Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

# Missione 4 Istruzione e Ricerca

## Rethinking, Understanding Modal Particles (RUM)

# Multimodal Profiles of Modal Particles
## Tracking *schon* in German Political Discourse

Paula Rebecca Schreiber (Università degli Studi di Bergamo)

2025-09-24

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

Die in diesem Vortrag präsentierten Inhalte und Ergebnisse sind Teil der Forschung, die Rebecca Schreiber im Rahmen des Projekts „Rethinking, Understanding Modal particles (RUM)“ durchgeführt hat. Das Projekt wurde finanziert durch PNRR – Missione 4 „Istruzione e ricerca“ – Componente C2 Investimento 1.1 „Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)“ – Direttoriale n. 104 del 02-02-2022 – CUP H53D23004410006 (Universität Venedig) und CUP F53D23004890006 (Universität Bergamo).

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

# Multimodal Profiles of Modal Particles
## Tracking *schon* in German Political Discourse

das habe ich eben **schon** mal erwähnt

*I've **already** mentioned that once before*

[ConcRUM_6097]

temporal reading

und da sind wir auf einem kleinen Weg irgendwo **schon** am Wegesrand stehen geblieben

*and at **some point** along a small path we ended up stopping by the side of the road*

[ConcRUM_2239]

scalar reading

das BSI schon irgendwie richten wird

*the BSI will **surely** sort it out somehow*

[ConcRUM_3493]

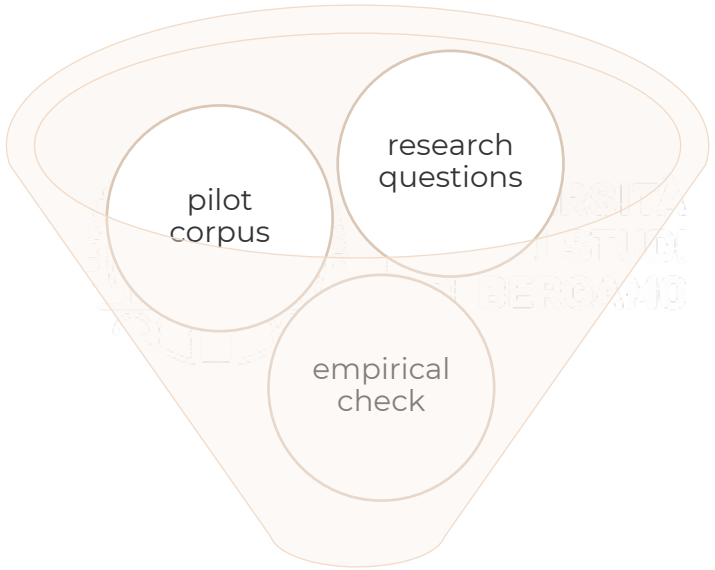
modal reading

The polyfunctionality of ***schon*** and its readings

# Outline

- The German Bundestag Dataset
- Corpus-Driven Analysis of Frequency & Collocation
- From Data to Multimodal Annotation
- Towards Cross-Linguistics Multimodal Analysis

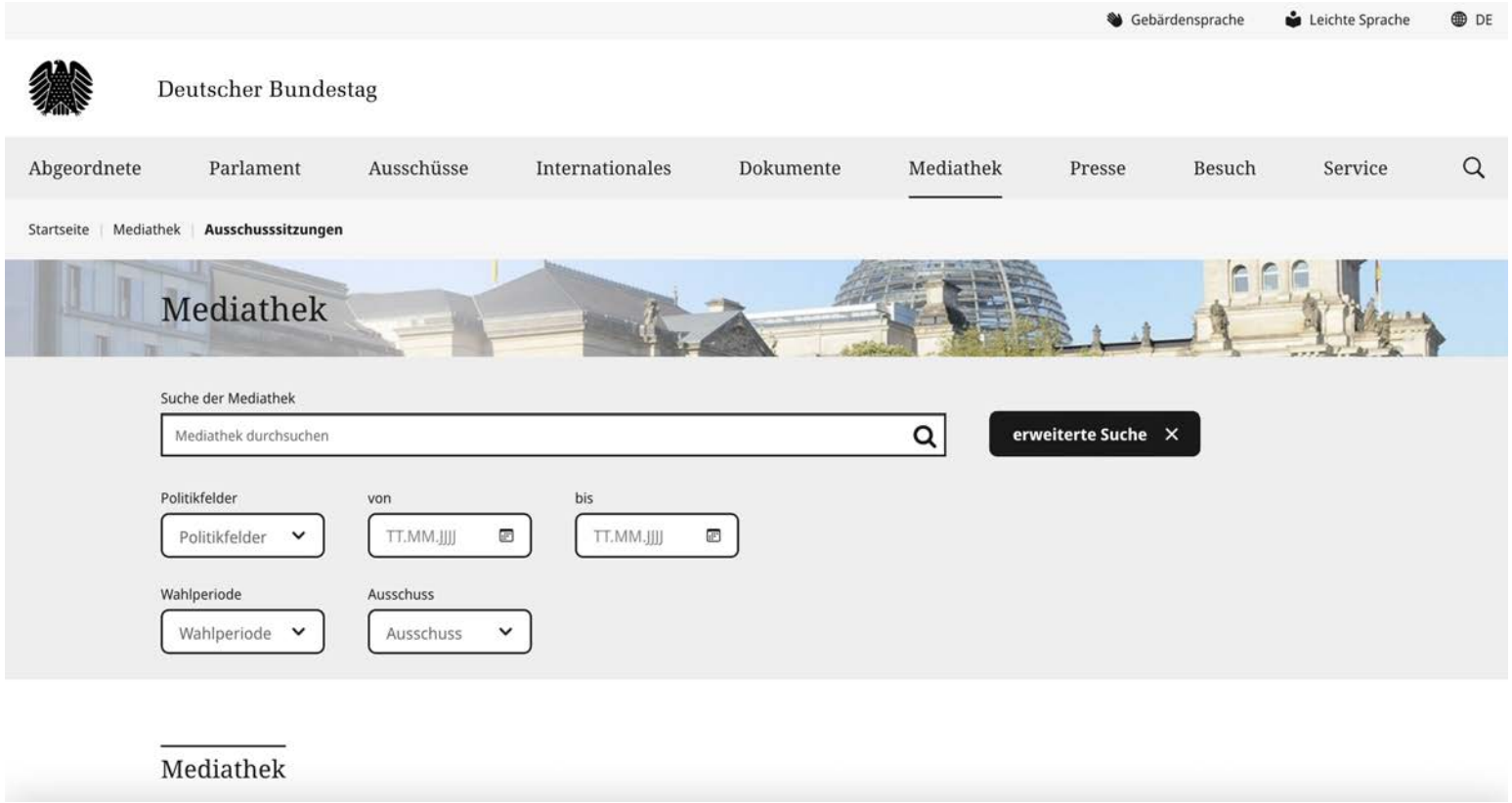
# The German Bundestag Dataset




constraints ⇄ corpus refinement ⇄ data sources

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

# The German Bundestag Dataset




- source: **German Bundestag Mediathek**
- openly accessible
- educational/research use permitted
- download formats: multiple qualities available
- chosen format: .mp4
  - stable
  - annotation-friendly
  - space-efficient

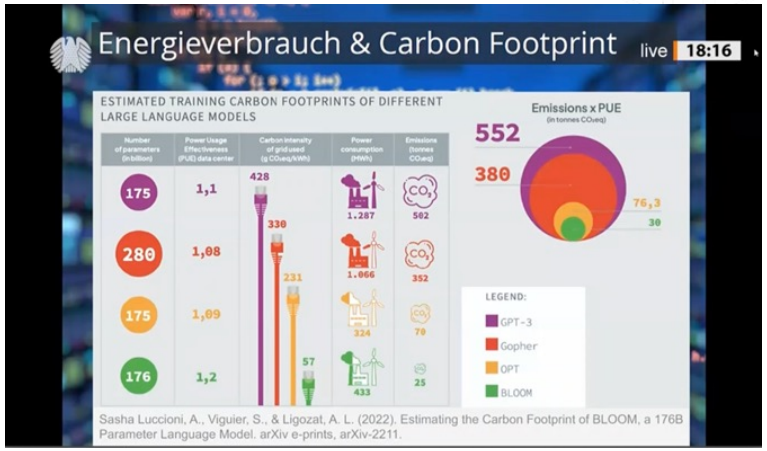
URL: https://www.bundestag.de/mediathek/ausschusssitzungen [2025-09-18]

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

# The German Bundestag Dataset


roundtable arrangement


auditorium setup




presentation mode


honorary guest arrangement

URL: https://www.bundestag.de/mediathek [2025-09-22]

# The German Bundestag Dataset

**Type**: Committee meetings
**Video material**: 186 recordings
**Duration**: 301 hours, 40 minutes, 05 seconds
**Timeframe**: January 1, 2024 – December 31, 2024
**Tokens**: 2.476.276
**Types**: 94.103
**Potential occurrences**: 85.170

**Automated Speech Recognition** (ASR)
**Model**: Whisper large-v3
**Layers**: 32
**Width**: 1280
**Heads**: 20
**Parameters**: 1550M

| auch | bloß | denn | doch | eben | halt | ja | mal | nur | schon |
|---|---|---|---|---|---|---|---|---|---|
| 44.712 | 37 | 2.884 | 1.417 | 4.837 | 409 | 8.335 | 9.092 | 5.981 | 7.466 |

**Table 1**: Token counts of particles with potential modal function

# The German Bundestag Dataset

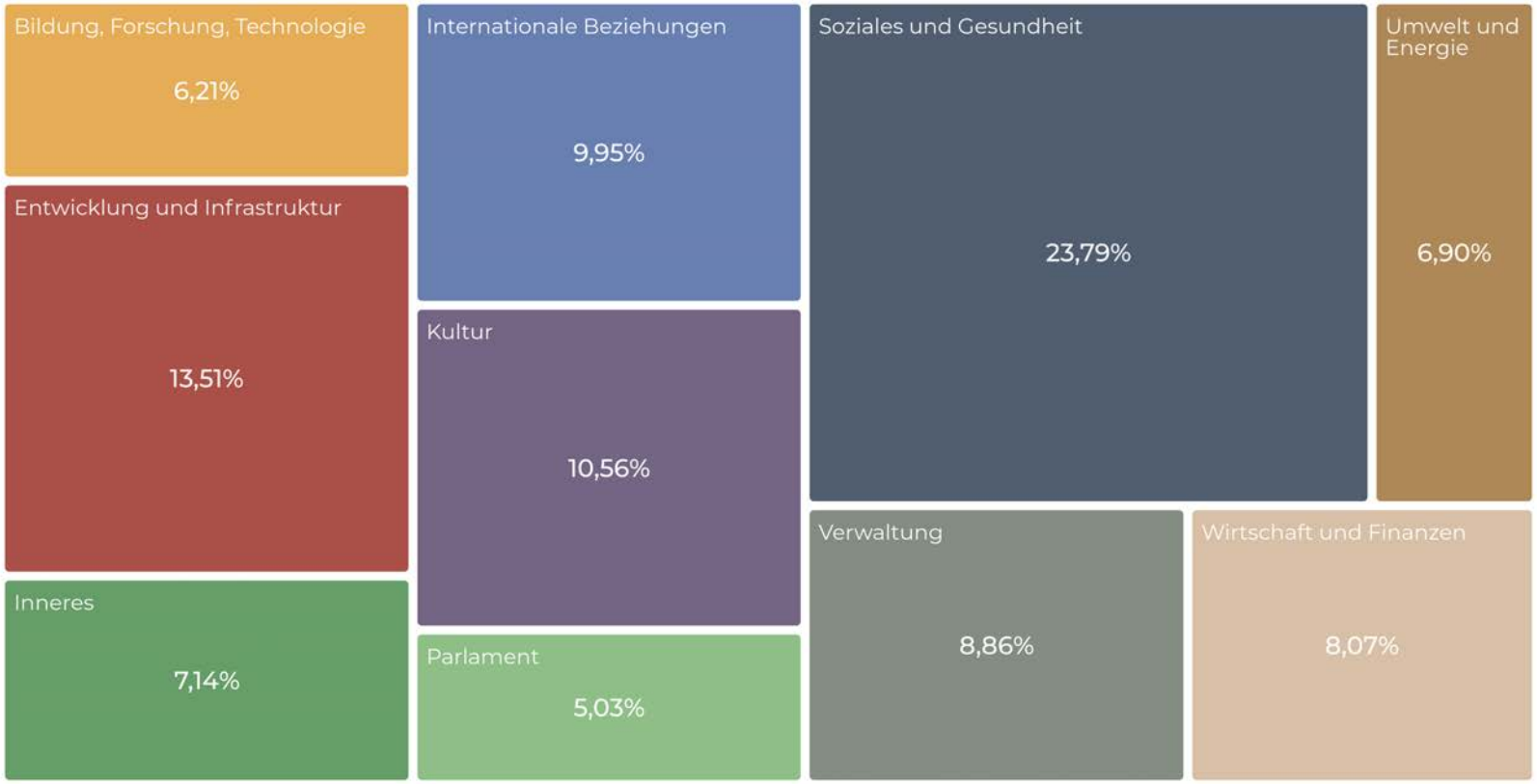



**Figure 1:** Distribution of committee meetings from the German Bundestag corpus across thematic domains (percentages)

# Corpus-Driven Analysis of Frequency & Collocation

- quantitative approach (Gries 2025)
- core measures
  - **frequency**
  - **association**
- bootstrapping to address statistical uncertainty (Gries 2022)
  - confidence intervals from repeated resampling
  - provides robust estimates, not just point values
- collocation profiles
  - advanced filtering
  - case analysis
  - multimodal annotation

# Corpus-Driven Analysis of Frequency & Collocation

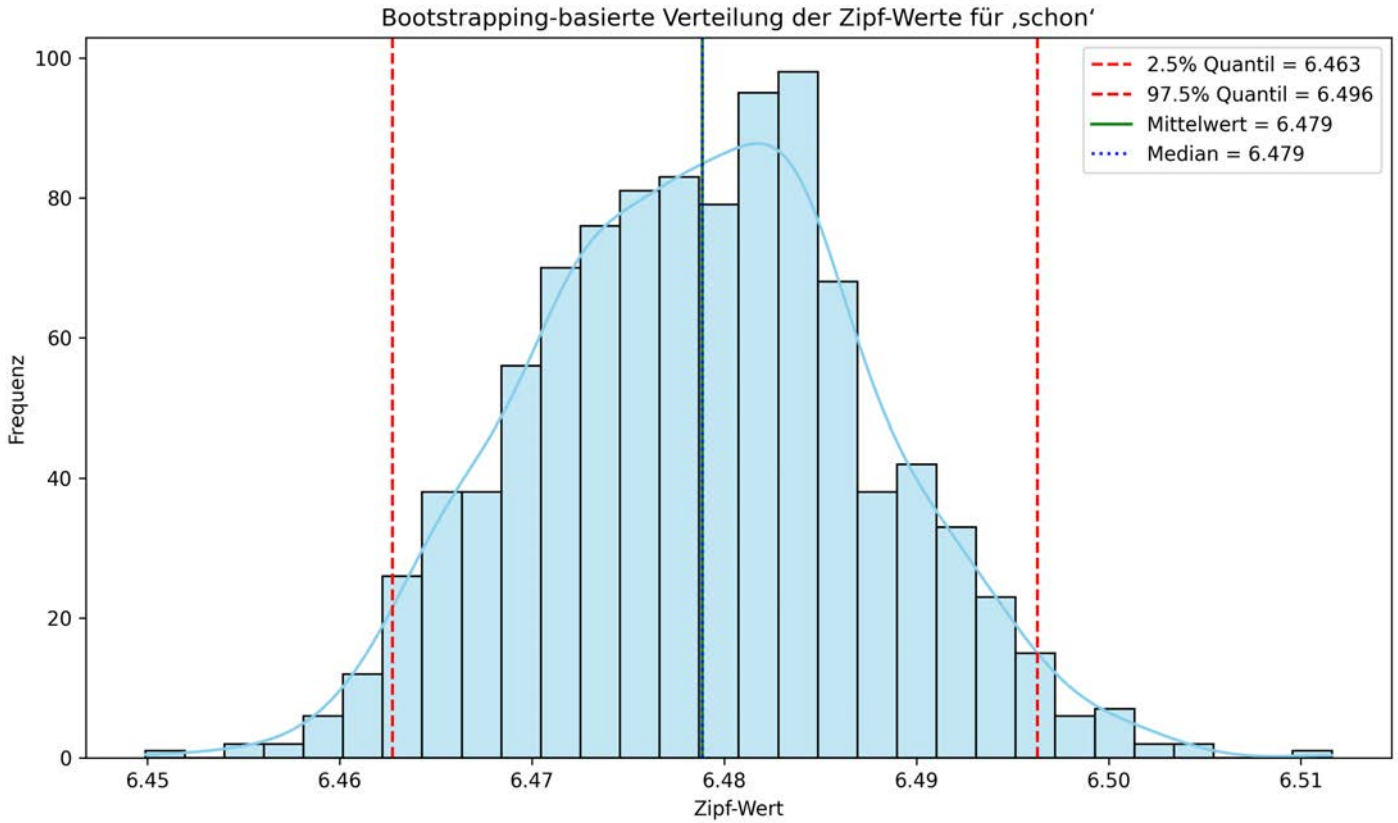



**Figure 2:** Zipf value distribution of ***schon*** with 95% confidence interval (1,000 resamples)

# Corpus-Driven Analysis of Frequency & Collocation

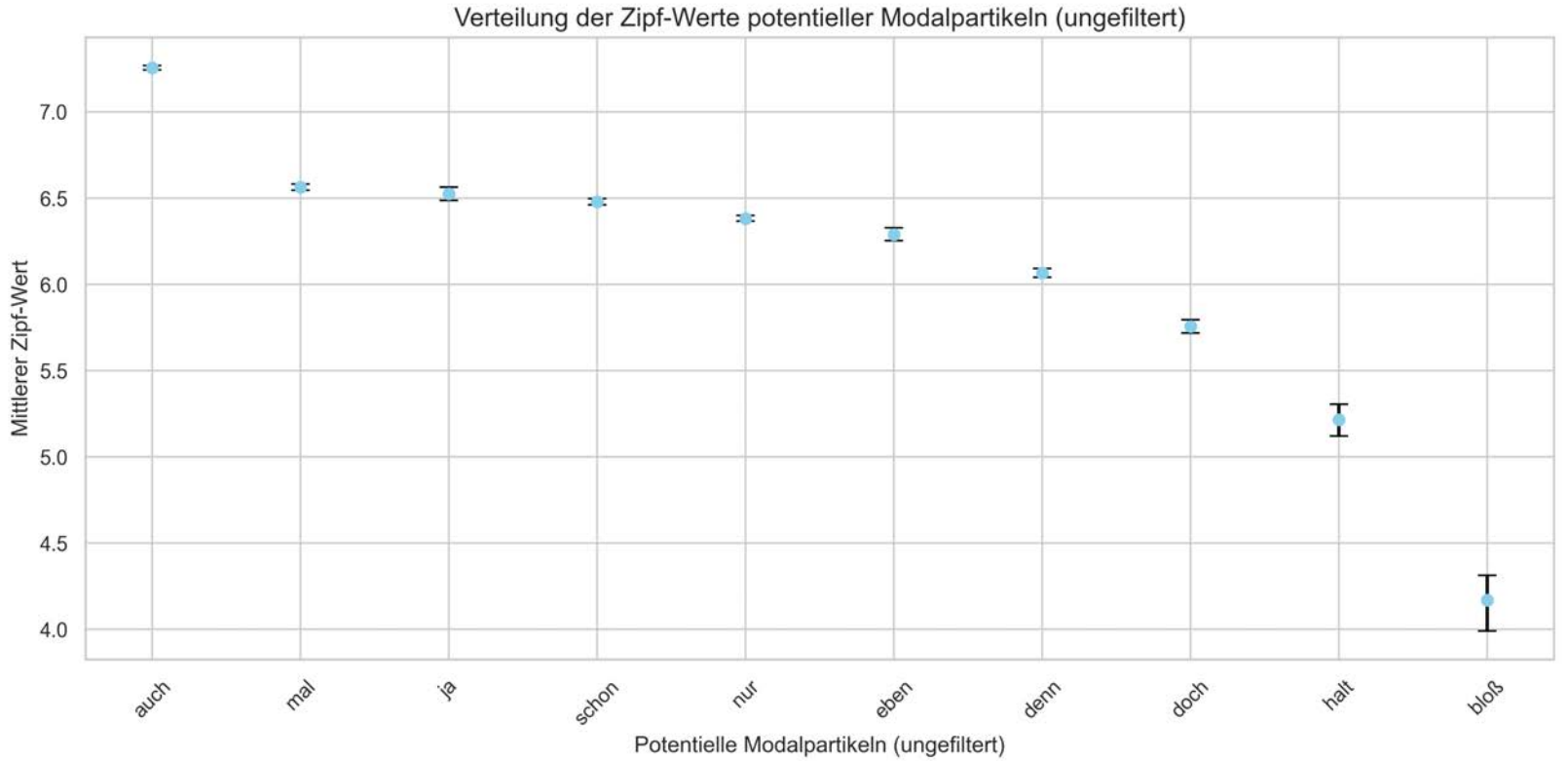



**Figure 3:** Mean Zipf values and 95% confidence intervals of potential modal particles (1,000 resamples)

## Corpus-Driven Analysis of Frequency & Collocation

- corpus segmented into bigrams (pairs of consecutive tokens)
- target word: ***schon***
  - L1 = immediately preceding collocate
  - R1 = immediately following collocate
- association strength measured with:
  - log-likelihood (LL) from 2×2 contingency tables (Gries 2016, 2025)
  - directional association measures (ΔP): ΔP(X|schon) and ΔP(schon|X)
- distinguishes symmetrical vs. asymmetrical collocational attraction
- null cells (0 values) explicitly marked as “not computable”

# Corpus-Driven Analysis of Frequency & Collocation

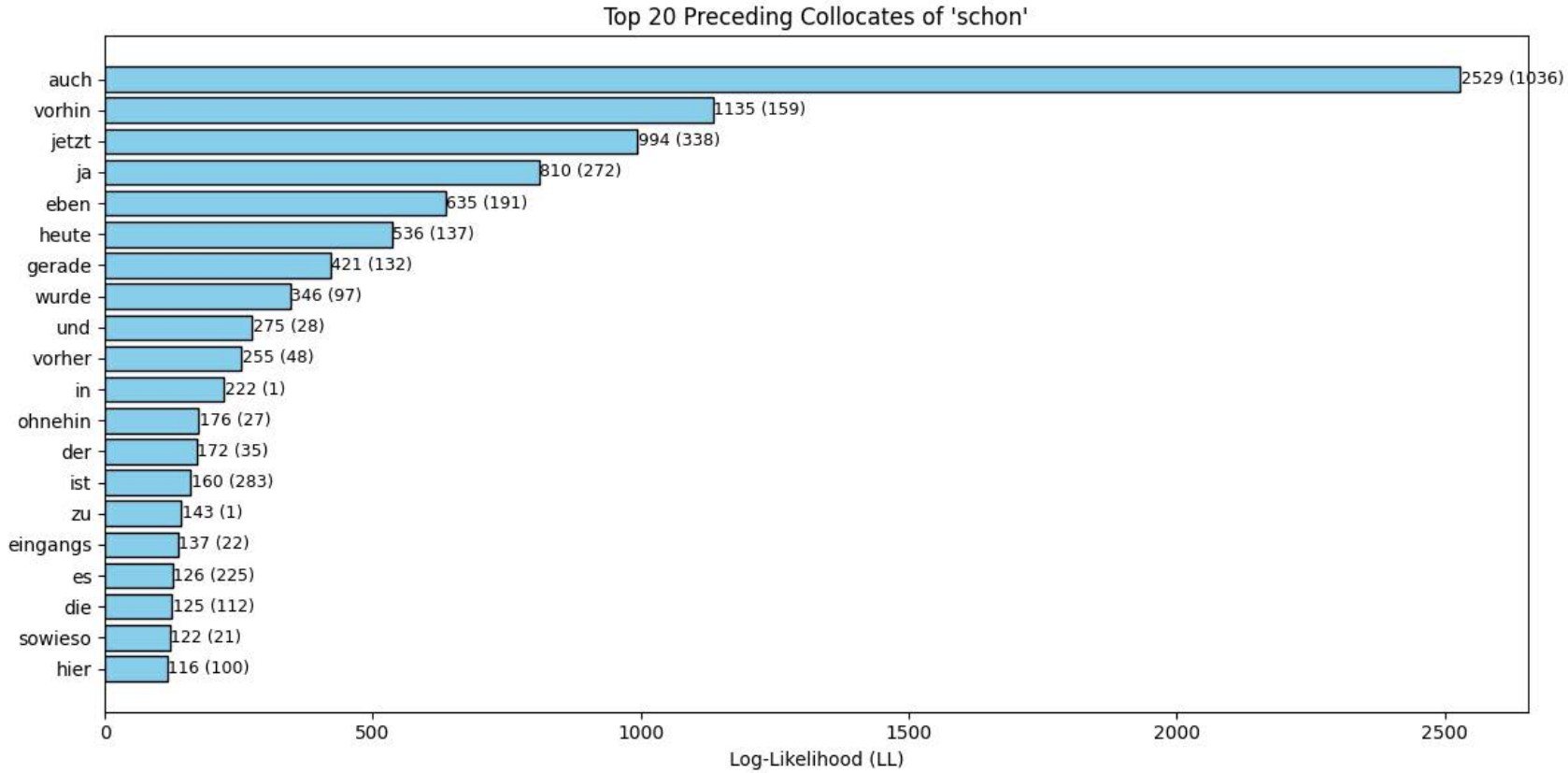



**Figure 4:** Preceding collocates of ***schon*** (ranked by LL values)

# Corpus-Driven Analysis of Frequency & Collocation

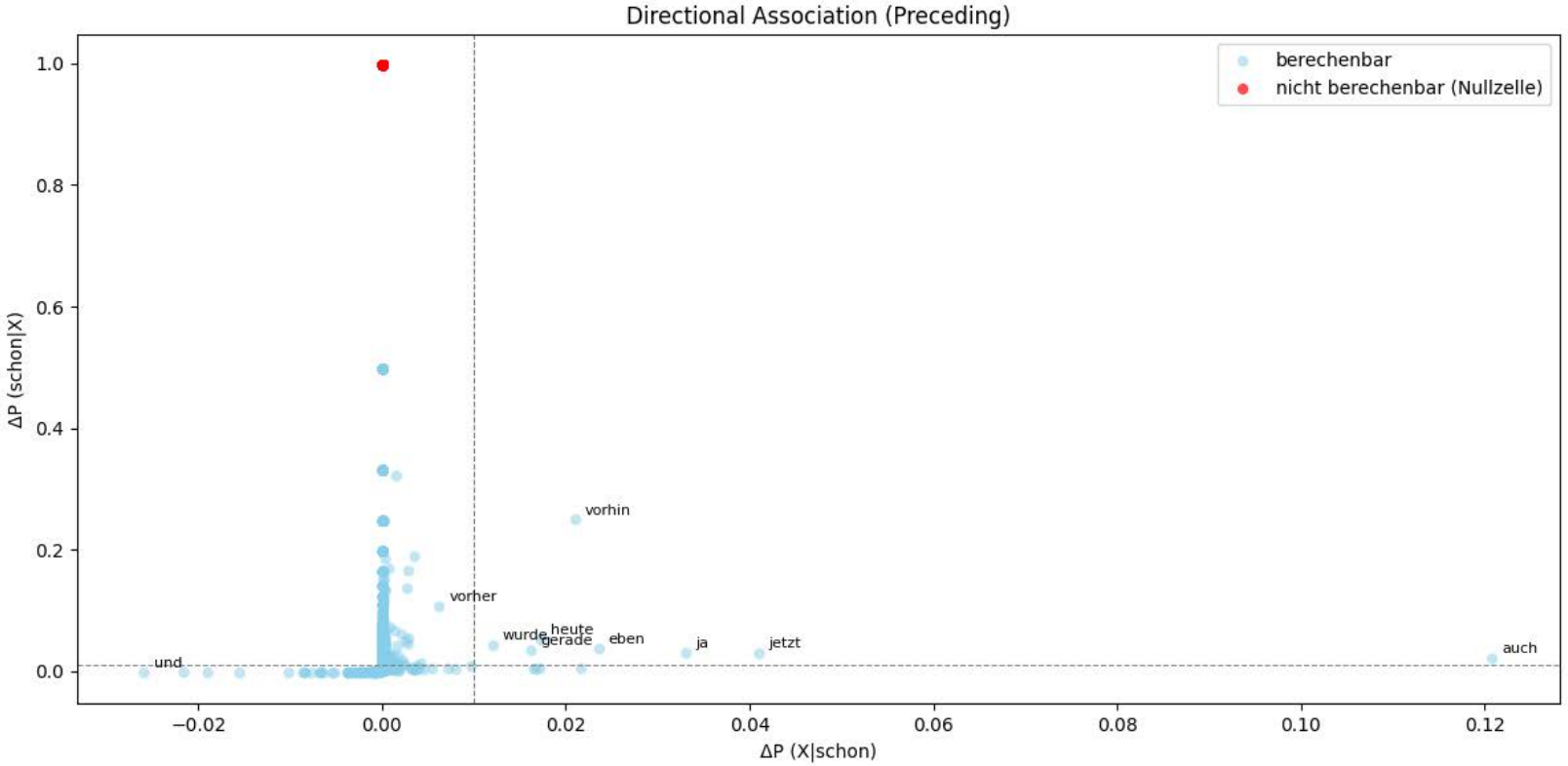



**Figure 5:** Association profile of ***schon*** and its preceding collocates

# Corpus-Driven Analysis of Frequency & Collocation

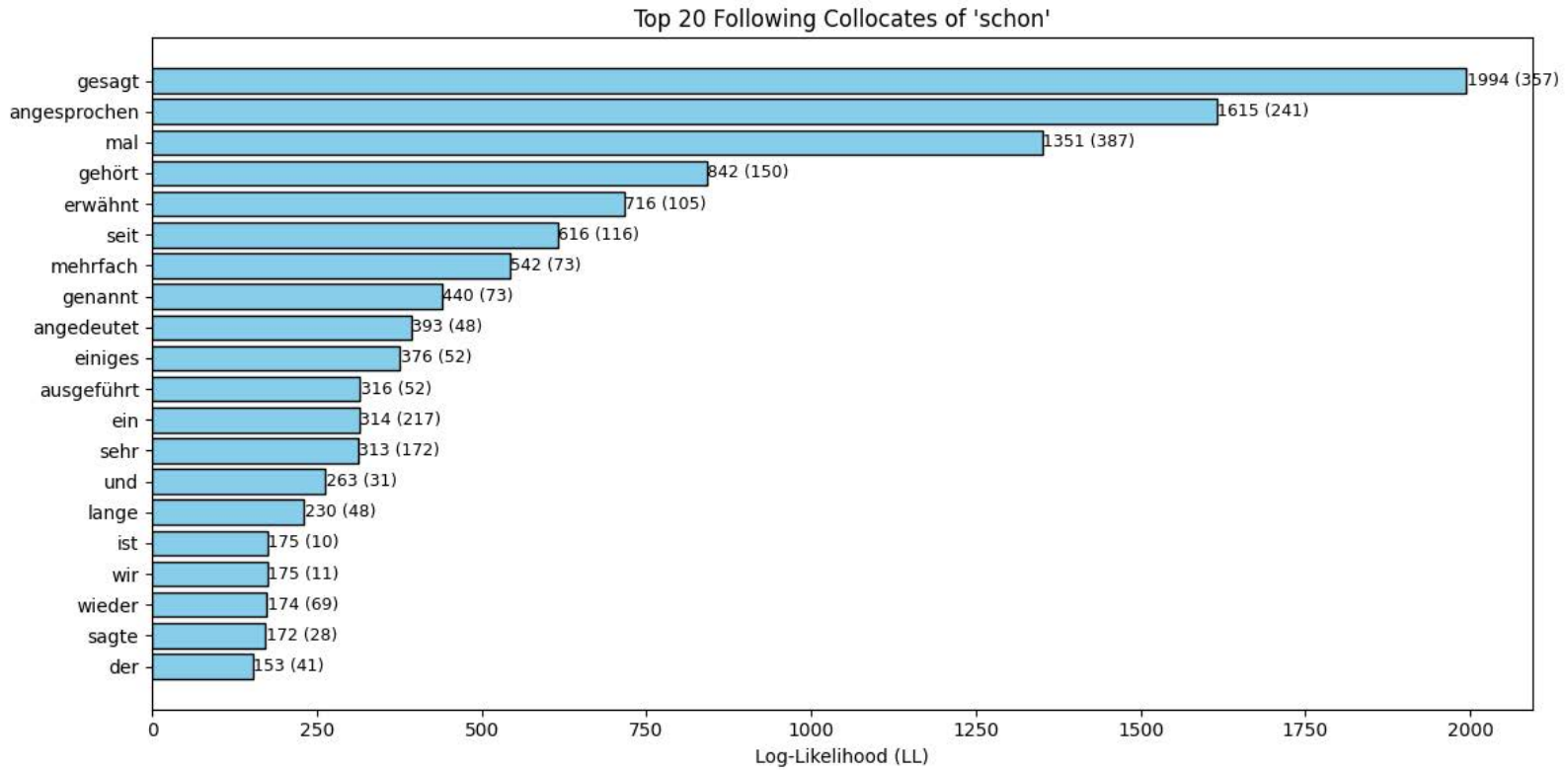



**Figure 6:** Following collocates of ***schon*** (ranked by LL values)

# Corpus-Driven Analysis of Frequency & Collocation

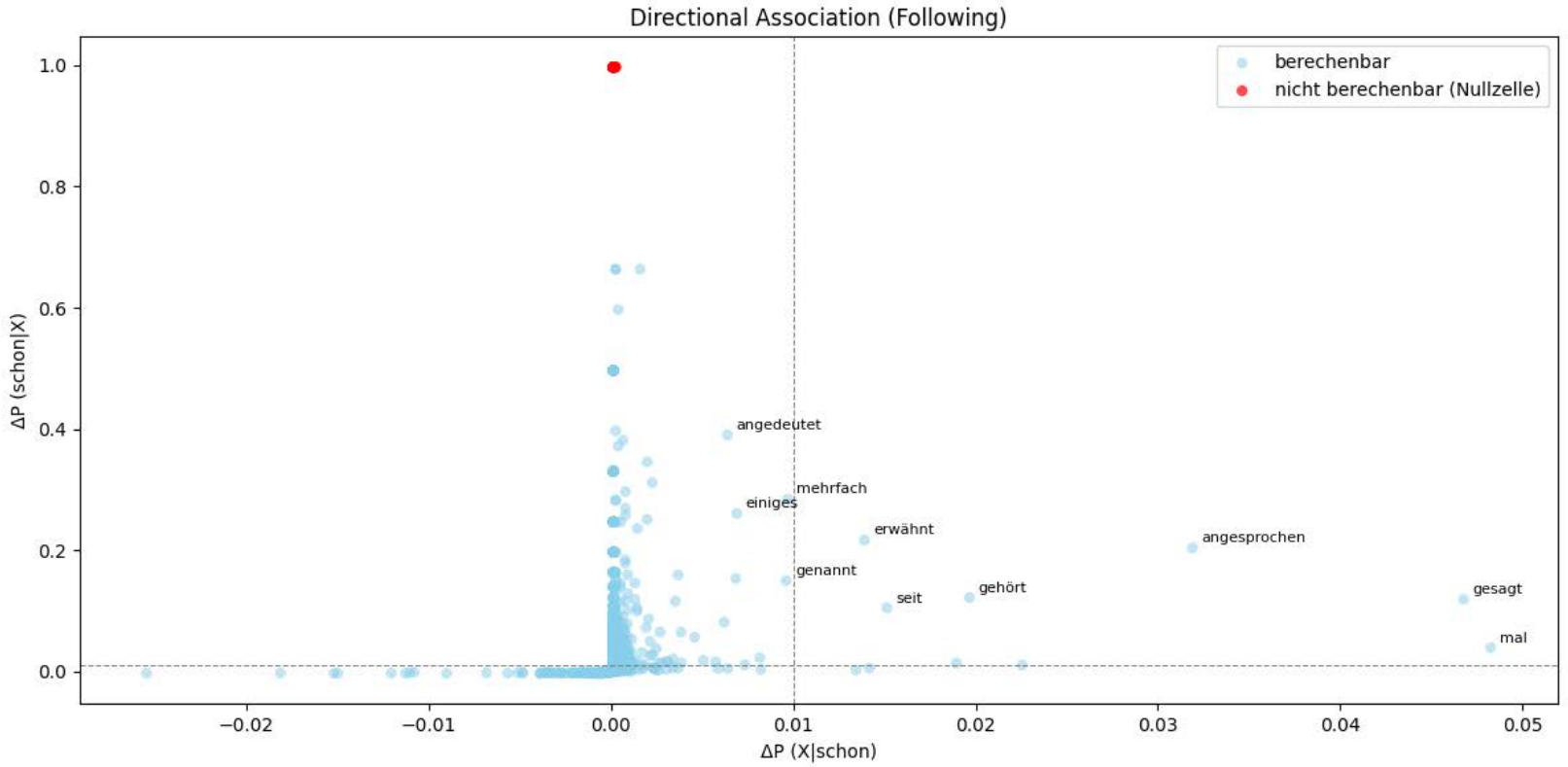



**Figure 7:** Association profile of ***schon*** and its following collocates

# Corpus-Driven Analysis of Frequency & Collocation

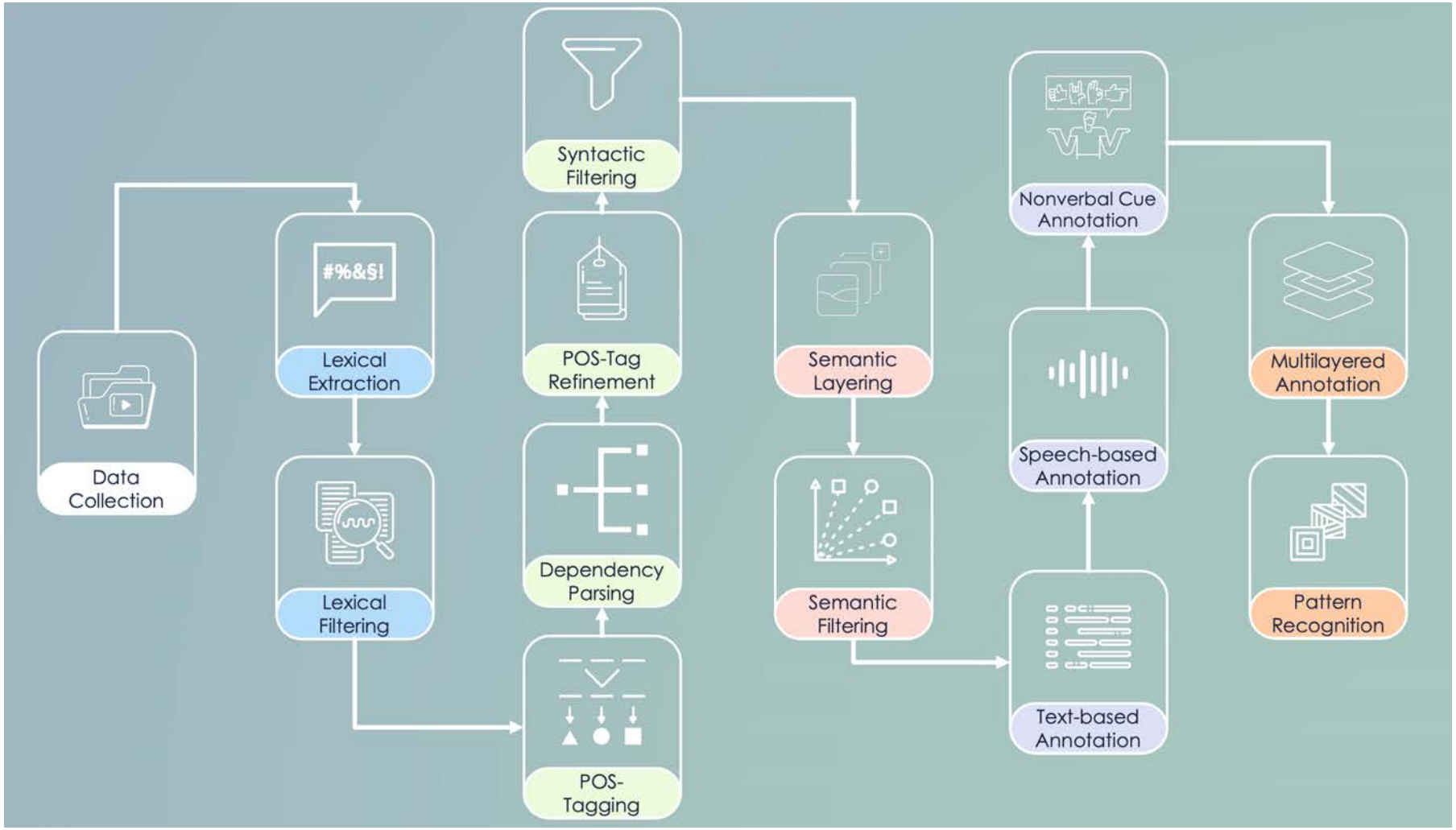
| ID_Collocation | Collocation | a (schon & X) | total_schon | total_X | LL | ΔP (X\|schon) | ΔP (schon\|X) |
|---|---|---|---|---|---|---|---|
| IDU1_0001 | ('auch', 'schon') | 1036 | 7463 | 44711 | 2529,22 | 12,1% | 2,0% |
| IDU2_0436 | ('schon', 'auch') | 110 | 7466 | 44711 | 5,80 | -0,4% | -0,1% |

**Table 2**: Association values for ***auch schon*** and ***schon auch*** presented in comparative table form

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI BERGAMO

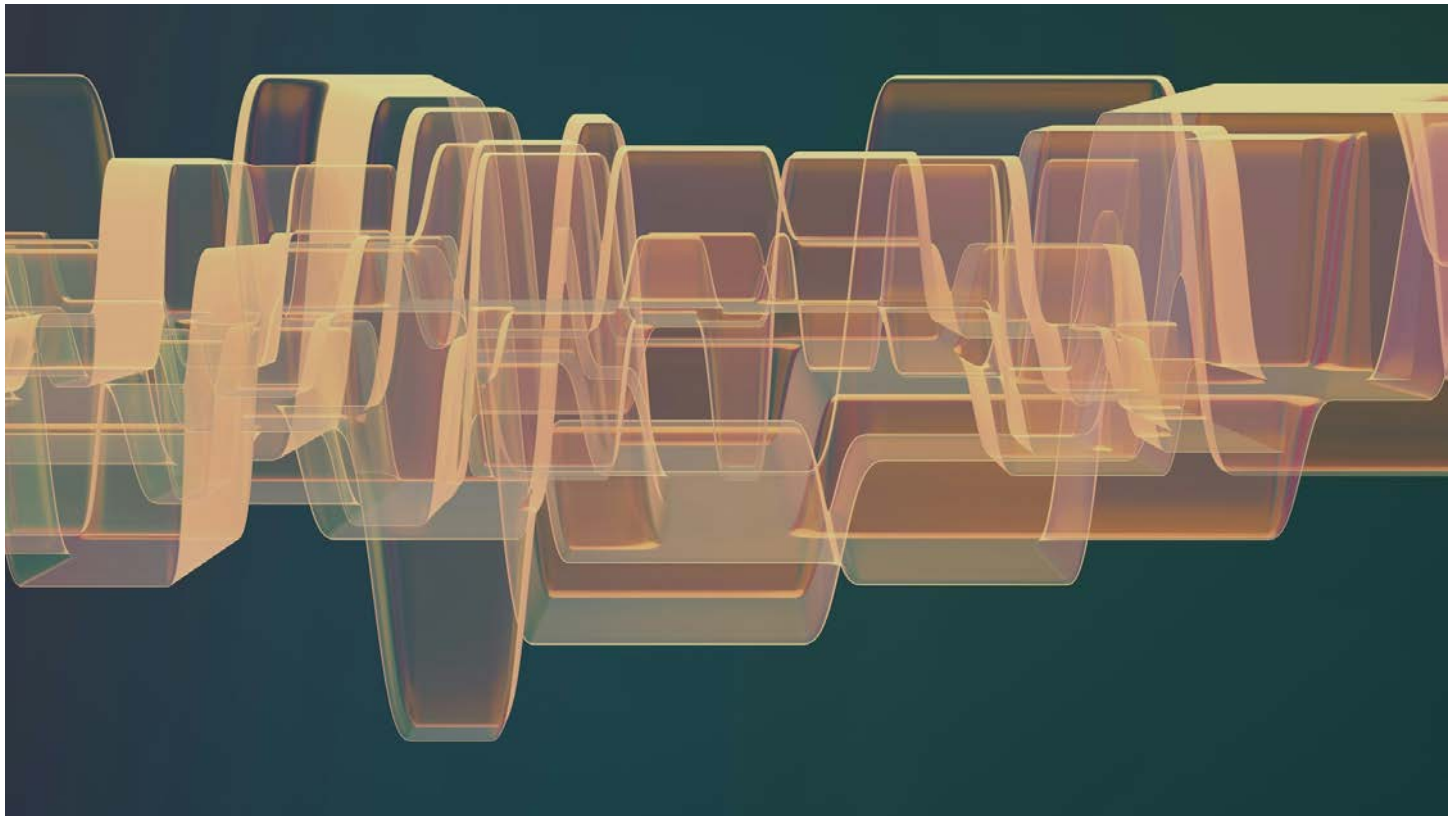
# From Data to Multimodal Annotation

Workflow Overview



# From Data to Multimodal Annotation

Data Complexity



# From Data to Multimodal Annotation

Managing Data Complexity – Filtering

- pipeline cycles
- data extraction & segmentation
- data modularisation
- feature evaluation
- operationalisation
- manage data density without loss

# From Data to Multimodal Annotation

## Establishing Taxonomies

**Cognola, Palilla & Petrocchi (2024), building on Schoonjans (2018)**

- Hand use – right, left, both
- Hand shape – e.g., ring, flat, plate, finger, carciofo
- Hand orientation – body-directed, outward, up, down, center
- Hand movement
  - Type – straight, arc, circle, zigzag, spiral
  - Direction – right, left, forward, back, up, down, diagonal
  - Character – big/small, fast/slow, long/short, flowing/jerky
  - Number – single, repeated
- Hand position – height (upper, lower, center) and depth (near, far, touching, behind)
- Head orientation – lift, tilt, direction (left, right, forward)
- Head movement
  - Direction – left, right, forward, back, up, down, diagonal
  - Character – big/small, fast/slow, long/short, flowing/jerky
  - Number – single, repeated

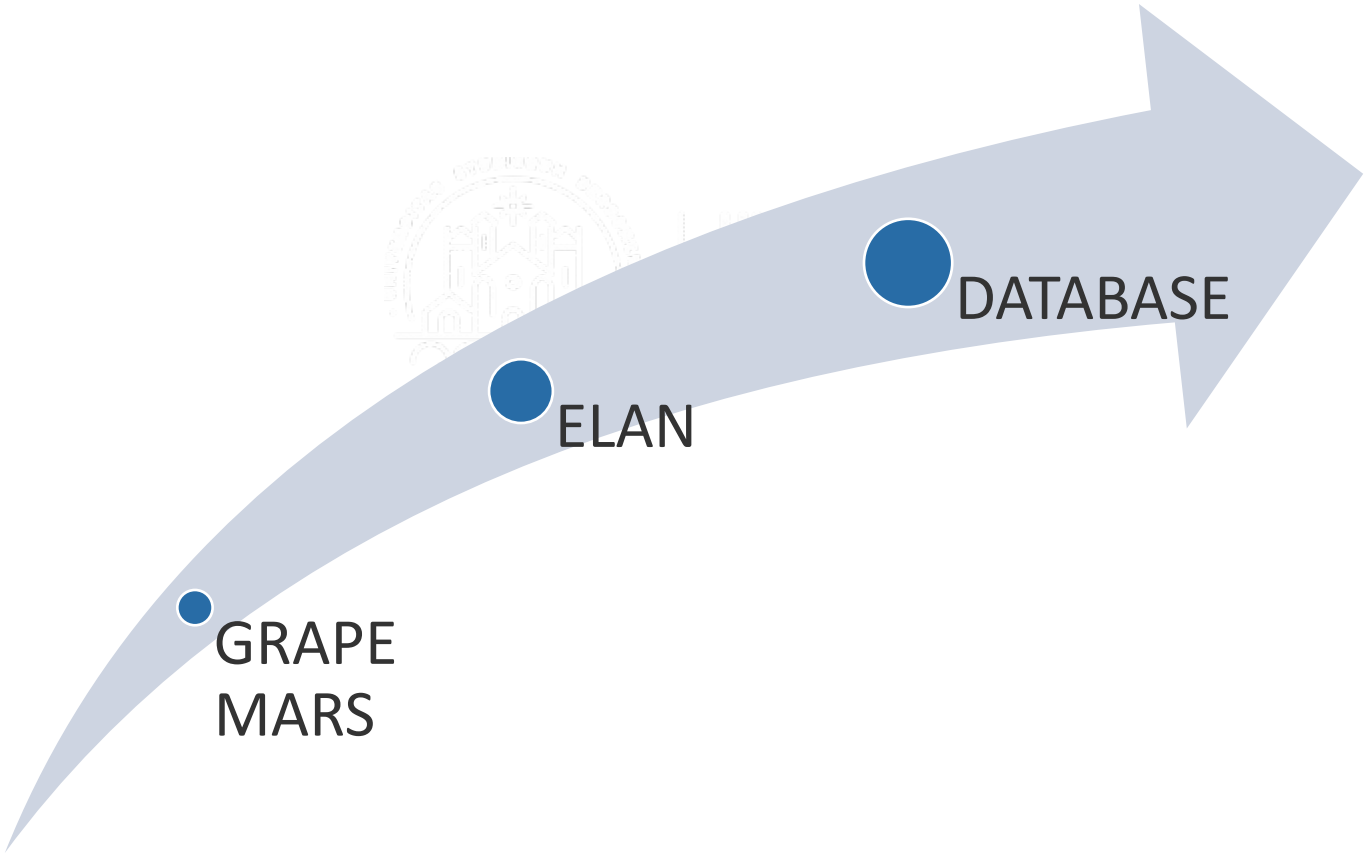
# From Data to Multimodal Annotation

Integrating Taxonomy of Cognola, Palilla & Petrocchi (2024) into the workflow

- Taxonomy of Cognola, Palilla & Petrocchi (2024)
  - fine-grained, systematic, qualitative depth
  - essential framework for multimodal analysis
- Mapped taxonomy (pipeline layer)
  - broader categories: head, facial, hand gestures
  - parameters: amplitude, duration, repetition, appearance
  - enables filtering & operationalisation
- Complementarity
  - Mapping = manageable in pipelines
  - Detailed taxonomy = interpretive richness
  - Both are indispensable in our corpus design

# Towards Cross-Linguistics Multimodal Analysis

Fine-Grained Annotation in ELAN – Exporting to Database



# Towards Cross-Linguistics Multimodal Analysis

- Methodological impact
  - workflow of filtering, taxonomies, and tool interoperability as a model for multimodal corpus construction
- Scientific impact
  - new insights into the polyfunctionality of modal particles like *schon* in institutional discourse
- Resource creation & scalability
  - towards a reliable multimodal database that balances manageability with descriptive depth
  - framework adaptable to other particles, constructions, and languages
- Collaboration & standards
  - basis for shared taxonomies and annotation protocols across research groups
  - integration into the broader goals of the RUM project and cross-linguistic analysis