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 text annotation to points in time of a video
- 1996 GlossLexer
  - Lexical database
- Project started in 2000
  - Lexical database & flexible transcription
Building on Past Experience

- 1994: syncWRITER for signed discourse transcription
- 1996: GlossLexer
- Lexical database
- Transcription limited to short phrases

Project started in 2000
- Lexical database & flexible transcription
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
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
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
- Type tiers
- Text tiers
- Numerical tiers
- Cross-reference tiers
- Value tiers

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

ARBEITEN1A

Gloss: ARBEITEN1A

HamNoSys: [...]

Mouth: arbeiten

Description:

Parent:

Gloss: MACHEN1A-$SAM

1999-05-03 (Dolly Blanck), 2005-12-05 (Lutz König)
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

<table>
<thead>
<tr>
<th>Timecodes</th>
<th>Kompl...</th>
<th>Gebärde</th>
<th>Syntax...</th>
<th>HamN...</th>
<th>Bedeut...</th>
<th>Mimik</th>
<th>Mund</th>
<th>XRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:18</td>
<td>behinder</td>
<td>BEHINDER NOUN</td>
<td>behinder</td>
<td>behinder</td>
<td>behinder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:02:16</td>
<td>MENSCH2 NOUN</td>
<td></td>
<td>Mensch</td>
<td>+</td>
<td>mensch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:02:16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:02:16</td>
<td>ARBEITEN VERB</td>
<td>Arbeit</td>
<td>Arbeit</td>
<td>-</td>
<td>arbeit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:04:03</td>
<td>TOLL1</td>
<td></td>
<td>toll</td>
<td>AB zusan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:06:15</td>
<td>Integratic</td>
<td>INTEGRAT NOUN</td>
<td>integratic</td>
<td>-</td>
<td>integratic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:07:09</td>
<td>AMT1A NOUN</td>
<td>Amt</td>
<td>Amt</td>
<td>-</td>
<td>bundesre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:09:03</td>
<td>HELFEN1</td>
<td></td>
<td>helfen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:11:19</td>
<td>Bundesre VERBAND</td>
<td>Bund</td>
<td>bundesre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:13:01</td>
<td>REPUBLIK</td>
<td></td>
<td>Republik</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:14:06</td>
<td>ES-GIB1</td>
<td></td>
<td>gibt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:15:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:15:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:15:02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:15:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Segment Name:
- Aufgaben

Theme:
- Integrationsamt

Transcript:
- Integrationsamt 1
Establishing References

<table>
<thead>
<tr>
<th>Timecodes</th>
<th>Kompl...</th>
<th>Gebäude</th>
<th>Syntax...</th>
<th>HamN...</th>
<th>Bedeut...</th>
<th>Mimik</th>
<th>Mund</th>
<th>XRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:01:17</td>
<td>behind</td>
<td>BEHINDER</td>
<td>NOUN</td>
<td></td>
<td>behind</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:01:22</td>
<td>MENSCH</td>
<td>NOUN</td>
<td></td>
<td></td>
<td>mensch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:02:16</td>
<td>ARBEITEN</td>
<td>VERB</td>
<td></td>
<td>Arbeit</td>
<td>arbeit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:03:24</td>
<td>TOLL1</td>
<td></td>
<td></td>
<td></td>
<td>ab Abzusa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:06:15</td>
<td>Integrati</td>
<td>NOUN</td>
<td></td>
<td></td>
<td>integrati</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:07:09</td>
<td>AMT1A</td>
<td>NOUN</td>
<td></td>
<td></td>
<td>bundesre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:09:03</td>
<td>HELFEN1</td>
<td></td>
<td></td>
<td></td>
<td>helfen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:11:19</td>
<td>Bundesre</td>
<td>VERB</td>
<td></td>
<td>Bund</td>
<td>bundesre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:13:01</td>
<td>REPUBLIK</td>
<td></td>
<td></td>
<td></td>
<td>Republik</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:00:14:06</td>
<td>ES-GIBT1</td>
<td></td>
<td></td>
<td></td>
<td>gibt</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Segment Name: Aufgaben

Theme: Integrationsamt

Transcript: Integrationsamt 1
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**

**DEFINITION**


**Empirische Sozialforschung**


Die Soziometrie kann bei der Analyse kleiner Gruppen, deren Unterschiede sich gut kennen, eingesetzt werden (→Gruppenarbeit). Da die Ergebnisse im Wesentlichen durch die Befragung der Gruppenmitglieder gewonnen werden, Aussagen auch von dem Bewusstsein der Gruppenmitglieder über ihre Beziehungen abhängig und von deren Wunschvorstellungen geprägt.
Avatar & Prescription = eSIGN

- Production of signed texts
- Evaluation of transcriptions by reproducing 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…

- IMDI Meta-data
- ELAN
- Anvil etc.
- iLex
- Sign-Stream
- Sync-WRITER
- eSIGN
- Quick-Time subtitled
- HTML w/thumbnails
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 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
Counts as Graphs

Movies by Languages

- Deutsche Gebärdensprache (58.02%)
- Österreichische Gebärdensprache (9.58%)
- British Sign Language (9.28%)
- Deutsch-Schweizerische Gebärdensprache (8.13%)
- American Sign Language (3.43%)
- Nederlandse Gebarentaal (2.49%)
- Langue de Signes Française (2.30%)
- Griechische Gebärdensprache (2.07%)
- Nihon Shuwa (0.90%)
- Dansk Tegnsprog (0.79%)
- Hrvatskog Znakovanog Jezika (0.76%)
- Svenskt teckenspråk (0.64%)
- Tschechische Gebärdensprache (0.60%)
- Hong-Kong Sign Language (0.51%)
- Koreanische Gebärdensprache (0.19%)
- Finnische Gebärdensprache (0.04%)
- Lengua de signos española (0.03%)
- Lingua Italiana dei Segni (0.03%)
- Lienga de Signes Catalana (0.02%)
- Kambodschanische Gebärdensprache (0.01%)
- Français (0.00%)
- Lautsprachbegleitende Gebärd (0.00%)
- Deutsch (0.00%)
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)