

# Charakteristika sprachtypologischer Daten und ihre projektübergreifende Integration mit Hilfe von digitalen Ontologien

Alexander Borkowski & Andrea C. Schalley  
Griffith University, Brisbane, Australien

Forschungsdaten in den Geisteswissenschaften (FORGE 2015)

2015-09-18 Universität Hamburg

# Überblick

- Einführung
  - Ontologien
  - Sprachtypologische Arbeitsweise
- Problemstellung
  - Charakteristika sprachtypologischer Daten im Bezug auf ~~projektübergreifende~~ forschübergreifende Integration und Nachnutzbarkeit
- Lösungsansatz
  - Digitale Ontologien im Typology Tool (TYTO)

# Ontologie

- ὄντο[ς] -λογία
- “Wissenschaft des Seienden”
  - Unterdisziplin der Philosophie
- “Wissen vom Seienden”
  - Netzwerk von Konzepten und deren Beziehungen
  - Mentales Abbild einer wahrgenommenen Realität
  - Hier: “Kognitive Ontologien”
- “Bericht vom Seienden”
  - Konkrete Ausprägung des Wissensstandes zu einem Thema
  - Hier: insbesondere “Digitale Ontologien”

# Digitale Ontologien

- Gruber (2009):

“In computer and information science, ontology is a technical term denoting an artifact that is *designed* for a purpose, which is to enable the modeling of knowledge about *some* domain, real or imagined.”
- Semantic Web (W3C) als Standard und technologischer Rahmen einer solchen Modellierung
- Insbesondere Web Ontology Language (OWL)

# Kognitive Ontologien und Linguistik

## Beispiel: Japanische Zählsuffixe

- Auswahl (aus Dutzenden):
  - *-mai* dünne, flache Gegenstände wie Papier und Fotografien, Kleidungsgegenstände, Bettlaken
  - *-hon* längliche, dünne, zylindrische Gegenstände wie Nägel, Schreibstifte, Flaschen
- Verpflichtend für Sprecher da grammatikalisiert
- Linguistische Analyse eines Japanischen Zählspechaktes erschließt (zwingend) Teile des mentale Abbilds der wahrgenommenen Realität – der kognitiven Ontologie - des Zählenden

# “Social Cognition and Language”

## Australian Research Council (ARC) Discovery Project

- Anwendung dieser Methodik auf soziale Kognition
- Erstellung eines detaillierten, umfassenden, sprachübergreifend gültigen Modells davon, wie soziale Kognition von einer breiten Auswahl der Grammatiken der Welt unterstützt wird
- Entwicklung eines Darstellungssystems, welches alle relevanten Kategorien so modelliert, daß diese in formalere linguistische Modelle integriert werden können
- Sprachtypologische Forschergruppe deren Ergebnisse in einer digitalen Ontologie zusammengeführt werden

# “Social Cognition and Language”

## Institute und Personen:

- Australian National University
  - Nicholas Evans (CI), Alan Rumsey (CI)
- University of Melbourne
  - Barbara Kelly (CI)
- Max-Planck-Institut für Psycholinguistik
  - Stephen Levinson (PI), Nick Enfield (PI)
- Griffith University
  - Andrea Schalley (CI), Alexander Borkowski

# (Sprach-)Typologie

- Unterdisziplin der Sprachwissenschaften
- Quelle sind sprachliche Äußerungen
- Ziel ist Vergleich und Klassifikation von Sprachen mittels struktureller Eigenschaften
- Kontrast: historisch-genetische Klassifikation in Sprachfamilien



# Typologische Arbeitsweise

- Korpora
  - Aufbauen
  - Aufnehmen
  - Auswerten
  - Aufbewahren
  - Auffinden
- Hauptziel die Auswertungsgrundlage

# Typologische Arbeitsweise

Beispiel eines “Einzelforscherkorpus”

- Sprache: Lamjung Yolmo (Nepal)
- Autorin: Lauren Gawne (2013)
- Institution: University of Melbourne
- Rahmen: Promotion

# Aufbauen

- Vorbereitung von sprachneutralen Elizitierungsaufgaben zur Anregung von Sprechern während der Aufnahme
  - Zauberkunststücke und Taschenspielertricks (Gawne 2013)
  - Geschmacksbaukasten (Senft 2012)
  - Bildmaterial
    - “Picture Task”-Kartensatz (San Roque et al 2012)
    - Optische Illusionen (Gawne 2013)
- Anfallende Daten in Vielzahl von Formaten
- Im Beispiel: Bilder im Appendix der Veröffentlichung (PDF)

# Aufbauen



Selbstportrait und Landschaft von Oleg Shuplyak (Gawne 2013, S. 631)

# Aufnehmen

- Photos (leider) oft nicht in Archivqualität (TIFF)
- Video in Vielfalt von Formaten
  - Frage nach Archivqualität noch nicht geklärt
  - Tonaufnahmen von Videokameras in der Regel nicht Archivqualität
- Eigenes Gerät mit Qualitätsmikrophon(en) für Audio in Archivqualität (WAV)
- Im Beispiel: WAV, MP4, MP3, JPEG, PDF (Scans von Feldnotizen)

# Auswerten

- Analyse
  - Linguistische Softwarewerkzeuge
    - Annotation der Aufnahmen
    - Erstellung von Lexika
  - Datenformat semistrukturierter Text
  - Projekte
    - Field Linguist's Toolbox
    - ELAN
    - EXMARaLDA

# Auswerten

- Veröffentlichung
  - Standardwerkzeuge für Verfassung wissenschaftlicher Texte
  - Datenformat un- bis semistrukturierter Text
- Im Beispiel: TXT (Field Linguist's Toolbox), EAF[XML] (ELAN), PDF (Microsoft Word)

# Aufbewahren

- Übergabe der Dateien an ein oder mehrere Repositories
- Im Beispiel:
  - Dissertation zum Download unter <http://www.academia.edu/3678533>
  - Alles Andere an das Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC, <http://catalog.paradisec.org.au/collections/LG1>)



# Auffinden bzw. auffindbar machen

- Im Beispiel:
  - PARADISEC-Katalog
  - Automatische Übernahme der PARADISEC-Metadaten in die Open Language Archive Community (OLAC, <http://www.language-archives.org/>)
  - Suchmaschinenindizierung

# Bestandsaufnahme

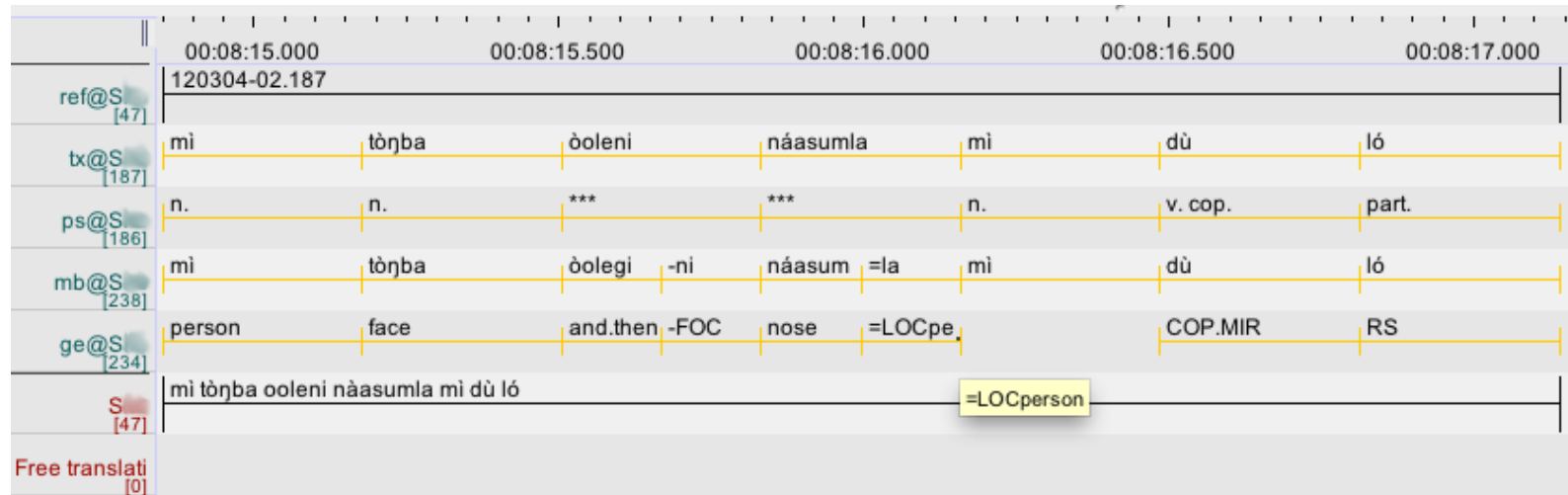
- Ergebnisse veröffentlicht, Lexika und Grammatiken in den Sprachcommunities, Daten im Langzeitarchiv
- The Language Archive (TLA) des MPI für Psycholinguistik in Nijmegen garantiert Aufbewahrung von 50 Jahren durch die MPG
- Disziplinspezifische Archive bieten Mehrwertdienste für die Nachnutzung an (hauptsächlich Suche und Präsentation von Metadaten und Daten z.B. Webplayer)

# Typologische Arbeitsweise

- Bisher: sprachdokumentarischer Aspekt typologischer Arbeit
- Ton-, Bild- und Filmmaterial nicht fachspezifisch
- Texte teilweise auch nicht
- Annotationsdaten fachspezifisch
- Fehlt: Möglichkeit die Analyseergebnisse von möglichst vielen Sprachen in die eigene Analyse einzubeziehen und das Ergebnis anderen geeignet zur Verfügung zu stellen

# Annotation in der Analyse

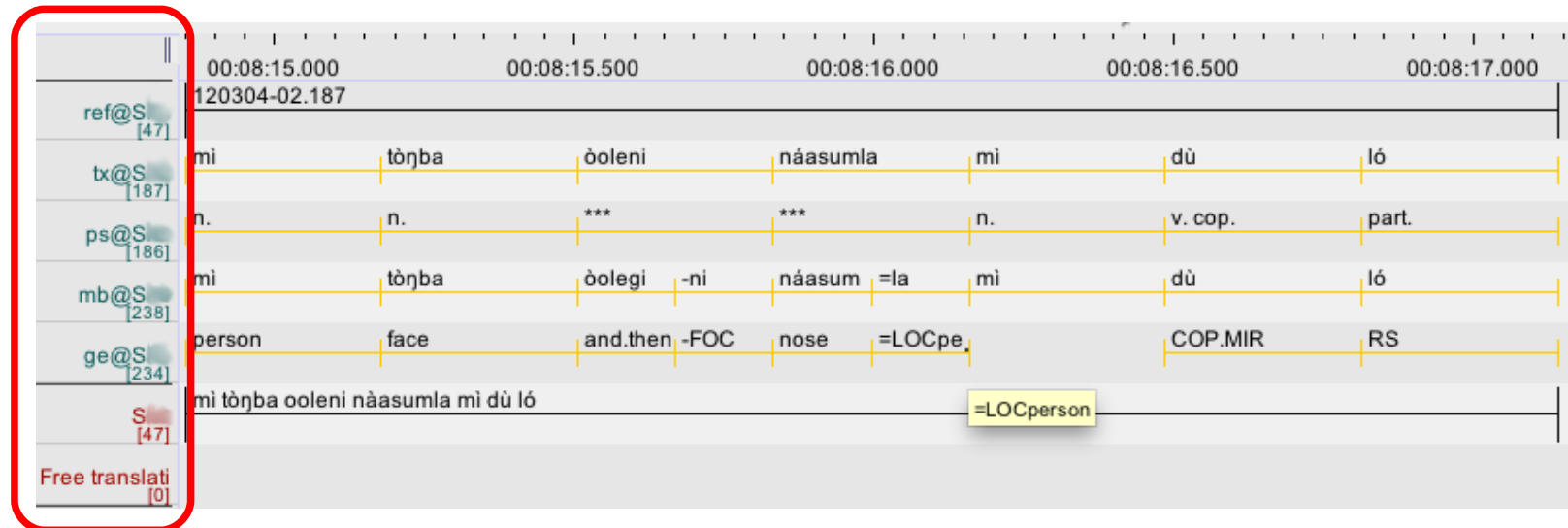
## Beispiel aus ELAN-Datei



- Diskussion in Lamjung Yolmo über die obige optische Illusion annotiert von Lauren Gawne

# Annotation in der Analyse

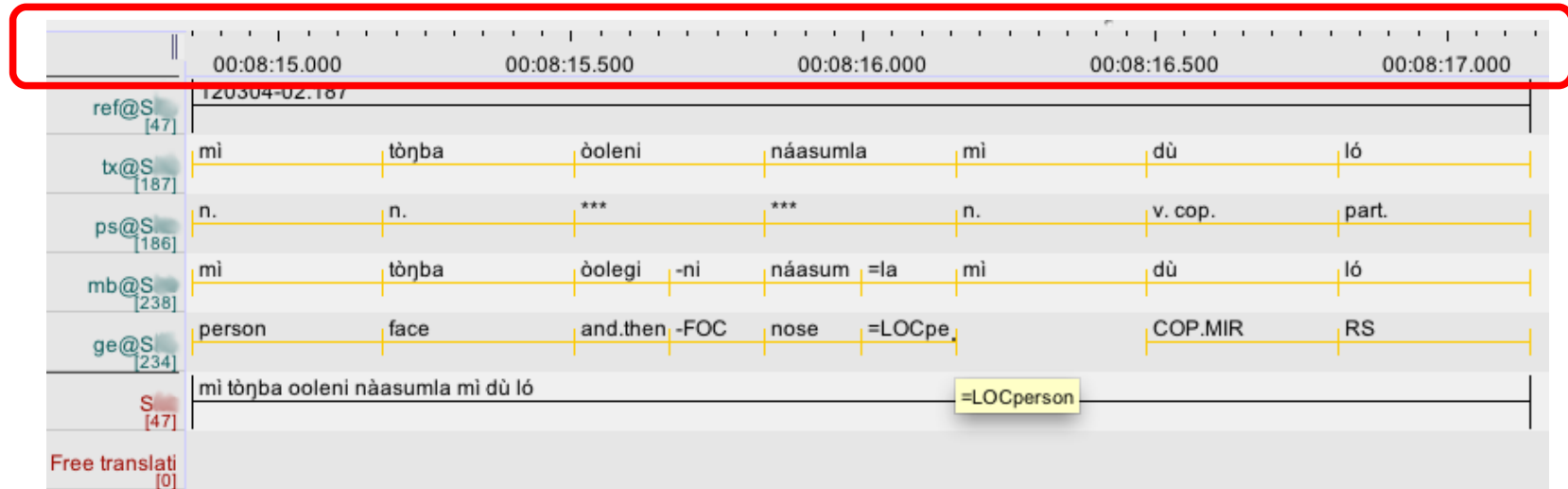
## Spurbezeichner



- Sprechernamen kombiniert mit
  - (konventioneller) Abkürzung des Spurtyps
  - Anderen textuellen Informationen aus der Analysesprache (z.B. “Free Translation”)
- Im Screenshot manuell anonymisiert, in ELAN nicht

# Annotation in der Analyse

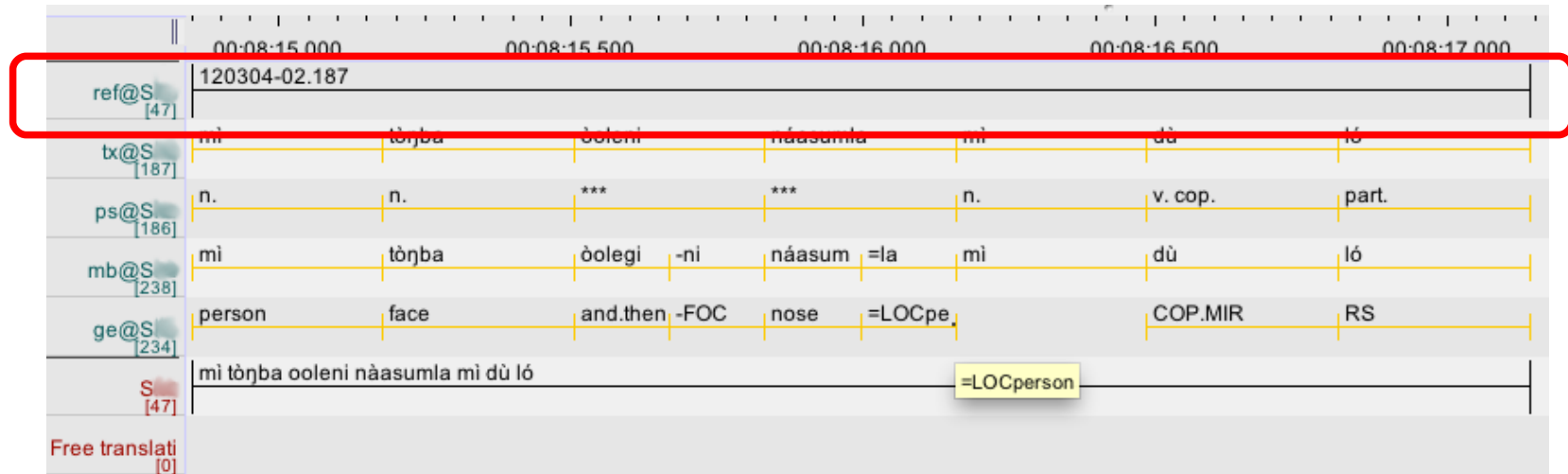
## Zeitachse



- Zeitverlauf seit Beginn der Aufnahme
- Referenz auf die Mediendateien

# Annotation in der Analyse

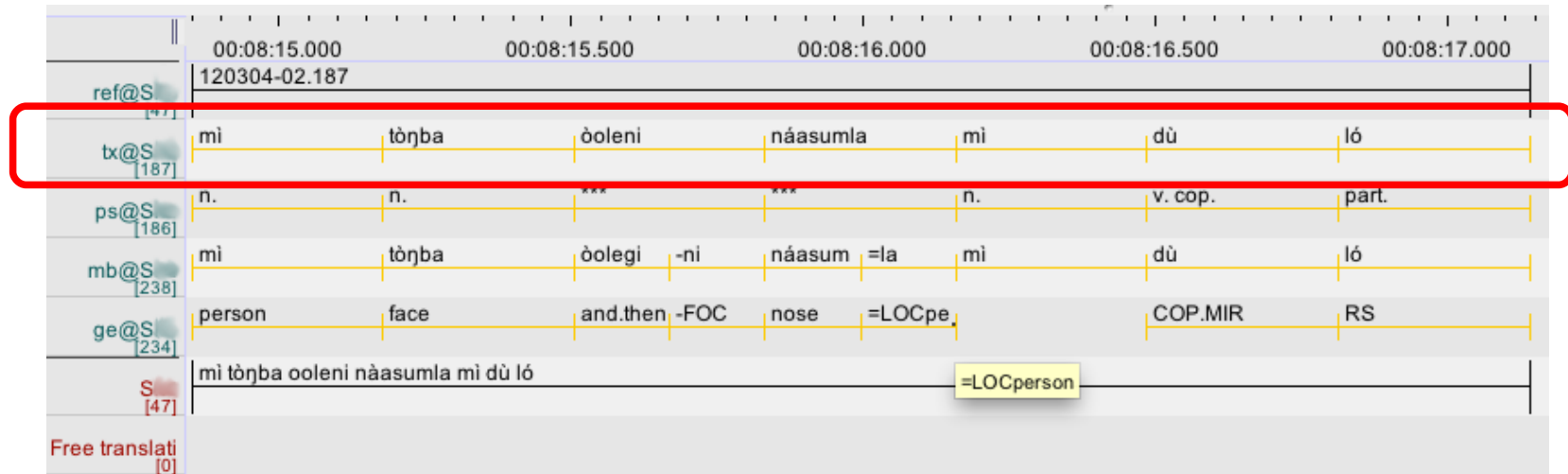
## Referenzspur ("ref" 1:1 Intervall auf Zeitachse)



- (Selbstvergebener) Bezeichner der sprachlichen Äußerung zusammengesetzt aus
  - Bezeichner der Aufnahmesitzung
  - Bezeichner innerhalb der Aufnahmesitzung

# Annotation in der Analyse

## Transkriptionsspur ("tx" n:1 "ref")

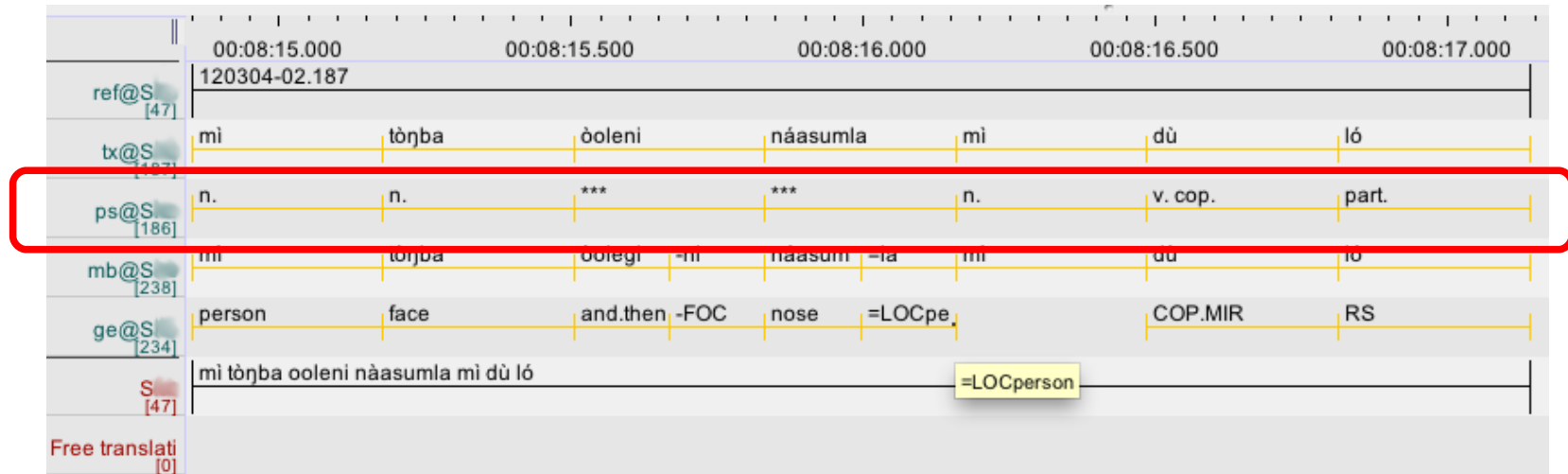


- Segmentierung in Wörter der Objektsprache



# Annotation in der Analyse

## Part-of-Speech-Spur (“ps” 1:1 “tx”)



- Abkürzungen aus der Analysesprache
- Linguisten geläufig oder aus anderen Spuren erschließbar, nicht formalisiert (z.B. “n.” = noun)

# Annotation in der Analyse

## Morphemisierungsspur (“mb” n:1 “tx”)

|                            | 00:08:15.000                      | 00:08:15.500 | 00:08:16.000  | 00:08:16.500 | 00:08:17.000     |
|----------------------------|-----------------------------------|--------------|---------------|--------------|------------------|
| ref@S <sub>[47]</sub>      | 120304-02.187                     |              |               |              |                  |
| tx@S <sub>[187]</sub>      | mi                                | tòŋba        | òoleni        | náasumla     | mi dù ló         |
| ps@S <sub>[188]</sub>      | n.                                | n.           | ***           | ***          | n. v. cop. part. |
| mb@S <sub>[238]</sub>      | mi                                | tòŋba        | òolegi -ni    | náasum =la   | mi dù ló         |
| ge@S <sub>[234]</sub>      | person                            | face         | and then -FOC | nose =LOCpe  | COP.MIR RS       |
| S <sub>[47]</sub>          | mi tòŋba ooleni náasumla mi dù ló |              |               |              | =LOCperson       |
| Free transl <sub>[0]</sub> |                                   |              |               |              |                  |

- Worte segmentiert in Morpheme (kleinste bedeutungstragende Einheiten)
- De-facto-Standard Leipzig Glossing Rules (LGR) (z.B. “-” kennzeichnet Affixe)

# Annotation in der Analyse

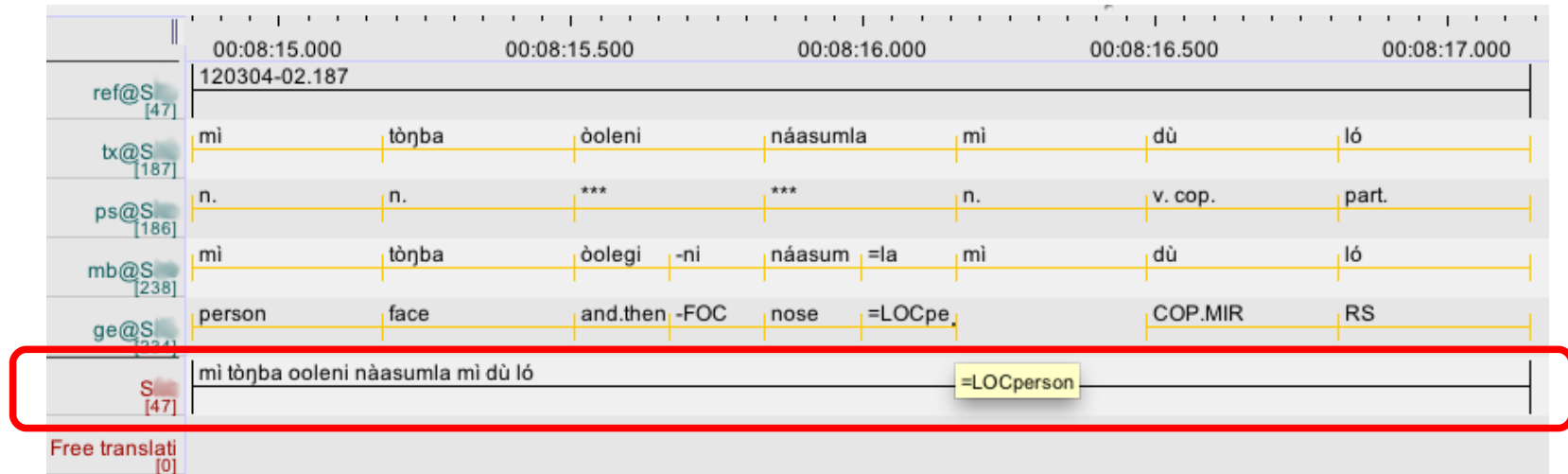
## Glossierungsspur (“ge” 1:1 “mb”)

|                            |                                   |              |              |                  |                  |
|----------------------------|-----------------------------------|--------------|--------------|------------------|------------------|
|                            | 00:08:15.000                      | 00:08:15.500 | 00:08:16.000 | 00:08:16.500     | 00:08:17.000     |
| ref@S <sub>[47]</sub>      | 120304-02.187                     |              |              |                  |                  |
| tx@S <sub>[187]</sub>      | mi                                | tòŋba        | òoleni       | náasumla         | mi dù ló         |
| ps@S <sub>[186]</sub>      | n.                                | n.           | ***          | ***              | n. v. cop. part. |
| mb@S <sub>[233]</sub>      | mi                                | tòŋba        | òolegi       | -ni náasum =la   | mi dù ló         |
| ge@S <sub>[234]</sub>      | person                            | face         | and then     | -FOC nose =LOCpe | COP.MIR RS       |
| S <sub>[47]</sub>          | mi tòŋba ooleni náasumla mi dù lo |              |              |                  |                  |
| Free transl <sub>[0]</sub> |                                   |              |              |                  |                  |

- Bedeutungen der jeweiligen Morpheme
  - Lexikalische in Analysesprache
  - Grammatikalische in großbuchstabigen Abkürzungen (z.B. “FOC” focus)
  - Interpunktion nach LGR (z.B. “.” als Trenner)

# Annotation in der Analyse

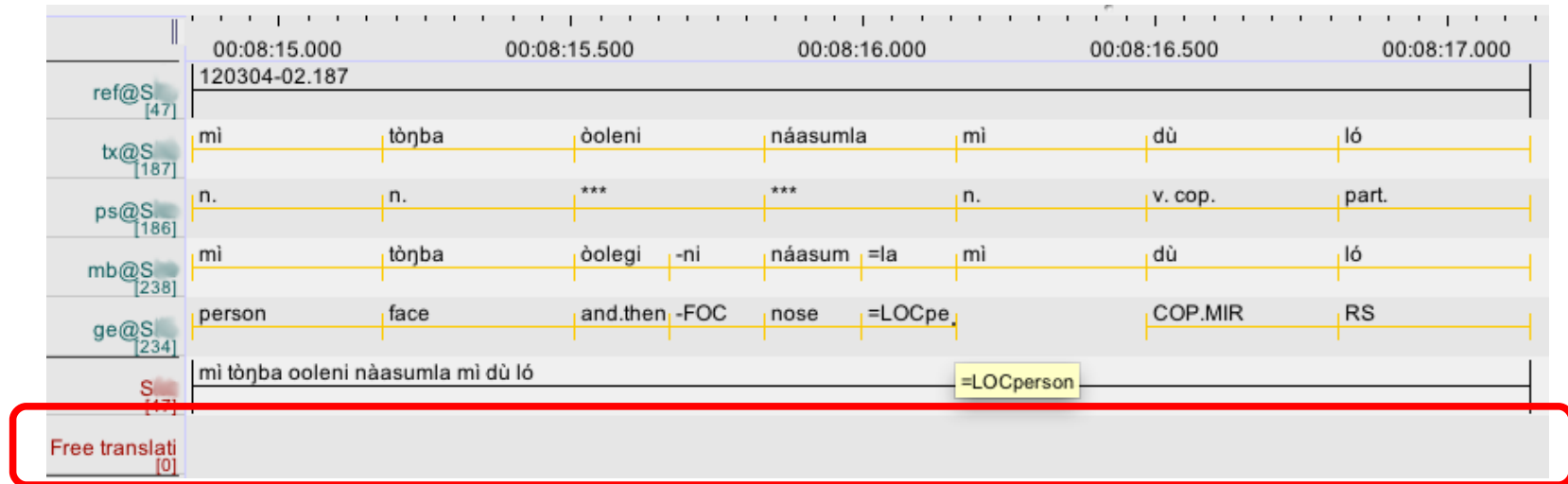
## Objektsprache (1:1 Intervall auf Zeitachse)



- (Gesamte) Äußerung in Objektsprache

# Annotation in der Analyse

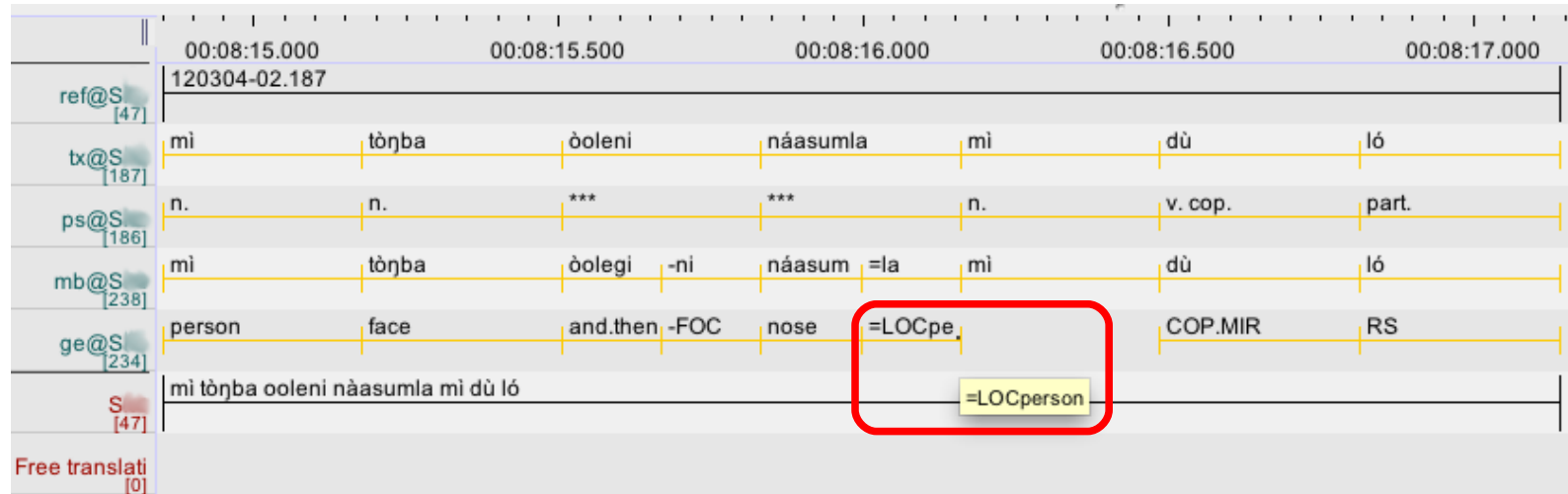
## Übersetzungsspur (1:1 Objektsprachspur)



- Übersetzung in Analysesprache
  - Fehlt
  - Aus Morphemen erahnbar
- Arbeitsdaten nicht Endergebnis

# Annotation in der Analyse

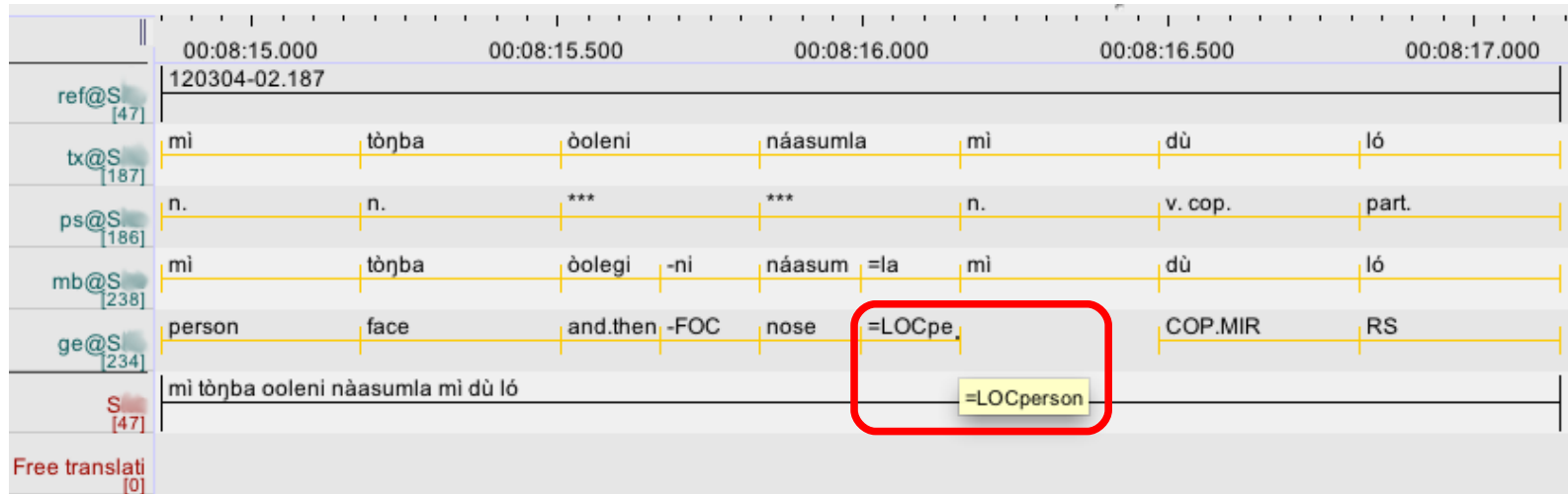
## Auffälligkeiten in den Arbeitsdaten



- Vermutlich Konvertierungsfehler, Spurbezeichner Hinweis auf Datenübernahme aus Toolbox

# Annotation in der Analyse

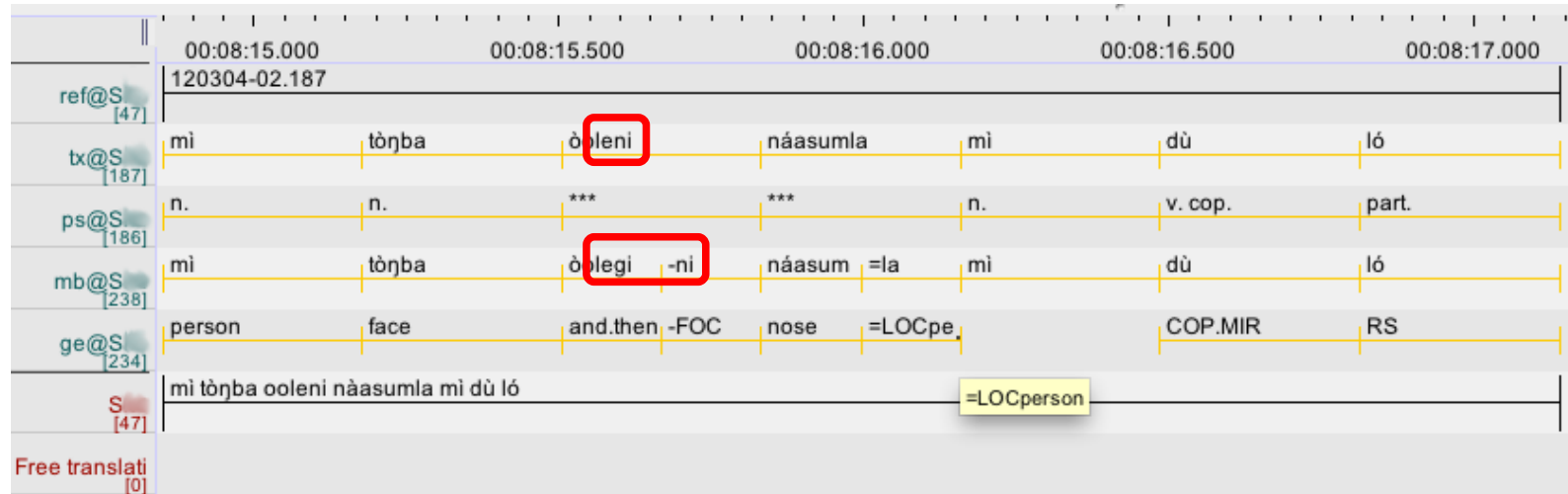
## Auffälligkeiten in den Arbeitsdaten



- Vermutlich Konvertierungsfehler, Spurbezeichner Hinweis auf Datenübernahme aus Toolbox

# Annotation in der Analyse

## Auffälligkeiten in den Arbeitsdaten

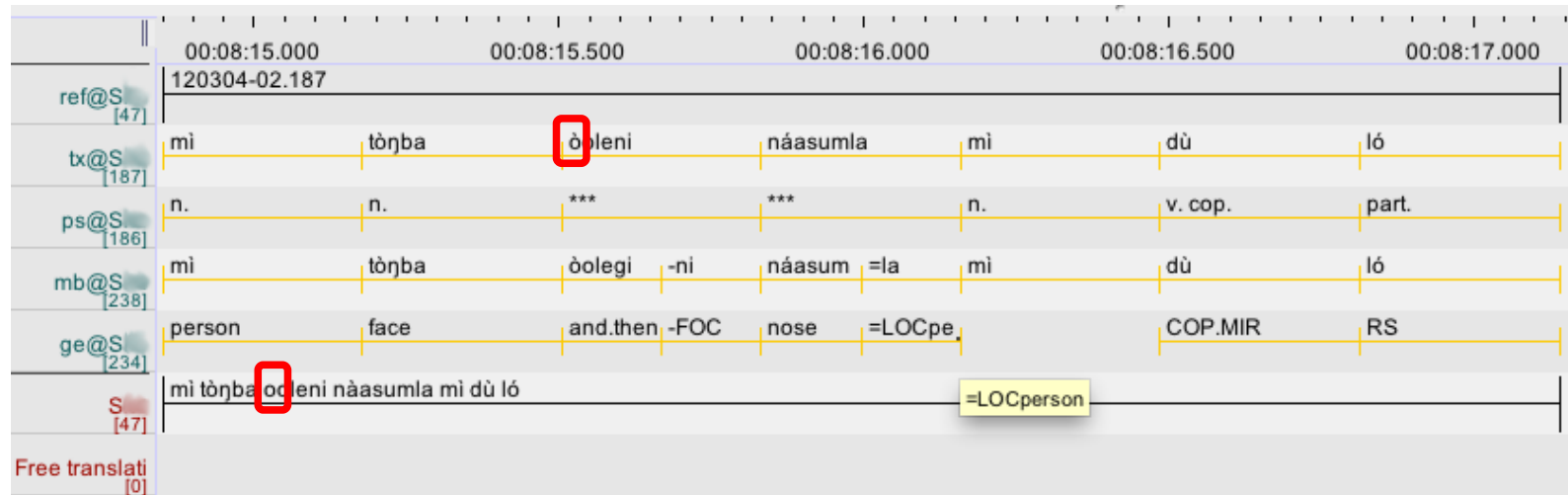


- Interpretationsmöglichkeiten
  - Fehler
  - Sprachliche Prozesse



# Annotation in der Analyse

## Auffälligkeiten in den Arbeitsdaten



- Vermutlich Eingabefehler (fehlender Akzent)

# Annotation in der Veröffentlichung

37) *mì tòŋba òolegi-ni náasum=la mì dù ló*  
 person face and.then-FOC nose=LOC person COP.PE RS  
 'a person's face, and then on the nose is a person (she said).'

(ST 120304-02 08:15)

She also repeats the same assertion at 08:28 in the interaction. Because I had also recorded the session with ST (120304-01) it was possible to go back and listen to what I had said to her in that interaction. (37) is not something I said at any point in that interaction - but ST herself did say similar things at various points (38).

(Gawne 2013, S. 340)

# Annotation in der Veröffentlichung

37) *mì tòŋba òolegi-ni náasum=la mì dù ló*  
 person face and.then-FOC nose=LOC person COP.PE RS  
 'a person's face, and then on the nose is a person (she said).'

(ST 120304-02 08:15)

She also repeats the same assertion at 08:28 in the interaction. Because I had also recorded the session with ST (120304-01) it was possible to go back and listen to what I had said to her in that interaction. (37) is not something I said at any point in that interaction - but ST herself did say similar things at various points (38).

# Weggefallen:

- Gesamtäußerung in Objektsprache
- Part-of-speech
- Analysedateiinterner Bezeichner der Äußerung

# Annotation in der Veröffentlichung

37)

*mì tòŋba òolegi-ni náasum=la mì dù ló*  
person face and.then-FOC nose=LOC person COP.PE RS

'a person's face, and then on the nose is a person (she said).'

(ST 120304-02 08:15)

She also repeats the same assertion at 08:28 in the interaction. Because I had also recorded the session with ST (120304-01) it was possible to go back and listen to what I had said to her in that interaction. (37) is not something I said at any point in that interaction - but ST herself did say similar things at various points (38).

Hinzugekommen:

- Veröffentlichungsinterner Bezeichner
- Übersetzung in die Analysesprache
- Ausführliche textuelle Erläuterungen in der Analysesprache (auch an anderen Stellen)

# Annotation in der Veröffentlichung

37)      *mì      tòŋba òolegi-ni      náasum=la mì      dù      ló*  
person face      and.then-FOC nose=LOC      person COMPE RS  
'a person's face, and then on the nose is a person (she said).'

[ST 20304-02 08:15)

She also repeats the same assertion at 08:28 in the interaction. Because I had also recorded the session with ST (120304-01) it was possible to go back and listen to what I had said to her in that interaction. (37) is not something I said at any point in that interaction - but ST herself did say similar things at various points (38).

## Verändert:

- Sprecheridentifikation anonymisiert
- Zeitintervall reduziert auf Startzeitpunkt
- Konvertierungsfehler behoben
- Änderung der Abkürzung oder Erkenntnisgewinn?

# Charakteristika von Annotationsdaten

- Textdaten (un- bis semi-strukturiert)
- Hochkondensiert
- Unvollständig
- Vielschichtig und von unterschiedlicher Relevanz für linguistische Teilbereiche
- Struktur und Inhalt stark beeinflußt von Konvention und Tradition

# Charakteristika von Annotationsdaten

- Qualität besser in der veröffentlichten Version, digital besser zugänglich in der Analyseversion
- Bedeutung nur schwer erschließbar ohne ähnliche kognitive Ontologie für den Fachbereich Linguistik
- Ausdruck der kognitiven Ontologie des Forschenden für den Fachbereich Linguistik
- Parallelen zur Untersuchungsmethodik im “Social Cognition and Language”-Projekt

# Integration

- Probleme:
  - Schon kleine Variationen diesen kognitiven Ontologien können zu Missverständnissen bis hin zum “Terminologiekrieg” führen
  - Vielfalt der (noch) 7000+ Sprachen
- Lösung
  - Erstellen einer digitalen Ontologie als Abbild eines (hinreichend großen Teilbereichs) einer kognitiven Ontologie der Linguistik



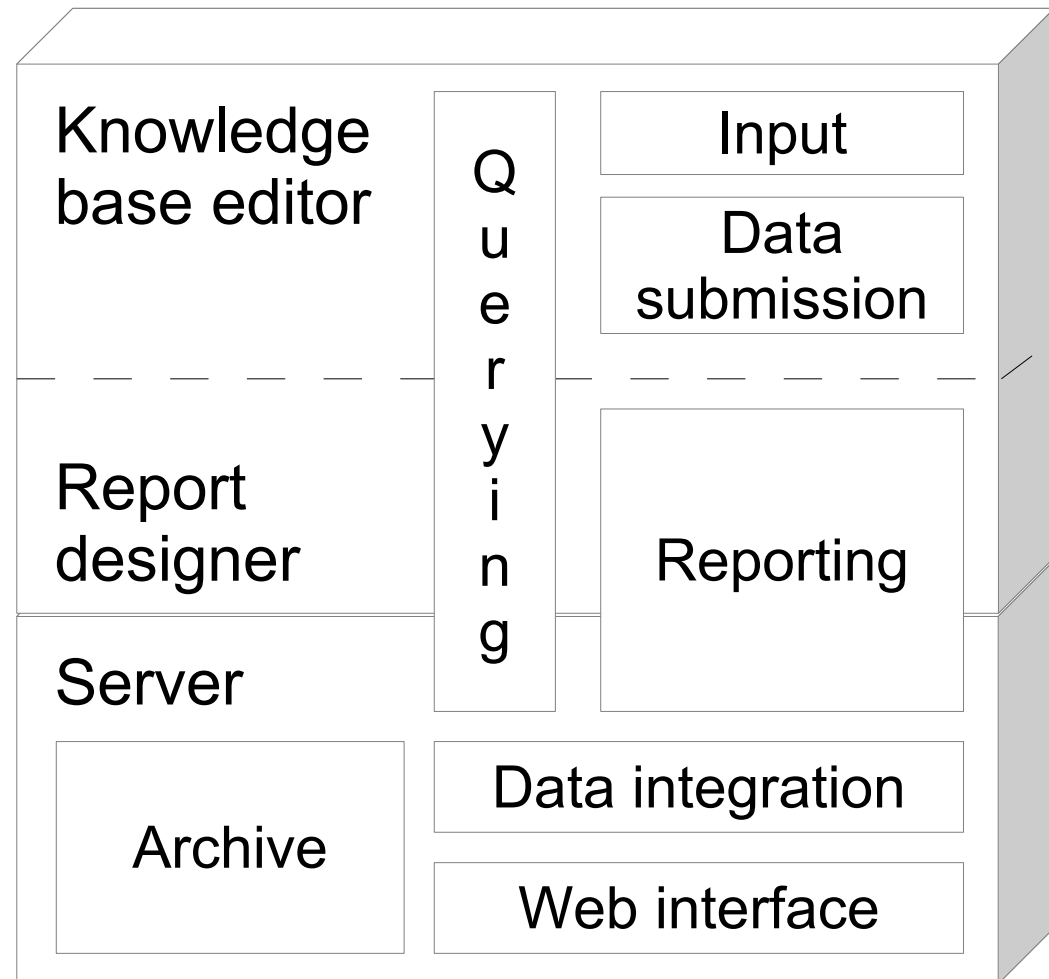
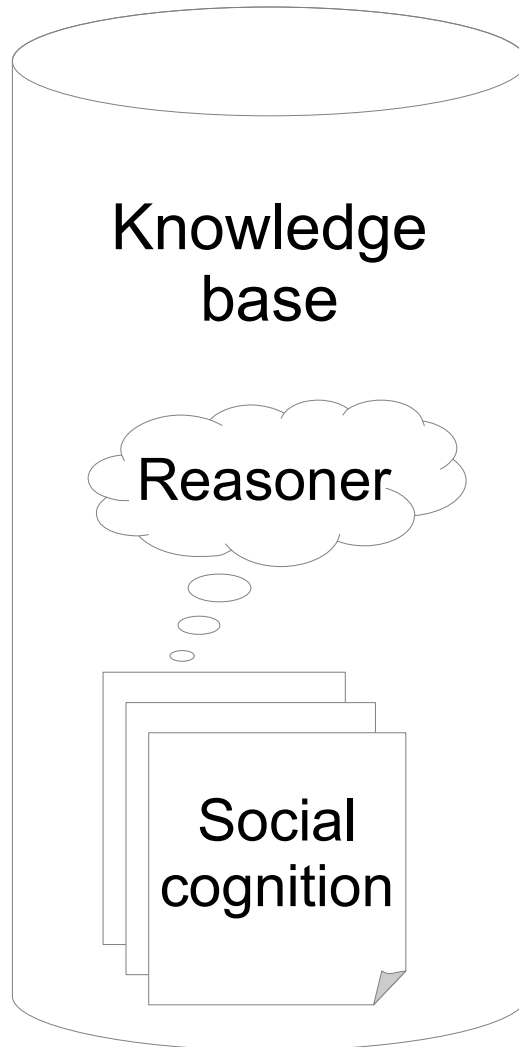
# Bisherige Lösungen

- Typological Database System (TDS)
  - Integration auf Ebene der Benutzerschnittstelle
  - Keine Zusammenführung von Daten
- World Atlas of Language Structures (WALS)
  - Nicht direkt erweiterbar
  - Beschränkte Abfrageschnittstelle

# Typology Tool (TYTO)

- Knowledge base (Modell)
  - Grundlegendes, hinzugewonnenes und herleitbares Wissen
  - Archiv linguistischer Beispiele/Analysen
- Zugriffssoftware unterstützt
  - Linguistenfreundliche Eingabe (ELAN, Leipzig Glossing Rules)
  - Gruppenarbeit
  - Offlinearbeit (z.B. Im Feld)
  - Zusatzdienste (Anfragesystem)
  - Archivprozess

# TYTO Systemüberblick



# Entwurf der TYTO Knowledge Base

Digitale Ontologien der zu untersuchenden Thematiken

Social  
cognition

?

?

Cross-  
Linguistic  
structure

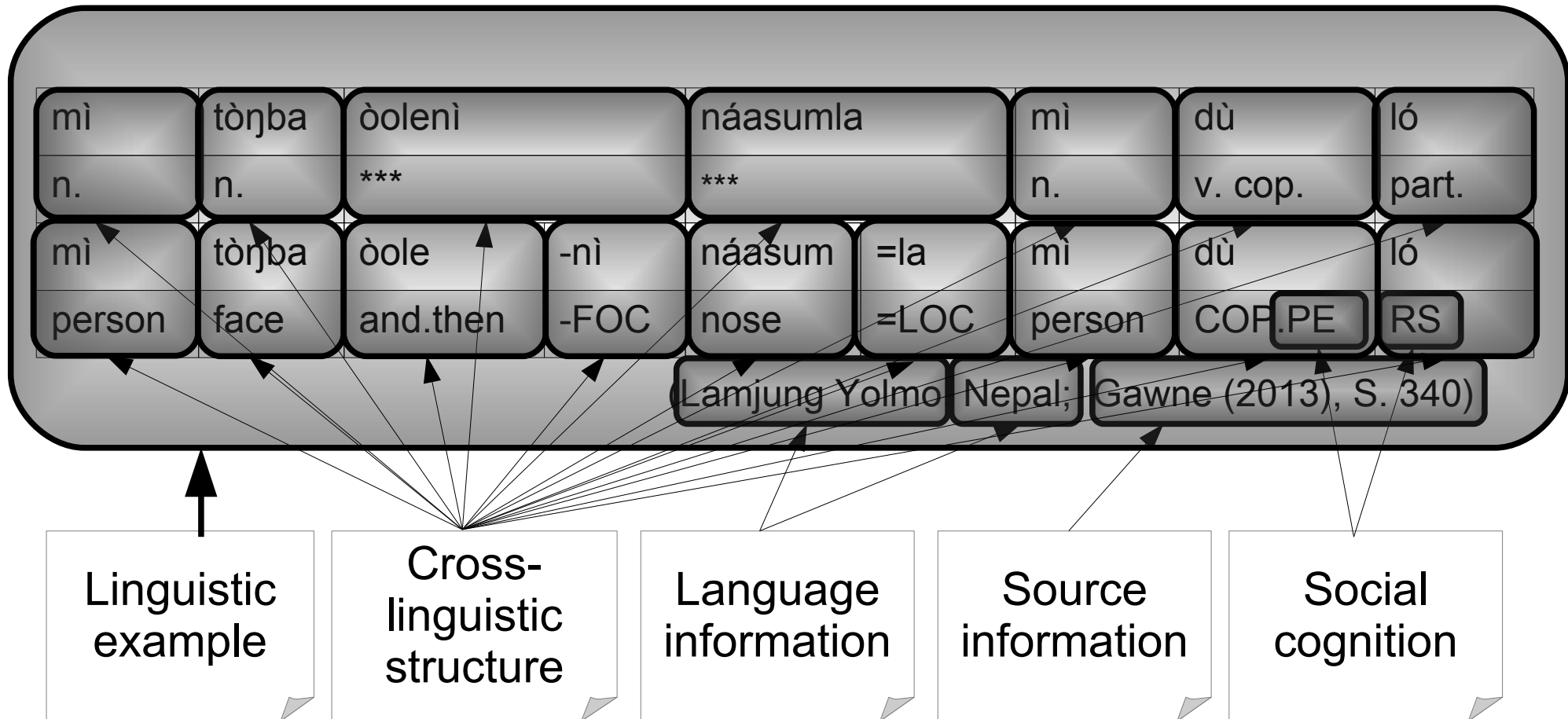
Linguistic  
example

Source  
information

Language  
information

Digitale Ontologien des linguistischen Fachwissens

# Annotation in TYTO



- Ontologien zur Spezifikation linguistischen Wissens
- Mühsam manuell einzugeben und schwer zu automatisieren

# Annotation in TYTO

|        |       |          |      |          |      |        |         |       |
|--------|-------|----------|------|----------|------|--------|---------|-------|
| mì     | tònba | òolenì   |      | náasumla |      | mì     | dù      | ló    |
| n.     | n.    | ***      |      | ***      |      | n.     | v. cop. | part. |
| mì     | tònba | òole     | -nì  | náasum   | =la  | mì     | dù      | ló    |
| person | face  | and.then | -FOC | nose     | =LOC | person | COP.PE  | RS    |

(Lamjung Yolmo, Nepal; Gawne (2013), S. 340)

Sprechakt,  
der nie statt-  
gefunden hat

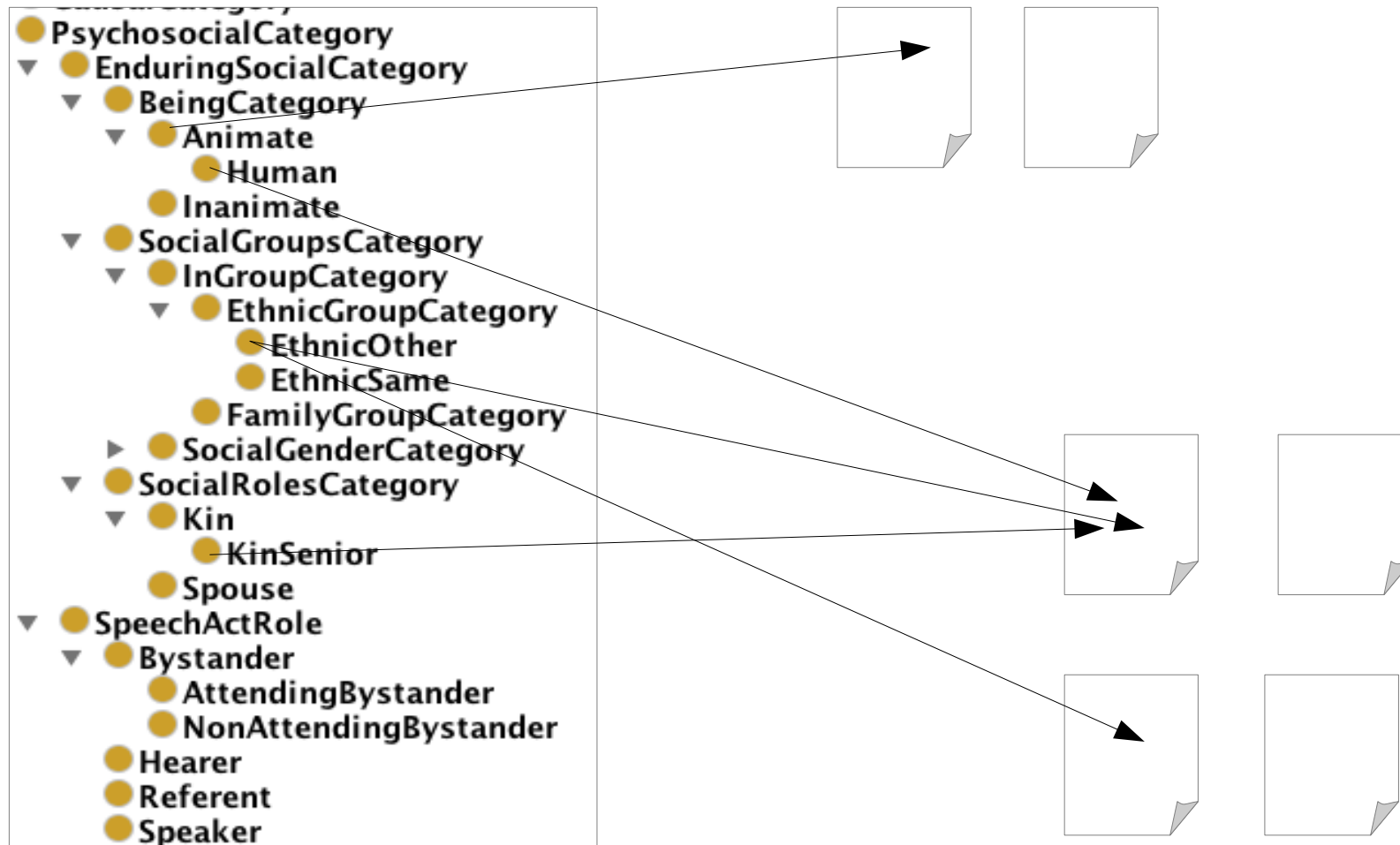
+

Indirekte-Rede-Partikel – Erfordert vom Sprecher, eine vorige Äußerung einer anderen Person auf eine Art zu modellieren, welche Relevant für die momentane Interaktion mit einer dritten Person ist

=

Versuch des Sprechers, seine Perspektive stärker zu legitimieren

# Integration in TYTO



# Nachnutzung / Dauernutzung

- Erhoffte Dauernutzung durch fortwährende Erweiterung des Datenbestandes
- Dienste wie Suche aber auch Berichtswesen zur Unterstützung bei der Veröffentlichung
- Stark gewünscht von den Typologen, man will seine Daten mit denen der anderen vergleichen können
- Problem der initialen kritischen Masse



# Nachhaltigkeit

- Gestuft
  - Zugangssoftware
  - SPARQL-Endpoint
  - RDF[XML]-Datei
- Finanzierung im Moment schwierig und, sollten wir Erfolg haben, ungeklärt.

# Danke für Ihre Aufmerksamkeit!

[a.borkowski@ontolinguistics.org](mailto:a.borkowski@ontolinguistics.org)  
[a.schalley@griffith.edu.au](mailto:a.schalley@griffith.edu.au)

# Literatur

- Gruber, Tom. 2009. Ontology. Entry in the *Encyclopedia of Database Systems*, Ling Liu and Özsu, M. Tamer (eds.). Springer.
- Gawne, Lauren. 2013. Lamjung Yolmo copulas in use: Evidentiality, reported speech and questions.

# Literatur (forts.)

- Senft, Gunter. 2012. Referring to colour and taste in Kilivila: Stability and change in two lexical domains of sensory perception. 71-98 In: Schalley, Andrea C. (ed.): Practical Theories and Empirical Practice. Human Cognitive Processing 40. John Benjamins. DOI:10.1075/hcp.40
- San Roque, Lila, et al. 2012. Getting the Story Straight: Language Fieldwork Using a Narrative Problem-Solving Task. Language Documentation & Conservation 6. 135-174. hdl:10125/4504