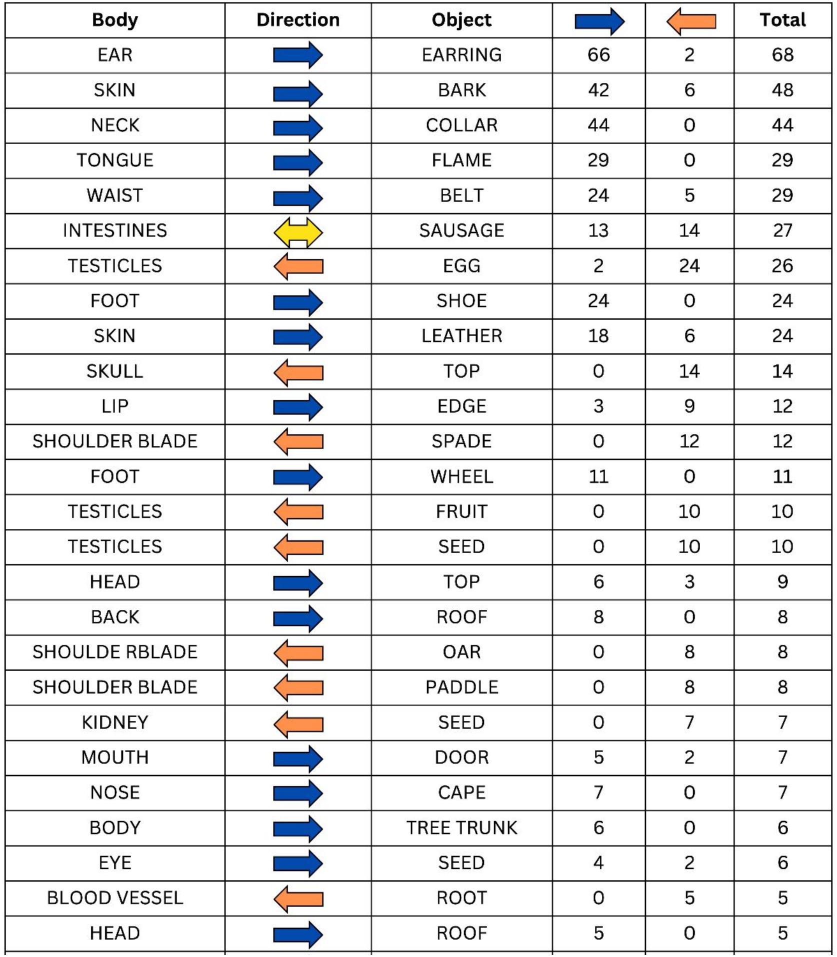


Materials & Methods

Seed list:	100 body-object colexifications from Tjuka (2024)
Target:	Weighted directed network from List (2023)
Overlap:	39 body-object colexifications

Tjuka & List (2024): Yearbook German Cog. Ling. Assoc.

Results



Body	Direction	Object	→	←	Total
EAR	→	EARRING	66	2	68
SKIN	→	BARK	42	6	48
NECK	→	COLLAR	44	0	44
TONGUE	→	FLAME	29	0	29
WAIST	→	BELT	24	5	29
INTESTINES	↔	SAUSAGE	13	14	27
TESTICLES	←	EGG	2	24	26
FOOT	→	SHOE	24	0	24
SKIN	→	LEATHER	18	6	24
SKULL	←	TOP	0	14	14
LIP	→	EDGE	3	9	12
SHOULDER BLADE	←	SPADE	0	12	12
FOOT	→	WHEEL	11	0	11
TESTICLES	←	FRUIT	0	10	10
TESTICLES	←	SEED	0	10	10
HEAD	→	TOP	6	3	9
BACK	→	ROOF	8	0	8
SHOULDE RBLADE	←	OAR	0	8	8
SHOULDER BLADE	←	PADDLE	0	8	8
KIDNEY	←	SEED	0	7	7
MOUTH	→	DOOR	5	2	7
NOSE	→	CAPE	7	0	7
BODY	→	TREE TRUNK	6	0	6
EYE	→	SEED	4	2	6
BLOOD VESSEL	←	ROOT	0	5	5
HEAD	→	ROOF	5	0	5

Results

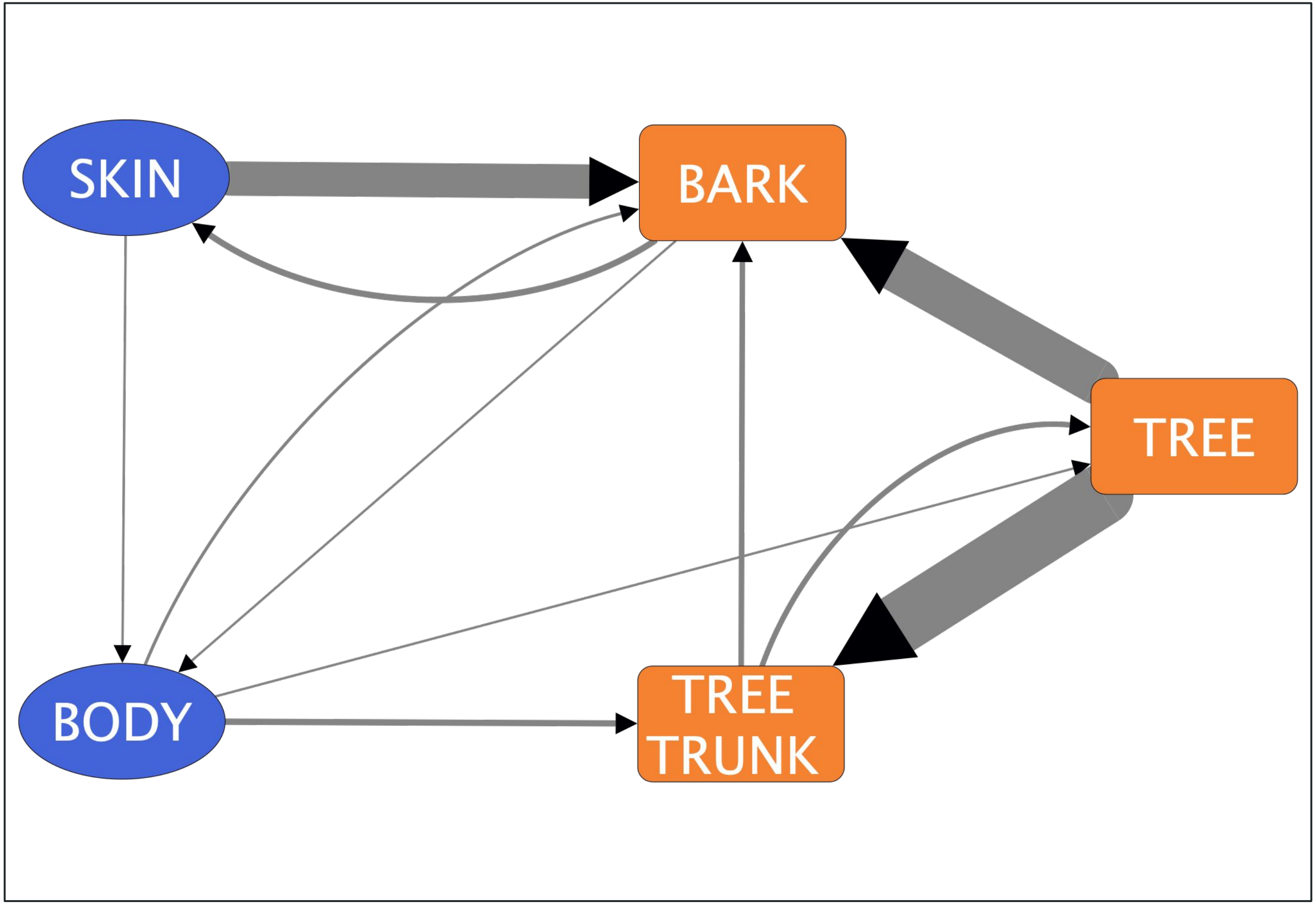
- 21 colexifications show a directional tendency from body to object
- 16 colexifications show a directional tendency from object to body
- 2 colexifications show no directional tendency

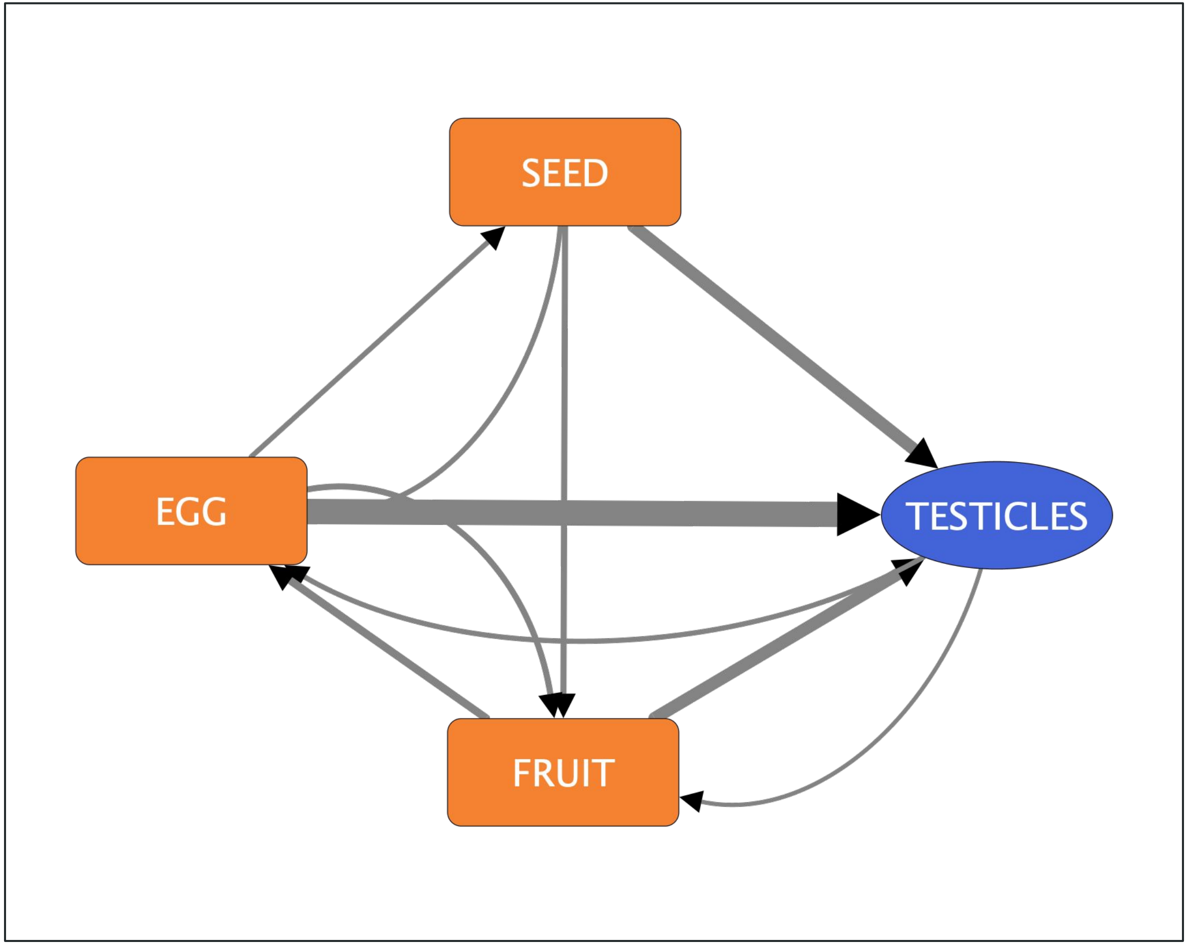
Tjuka & List (2024): Yearbook German Cog. Ling. Assoc.

Examples

- EAR-EARRING
  - *kula-pepeiao* lit. ‘gold-ear’ in Hawaiian (Austronesian)
  - *sau faliŋa* lit. ‘king ear’ in Rotuman (Austronesian)
- SKIN-BARK
  - *ror kulun* lit. ‘tree/wood skin’ in Kalamang (West Bomberai)
- NECK-COLLAR
  - *sɨpluw tor̃* lit. ‘neck cloth’ in Mansi (Uralic)
  - *ynĩ teʔ* lit. ‘neck clothing’ in Chatino (Otomanguean)

Tjuka & List (2024): Yearbook German Cog. Ling. Assoc.


SKIN
BARK
TREE
BODY
TREE TRUNK


SEED
EGG
TESTICLES
FRUIT

Conclusions

The domain of the human body serves as the source for the target domain of everyday objects.

However, certain concepts such as TESTICLES and SHOULDER BLADE were named after object concepts more frequently.

Agenda

I Introduction

II Lexical Databases

III Study I — Colexifications of Human Body Parts

IV Study II — Perception of Human Body Parts and Objects

V Study III — Directional Tendencies of Body-Object Colexifications

VI Summary

Summary

There is a great deal of linguistic diversity, but there are also general tendencies that arise.

Summary

Different factors can cause words to have multiple meanings, but similar perceptual features, especially visual similarity, lead to widespread colexifications.

Summary

Computer-assisted methods allow us to build databases and analyse data on a large scale.

Summary

There is a great deal of linguistic diversity, but there are also general tendencies that arise.

Different factors can cause words to have multiple meanings, but similar perceptual features, especially visual similarity, lead to widespread colexifications.

Computer-assisted methods allow us to build databases and analyse data on a large scale.

Thank you

Publications

Tjuka, Annika & Johann-Mattis List. 2024. Partial Colexifications Reveal Directional Tendencies in Object Naming. *Yearbook of the German Cognitive Linguistics Association* 12(1). 95–114. https://doi.org/10.1515/gcla-2024-0005.

Tjuka, Annika. 2024. Objects as Human Bodies: Cross-Linguistic Colexifications Between Words for Body Parts and Objects. Linguistic Typology 28(3). 379–418. https://doi.org/10.1515/lingty-2023-0032.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties. *Behavior Research Methods* 54. 864–884. https://doi.org/10.3758/s13428-021-01650-1.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2023. Curating and Extending Data for Language Comparison in Concepticon and NoRaRe. *Open Research Europe* 2(141). 1–13. https://doi.org/10.12688/openreseurope.15380.3.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2024. Universal and Cultural Factors Shape Body Part Vocabularies. *Scientific Reports* 14(1). 1–12. https://doi.org/10.1038/s41598-024-61140-0.

Tutorials and Blog Posts

Tjuka, Annika. 2020a. A List of 171 Body Part Concepts. Computer-Assisted Language Comparison in Practice 3(10). 1–3. https://calc.hypotheses.org/3023

Tjuka, Annika. 2020b. Adding Concept Lists to Concepticon: A Guide for Beginners. Computer-Assisted Language Comparison in Practice 3(1). 1–5. https://calc.hypotheses.org/2225

Tjuka, Annika. 2021a. Comparing NoRaRe Data Sets: Calculation of Correlations and Creation of Plots in R. Computer-Assisted Language Comparison in Practice 4(11). 1–5. https://calc.hypotheses.org/3109

Tjuka, Annika. 2021b. A List of Color, Emotion, and Human Body Part Concepts. Computer-Assisted Language Comparison in Practice 4(11). 1–4. https://calc.hypotheses.org/3023

Tjuka, Annika. 2021c. Adding Data Sets to NoRaRe: A Guide for Beginners. Computer-Assisted Language Comparison in Practice 4(8). 1–5. https://calc.hypotheses.org/2890

Tjuka, Annika. 2022a. A Concept List for the Study of Semantic Extensions from Body to Objects. Computer-Assisted Language Comparison in Practice 5(4). 1–6. https://calc.hypotheses.org/3840

Tjuka, Annika. 2022b. Extending the List of Color, Emotion, and Human Body Part Concepts. Computer-Assisted Language Comparison in Practice 5(2). 1–3. https://calc.hypotheses.org/3913

Tjuka, Annika. 2024. How to Visualize Colexification Networks in Cytoscape (How to Do X in Linguistics 14). Computer-Assisted Language Comparison in Practice 7(1). 7–16. https://doi.org/10.15475/calcip.2024.1.2.

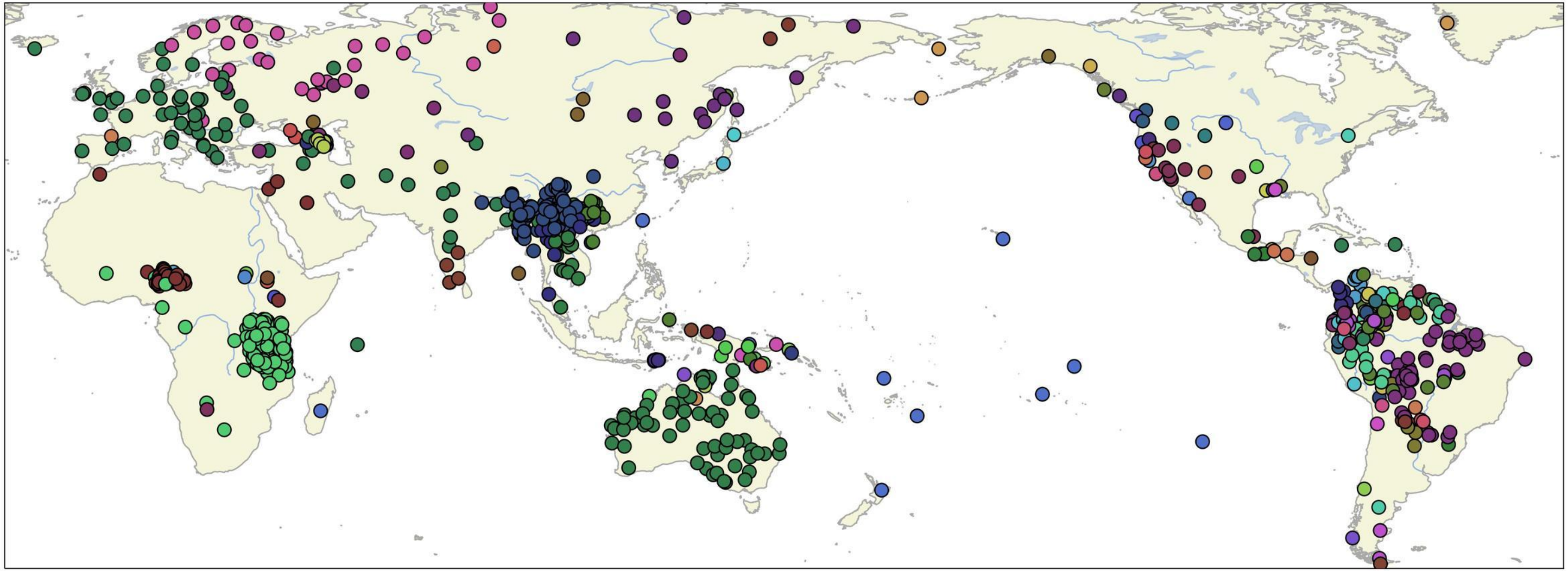
Step 3: Create your source list

ID	CONCEPTICON_ID	CONCEPTICON_GLOSS	ENGLISH	GROUP	SEMANTIC_FIELD
Tjuka-2022-784-1	802	ADAM'S APPLE	ADAM'S APPLE	body	human body part
Tjuka-2022-784-2	678	BEARD	BEARD	body	human body part
Tjuka-2022-784-3	1402	BREAST	BREAST	body	human body part
Tjuka-2022-784-4	834	BUTTOCKS	BUTTOCKS	body	human body part
Tjuka-2022-784-5	498	CALF OF LEG	CALF OF LEG	body	human body part
Tjuka-2022-784-135	20	SCYTHE	SCYTHE	object	tool
Tjuka-2022-784-136	124	THORN	THORN	object	plant
Tjuka-2022-784-137	146	SUGAR CANE	SUGAR CANE	object	plant
Tjuka-2022-784-138	159	SWEET POTATO	SWEET POTATO	object	food
Tjuka-2022-784-139	217	BETELNUT	BETELNUT	object	food

Step 3: Select your target word lists from Lexibank

ID	Organisation	Dataset	Zenodo
abrahammonpa	lexibank	abrahammonpa	10.5281/zenodo.5115885
allenbai	lexibank	allenbai	10.5281/zenodo.5115649
bantubvd	lexibank	bantubvd	10.5281/zenodo.5115982
beidasinitic	lexibank	beidasinitic	10.5281/zenodo.5119295
bodtkhobwa	lexibank	bodtkhobwa	10.5281/zenodo.5119330
bowernpny	lexibank	bowernpny	10.5281/zenodo.5119341
chenhmongmien	lexibank	chenhmongmien	10.5281/zenodo.5118744
chindialectsurvey	lexibank	chindialectsurvey	10.5281/zenodo.5121280
halenepal	lexibank	halenepal	10.5281/zenodo.5121540

(List et al. 2022)



Convenient (opportunistic) sample of 931 language varieties

(Bakker 2010)

Step 5: Find matches in Lexibank lists

✓	Allen-2007-500-119	119	fruit	果	1507	FRUIT
✓	Allen-2007-500-120	120	pit,stone	核	1762	STONE (OF FRUIT)
✓	Allen-2007-500-121	121	peel,husk	皮	275	PEEL
✓	Allen-2007-500-122	122	thorn	刺	124	THORN
✓	Allen-2007-500-123	123	body	身体	1480	BODY
✓	Allen-2007-500-124	124	head	头	1256	HEAD
✓	Allen-2007-500-125	125	hair	头发	2648	HAIR (HEAD)
✓	Allen-2007-500-126	126	face	脸	1560	FACE
✓	Allen-2007-500-127	127	eye	眼	1248	EYE
✓	Allen-2007-500-128	128	nose	鼻子	1221	NOSE

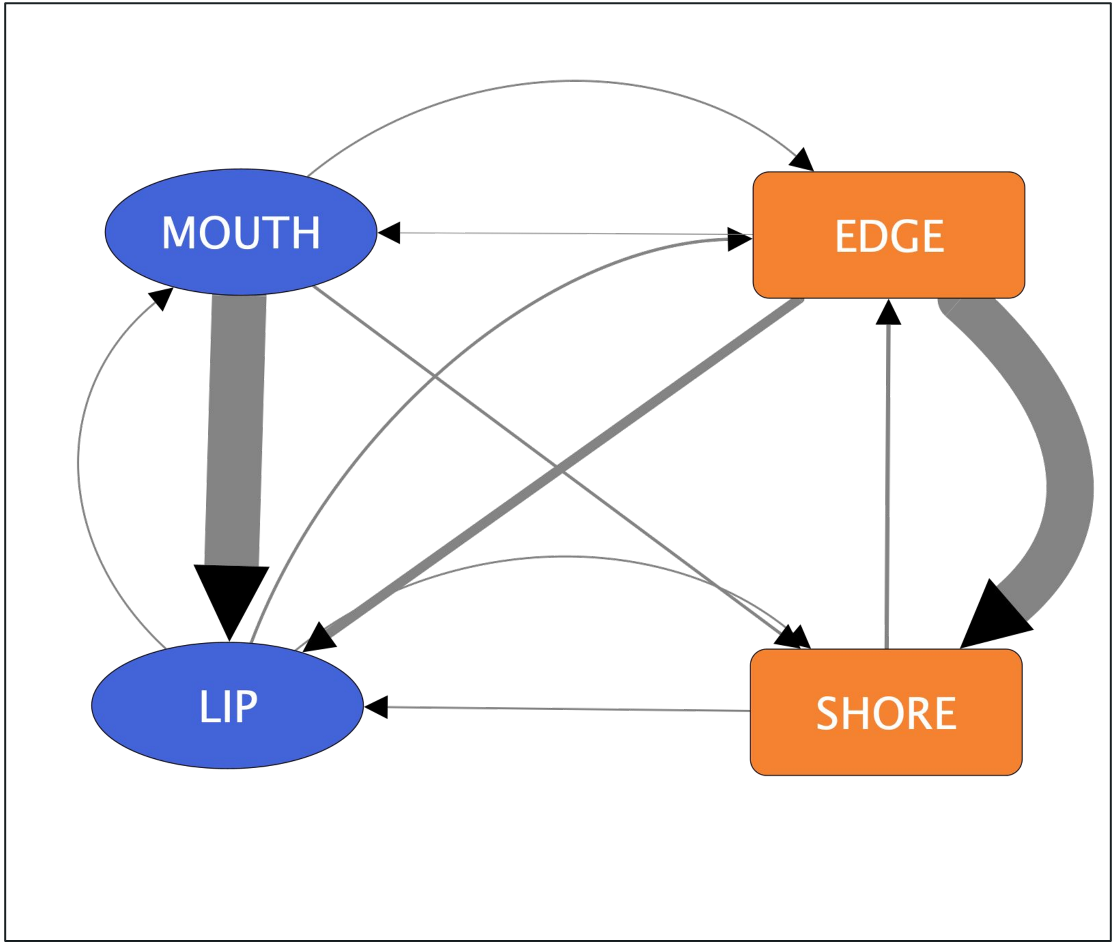
Lexibank list

Source list

ID	CONCEPTICON_ID	CONCEPTICON_GLOSS	ENGLISH	GROUP	SEMANTIC_FIELD
Tjuka-2022-784-1	802	ADAM'S APPLE	ADAM'S APPLE	body	human body part
Tjuka-2022-784-2	678	BEARD	BEARD	body	human body part
Tjuka-2022-784-3	1402	BREAST	BREAST	body	human body part
Tjuka-2022-784-4	834	BUTTOCKS	BUTTOCKS	body	human body part
Tjuka-2022-784-5	498	CALF OF LEG	CALF OF LEG	body	human body part
Tjuka-2022-784-135	20	SCYTHE	SCYTHE	object	tool
Tjuka-2022-784-136	124	THORN	THORN	object	plant
Tjuka-2022-784-137	146	SUGAR CANE	SUGAR CANE	object	plant
Tjuka-2022-784-138	159	SWEET POTATO	SWEET POTATO	object	food
Tjuka-2022-784-139	217	BETELNUT	BETELNUT	object	food

Step 6: Extract colexifications automatically

EXAMPLE_ID	GLOTTOCODE	CLTS_FORM	CONCEPTICON_GLOSS
allenbai-Eryuan-149_heart-1	eryu1239	ɕi^{55}	HEART
allenbai-Eryuan-222_firewood-1	eryu1239	ɕi^{55}	FIREWOOD
allenbai-Jianchuan-139_bellystomach-1	jian1239	fɤ^{44}	BELLY
allenbai-Jianchuan-235_pen-1	jian1239	fɤ^{44}	PEN


MOUTH
EDGE
LIP
SHORE