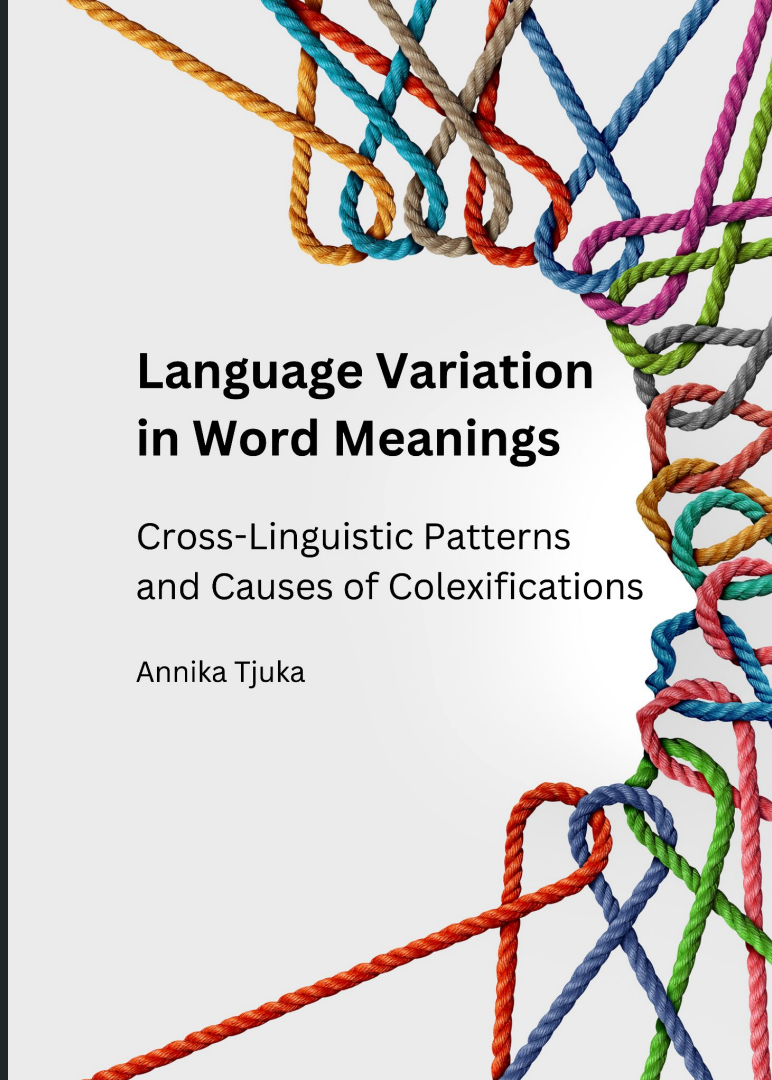
Dissertation Defense

Annika Tjuka

Friedrich-Schiller-Universität Jena
11/03/2024



Agenda

I Background

Word Meaning and Language Comparison

II Database

Cross-Linguistic Norms, Ratings, and Relations

III Studies

Colexifications of Body Part and Object Concepts

Colexifications of Body Part Concepts

IV Conclusions & Outlook

Agenda

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.

Aim

Finding regularities in word meanings and causes for language variation.

Research Questions

Method

How can lexical data be made comparable?

Research Questions

Method

How can lexical data be made comparable?

Theory

Why do words have multiple meanings?

Contributions

Method

New workflows for curating lexical data across research fields.

Facilitation of analyses for cross-linguistic comparison.

Contributions

Method

New workflows for curating lexical data across research fields.

Facilitation of analyses for cross-linguistic comparison.

Theory

Differentiating factors that cause words to have multiple meanings.

Testing universal claims with large data sets.

Agenda

I Background

Word Meaning and Language Comparison

II Database

Cross-Linguistic Norms, Ratings, and Relations

III Studies

Colexifications of Body Part and Object Concepts

Colexifications of Body Part Concepts

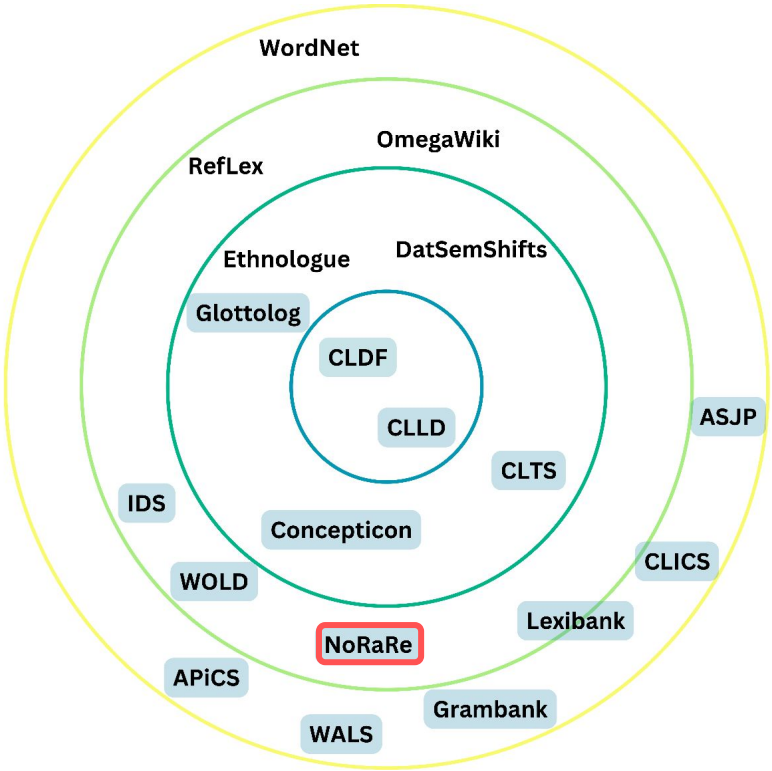
IV Conclusions & Outlook

Lexical Databases

Progress: more linguistic data

Challenge: FAIR data (Wilkinson et al. 2016)

Solution: Cross-Linguistic Data Formats
(CLDF, Forkel et al. 2018)

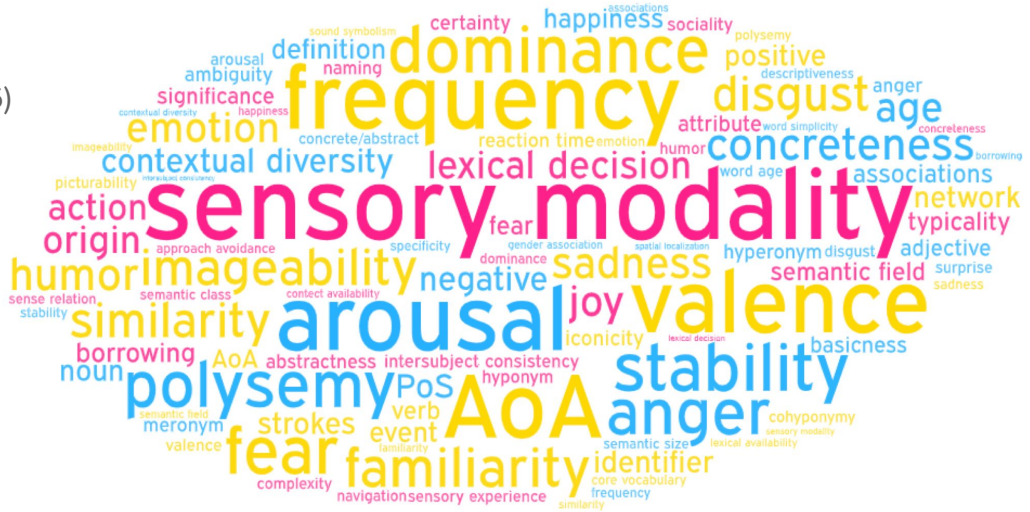


Aim

Extension of the Concepticon (List al. 2016)

Integration of Norms, Ratings, and Relations (NoRaRe)

Interdisciplinary resource for linguistics and psychology



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

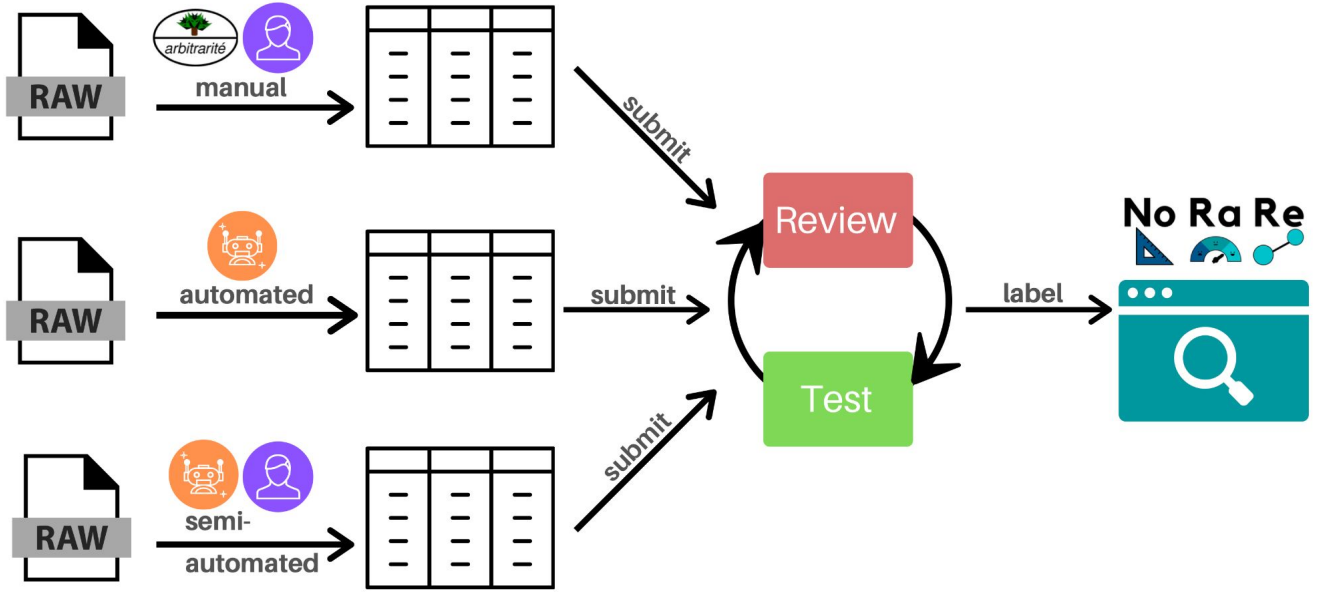
Tjuka et al. (2023): *Open Science Europe*

Materials & Methods

- 113 data sets, 39 languages, 75 data types in NoRaRe v1.0
- Manual, automated, semi-automated workflow
- Test-driven data curation
- Convenient access of data in web app

Tjuka et al. (2022): *Behavior Research Methods*
Tjuka et al. (2023): *Open Science Europe*

Materials & Methods



Tjuka (2020a; 2021a; 2021c): Tutorials in *Computer-Assisted Language Comparison in Practice*

Application

Material: Word frequency norms for English (Brysbaert & New 2009), German (Brysbaert et al. 2011), and Chinese (Cai & Brysbaert 2010).

Method: Automated mapping, Pearson correlation.

Hypothesis: Genealogically related languages have more similar word frequency distributions than unrelated languages.

Result: $\log_{10}$ word frequencies are more similar in English and German (r = .76) versus Chinese-English (r = .71) and Chinese-German (r = .68).

Tjuka (2020c): *Proceedings Aspects of the Lexicon at ACL*

Agenda

I Background

Word Meaning and Language Comparison

II Database

Cross-Linguistic Norms, Ratings, and Relations

III Studies

Colexifications of Body Part and Object Concepts

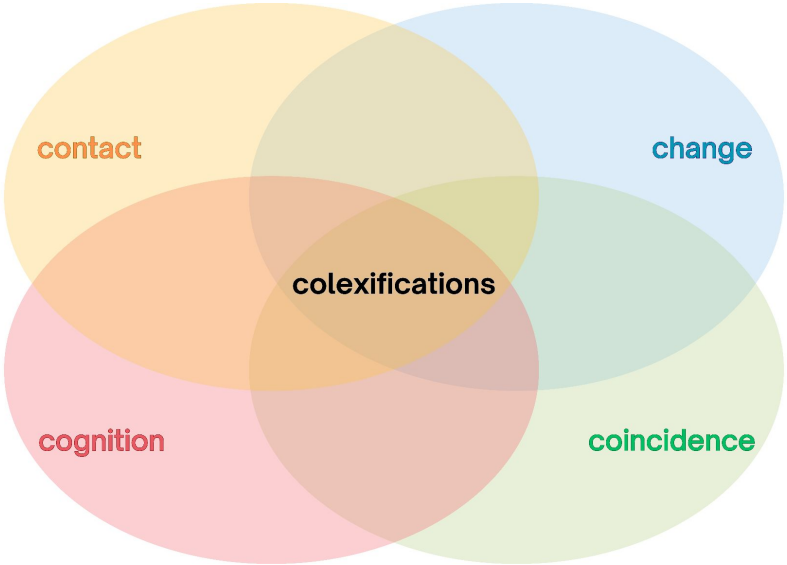
Colexifications of Body Part Concepts

IV Conclusions & Outlook

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.

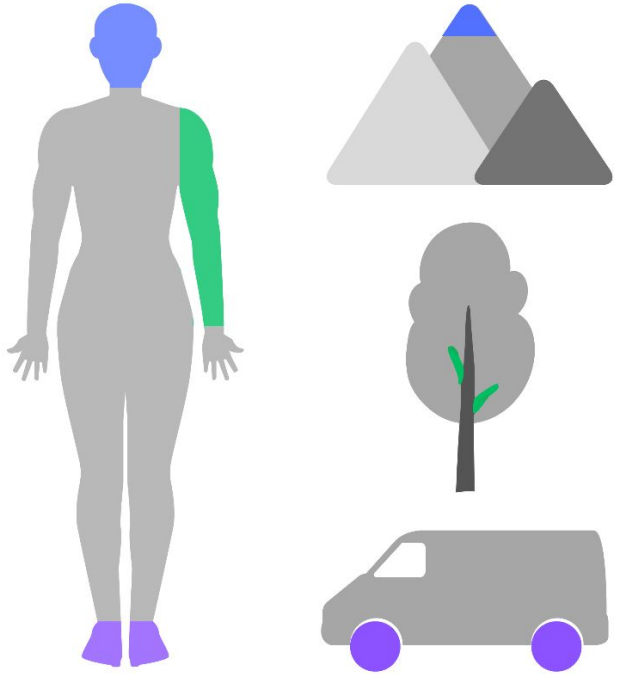


Aim

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

Quantitative study on perceptual features (vision and touch)

Qualitative study on partial colexifications in Vietnamese

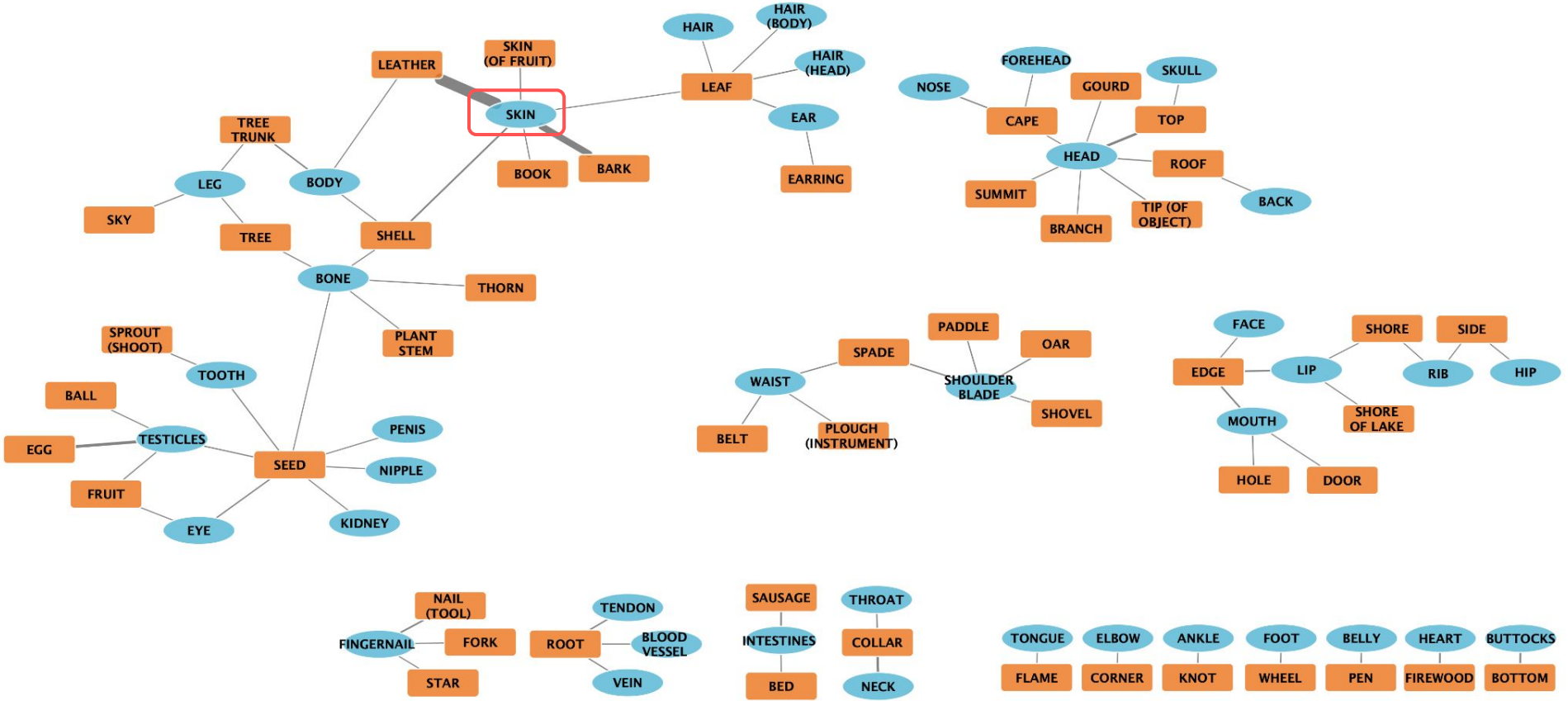


Tjuka (forthcoming): *Linguistic Typology*

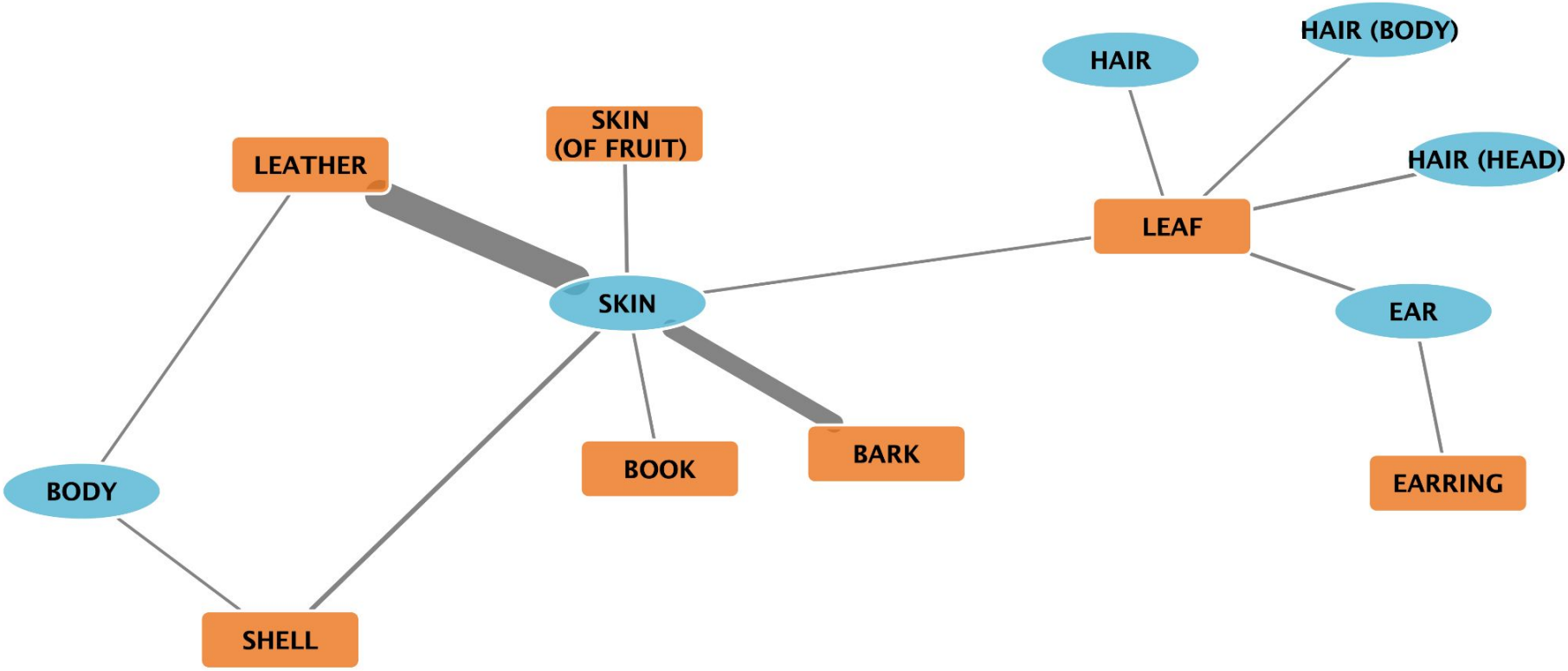
Materials & Methods

- 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 (forthcoming): *Linguistic Typology*



Tjuka (2024): Tutorial in
Computer-Assisted Language Comparison in Practice

Perceptual Features: Vision & Touch

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

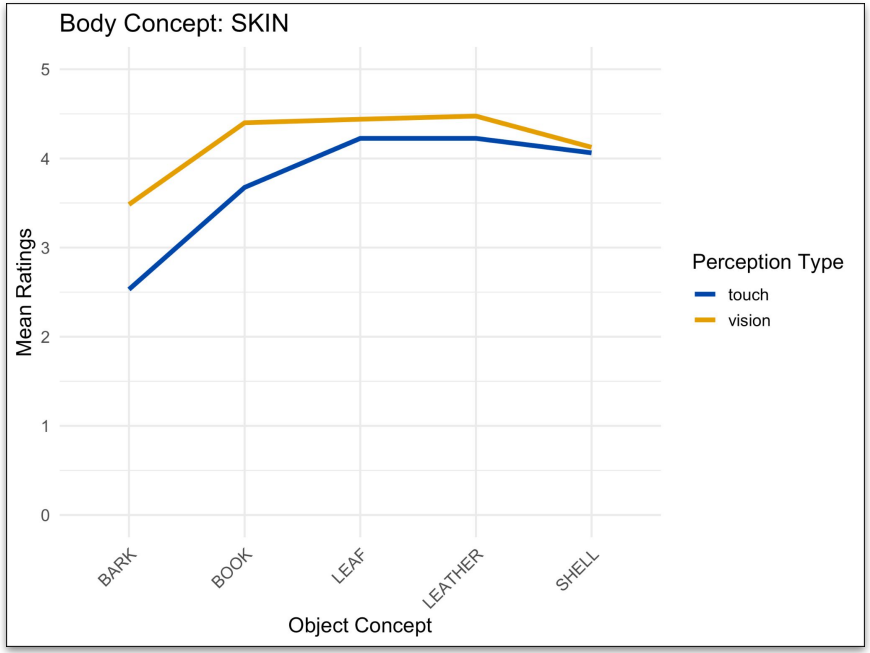
Method: Bayesian linear regression model with perception type as varying residuals.

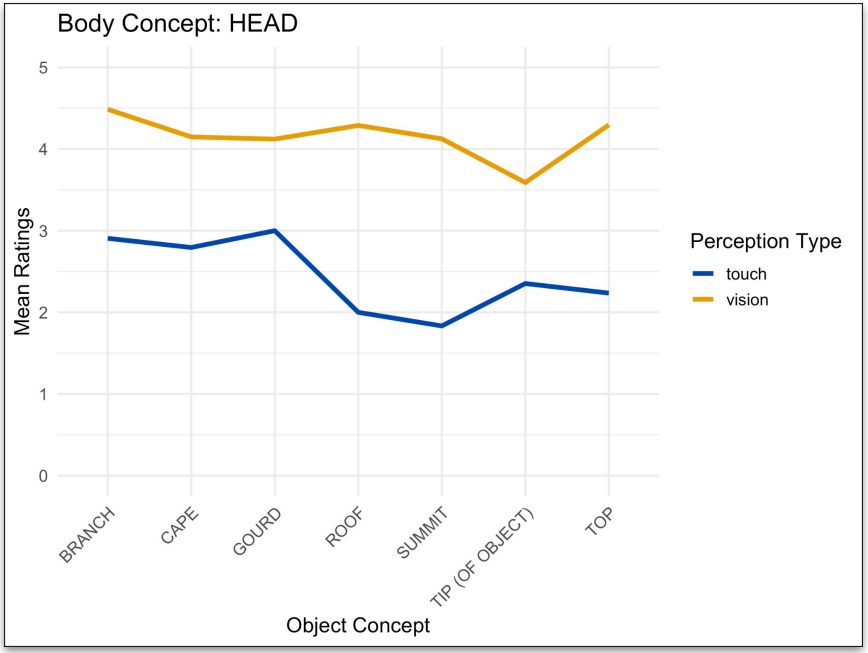
Question: Are body and object concepts perceived similarly across speakers?

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

Tjuka (forthcoming): *Linguistic Typology*

Perceptual Features: Vision & Touch





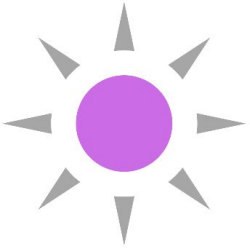
Tjuka (forthcoming): *Linguistic Typology*

Vietnamese

Material:	Partial colexifications of 4 human body part terms: *đầu* ‘head’, ***mặt*** **‘face’**, *mũi* ‘nose’, *tay* ‘hand, arm’, and *chân* ‘foot, leg’.
Method:	Analysis of examples and comparison with cross-linguistic sample.
Question:	Is the same perceptual feature used to establish body-object colexifications across different objects?
Result:	The features of shape and function establish most partial colexifications with *mặt* ‘face’. Similar patterns were found in the cross-linguistic sample.

Tjuka (2023): *Embodiment in Cross-Linguistic Studies: The ‘Face’*

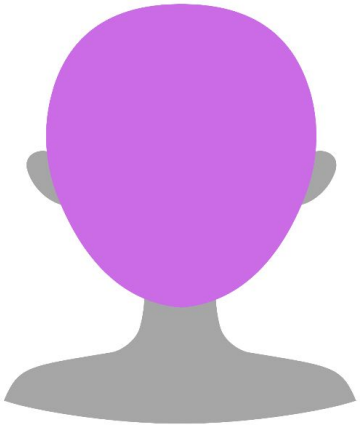
mặt trời

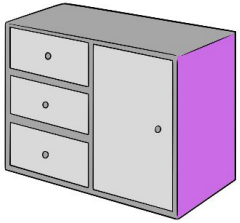


mặt trăng

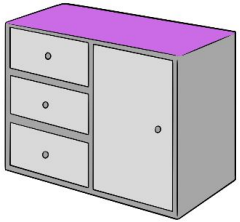
mặt đồng hồ

mặt

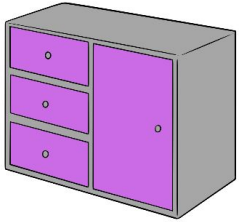




mặt bên phải tủ quần



mặt trên tủ quần



mặt tủ quần

Tjuka (2023): *Embodiment in Cross-Linguistic Studies: The 'Face'*

Agenda

I Background

Word Meaning and Language Comparison

II Database

Cross-Linguistic Norms, Ratings, and Relations

III Studies

Colexifications of Body Part and Object Concepts

Colexifications of Body Part Concepts

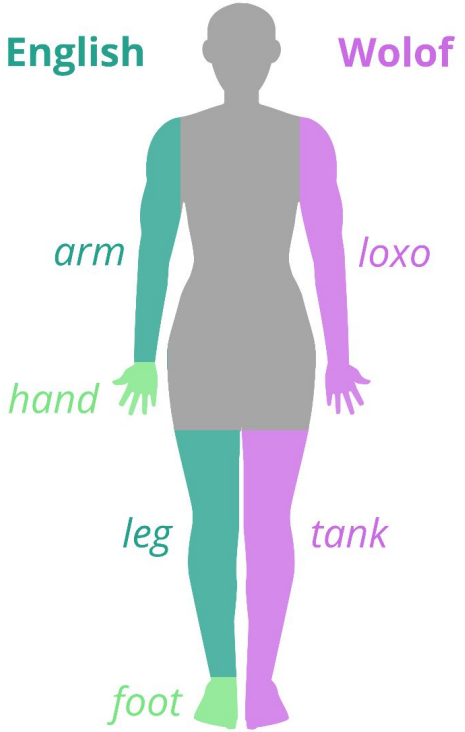
IV Conclusions & Outlook

Aim

Exploration of variation in human body part vocabularies across languages

Analysis of perceptual features (contiguity, function, shape)

Comparison with the semantic domains of colour and emotion

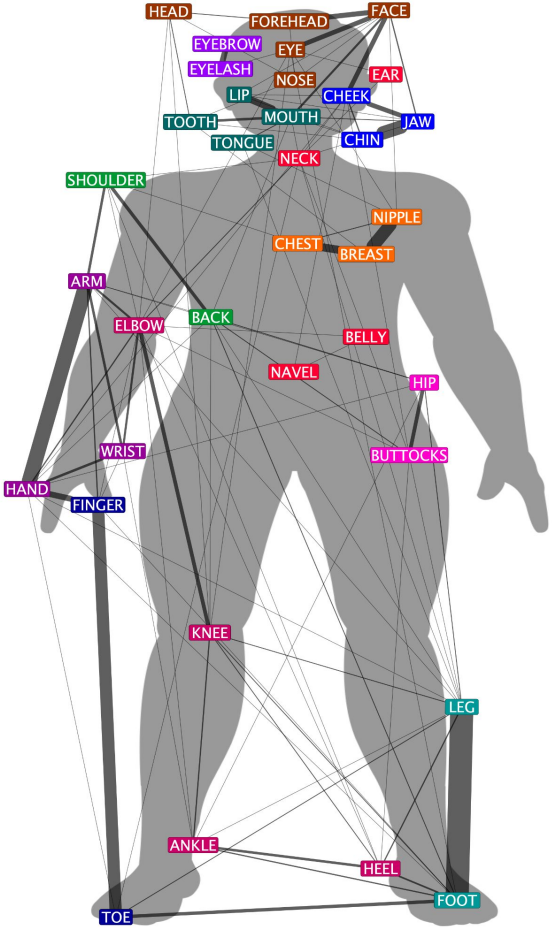


Tjuka et al. (in review): *Scientific Reports*

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. (in review): *Scientific Reports*

Perceptual Features: Contiguity, Shape, Function

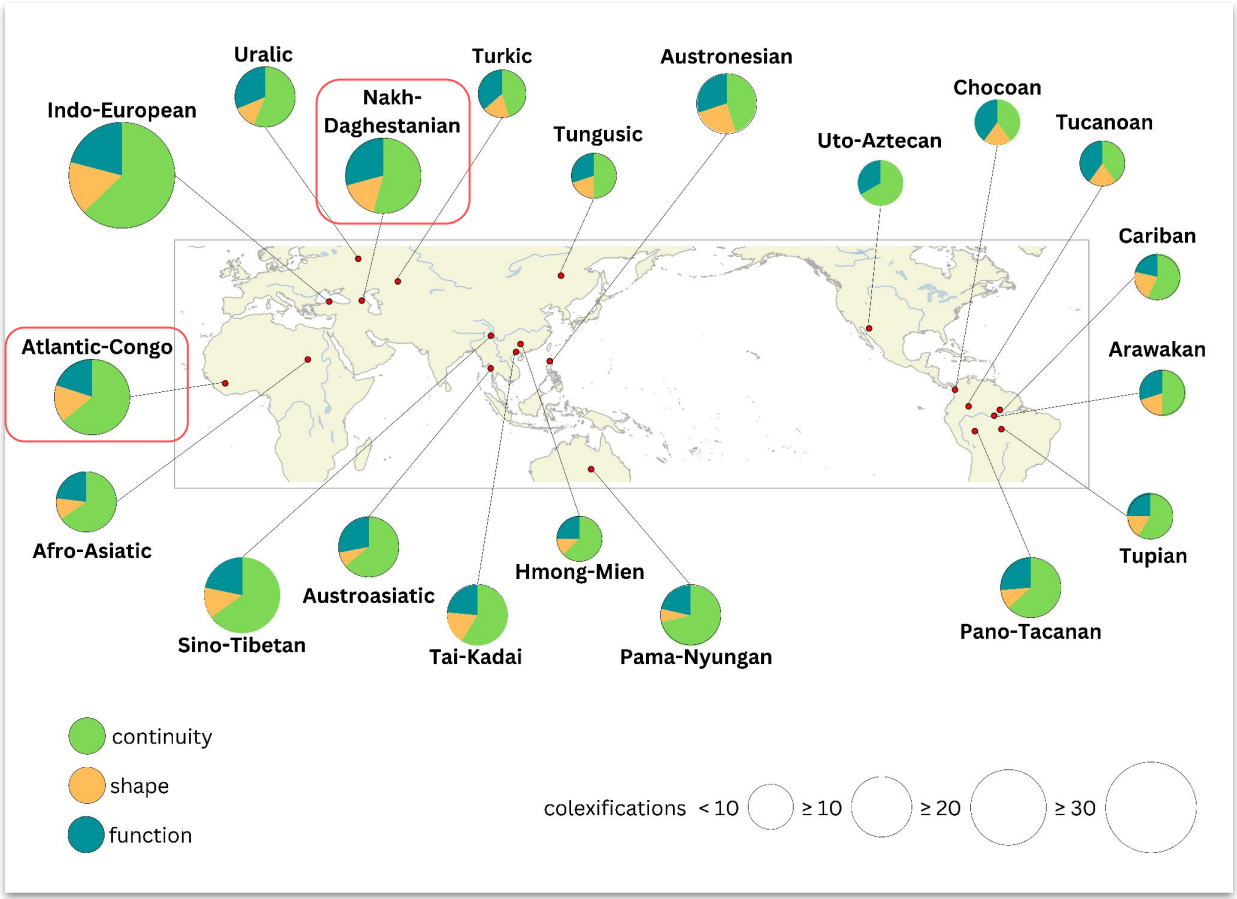
Material: Frequency of 110 body part colexifications across 20 language families.

Method: Coding body part colexifications for contiguity, shape, function.

Question: Do languages in a language family prefer body part colexifications based on a particular perceptual feature?

Result: Body parts that are adjacent to one another are more likely to be colexified. Slight differences in proportions, but function outweighs shape in all language families.

Tjuka et al. (in review): *Scientific Reports*



Tjuka et al. (in review): *Scientific Reports*

Semantic Domains

Material: Networks of 36 body part concepts, 22 colour concepts, and 62 emotion concepts.

Method: Replication and extension of Jackson et al. (2019).

Question: How similar are the network clusters of the three semantic domains?

Result: Body part colexification networks (*ARI* = .3) differ significantly from colour (*ARI* = .16) and emotion networks (*ARI* = .14), while colour and emotion networks are similar.

Tjuka et al. (in review): *Scientific Reports*

Agenda

I Background

Word Meaning and Language Comparison

II Database

Cross-Linguistic Norms, Ratings, and Relations

III Studies

Colexifications of Body Part and Object Concepts

Colexifications of Body Part Concepts

IV Conclusions & Outlook

Conclusions

?

Conclusions

Standardised data sets facilitate systematic comparison across languages.

Conclusions

Standardised data sets facilitate systematic comparison across languages.

New reproducible workflows are applicable to other semantic domains.

Conclusions

Standardised data sets facilitate systematic comparison across languages.

New reproducible workflows are applicable to other semantic domains.

Large-scale approaches enable further analyses and data collection.

Conclusions

Standardised data sets facilitate systematic comparison across languages.

New reproducible workflows are applicable to other semantic domains.

Large-scale approaches enable further analyses and data collection.

Visual perception is one factor behind colexifications.

Outlook

Method

Adding more data to NoRaRe

Integrating partial colexifications

Conducting targeted studies

Thank you

Publications

Tjuka, Annika. 2020c. General Patterns and Language Variation: Word Frequencies Across English, German, and Chinese. In Michael Zock, Emmanuele Chersoni, Alessandro Lenci & Enrico Santus (eds.), *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, 23–32. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.cogalex-1.3.

Tjuka, Annika. 2023. Body Part Extensions with *Mặt* 'Face' in Vietnamese. In Kelsie E. Pattillo & Małgorzata Waśniewska (eds.), *Embodiment in Cross-Linguistic Studies: The 'Face,'* 237–255. Leiden: Brill. https://doi.org/10.1163/9789004521971_012.

Tjuka, Annika. 2024. Objects as Human Bodies: Cross-Linguistic Colexifications Between Words for Body Parts and Objects. Linguistic Typology.

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. *PsyArXiv*. https://osf.io/tu74k.

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.