Towards more Linguistic Modelling in a Sign Language Transcription Environment

Thomas Hanke, Susanne König,
Reiner Konrad, Gabriele Langer
University of Hamburg
Institute of German Sign Language…
mailto:dgs-korpus@sign-lang.uni-hamburg.de
http://www.sign-lang.uni-hamburg.de/dgs-korpus/

iLex- Why another Transcription Environment?

- Database approach instead of tons of XML files
- Not general-purpose, but specific to sign language
  - Specific support to compensate the major difference between sign languages and many spoken languages: The lack of a writing system with a standard orthography
  - The essential step is lemmatisation, i.e. identification of the type a token is related to
  - Transcription and growing the lexical database is one single process

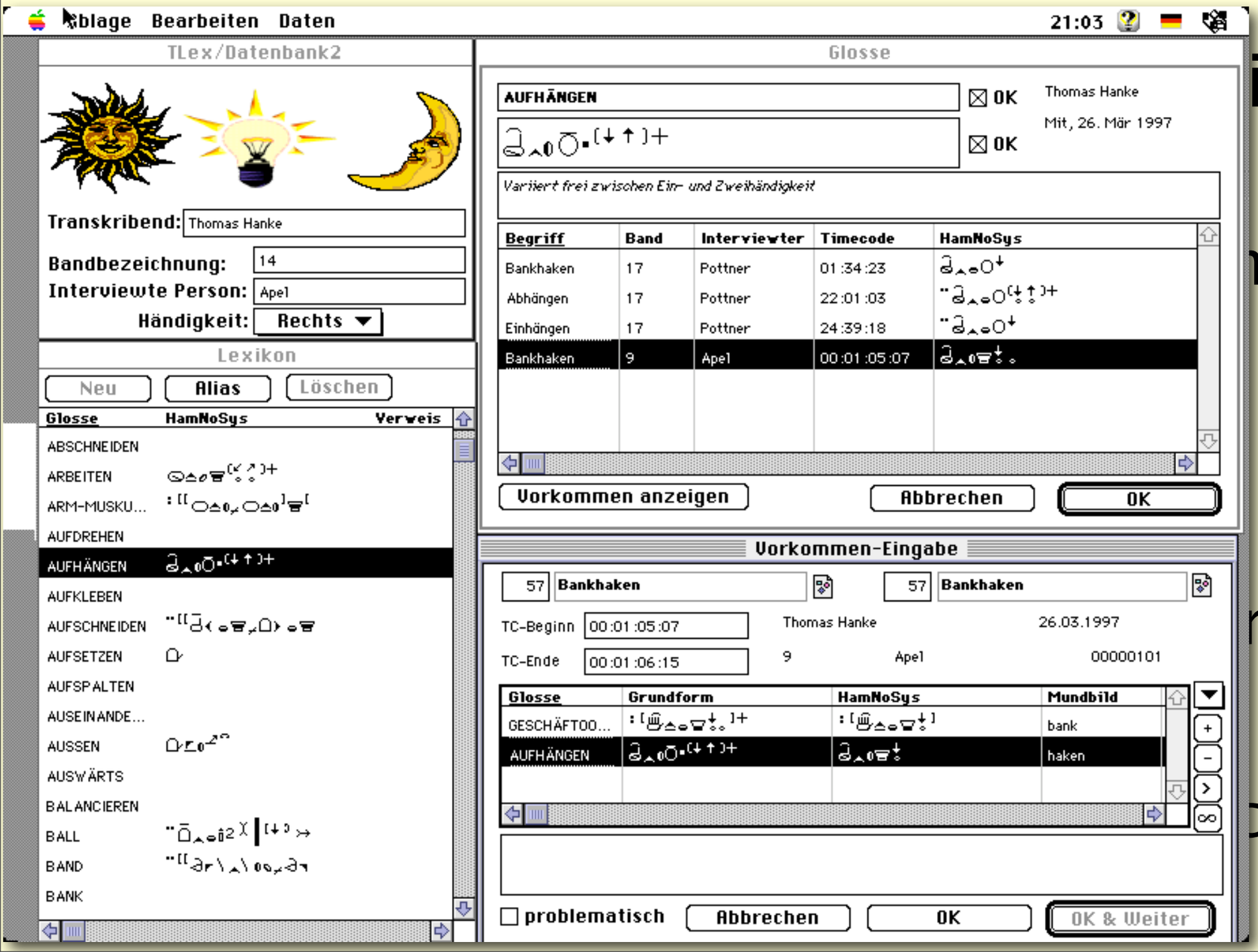
Building on Past Experience

- 1994 syncWRITER for signed discourse transcription
  - multi-tier text annotation to points in time of a video
- 1996 GlossLexer
  - Lexical database
  - Transcription limited to short phrases
- Project started in 2000
  - Lexical database & flexible transcription

Building on Past Experience

- 1994 syncWRITER for signed discourse transcription
  - multi-tier tex
    of a video
- 1996 GlossLe
  - Lexical datab
  - Transcriptior
- Project starte
  - Lexical database & flexible transcription

Building on Past Experience



iscourse

nts in time

rases

cription

iLex Development

- 2000–2008 inhouse tool for various projects
- 2009–2023 further development secured by means of the DGS corpus project
  - resources available to make the tool available to other research groups
    - 2 persons working on iLex
- 2009–2012 one of the target platforms of the Dicta-Sign EU project
  - video recognition to be integrated

In-house Use of iLex

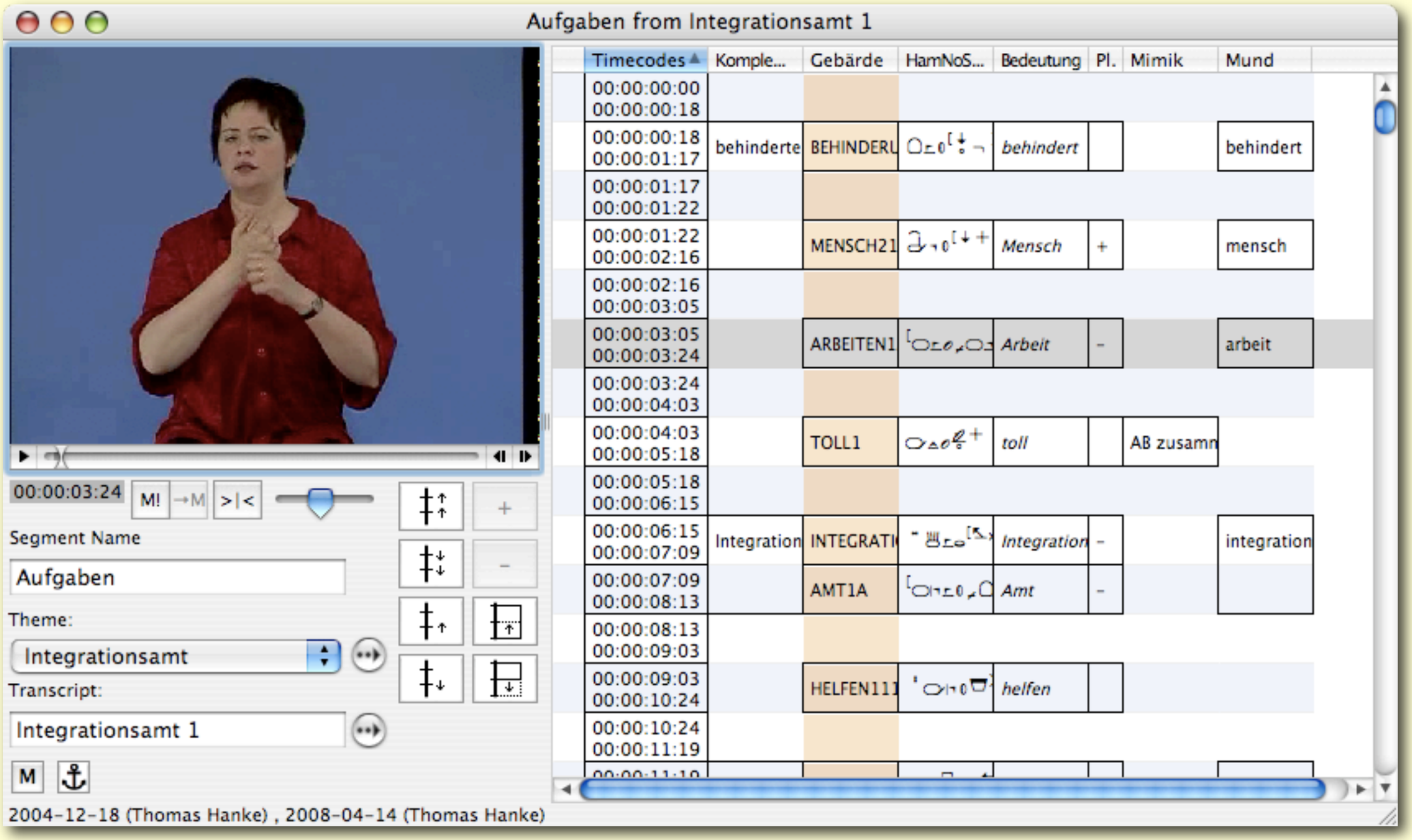
- DGS Corpus project
- Special terminology dictionary projects
- Individual research & exam paper work

- peak-time user count > 20
- 26700 transcripts (for 360 hours of video) in the database with some 700000 tags
- nothing gets lost on someone's harddisk
- central backup

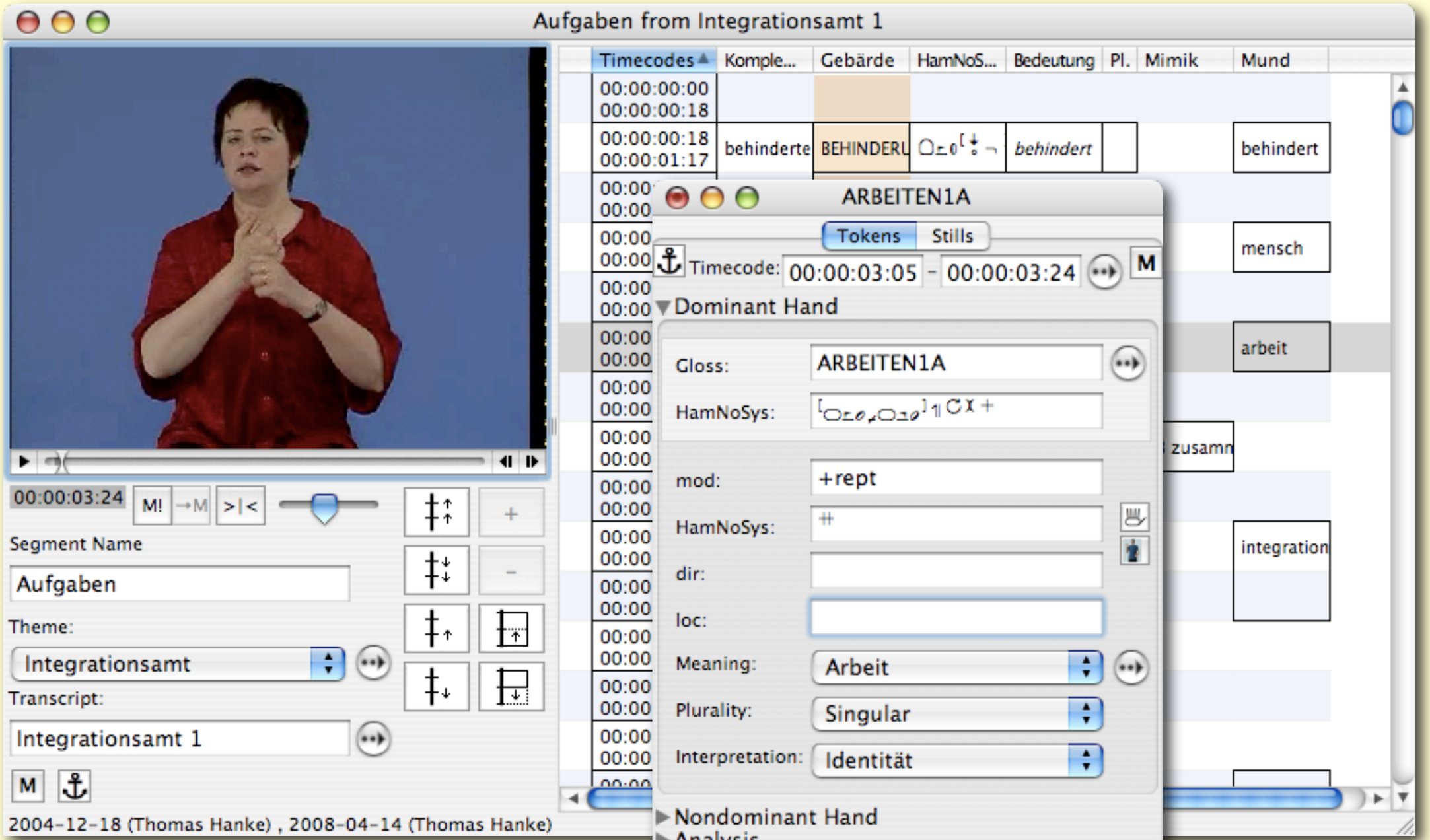
Transcription in iLex

- Building the corpus and (extending) the lexicon at the same time
- Type-token matching is the essential step
  - Relational database!

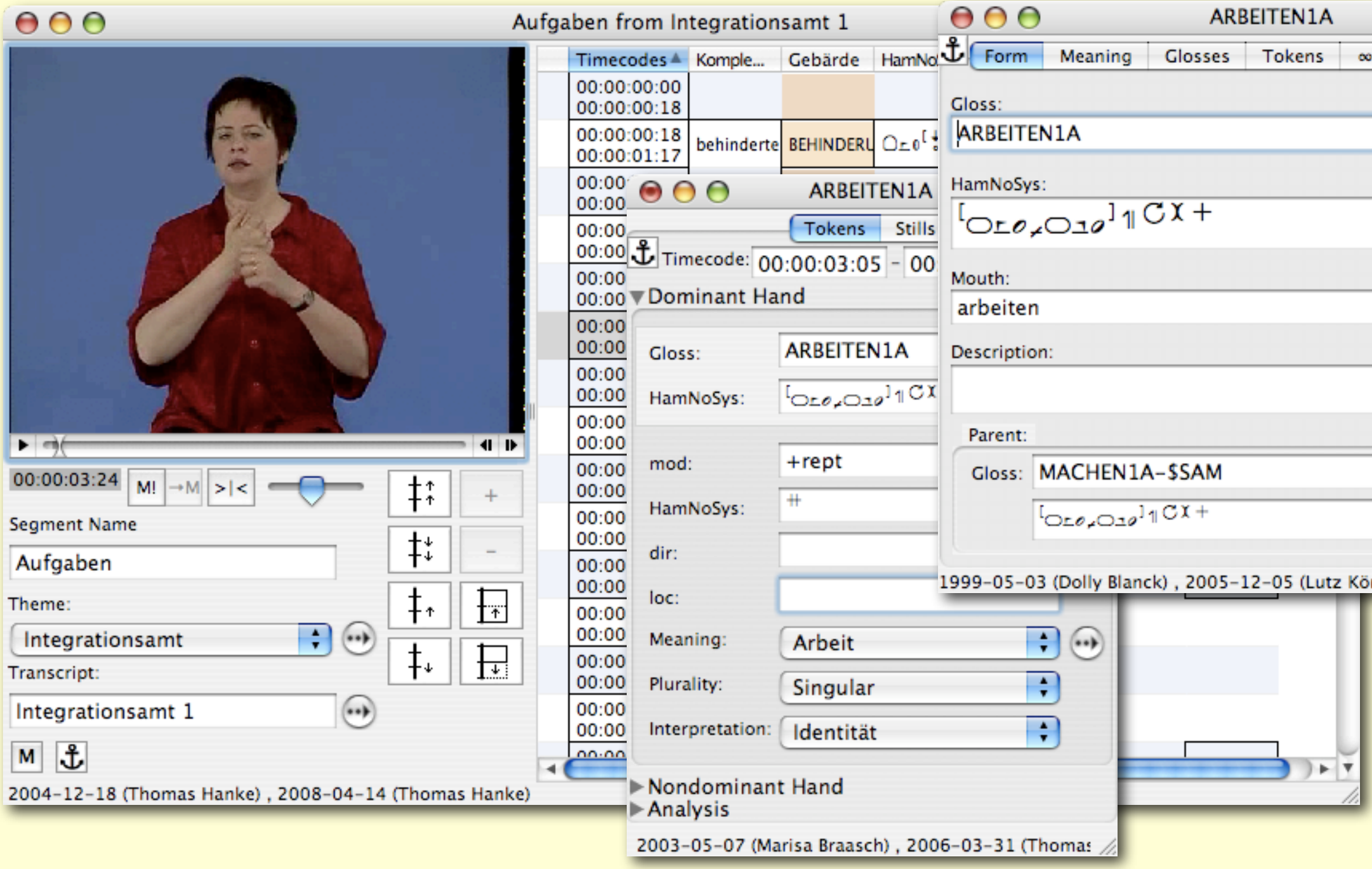
iLex: Type-token matching is the essential step in transcription



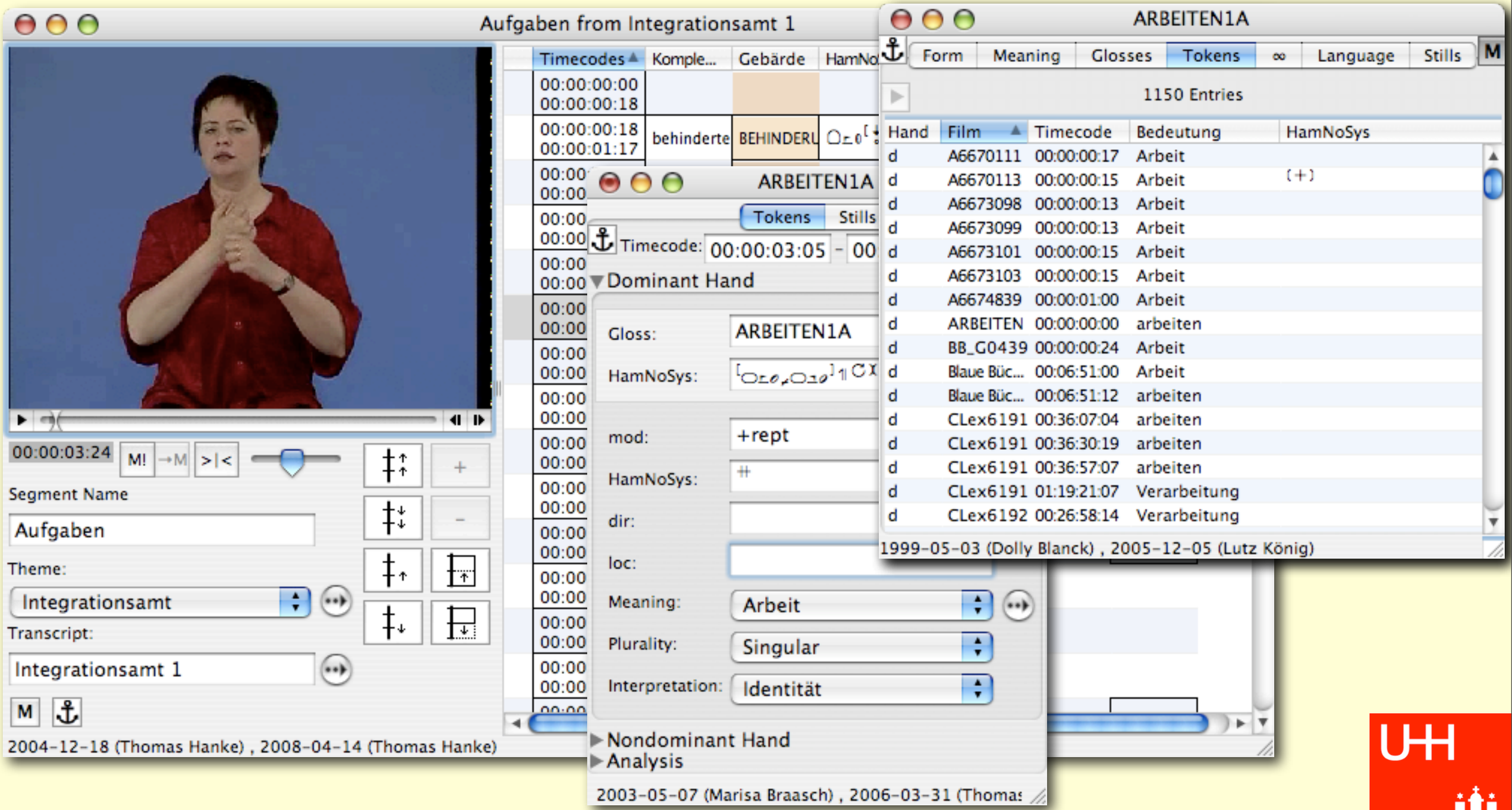
iLex: Type-token matching is the essential step in transcription



iLex: Type-token matching is the essential step in transcription



iLex: Type-token matching is the essential step in transcription

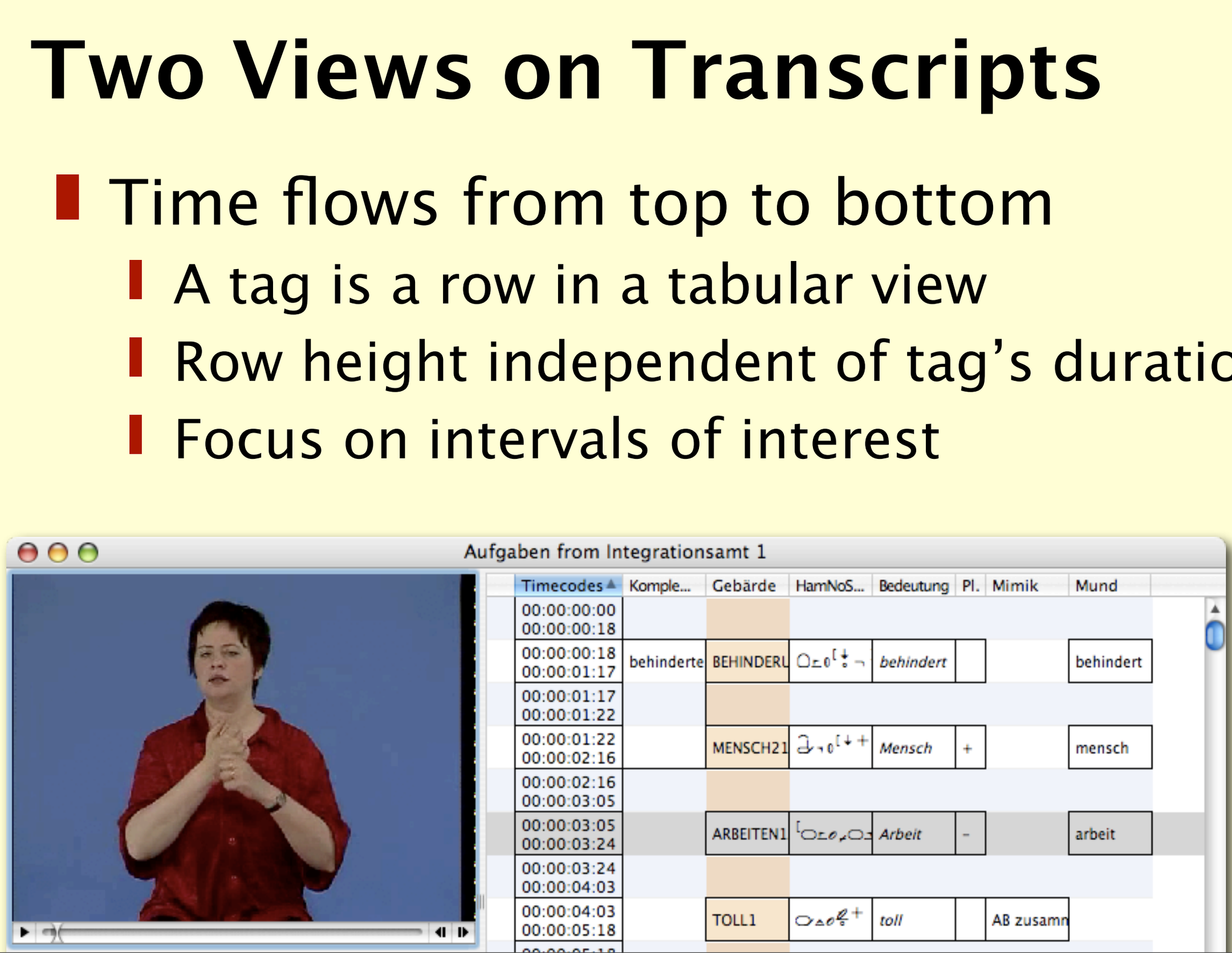


Type & Token

- The identified type is dragged into the transcript to become a token
- If no type is found: A new type is created and then assigned

Two Views on Transcripts

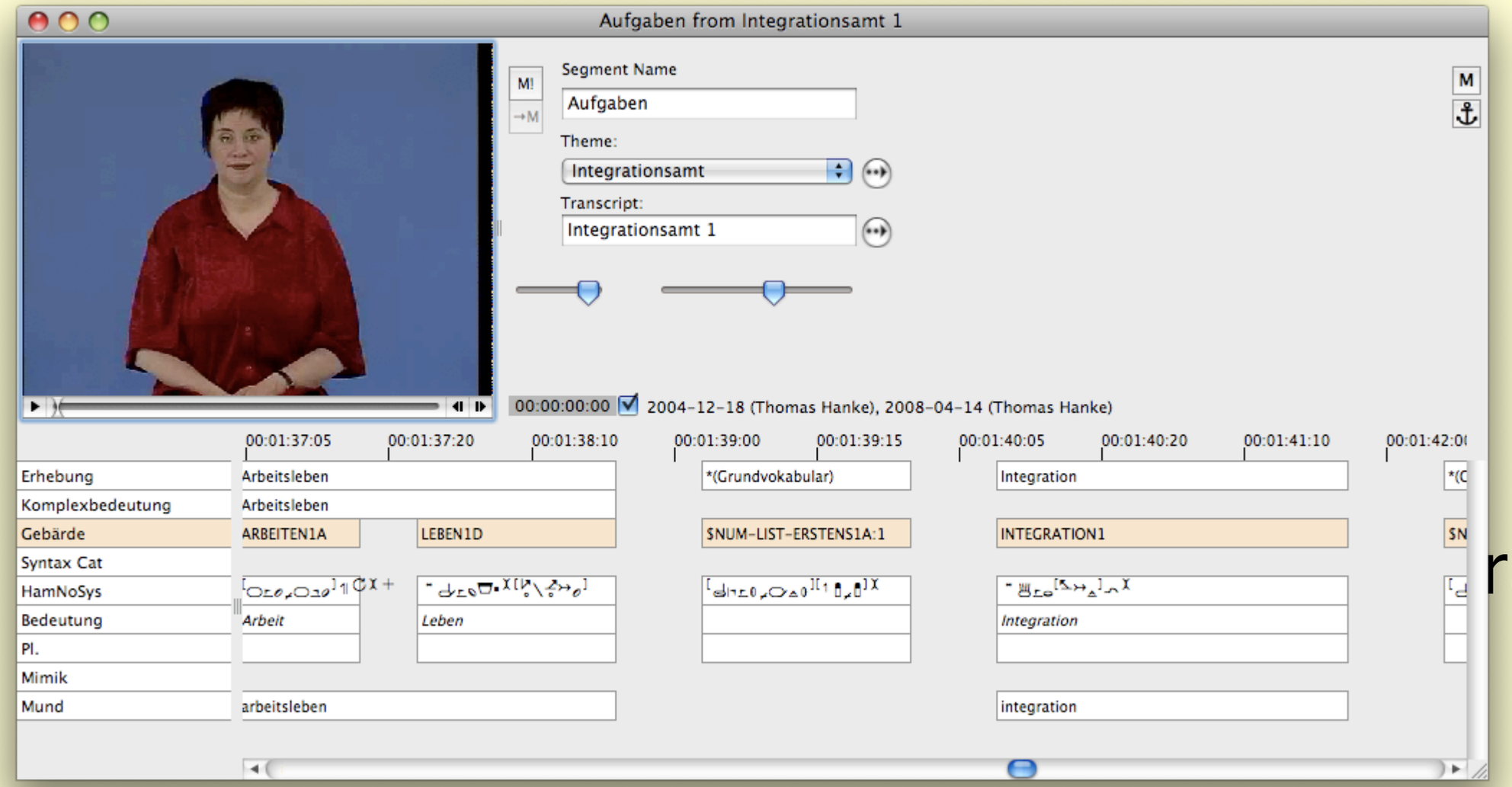
- Time flows from top to bottom
  - A tag is a row in a tabular view
  - Row height independent of tag's duration
  - Focus on intervals of interest



Two Views on Transcripts

- Time flows from top to bottom
  - A tag is a row in a tabular view
  - Row height independent of tag's duration
  - Focus on intervals of interest

- Time flows from left to right
  - A tag's length is proportional to its duration



ration

- Time flows from left to right
  - A tag's length is proportional to its duration

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

A tag may consist of one or two tokens (dom+nondom hand) or

A tag consists of one token only, and you have two or three token tiers per informant

Token Tiers

- System can make sure that cotemporal tokens for one informant take a maximum of two hands
- Or only one hand once video processing is able to tell us
- What then?
  - Wrong type assigned to token?
  - Forgot to mark as weak drop?

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers

Tags group several token tags to a phrase or whatever with a certain meaning

- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

Describe the prompts presented, allows access to multimedia data presented to user

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

References to types that do not count as tokens, such as corrections

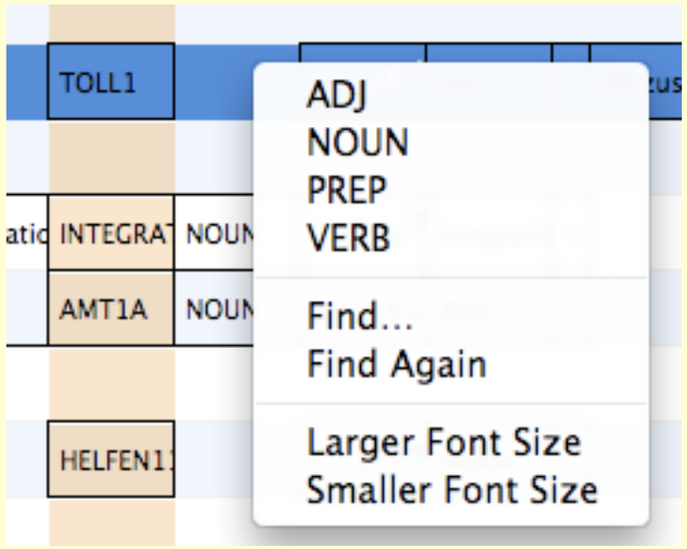
Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

As you know them from ELAN etc.
Can have open or closed vocabularies associated

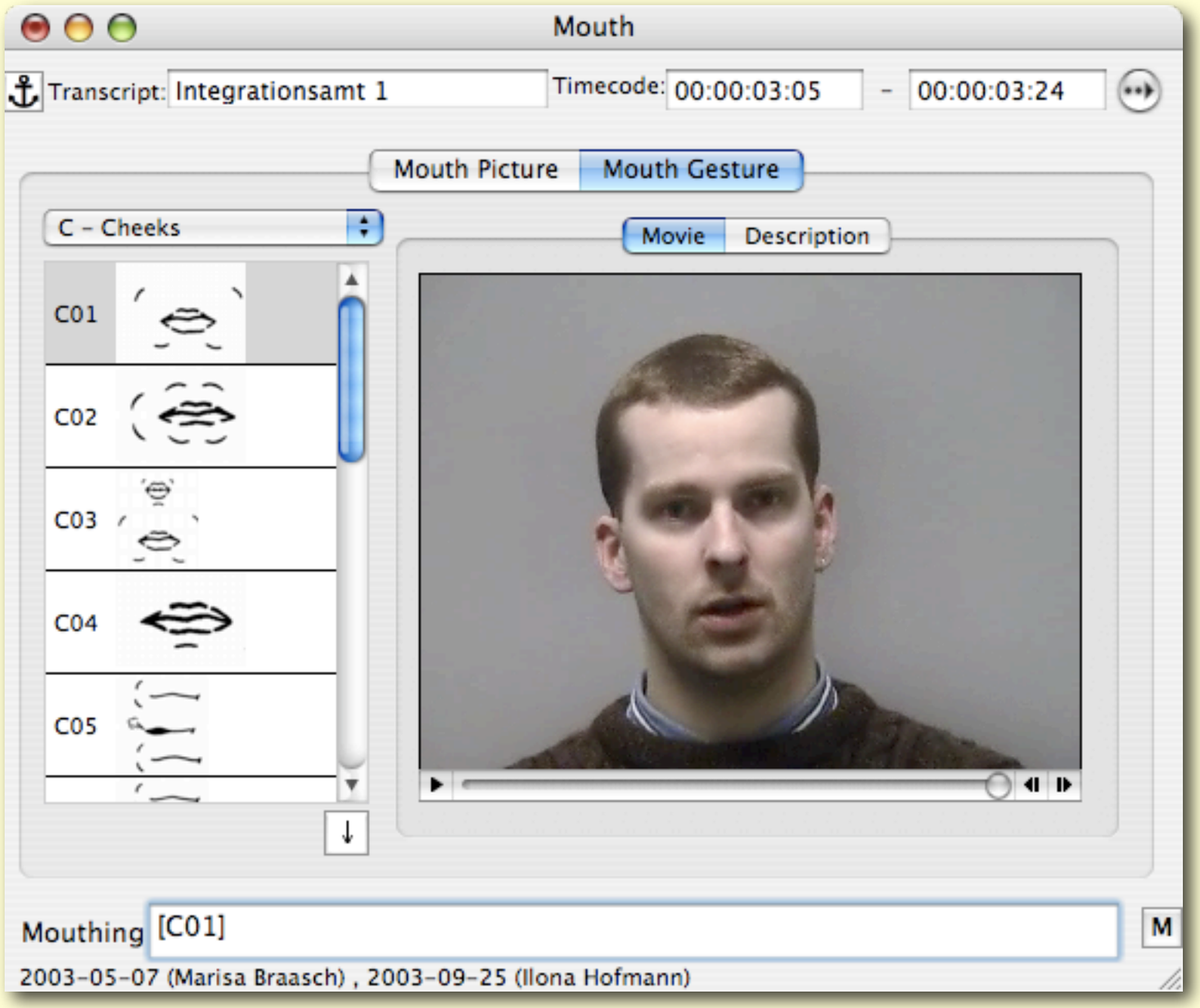
Text Tiers with Vocabularies

- Closed or open: Texts to choose from



- Mouthings: default mouthing associated to type plus signer's sign/mouthing combinations
- Built-in vocabularies with graphical editors

Text Tiers with Special Editors, e.g. Mouth gesture encoding



Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

E.g. to describe positions in the video

Tier Kinds in iLex

- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

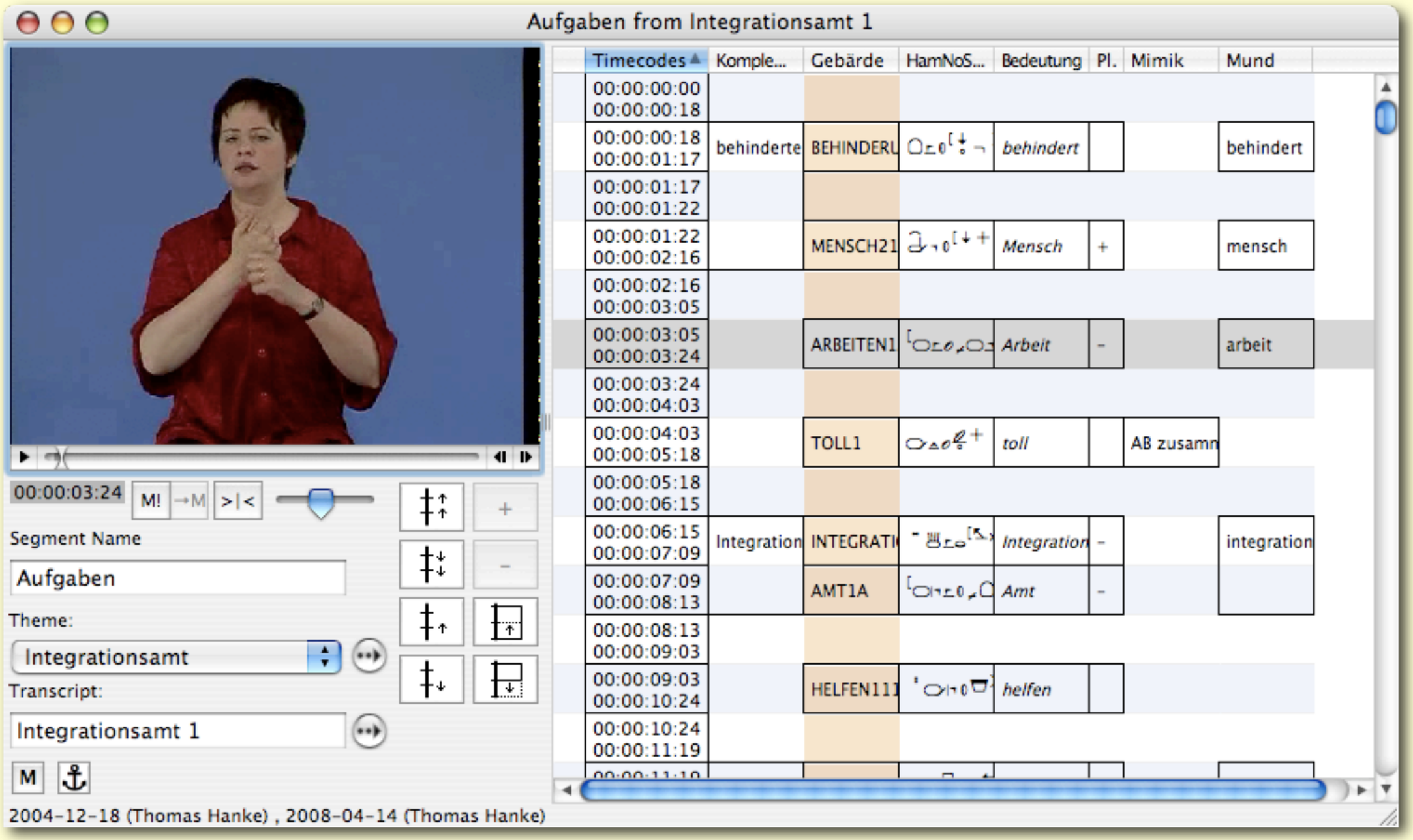
E.g. to relate anaphoric elements to their referents

Tier Kinds in iLex

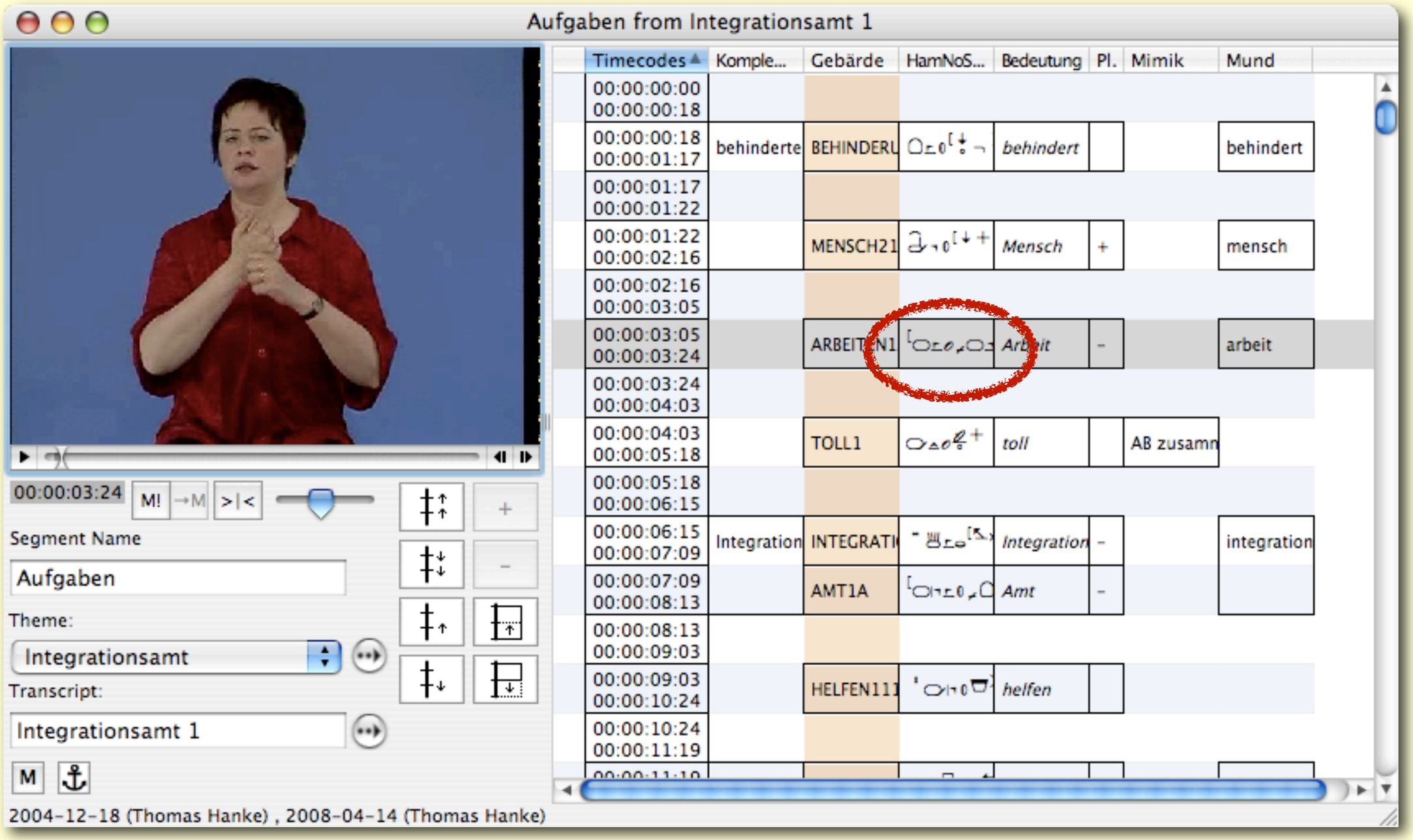
- Tokens tiers
- Phrasal units
- Elicitation tiers
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

To reflect other features of complex tag data than displayed in the respective tier, e.g. citation form in HamNoSys of the type assigned to a token

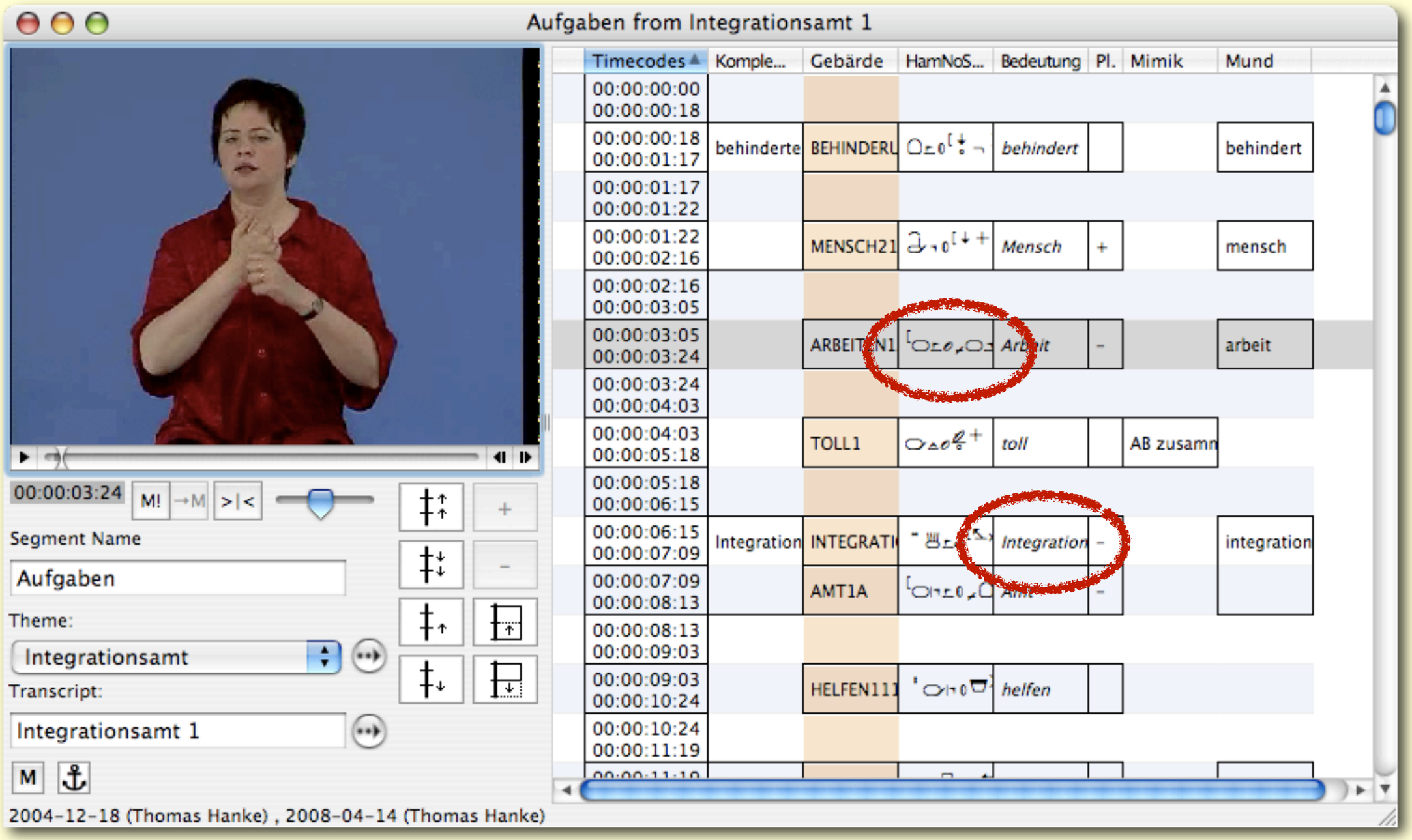
Value Tiers: Copy values from complex objects in other tiers



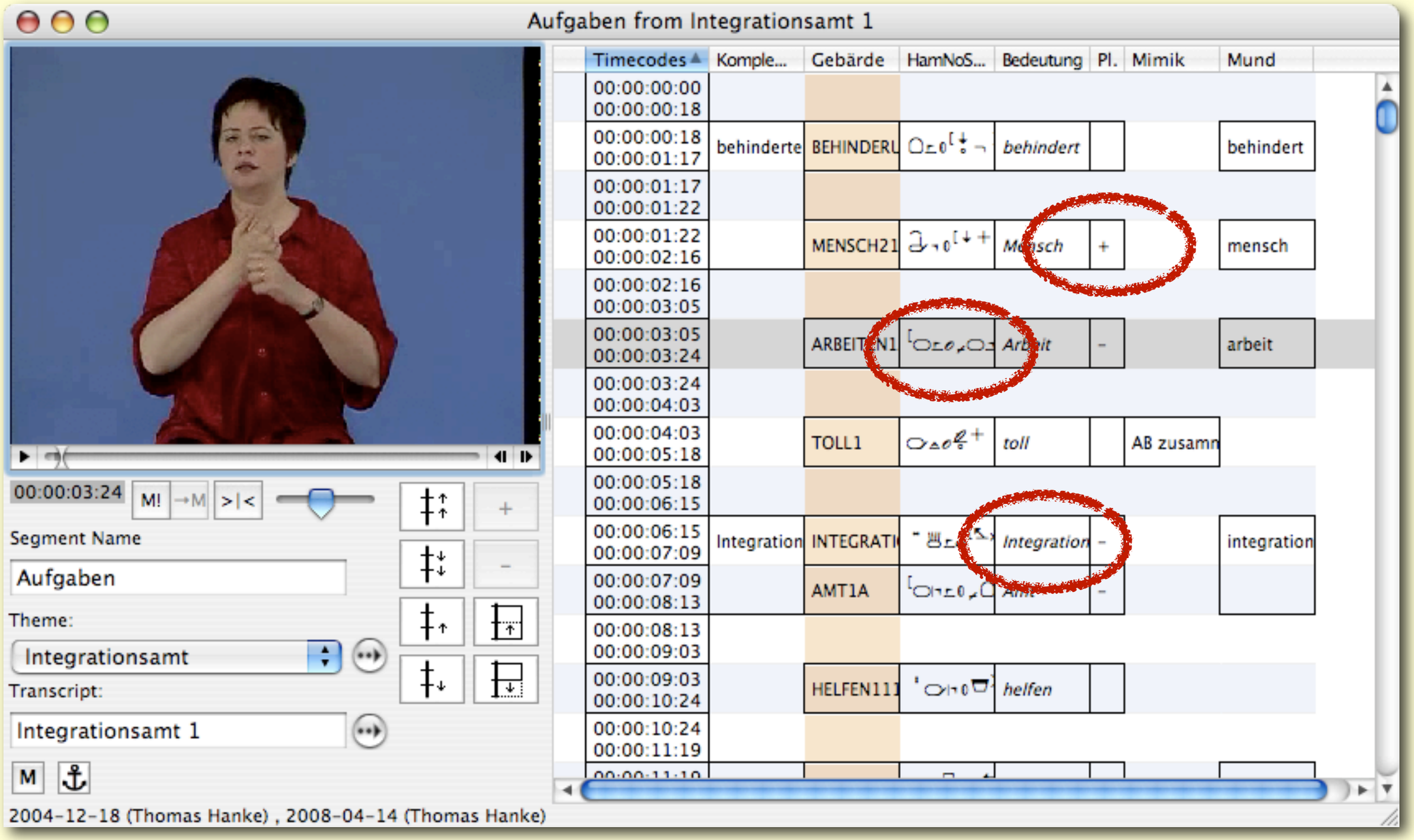
Value Tiers: Copy values from complex objects in other tiers



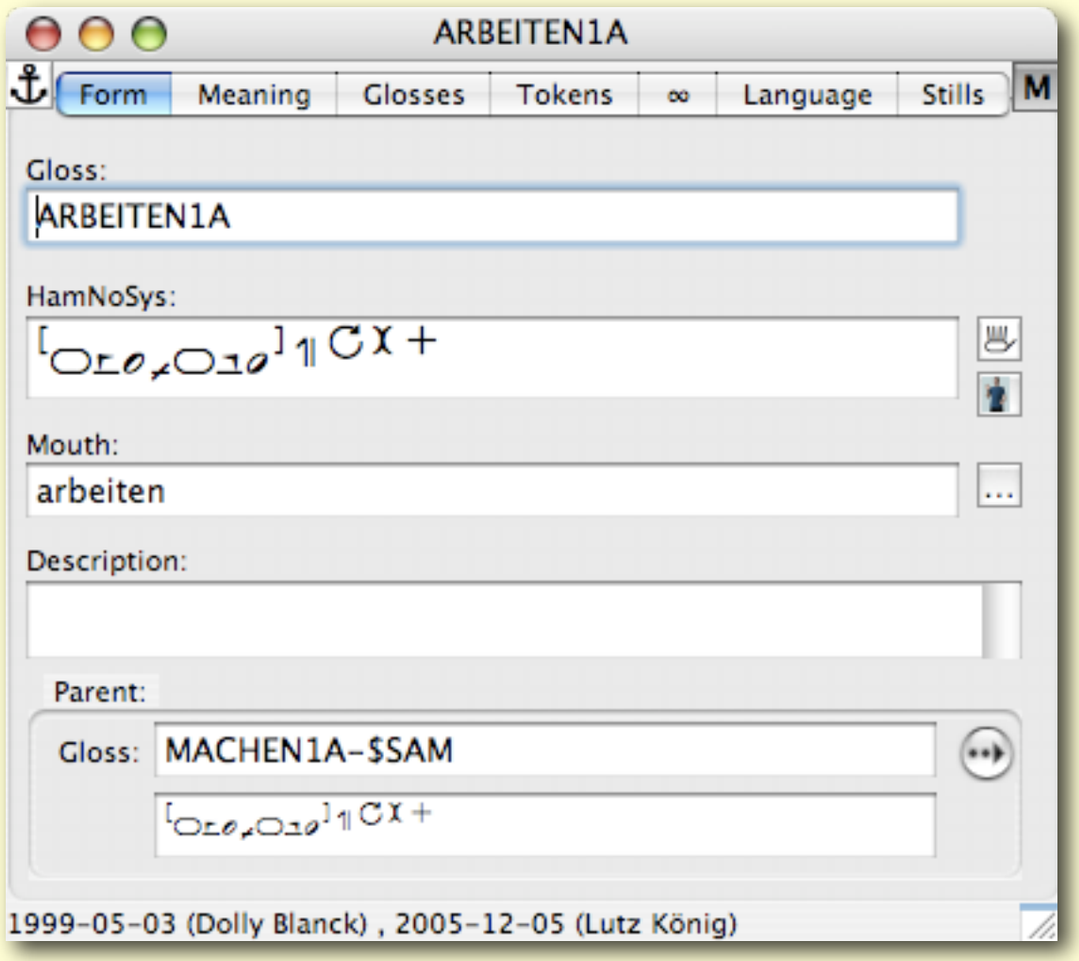
Value Tiers: Copy values from complex objects in other tiers



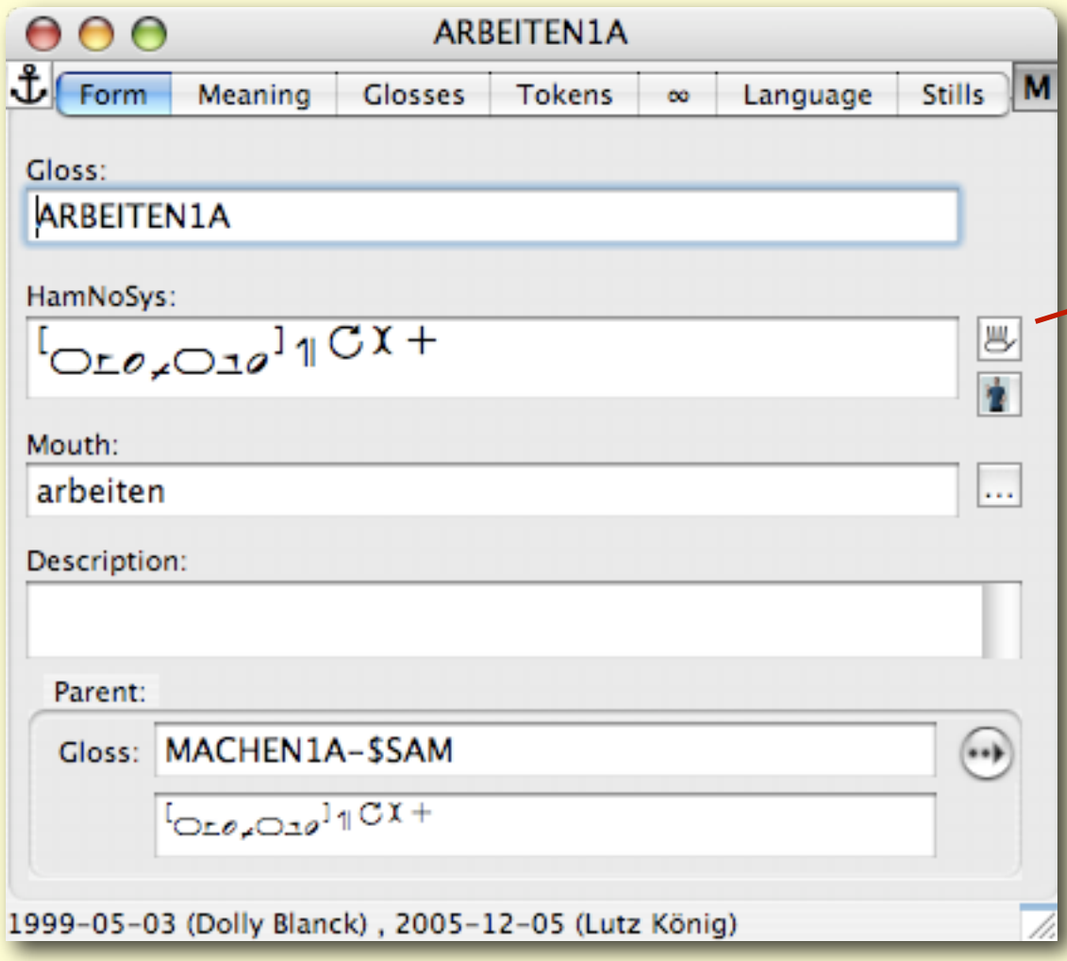
Value Tiers: Copy values from complex objects in other tiers

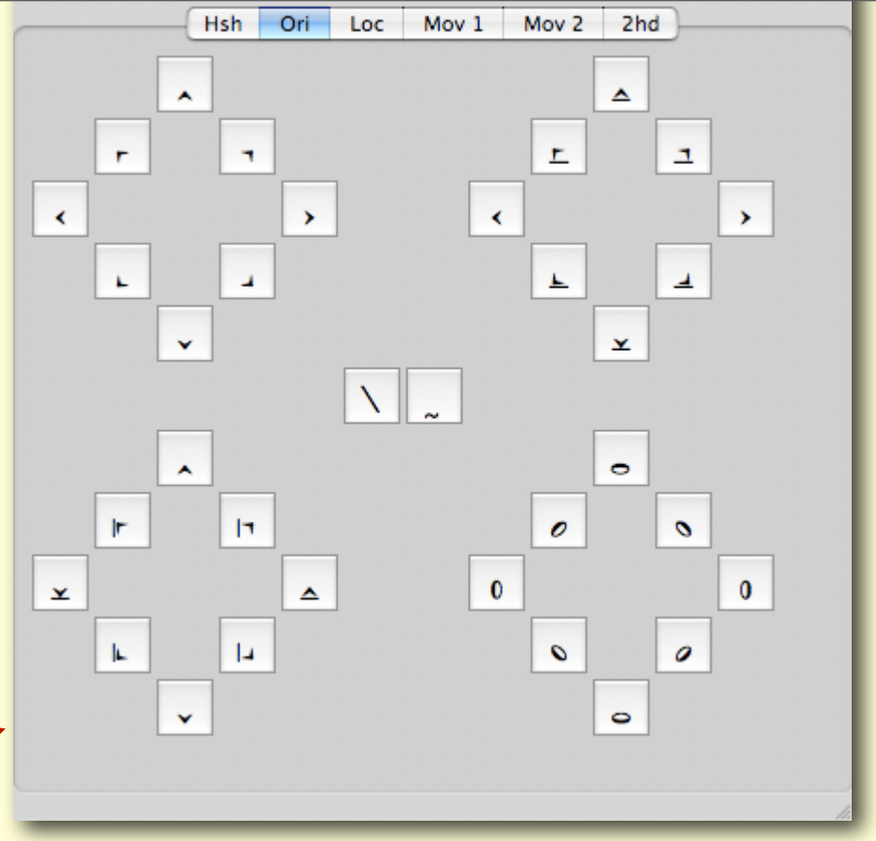


Immediate input verification: HamNoSys & avatar

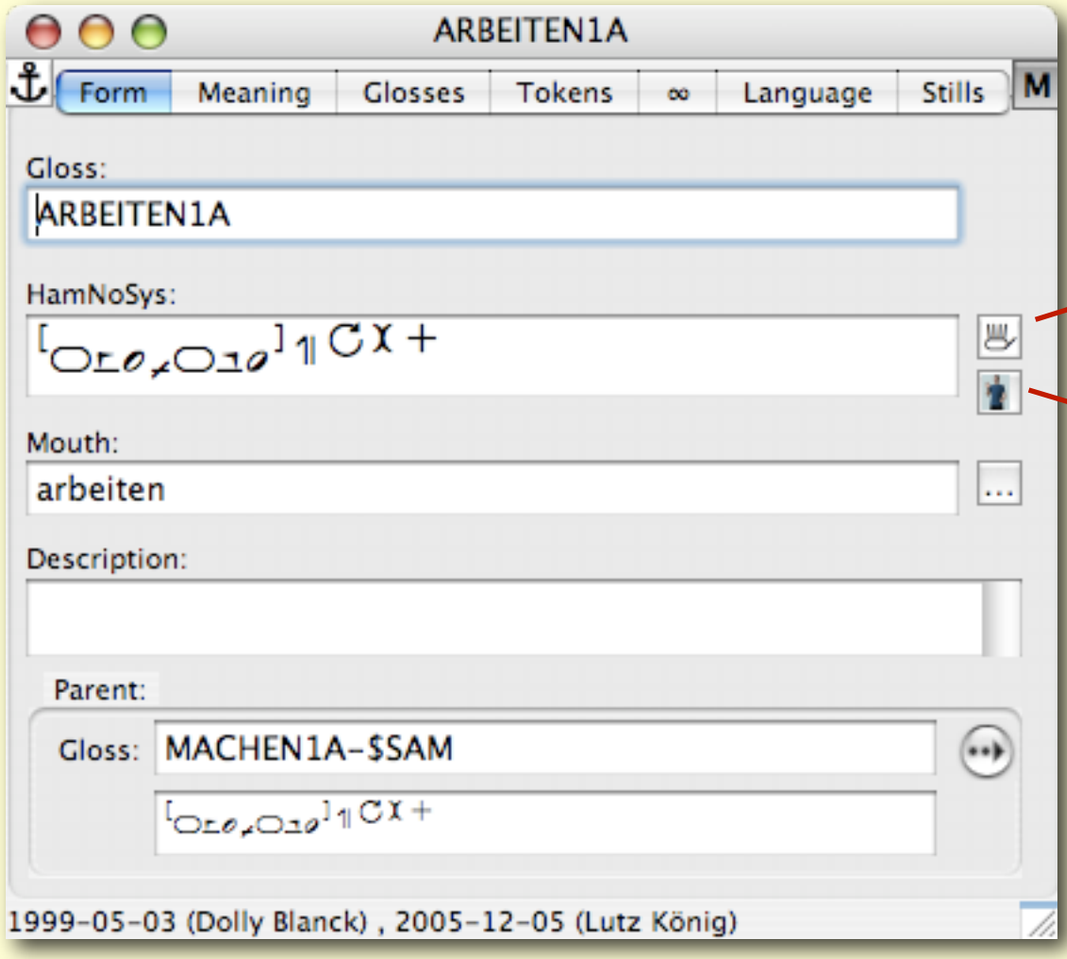


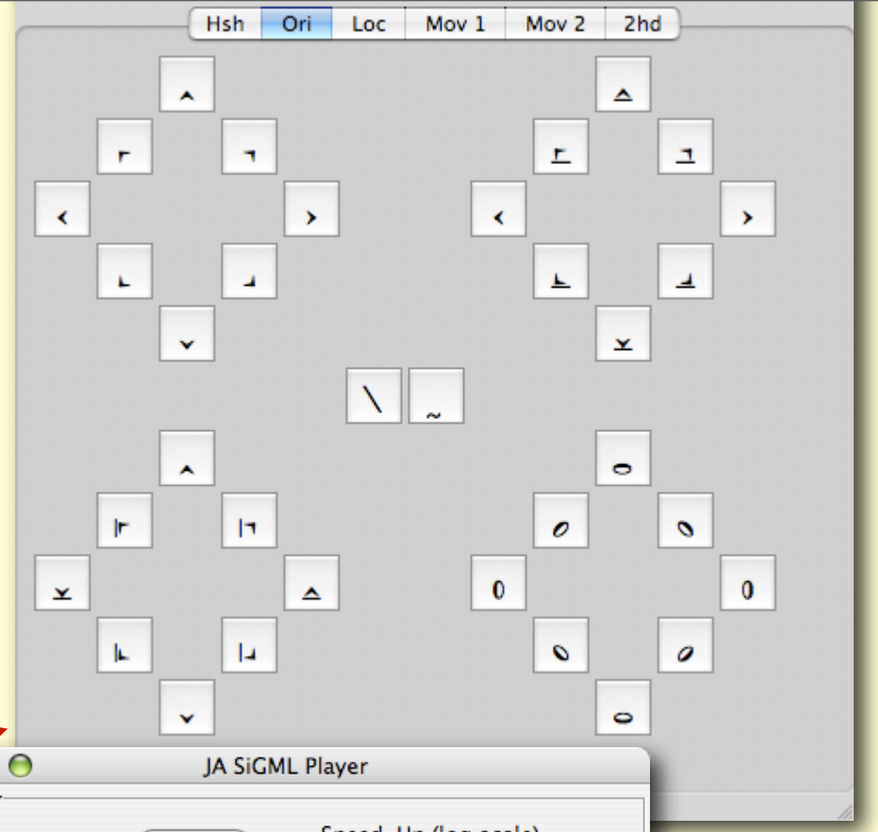
Immediate input verification: HamNoSys & avatar

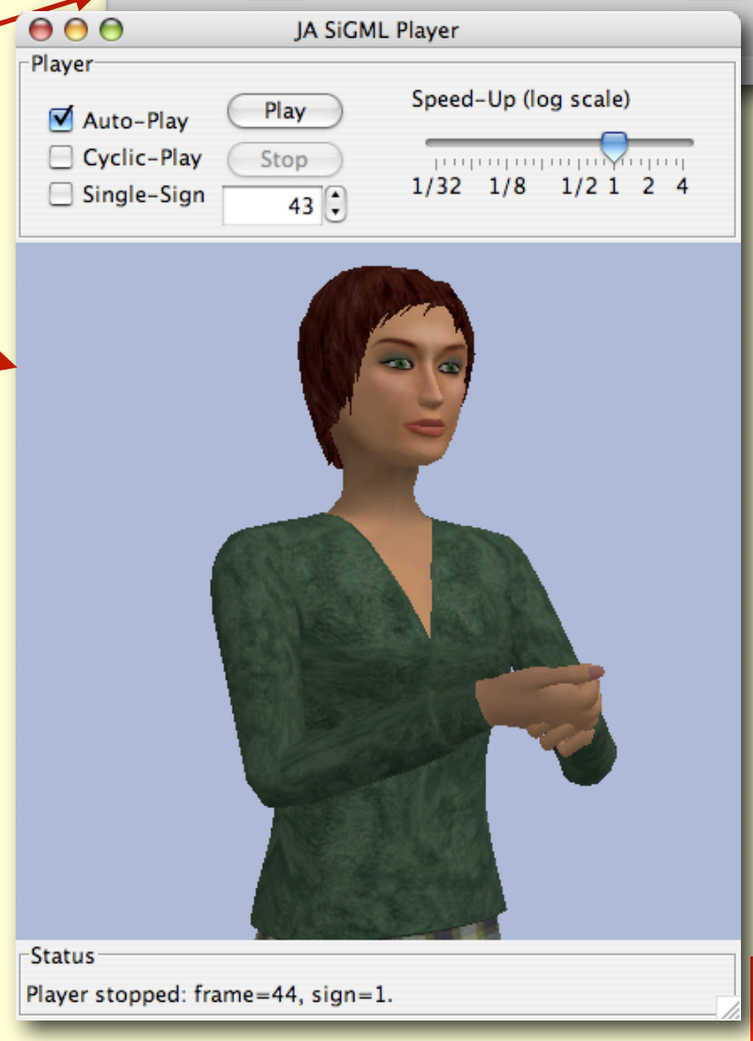




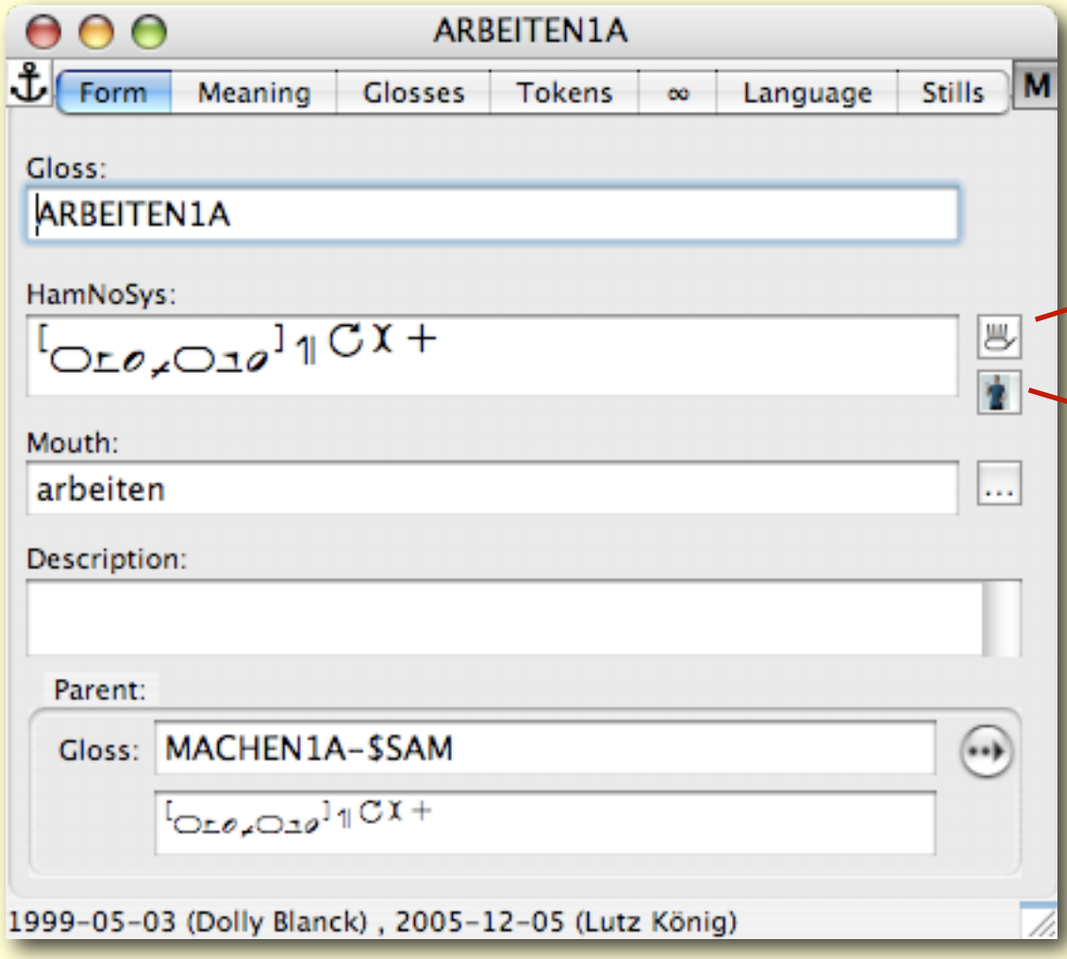
Immediate input verification: HamNoSys & avatar

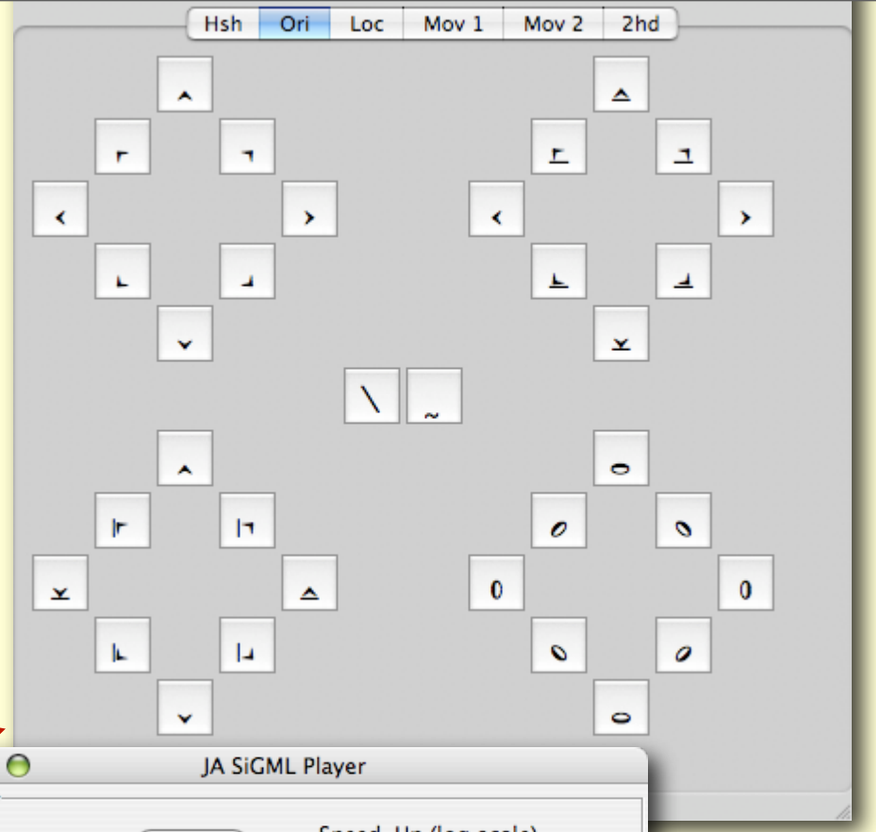


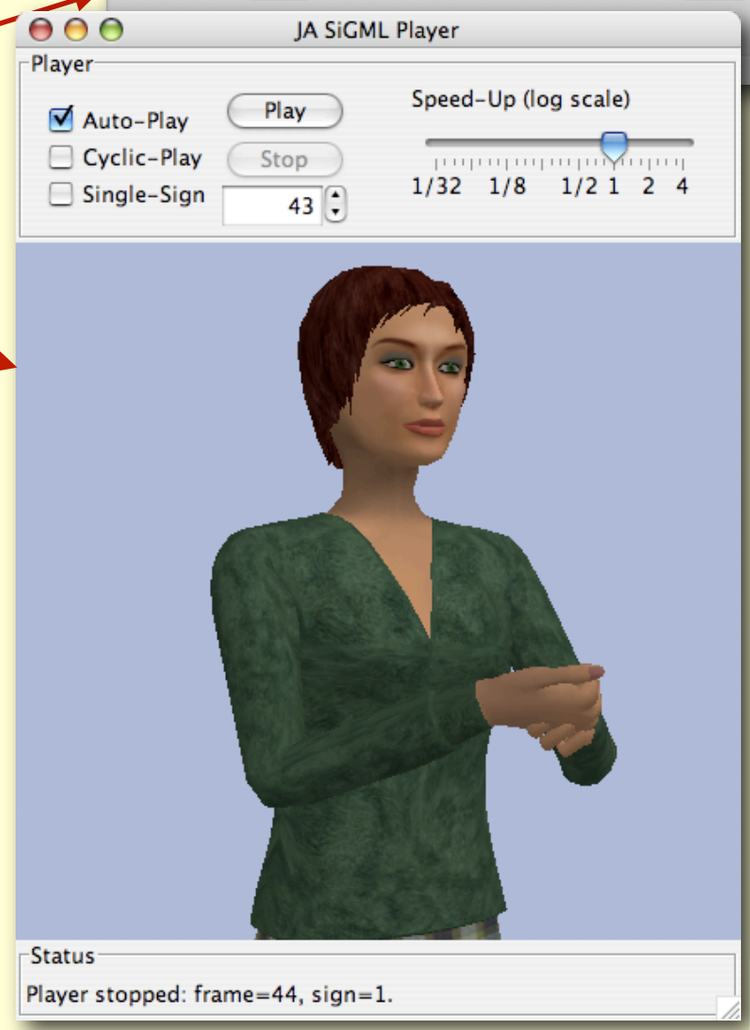




Immediate input verification: HamNoSys & avatar







- Side effect: anonymisation

How do you annotate the form differences between type and token?

- Just ignore them
  - Refer to the citation form
- Describe the form deviation in the token
- Have separate types to describe citation form and inflected form
- Have a separate text tier describing the kind of modification… you encounter

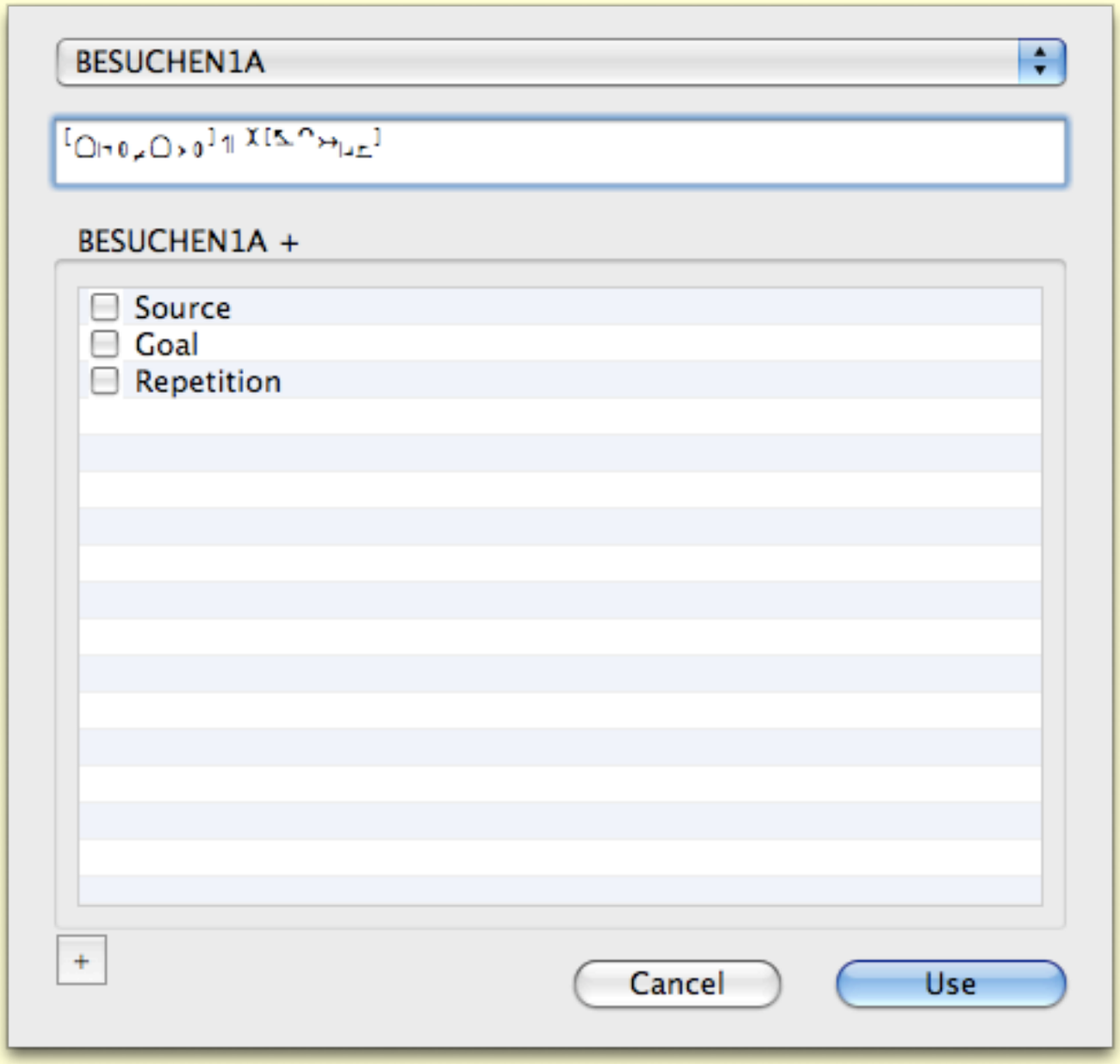
Qualifications

- Inflection
  - Agreement
  - Plural repetition
- Modification
  - Extension of the sign's image to convey extra meaning
- Variation

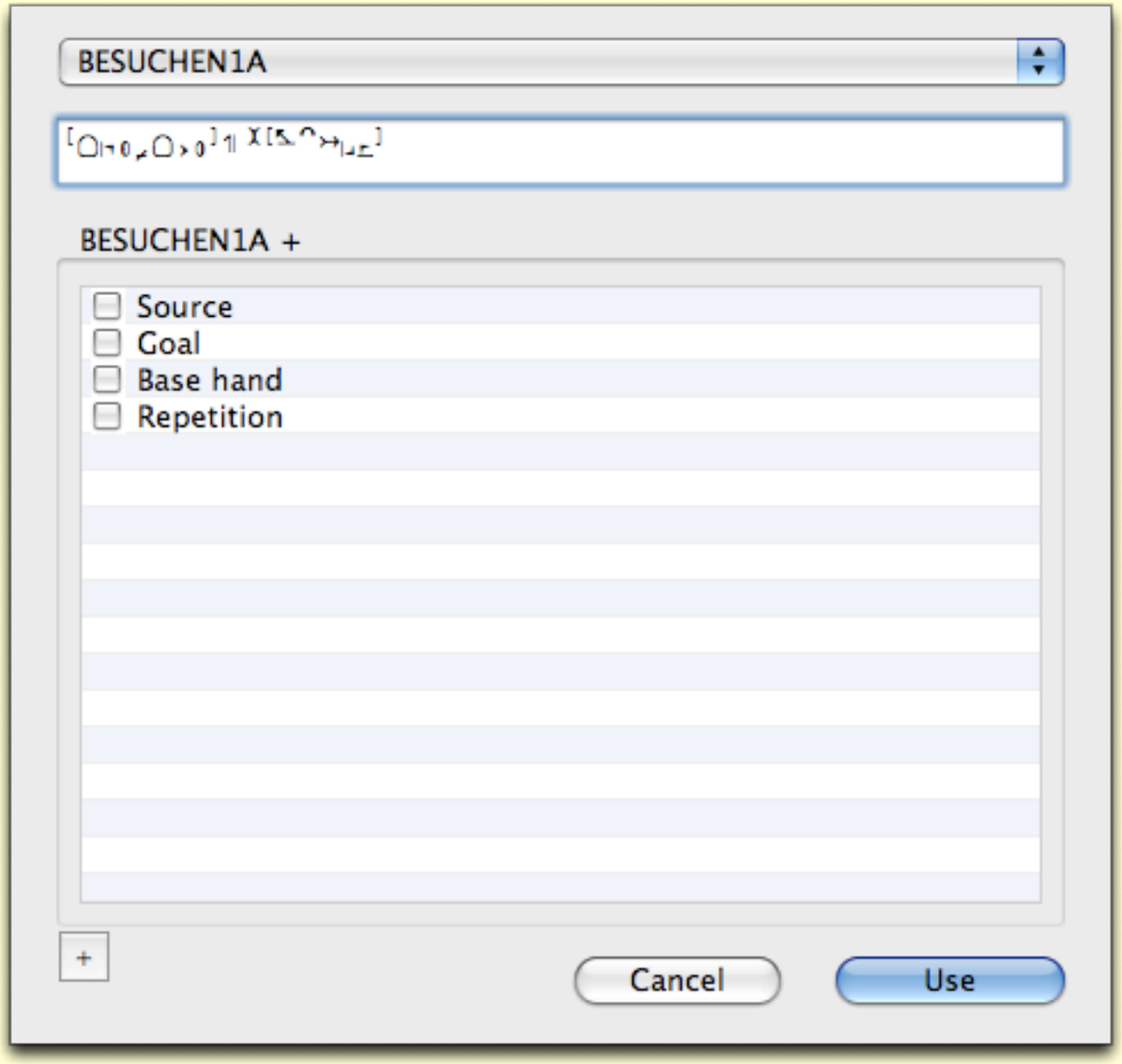
Qualifiers

- “Inflection paradigms” as controlled vocabularies
- For each type, previously used qualifiers are readily available
- Others require an extra step

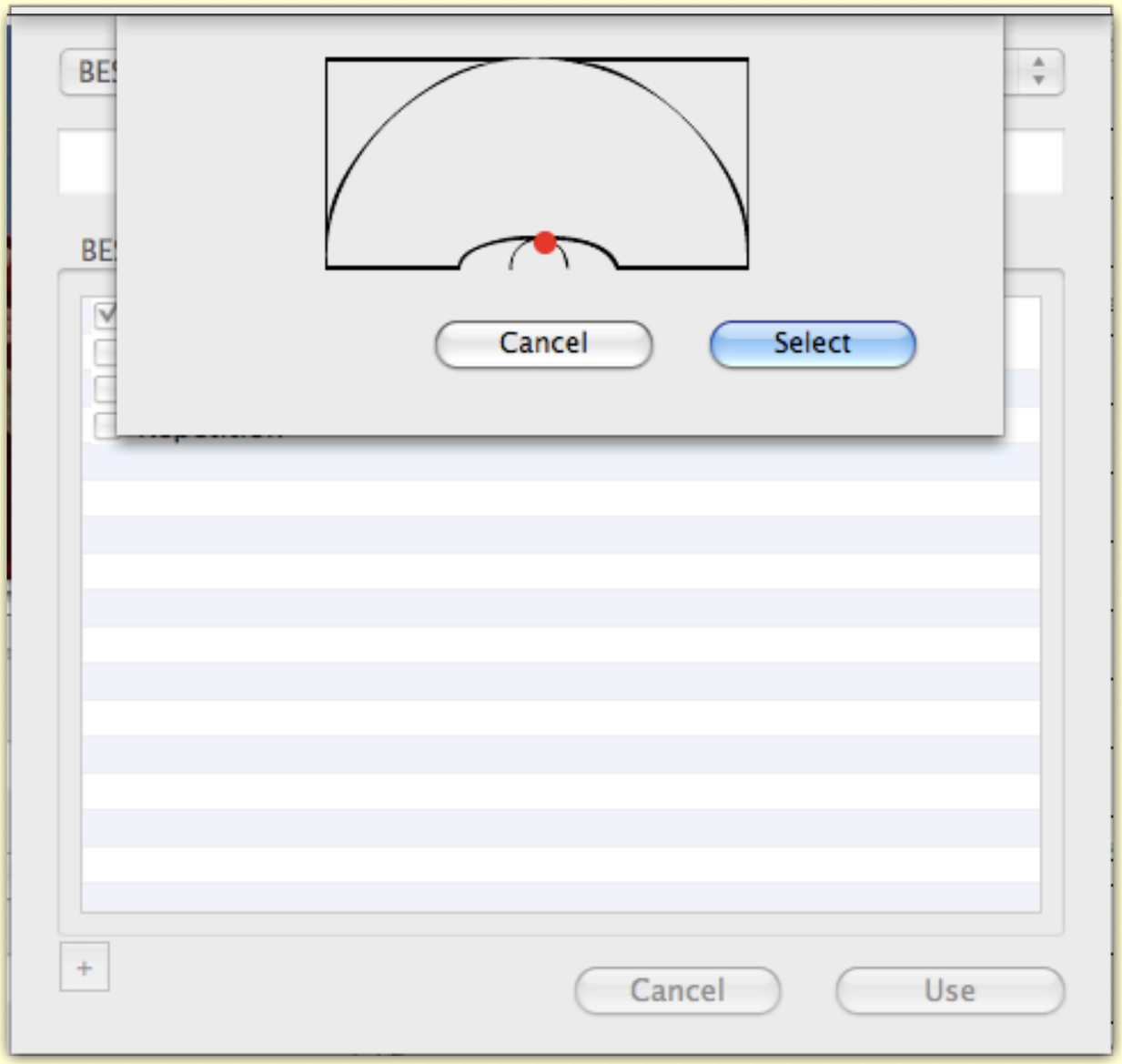
Qualifiers step by step



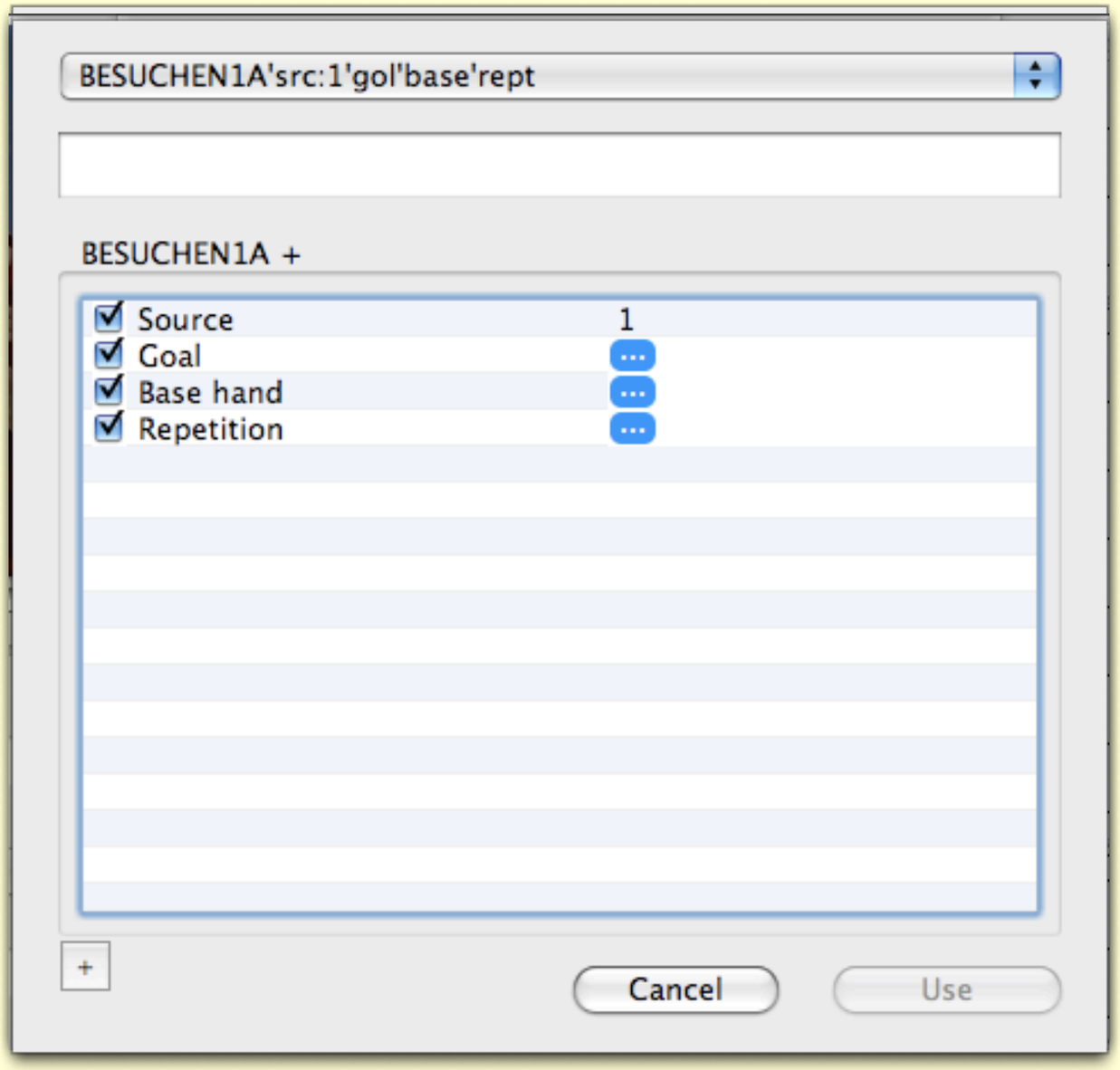
Qualifiers step by step



Qualifiers step by step



Qualifiers step by step



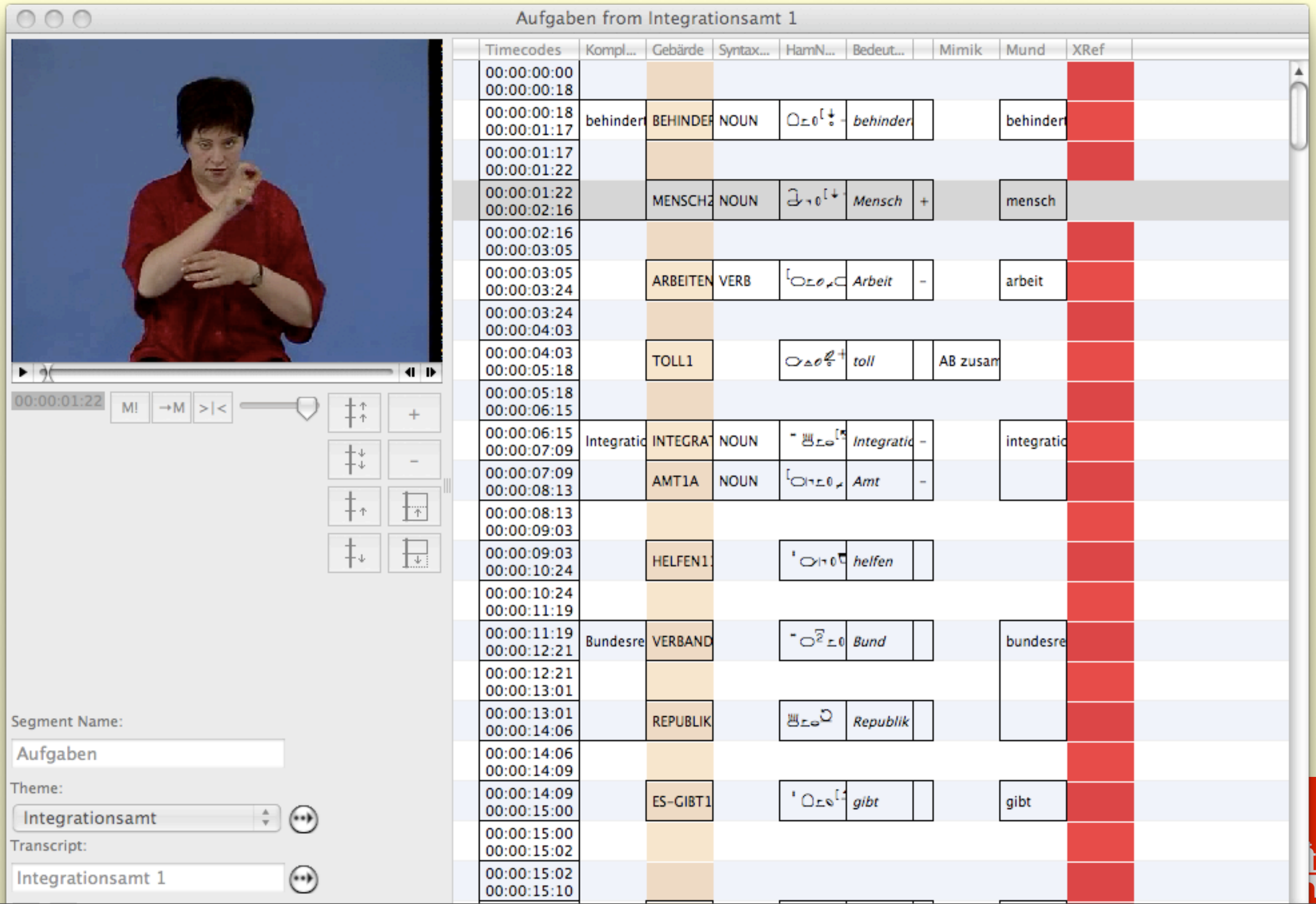
Qualified Types: Advantages

- Per-type controlled vocabularies
- Any statistics easily take qualifications into account ("word-form") or ignore them ("type-level")
- Lexicographic view supported:
  - Qualified types can "freeze" into lexemes
  - Relations between types can include qualified types

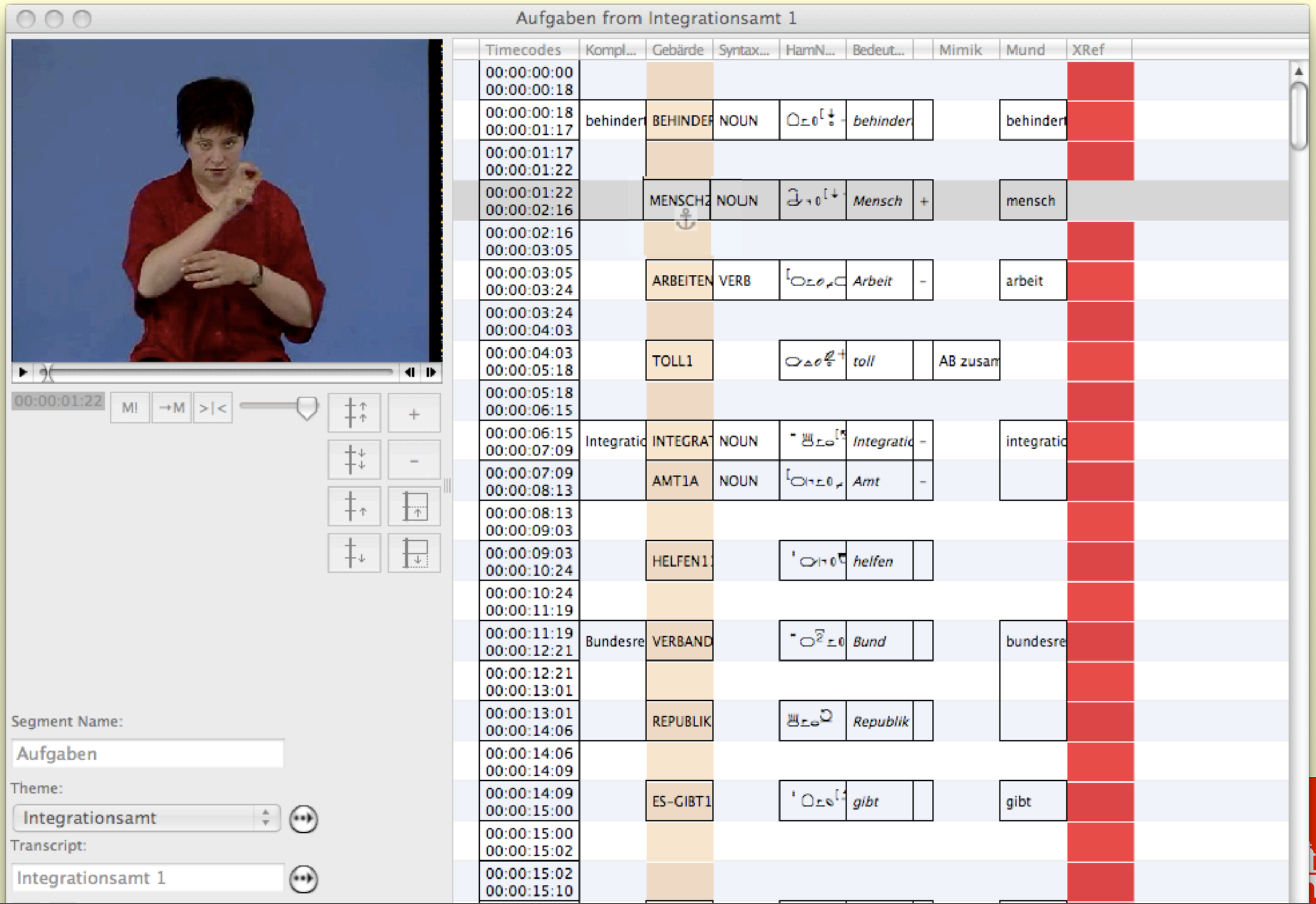
Cross References

- A new tier kind with tags establishing relations to tags in other tiers
- E.g. anaphora linking to their referents
  - Centering theory
- E.g. source/goal/location linking to where the location was established

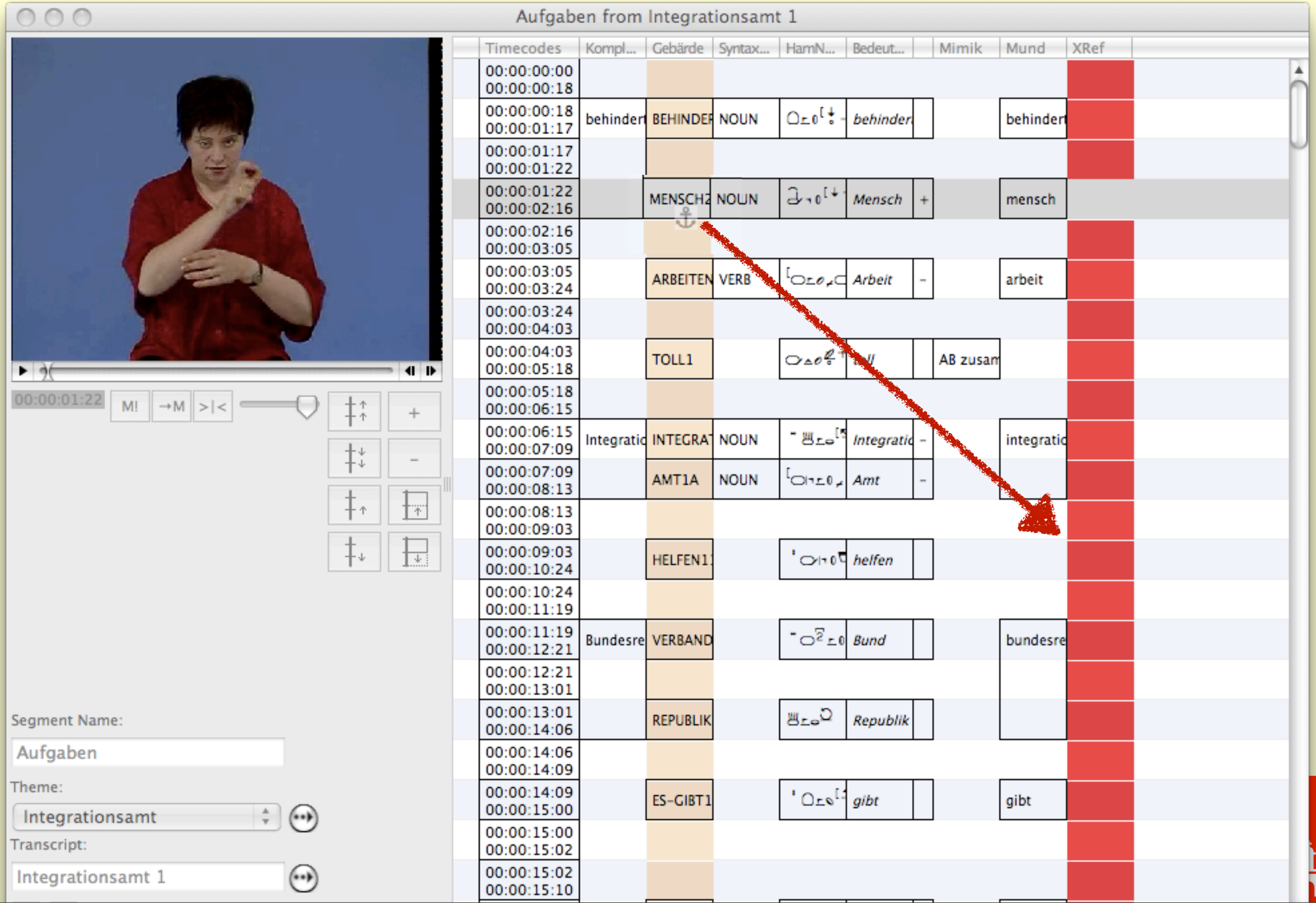
Establishing References



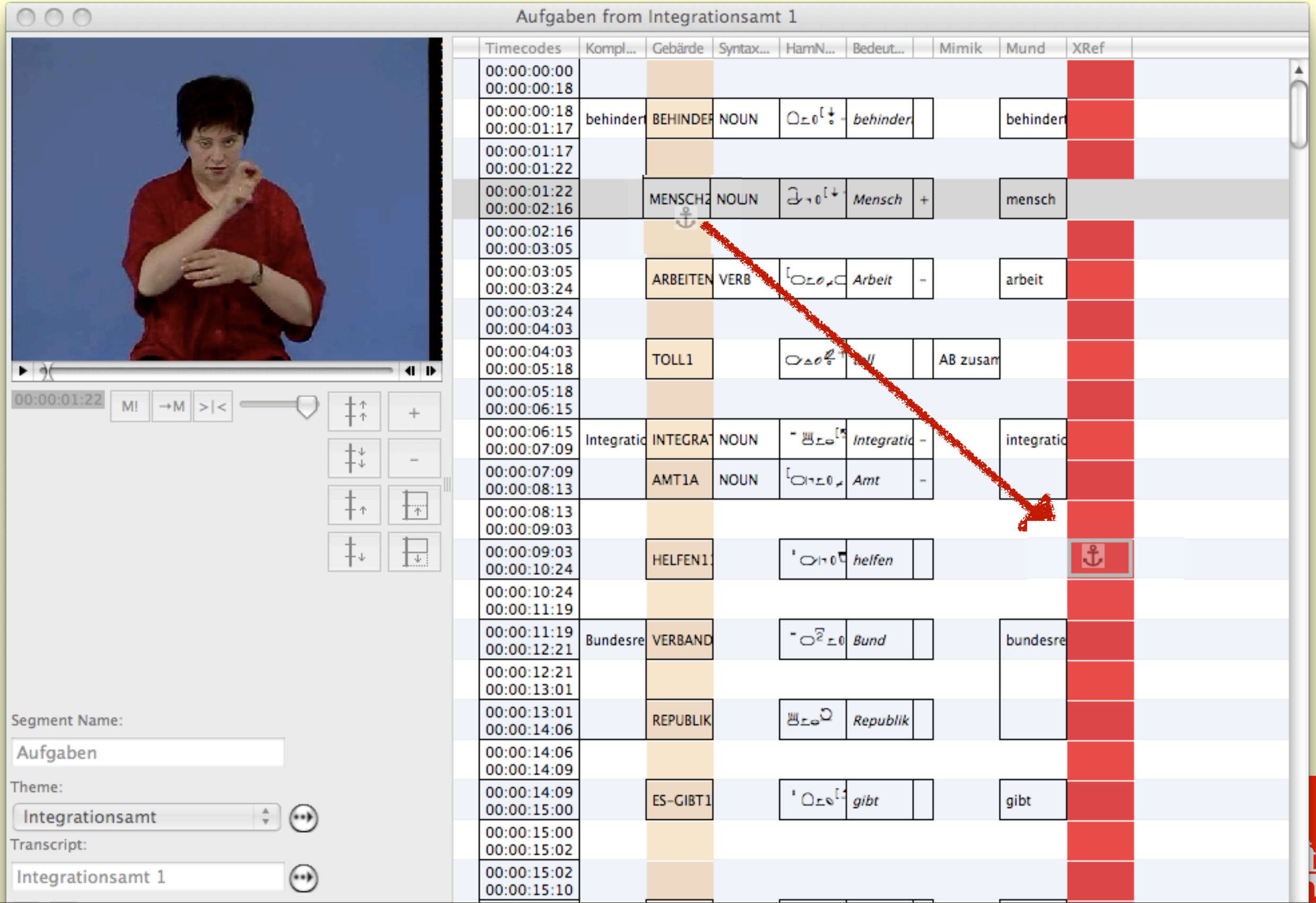
Establishing References



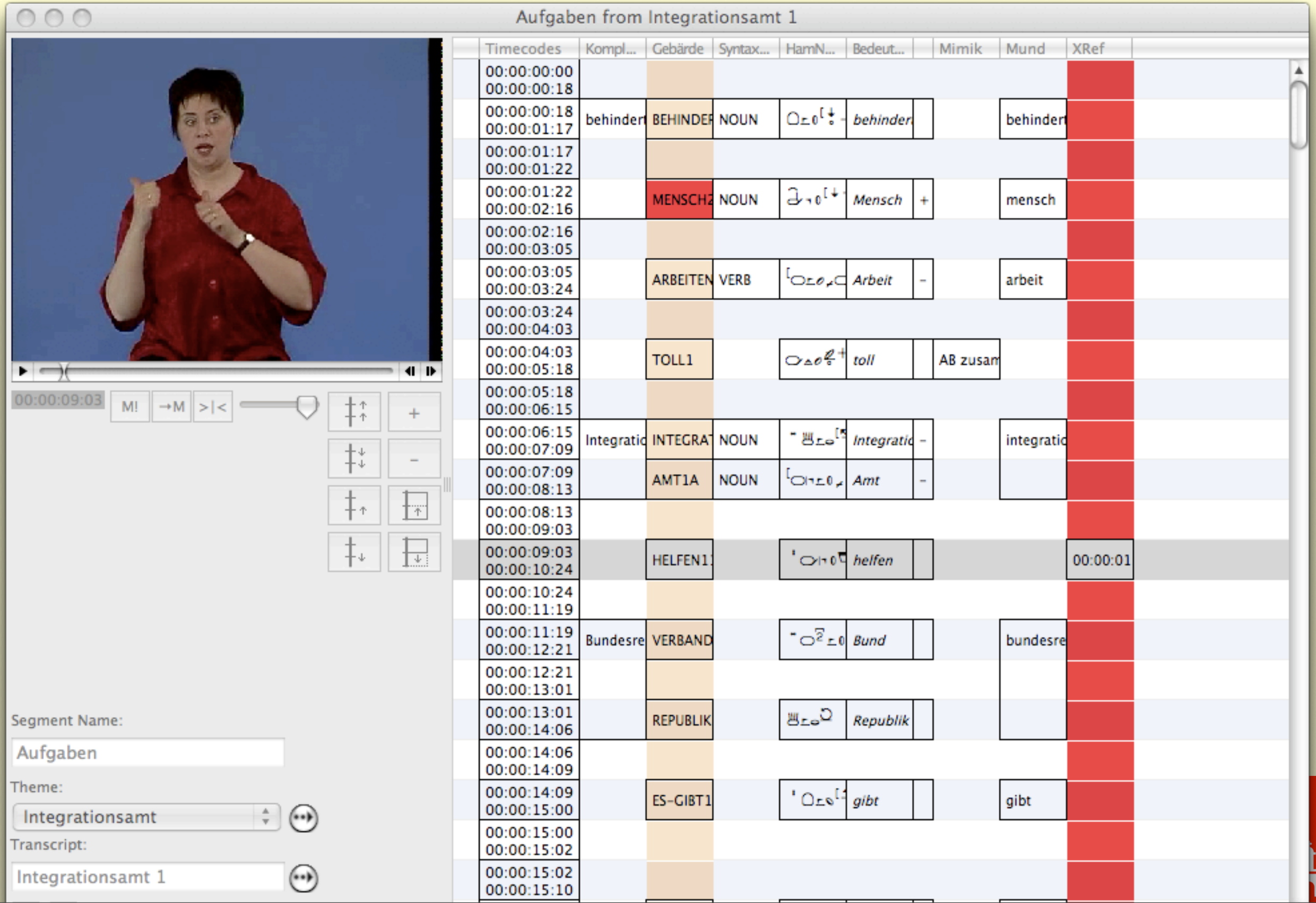
Establishing References



Establishing References



Establishing References



In Combination with Qualifications: Chances for Automatic Linking

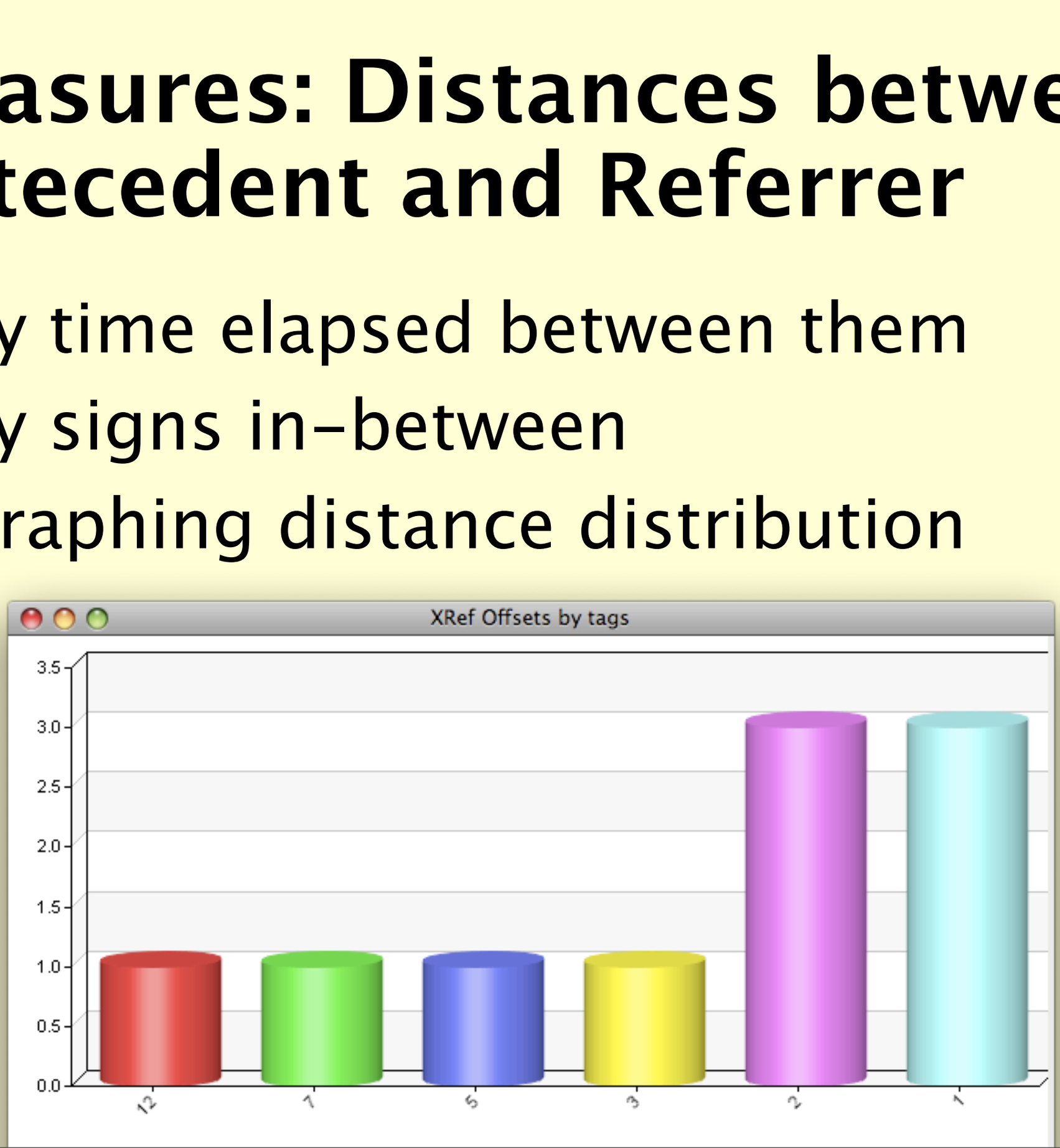
- For tokens with a specific goal:
  - Search for predecessor tags that establish that location in signing space
  - Scripting language available to meet individual needs

Measures: Distances between Antecedent and Referrer

- by time elapsed between them
- by signs in-between
- Graphing distance distribution

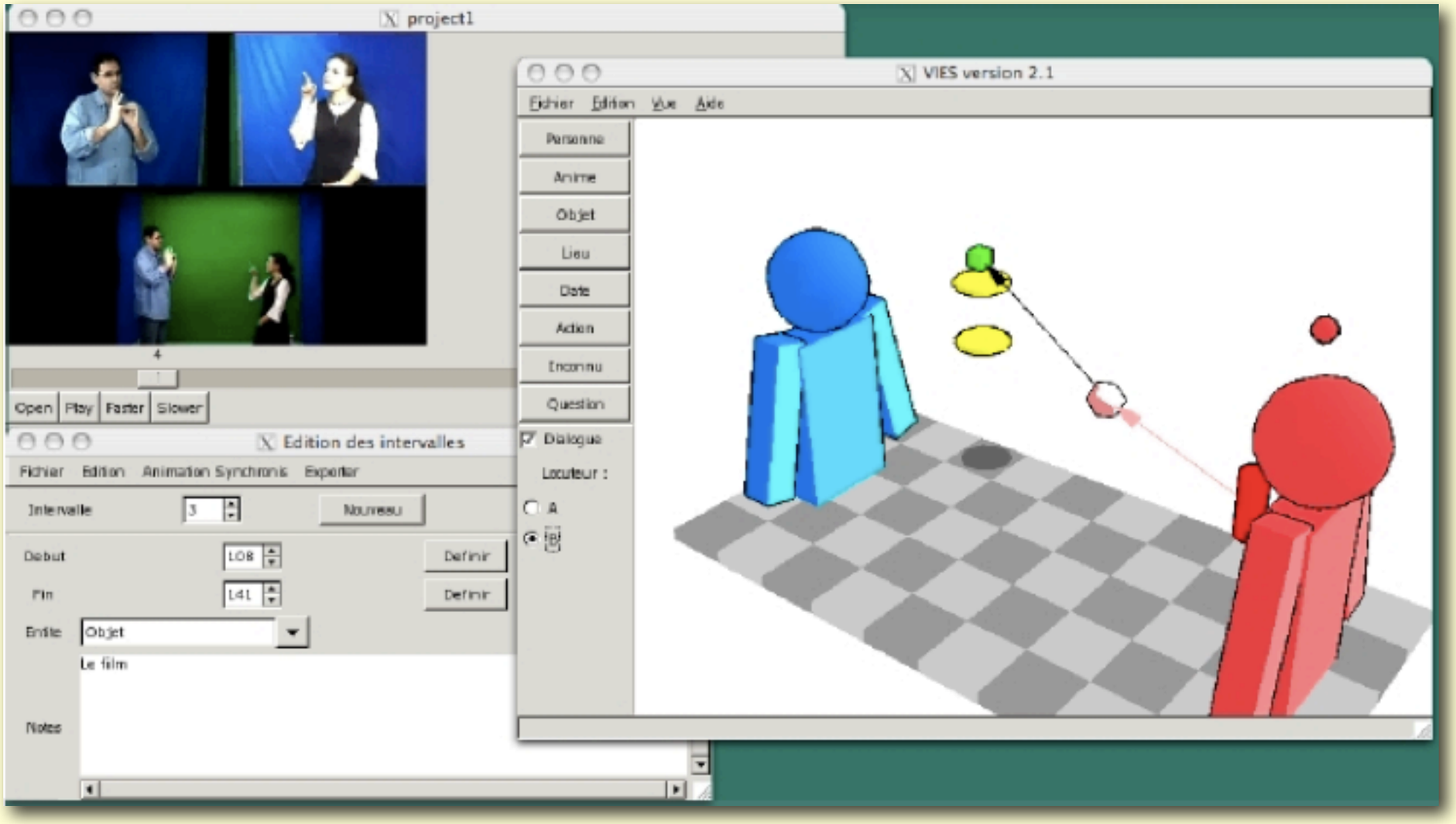
Measures: Distances between Antecedent and Referrer

- by time elapsed between them
- by signs in-between
- Graphing distance distribution



Improved visualisation of spatio-temporal data & semantic relations

- IRIT Toulouse

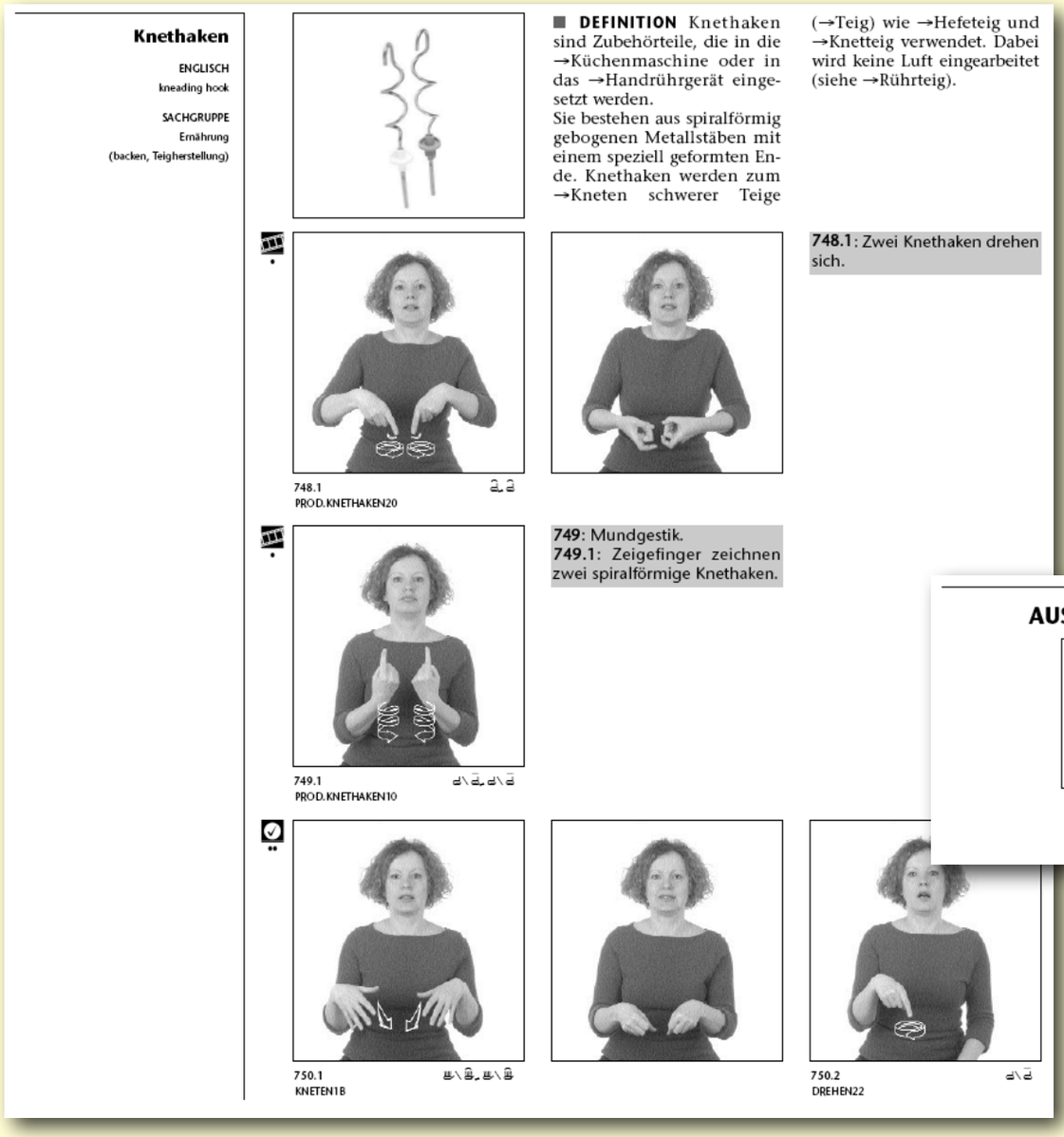


Applications of iLex

- Documentation
- Lexicography
- Animation
- Teaching
  - eLearning
  - Language lab materials

Dictionary Production

- Book
- Internet/DVD
- DVD Video



Knethaken

ENGLISCH
kneading hook

SACHGRUPPE
Ernährung
(backen, Teigherstellung)

■ **DEFINITION** Knethaken sind Zubehörteile, die in die →Küchenmaschine oder in das →Handrührgerät eingesetzt werden.
Sie bestehen aus spiralförmig gebogenen Metallstäben mit einem speziell geformten Ende. Knethaken werden zum →Kneten schwerer Teige (→Teig) wie →Hefeteig und →Knetteig verwendet. Dabei wird keine Luft eingearbeitet (siehe →Rührteig).

748.1: Zwei Knethaken drehen sich.

748.1
PROD.KNETHAKEN20

749: Mundgestik.
749.1: Zeigefinger zeichnen zwei spiralförmige Knethaken.

749.1
PROD.KNETHAKEN10

750.1
KNETEN1B

750.2
DREHEN22

Soziometrie

- **DEFINITION**

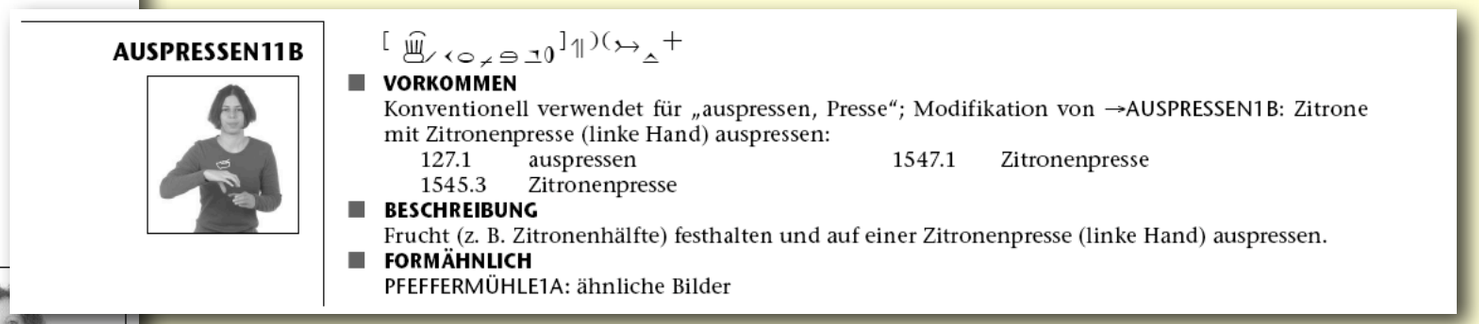
empirische Sozialforschung

Die Soziometrie ist ein Verfahren der empirischen Sozialforschung, mit dem die Beziehungen zwischen den M
Gruppe festgestellt, dargestellt und untersucht werden können. Es wird deshalb nach Sympathie und Antipath
Mitgliedern einer Gruppe gefragt, man untersucht die Kontaktwünsche (subjektive Interaktionspräferenzen) u
Kontakte (faktische Interaktionsbeziehungen). Das erfolgt im Wesentlichen durch die Befragung der Gruppen
(Interaktion).
Soziometrische Ergebnisse werden in einem Soziogramm grafisch dargestellt. Eine mathematische Darstellung
unter den Mitgliedern einer Gruppe erfolgt in Soziomatrizen. Seltener sind Darstellungen, in denen eine Kenn
Beziehungen der Gruppenmitglieder informiert.
Die Soziometrie kann bei der Analyse kleiner Gruppen, deren Mitglieder sich gut kennen, eingesetzt werden (
Gruppenarbeit). Da die Ergebnisse im Wesentlichen durch die Befragung der Gruppenmitglieder gewonnen w
Aussagen auch von dem Bewusstsein der Gruppenmitglieder über ihre Beziehungen abhängig und von deren
Wunschvorstellungen geprägt.

- **THEMATISCH VERWANDTE FACHBEGRIFFE** Psychodrama
- **SACHGRUPPEN**
  - Methoden der Sozialarbeit/Sozialpädagogik
  - Psychologie

1 (803) VERBAND21A

2 (804) SOZIAL2A MUSTER1A



AUSPRESSEN11B

[⟨ ⟩]

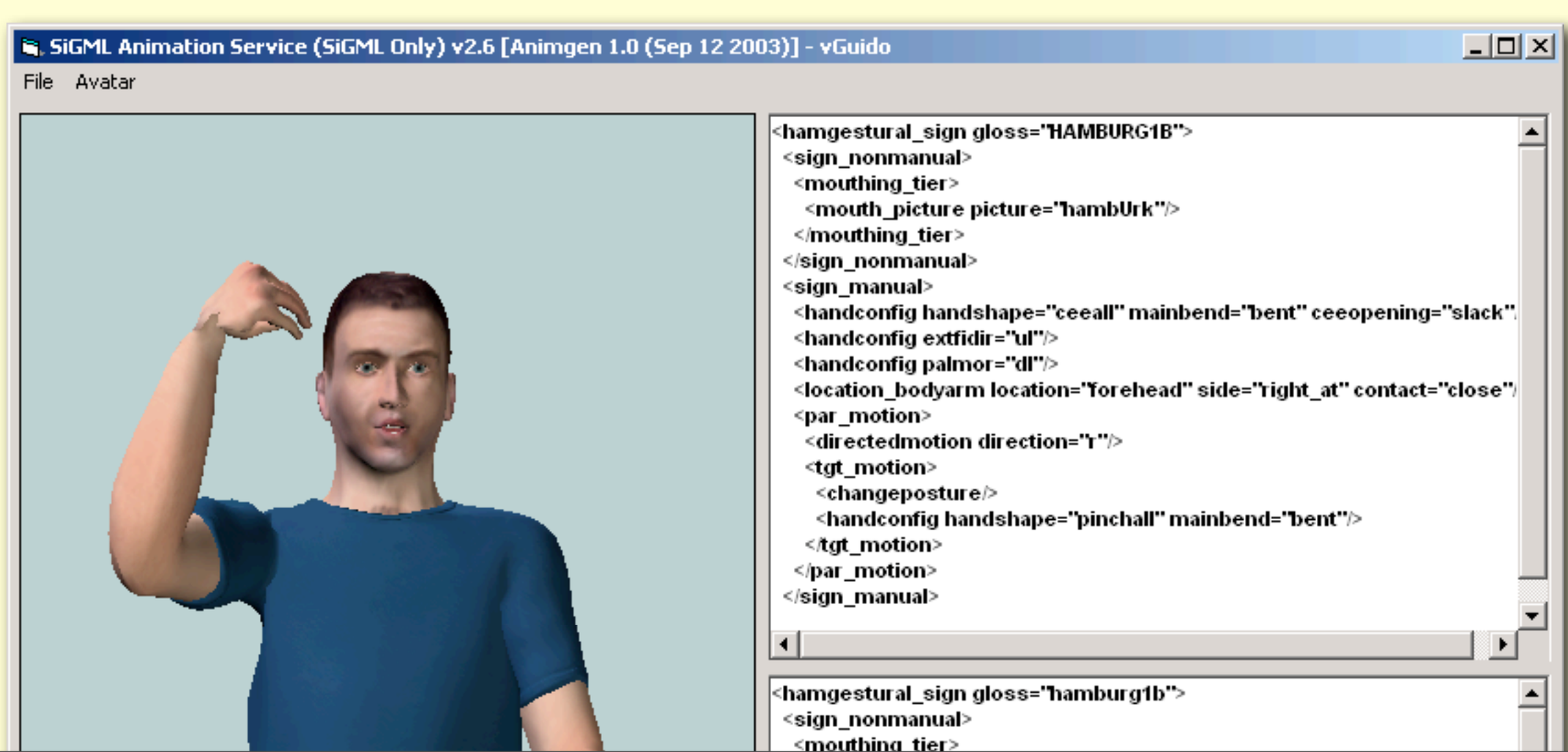
■ **VORKOMMEN**
Konventionell verwendet für „auspressen, Presse“; Modifikation von →AUSPRESSEN1B: Zitrone mit Zitronenpresse (linke Hand) auspressen:
127.1 auspressen
1545.3 Zitronenpresse
1547.1 Zitronenpresse

■ **BESCHREIBUNG**
Frucht (z. B. Zitronenhälfte) festhalten und auf einer Zitronenpresse (linke Hand) auspressen.

■ **FORMÄHNLICH**
PFEFFERMÜHLE1A: ähnliche Bilder

Avatar & Prescription= eSIGN

- Production of signed texts
- Evaluation of transcriptions by re-producing the signing

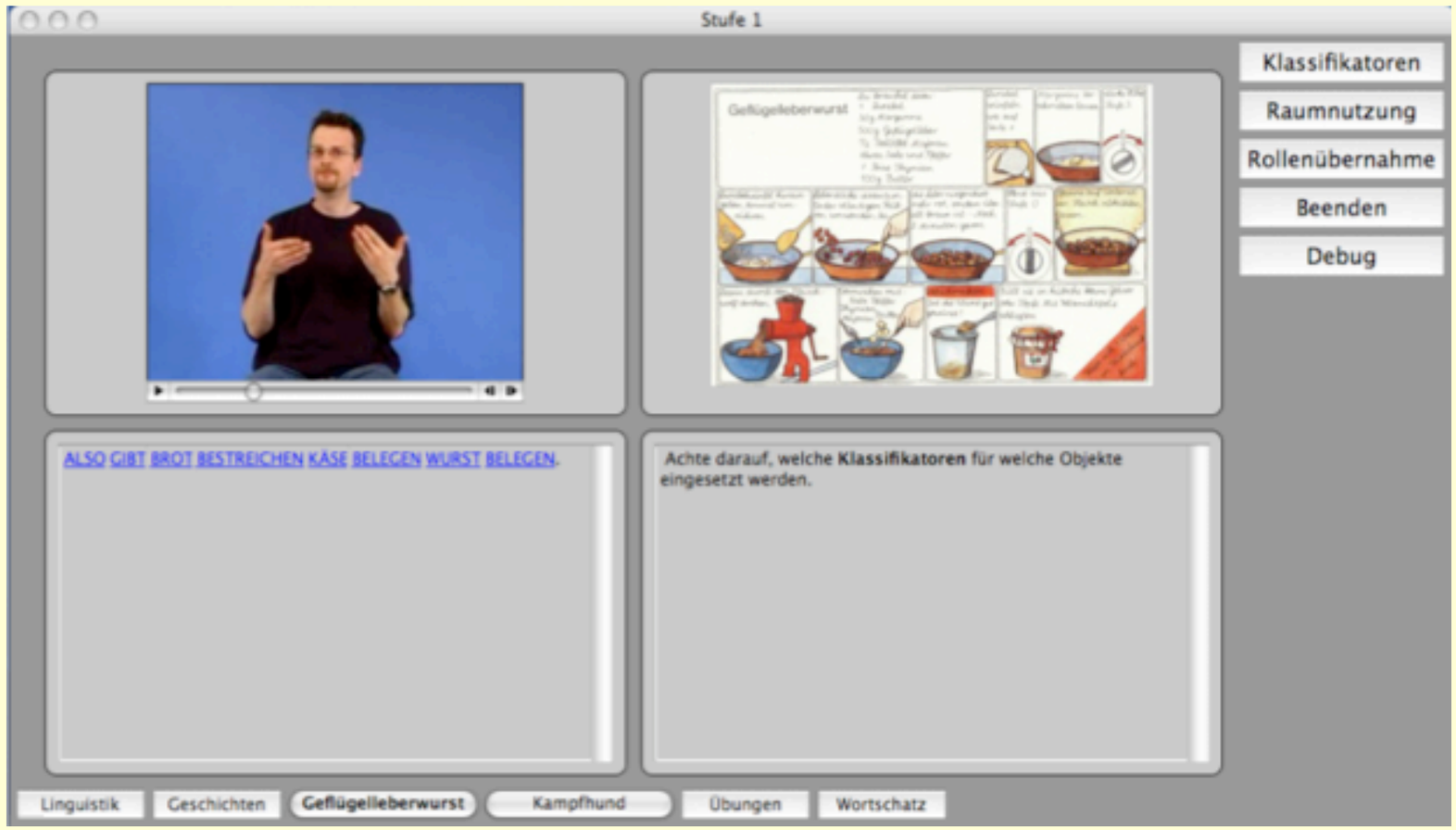


Teaching

- Find examples to illustrate what you talk about: Search iLex and have instant access to the data, copy the example into your presentation (if permission available)
- Prepare contact sheets to be handed out to students
- Create (quick & dirty) eLearning materials

eLearning

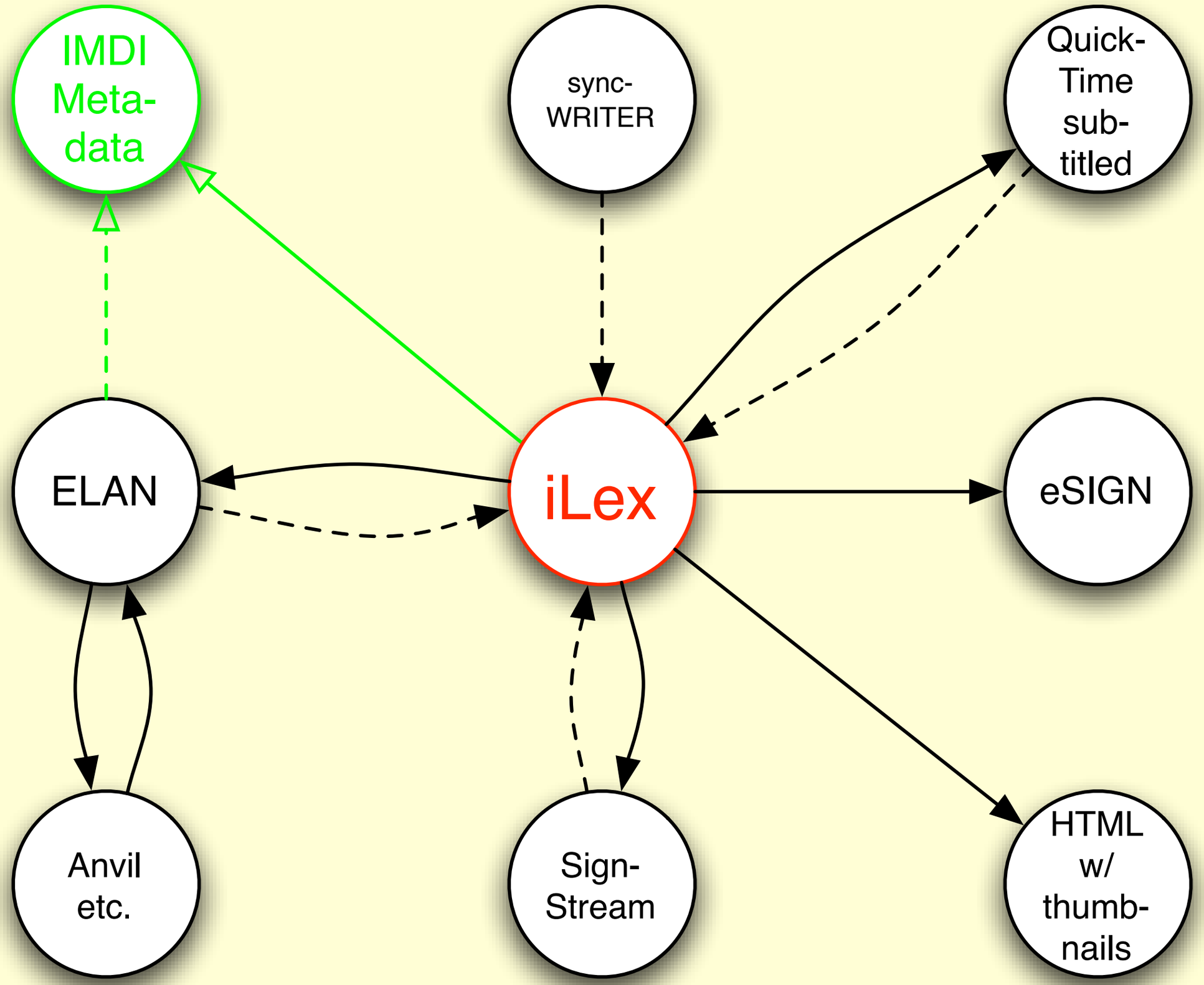
- Idea: Transcripts are labour-intensive, so make double use of them:
  - CourseBuilder



iLex & the Rest of the World

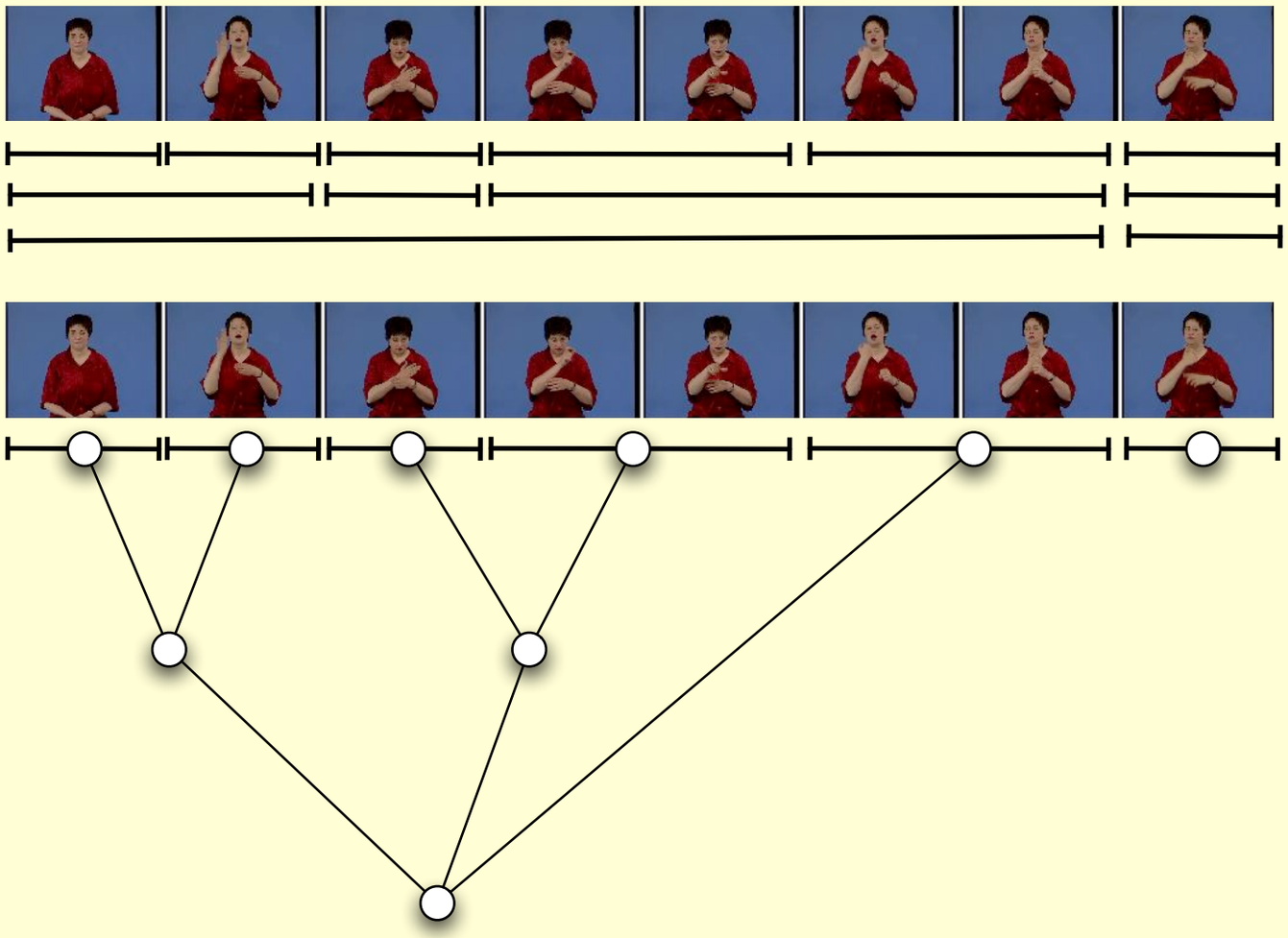
- SignStream
- ELAN
  - Anvil
  - TASX
  - …
- AnColin

Data exchange is possible…

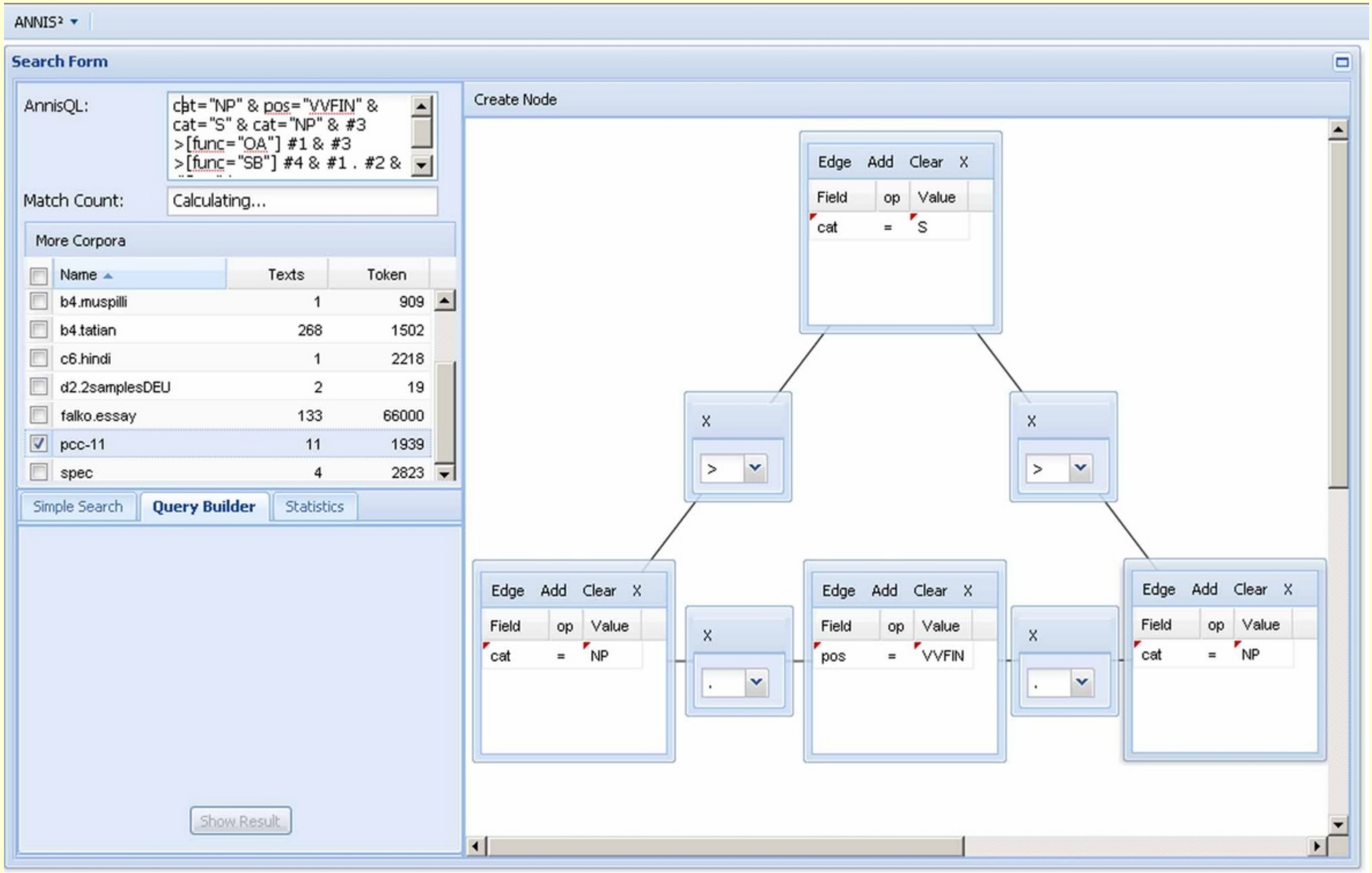


Further Developments?

- Treebanks

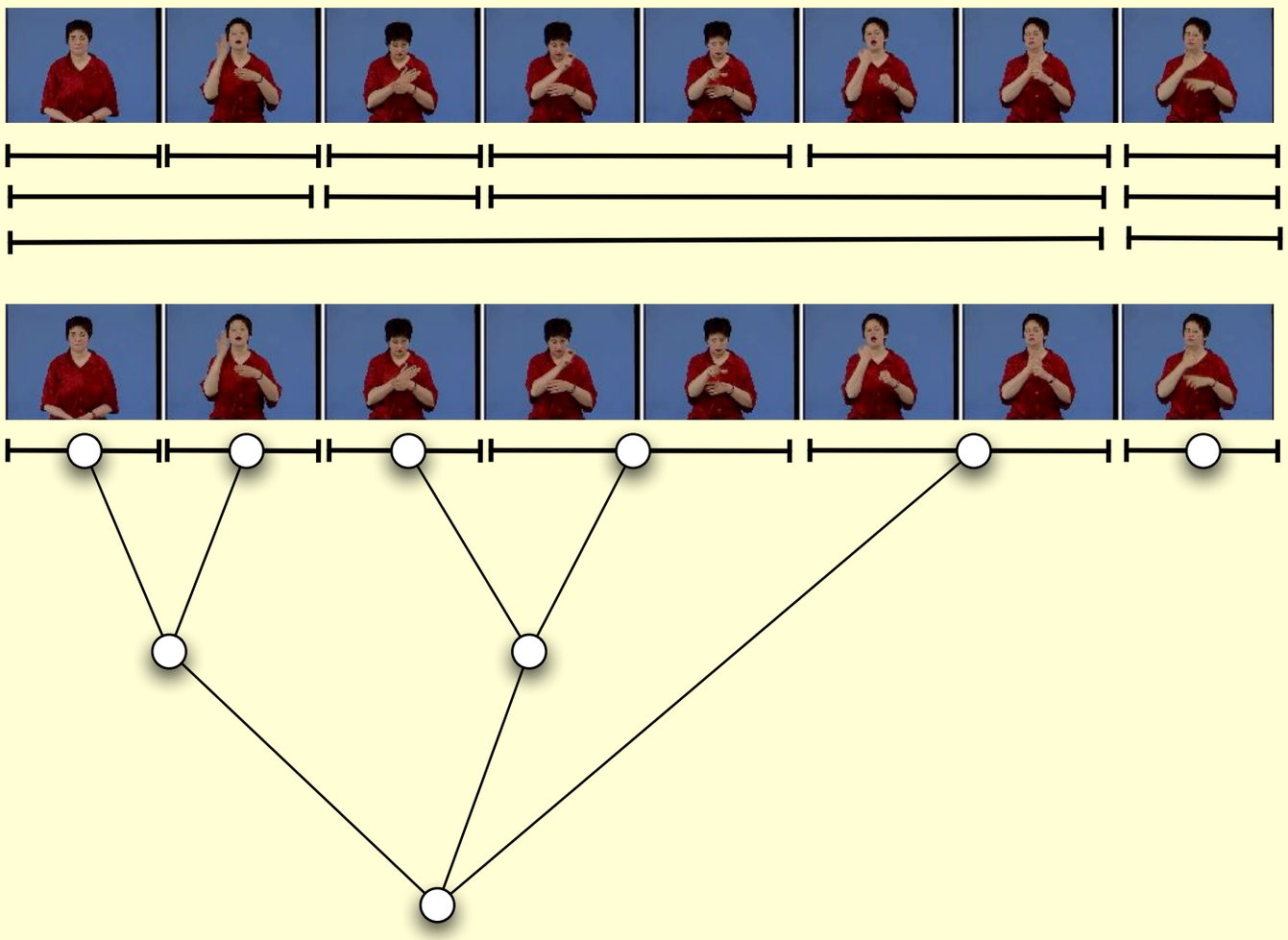


Further Developments?



Further Developments?

- Treebanks



- Frame Semantics (Fillmore's FrameNet)

Statistics

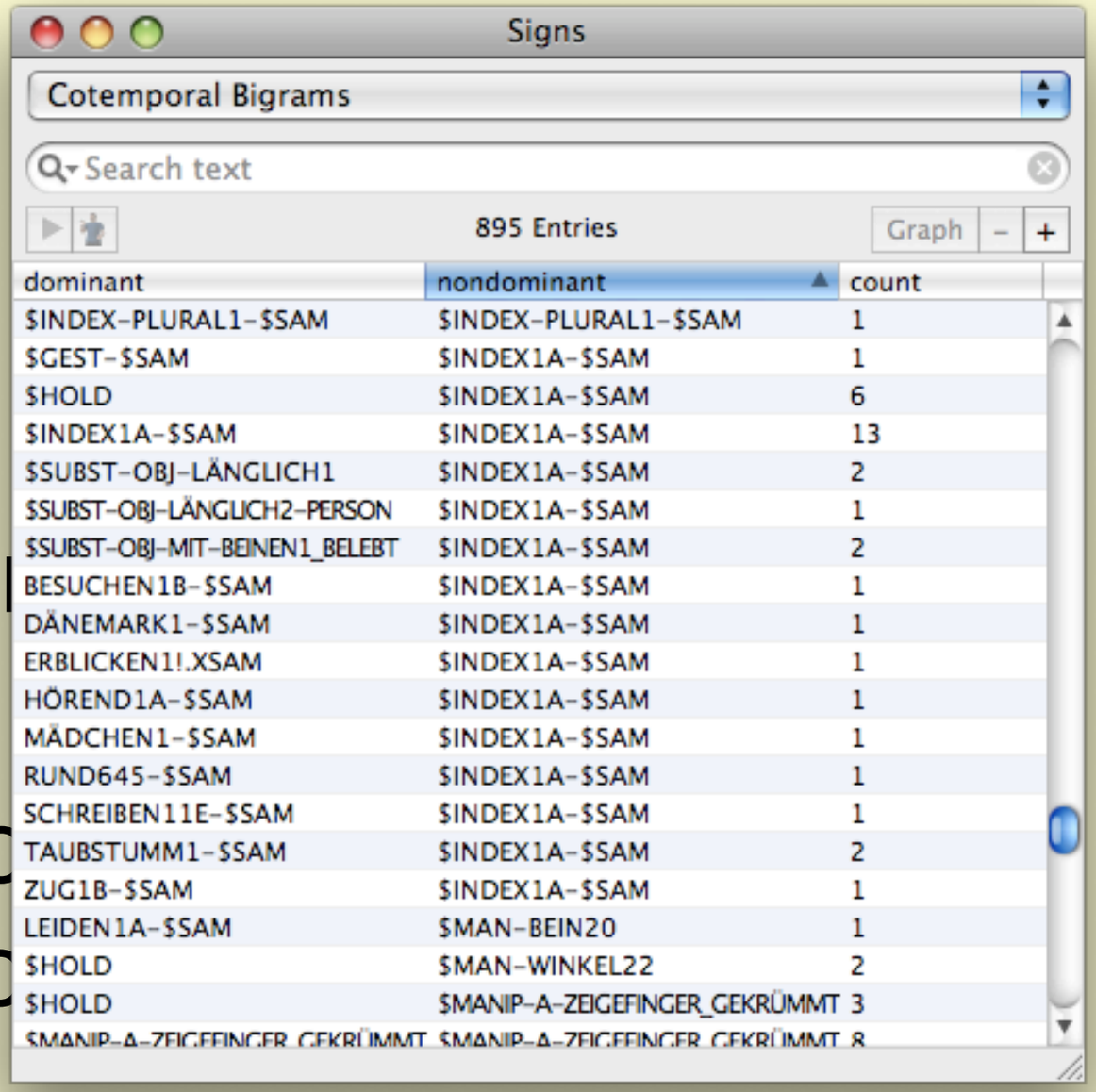
- For the time being, we are quite happy if there is anything worth counting
- How large and balanced should a corpus be before statistics really make sense?
  - Sheer size never an option for sign language

Counting

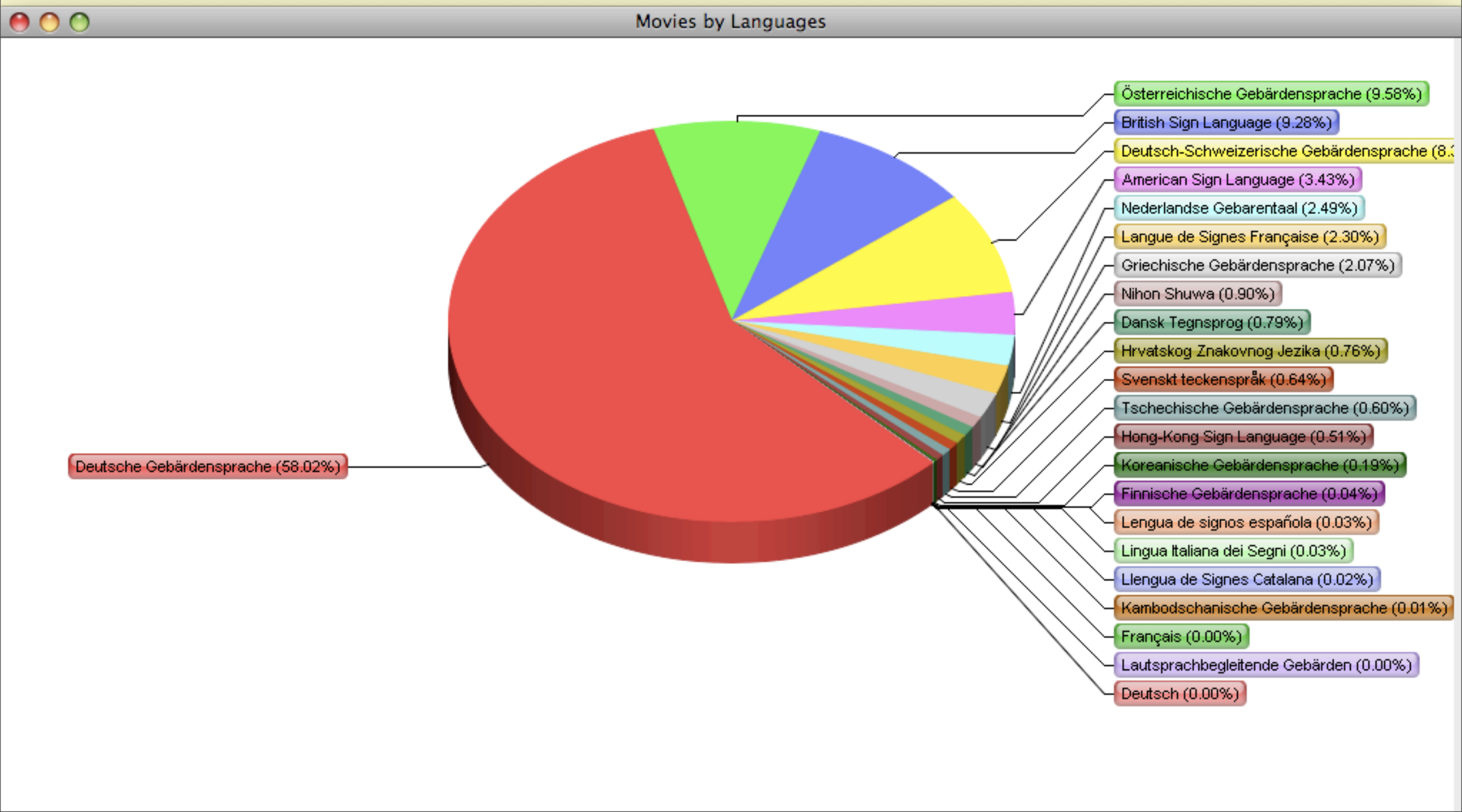
- Counting is easy in a relational database
- Views customisable to specific subsets of your database or to show linked data you are interested in
- Bigram analyses adapted to sign language
  - cotemporal bigrams
  - sequential bigrams

Counting

- Counting is easy in a r
  database
- Views customisable to
  of your database or to
  you are interested in
- Bigram analyses adapted to sign language
  - cotemporal bigrams
  - sequential bigrams



Counts as Graphs



Data Quality

- The deeper the analysis, the higher the costs for inter-transcriber measures
- Consistency
  - Relational database guarantees referential integrity
  - No typos
- Different perspectives on the data
  - Lemma reviewing for a dictionary compilation process reviews transcription data from a different point of view

Quality

- If we can never reach “big” sign language corpora:
- Other measures about the quality of a corpus needed
  - such as compatibility of results with other methods
  - e.g. cognitive models
- On the lexical level: Which corpus linguistics measures are good predictors for lexical item retrieval time?
  - For spoken language: frequency measures rather uncorrelated

Thank you for your attention!

- The work described here is partially supported by
  - German Academies of Science programme (DGS-Korpus)
  - German Federal Ministry of Labour & State of Hamburg (GaLex)
  - European Commission, 7th framework IST (Dicta-Sign)