

The slide is framed by a black border. At the top right is the logo of the Institut für Deutsche Sprache (IDS) with the text "INSTITUT FÜR DEUTSCHE SPRACHE". Below the logo is a photograph of a stack of blue books. To the right of the stack is a close-up of a book cover with the title "Pragmalinguistik" and "Pragmatik - Linguistik - Laut" visible. A yellow rectangular banner with the text "STANDARDS IN SPOKEN CORPORA" is positioned in the center. Below the banner is a small logo for Leibniz-Gemeinschaft.

Thomas Schmidt, Programmbericht „Mündliche Korpora“

STANDARDS IN SPOKEN CORPORA

Mitglied der

Leibniz-Gemeinschaft

The slide is framed by a black border. At the top right is the logo of the Institut für Deutsche Sprache (IDS) with the text "INSTITUT FÜR DEUTSCHE SPRACHE". Below the logo is a yellow rectangular section containing the word "OUTLINE" in orange. The main content consists of a numbered list of five points:

- (1) Case study: Spoken corpora at the SFB 538
- (2) Interoperability for spoken language corpora
- (3) Standards for spoken language corpora
 - Transcription and Annotation
 - Audio and Video
 - Metadata
- (4) Good practices for spoken language corpora
- (5) Outlook: (More) common ground?

2

INSTITUT FÜR
DEUTSCHE SPRACHE

SFB 538

Research Centre on Multilingualism 1999-2011
 Over 20 projects organised into four groups

- E: Multilingual Acquisition
- K: Multilingual Communication
- H: Historical Multilingualism
- T: Transfer

Empirical approach

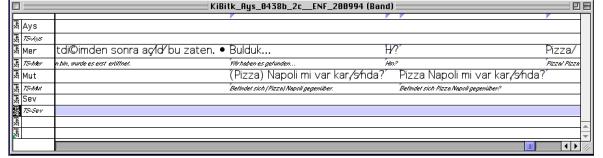
spoken language corpora
 written language corpora (historical and modern)

INSTITUT FÜR
DEUTSCHE SPRACHE

SFB 538

Situation in 2000:

- Many larger corpora already in existence, e.g.:
 - DUFDE (French/German bilingual children)
 - SKOBI (Turkish/German bilingual children)
- Very different technical realisations:
 - dBase/Lapsus
 - 4th dimension/WordBase
 - syncWriter
 - HIAT-DOS

4

SFB 538

Situation in 2000:

- All data dependent on the software they were created with
- No data exchange between software tools
- No data exchange between operating systems
- No common environment for maintaining data
- No possibility of cross-corpus analyses
- No possibility of improving tools
- No digital audio and video

→ Acute danger of „**data death**“



5

SFB 538

Project „Computer assisted methods for the creation and analysis of multilingual data“

Development of corpus technology (EXMARaLDA)

Support for corpus building and analysis

Prepare/Develop a solution for archiving and sharing corpora beyond the Research Centre's lifetime

Corpus curation

INSTITUT FÜR
DEUTSCHE SPRACHE

SFB 538

Situation in 2011:

31 corpora, most of them available for reuse on request, 5 written, 26 spoken
about 6 million transcribed words
about 2000h of digital audio and video recordings
20 languages involved

→ Hamburg Centre for Language Corpora (HZSK)
since January 2011
part of the CLARIN infrastructure

INSTITUT FÜR
DEUTSCHE SPRACHE

hzsk hamburg zentrum
für sprachkorpora



Korpus des Sonderforschungsbereichs 538 Mehrsprachigkeit

Das folgende Tablett gibt einen Überblick über alle Korpora, die zwischen 1999 und 2011 am Sonderforschungsbereich 538 Mehrsprachigkeit erstellt wurden und nur vom Hamburger Zentrum für Sprachkorpora (HZSK) gepflegt werden. Klicken Sie auf [Mehr information](#), um detaillierte Angaben zur Größe eines Korpus, zu den Zugangsbedingungen etc. zu erhalten.

Der Großteil der Korpora sind EXMARaLDA-Korpora. Um mehr über EXMARaLDA-Korpora zu erfahren, empfehlen wir Ihnen den folgenden Schritte:

- lesen Sie das Dokument [How to use an EXMARaLDA corpus](#)
- sehen Sie sich die [EXMARaLDA Demokorpus](#) an, das ohne Passivschutz zugänglich ist
- sehen Sie sich die [Hamburg HLT Test Corpus](#) an, das exemplarisch die Möglichkeiten eines mehrbebenannotierten EXMARaLDA-Korpus veranschaulicht
- besuchen Sie die [EXMARaLDA-Website](#)

Korpsname Körpersprache / Domäneninhalte Sprache	Kurzbeschreibung	Schlüsselwörter	Sprache(n)
Hamburg Adult Bilingual Language (HABA) Mehr information	Audiodaten (schreitende Interviews) mit deutscher/Standard- und Hochdeutschvariante sowie französischer/Standard- und Hochdeutschvariante. Die Daten stammen von 15 bis 52 Jahren. Die Daten simulieren Zweisprachigkeit mit Deutsch und Französisch/Hochdeutsch als L1 werden Sprachversionen für jede Sprache einzeln aufgenommen. Die konsekutiven Zweisprachigkeit der Daten ist nicht untersucht. Die Daten sind als L1 und Hochdeutsch als L2, haben alle eine Erwerbszeit von 11 bis 38 Jahren und werden in ihnen 12 aufgezeichnet.	L2-Daten Erwerbsdaten Zweisprachigkeit bei Erwachsenen Zweisprachigkeit bei Erwachsenen konsekutive Zweisprachigkeit simulierte Zweisprachigkeit	German, Standard (de), French (fr), Italian (ita)
Hamburg Corpus of Polish in Germany (HamCoPolig) Mehr information	Gesprochene Daten (3 Themen) und Erstellung einer Bildergeschichte (auch Vater und Sohn) von bilingualen (Polnisch und Deutsch) und monolingualen (Polnisch) Erwachsenen (20-40 Jahre).	L2-Daten Querschnittsstudien Zweisprachigkeit bei Erwachsenen Zweisprachigkeit bei Erwachsenen konsekutive Zweisprachigkeit simulierte Zweisprachigkeit	Polish (pol)
Hamburg Corpus of Argentinean Spanish (HaCaSpa) Mehr information	Audio- und Videoschnüsse von spontanmenschlichen und labortypologischen Kontextenvariablen. Daten von erwachsenen Sprechern des Porteño-Spanischen (Alter 18-69) aus zwei Querschnittsstudien (gesprochene Frage-Antwort-Pause, Interviewabfragebogen, freie Interviews und Map-Talk); insgesamt 17 Teilnehmerteile. Das in den Experiments als Stimuli verwendete Material besteht aus 1000 Wörtern, die in 100 Kategorien unterteilt sind.	Spanish (spa)	
Dolmetschen im Krankenhaus (DIK) Mehr information	Auslaufnahmen verschiedenster Arten von Arzt-Patienten-Kommunikation im Krankenhaus. Monologische Gespräche auf Deutsch, Portugiesisch und Türkisch im Kontext der medizinischen Versorgung. Die Daten sind in drei Gruppen unterteilt: Türkisch, Deutsch-Portugiesisch und Deutsch-Portugiesisch/Spanisch. In Deutschland leben ca. 100.000 Menschen mit Migrationshintergrund. Die doktorierenden Personen sind zweisprachige Pflegekräfte oder Familienangehörige der Patienten, alle leben in Deutschland, vertagen jedoch über kommunikative Kompetenzen im Deutschen.	Arzt-Patienten-Kommunikation Kommunikation in Institutionen Kommunikation in sozialen Einheiten Laiendoktoren gelärmsoziale Kommunikation	German, Standard (de), Portuguese (prt), Turkish (tur)

**INSTITUT FÜR
DEUTSCHE SPRACHE**

EXMARaLDA

- Data-Centric: data are more valuable than the software (in the long run)
- Abstract data model: Annotation Graphs (Bird/Liberman)

„One fundamental action: to associate a label with a stretch of time in a recording“
- Data formats: Open standards: Unicode and XML
- Tools for working with these formats → Partitur-Editor, Corpus Manager, EXAKT
- Guidelines for working with these formats → HIAT transcription conventions, specific annotation guidelines, ...

9

**INSTITUT FÜR
DEUTSCHE SPRACHE**

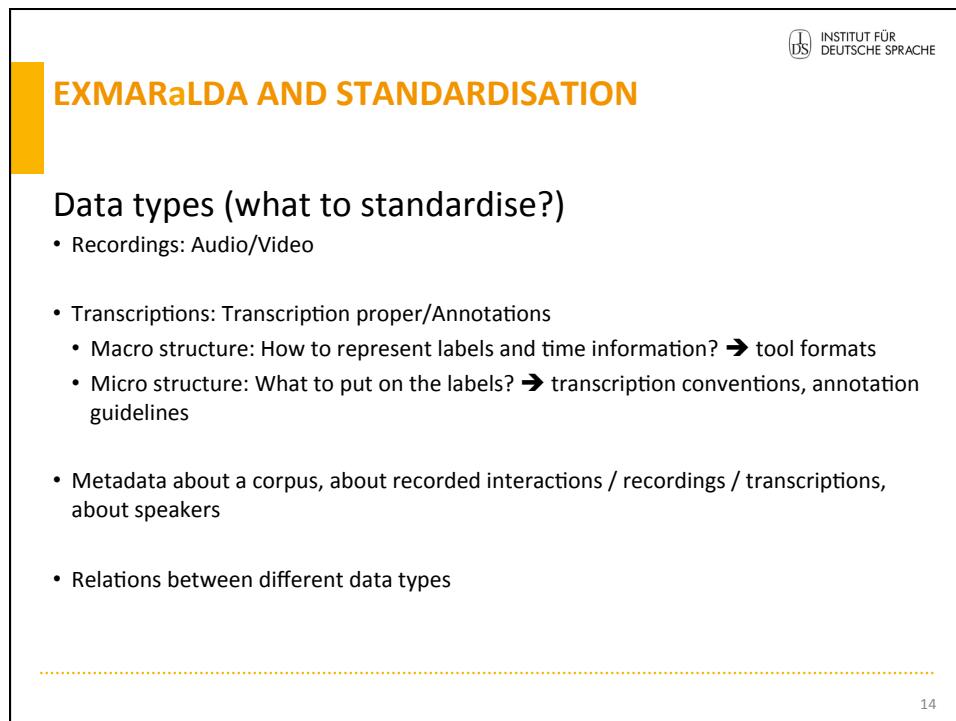
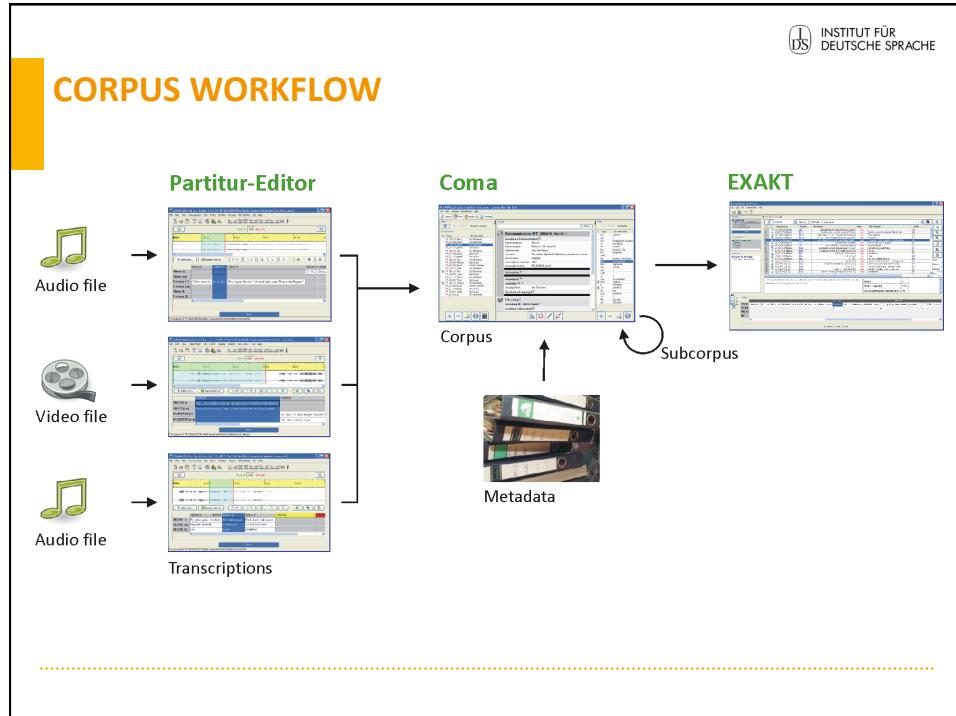
PARTITUR-EDITOR

COMA

The screenshot shows the EXMARaLDA COMA software interface. The main window displays a communication log for 'Communication MT_280410_Hamit'. The log includes fields such as Description (Communication), Aufnahmedatum (280410), Aufnahmegerät (M-Audio - Microtrack II), Aufnehmenende (Kim Chi Hamze), comment (Die beiden Map-Task-Teilnehmer kannten sich vorher.), project-name (Maptask), transcription-convention (HIAT), transcription-name (MT_280410_Hamit), and No Location. On the right side, there is a filter panel for 'Speaker' with 24 entries listed, including Arabic, French, Turkish, Russian, Korean, German, Vietnamese, Afghani, etc. At the bottom, there is a recording section with one recording labeled 'Recording: MT_280410_Hamit'.

EXAKT

The screenshot shows the EXMARaLDA EXAKT 0.8 software interface. The main window displays a search results table for the 'HAPTASK' corpus, specifically for the query 'Vb[Vw](er|le|as|arum))b'. The table includes columns for #, S, Communication, Speaker, Left Context, Match, Right Context, and Age[S]. There are 95 rows of results. Below the table, a 'Partitur' window shows a transcription of a speech act, and a small window shows participant details: Age[S] 31, Mother tongue[S] Arabic, and Are the participants acquainted? [C] Yes.



STANDARDISATION

- First approximation:
 - Data model + XML/Unicode for textual data
 - Industry standards for digital audio/video (WAV, MPEG etc.)
 - Not a standard, but a basis for exchange and sustainability
 - SFB 538 / HZSK → EXMARaLDA
 - MPI Nijmegen / DOBES / TLA → ELAN
 - IDS / AGD /FOLK → FOLKER
 - Transcriber (many speech and spoken language corpora)
 - ANVIL (multimodal corpora)
 - (Praat) / (CHILDES / Talkbank → CHAT)
- Second approximation: tool interoperability

15

MULTIMODAL EXCHANGE FORMAT

International Society for Gesture Studies (ISGS)

2005 Conference in Lyon (‘Interacting Bodies’)

User workshop on ‘Multimodal Annotation Tools’

→ Rohlfing et al. (2005): Comparison of Multimodal Annotation Tools

2007 Conference in Chicago (‘Integrating Gestures’)

Developer workshop on ‘Annotation Interchange among Multimodal Annotation Tools’

Goal: Interoperability between existing tools

2008 LREC Workshop (‘Multimodal Corpora’)

→ Thomas Schmidt, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Magnus Magnusson, Travis Rose, Han Sloetjes (2009). An Exchange Format for Multimodal Annotations. In Jean-Claude Martin P. Paggio Michael Kipp, D. Heylen, eds., *Multimodal Corpora* (pp. 207-221). Springer.

TOOLS (1): ANVIL

The screenshot shows the Anvil 3.6 interface. At the top, there's a video player window showing a man in a suit gesturing with his hands. To the left of the video is a status bar with file information: "Loading video: lv50, 384x288, FrameRate=25", "Video frame rate: 25.0", "Video format: UNEAR, 22050.0 Hz, 8-bit, Mono", "Duration: 03:43:52 (6597 frames)", and "Open first player wrote file lq1-7-reich anvil". Below the video player is a toolbar with icons for file operations like open, save, and zoom. A central panel displays the current specification: "D:\Research\anvil-specification2.xml" and the timestamp "03:08:76 modified frame 4719". To the right of the video player is a panel titled "Track: gesture.phase" with details: "Referenced track: gesture phase", "Time: 03:07:63 - 03:10:27 (66 frames)", and a list of attributes: category: iconic, iconic type: smash, handtype: 2H, cec: 2, function: emblematic, timing: direct. Below this is a "Comment" field with the text "Compare with lq1-8 at 0:20". At the bottom of the interface is a timeline from 03:06 to 03:11, with several annotation tracks visible, including "wave", "prat", "ter1", "ter2", "ling", "posture", "pose", "shift", "phase", and "phrase". The "phrase" track contains labels like "übersetzen", "ich spüre", "den poeltischen Stil", "den ich bei Rivas", "na nur die Bemühun", and "metaphoric, heart, 2H", "iconic smash, 2H", and "emblem, so-what, 2H".

Developer: Michael Kipp, DFKI Saarbrücken

TOOLS (2): C-BAS

The screenshot shows the C-BAS software interface. At the top, there's a video player window showing a man in a dark shirt gesturing with his arms raised. To the right of the video player is a panel titled "Video Information" with details: "Tat: Boundary, John", "Time: 00:07:20", and "Count: 243,205". Below the video player is a list of gesture codes. At the bottom of the interface is a table titled "Event Log" with columns: Start Frame/End Frame, Duration, Event, and End Frame. The table lists numerous events, such as "Frm: 1 Cue Right Hand to Face", "Frm: 1 Cue Left Hand to Face", "Frm: 1 Both Hands to Face", "Frm: 1 Both Hands Together", "Frm: 1 Both Hands Up", "Frm: 1 Both Hands Down", "Frm: 1 Both Hands Side", "Frm: 1 Both Hands Side Up", "Frm: 1 Both Hands Side Down", "Frm: 1 Both Hands In", "Frm: 1 Both Hands Out", "Frm: 1 Both Hands Moving", "Frm: 1 Both Hands Stop", "Frm: 1 Both Hands Resting", "Frm: 1 Both Hands Head Up", and "Frm: 1 Both Hands Head Down". The table continues with many more entries, each with specific start and end frame times.

Developer: Kevin Moffit, University of Arizona

TOOLS (3): ELAN

The screenshot shows the ELAN software interface. At the top, there's a menu bar with File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help. Below the menu is a toolbar with Grid, Text, Subtitles, Controls buttons. A video frame is visible on the left, showing two people in a hut. To the right of the video is a transcript table titled 'Gloss2' with columns for 'Nr.', 'Annotation', 'Begin Time', 'End Time', and 'Duration'. The transcript contains numbered annotations from 1 to 11, such as 'he's speaking into his 1/2 bush tente' and 'you sing those 2 parts of it'. Below the transcript is a timeline with a waveform and numerical markers from 00:01:00.000 to 00:01:30.000. On the far left, there's a vertical list of tiers: Gloss1, Gloss2, Ke, Plus, K, Riky, man outside vie, child 1, and loudspeaker.

Developer: Han Sloetjes, MPI Nijmegen

TOOLS (4): EXMARALDA EDITOR

The screenshot shows the EXMARALDA Partitur-Editor interface. The main window displays a transcription grid with four columns labeled 0|0, 1, 2, 3, and 4. The grid contains several entries with text and color-coded highlights. An audio/video panel is open on the right side, showing a video of two people in a room. Below the video are controls for playback, including Start, Stop, Position, and a timeline from 0.0 to 43.8 seconds. The overall interface includes various toolbars and a menu bar at the top.

TOOLS (5): MACVISSTA

The screenshot shows the MACVISSTA software interface. At the top, there are three video frames labeled "AIFT-Jan7_01-cam1.mp4", "AIFT-Jan7_01-cam5.mp4", and "AIFT-Jan7_01-cam4.mp4". Below the frames is a timeline with time markers from 23191 to 23310. The main area displays several tracks of linguistic annotations. The tracks include:

- C-word-present.owl**: Shows annotations for words like "DAD", "MOM", "BROT", "MILDE", "NIGHT", "ZWEI", "HUND", "SIEZE", "ZWEI", "DREI", "AND", "IN", and "PON".
- E-word-present.owl**: Annotations for words like "DAD", "MOM", "BROT", "MILDE", "NIGHT", "ZWEI", "HUND", "SIEZE", "ZWEI", "DREI", "AND", "IN", and "PON".
- dword-present.owl**: Annotations for words like "DAD", "MOM", "BROT", "MILDE", "NIGHT", "ZWEI", "HUND", "SIEZE", "ZWEI", "DREI", "AND", "IN", and "PON".
- ges_phrase_C**: Annotations for "ZWEI", "HUND", "SIEZE", and "ZWEI".
- ges_phrase_E**: Annotations for "ZWEI".
- ges_phrase_G**: Annotations for "ZWEI", "HUND", "SIEZE", and "ZWEI".
- GAZE C**: Annotations for "at E", "at T", and "at X" with labels "mental space mental space" and "at object".
- GAZE E**: Annotations for "at E", "at T", and "at X" with labels "mental space mental space" and "at object".
- GAZE G**: Annotations for "at E", "at T", and "at X" with labels "mental space mental space" and "at object".

Developer: Travis Rose, Virginia Tech

TOOLS (6): TRANSFORMER

The screenshot shows the Transformer software interface. At the top, there is a toolbar with options like "Save as", "Open in", "Workspace", "lock", "Lock display update", "Position", "Selection", "Start", "End", "Duration", "Select all", "Text", "Format options", "Modify data", and "Bookmarks". The main area has two panes: a video player showing a scene between two people and a transcript editor. The transcript editor pane contains the following text:

```

Start editing | Select all | Redraw TextStyles | Copy to clipboard | X Delete | Save changes to DB | Autoselect columns
# Tier Content
14 Jhn: <hinterher>
15 Jhn: <hinterher>
16 Jhn: <hinterher>
17 Jhn: <hinterher>
18 Sequenz: Konstellation 6
19 Sbr: willkommen,
20 zwanzig Uhr fünfzehn bei beeBEE,
21 die Sitzung.
22 Sequenz: Konstellation 6
23 Sbr: <cfliuert> ich seh nichts<br>
24 <cfliuert> ich seh nichts<br>
25 <cfliuert> wir sehn nichts<br>
26 <cfliuert> interessant<br>
27 <hinterher>

```

Developer: Oliver Ehmer, University of Freiburg

TOOLS (7): THEME

INSTITUT FÜR
DEUTSCHE SPRACHE

(S1) angry_worker.gaze_at.monitor
(S2) angry_worker.isTyping.keyboard
(S3) angry_worker.eTyping.Keyboard
(S4) angry_worker.eTyping.Keyboard
(S5) annoyed_colleague.gaze_at.colleague

Episode 1

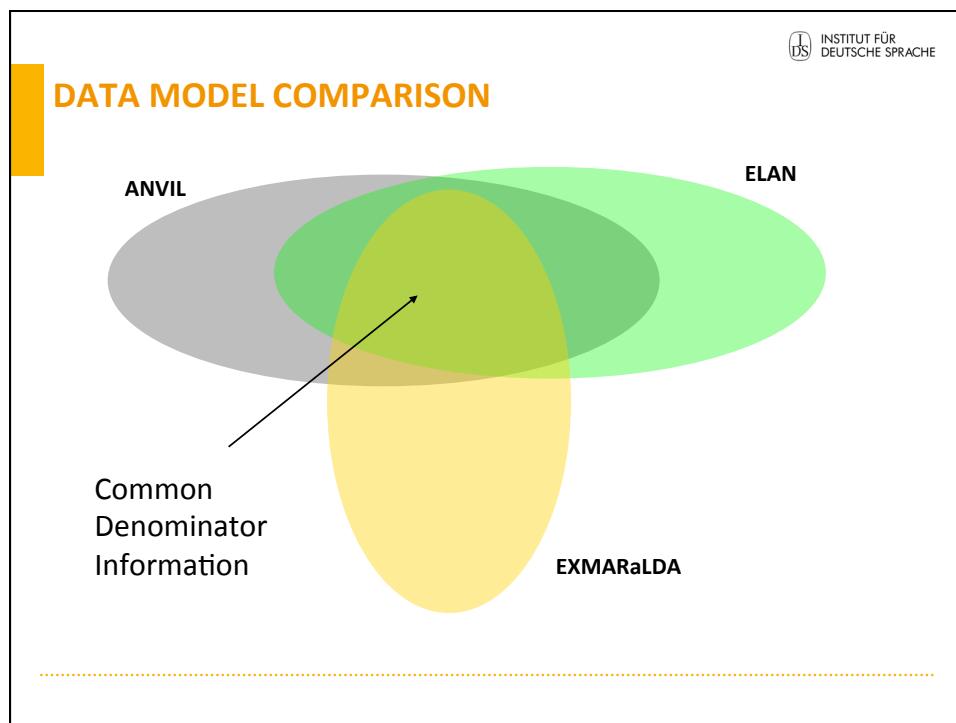
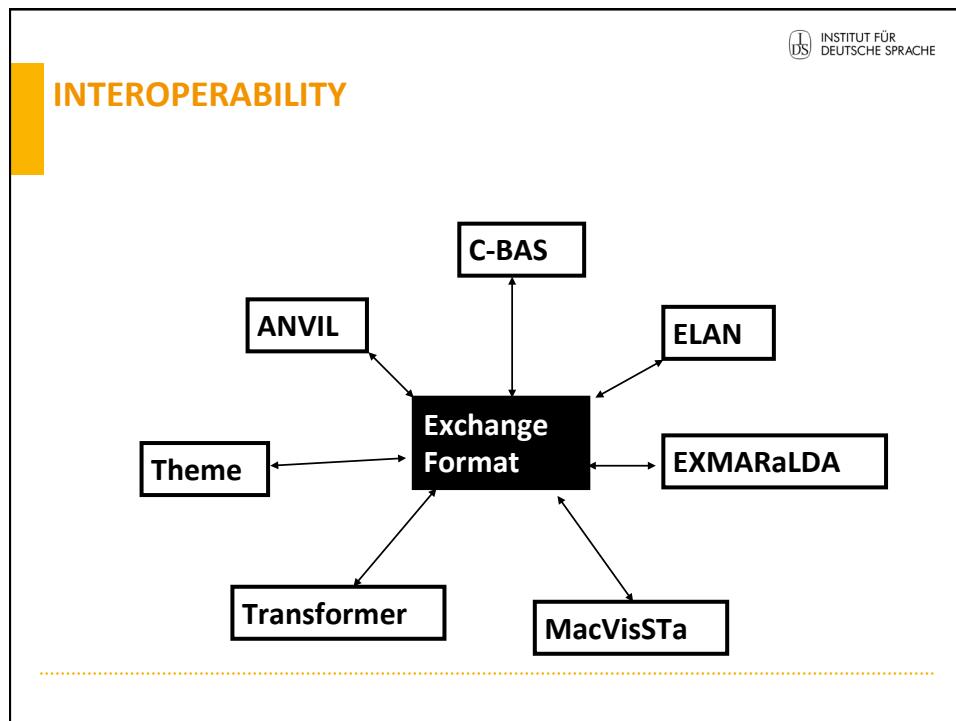
Developer: Magnus Magnusson, NOLDUS

INSTITUT FÜR
DEUTSCHE SPRACHE

INTEROPERABILITY

```

    graph TD
      C_BAS[C-BAS] <--> ANVIL[ANVIL]
      C_BAS <--> ELAN[ELAN]
      C_BAS <--> EXMARaLDA[EXMARaLDA]
      C_BAS <--> Transformer[Transformer]
      C_BAS <--> MacVisSTA[MacVisSTA]
      ANVIL <--> ELAN
      ANVIL <--> EXMARaLDA
      ANVIL <--> Transformer
      ANVIL <--> MacVisSTA
      ELAN --> EXMARaLDA
      EXMARaLDA --> Transformer
      EXMARaLDA --> MacVisSTA
      Transformer --> MacVisSTA
  
```



MULTIMODAL EXCHANGE FORMAT

- Proof of Concept
- Better understanding of differences and commonalities
- Not used in practice, not a Standard
- Why?
 - No added value
 - Macro structure only
 - No reference document
 - No standardising body behind it („grass roots effort“)
 - Not implemented in all tools
 - Maintenance? Distribution?

➔ Third approximation: ISO standard based on TEI

27

TEI/ISO STANDARD FOR SPOKEN LANGUAGE TRANSCRIPTION

- TEI: Text Encoding Initiative
 - Guidelines for electronic text encoding since the 90s
 - Based on XML
 - Widely used by libraries, museums, archives, individual scholars for editions of historical texts, written corpora
 - Little used for spoken language transcription
 - No relation to transcription tools
- ISO: International Standardisation Organisation
 - Technical Committee 37 (TC37, Terminology and Other Language Resources)
- ➔ Define a TEI based standard, compatible with Multimodal Exchange Format, ratified in an ISO process, related to other TC37 standards
- ➔ Current status: Draft International Standard (almost there!)

28

INSTITUT FÜR
DEUTSCHE SPRACHE

TEI/ISO STANDARD FOR SPOKEN LANGUAGE TRANSCRIPTION

```

174      <!-- **** -->
175      <!-- The actual transcription, see 5 and 6 -->
176      <!-- **** -->
177  <body>
178      <!-- annotationGrp grouping u with dependent annotations, see 5.4 -->
179  <annotationGrp xmlns="http://standoff.proposal" who="#SPKO" start="#T0" end="#T9"
180      xml:id="a1">
181          <!-- utterance, see 5.2 -->
182  <u xml:ns="http://www.tei-c.org/ns/1.0" xml:id="u1">
183      <!-- unit above the token and below the u level, see 6.6 -->
184  <seg xml:id="seg0" type="utterance" subtype="declarative">
185      <!-- (word) token, see 6.1 -->
186      <w xml:id="w1">And</w>
187      <w xml:id="w2">what</w>
188      <w xml:id="w3">comes</w>
189      <!-- uncertainty on the transcriber's part, see 6.5 -->
190  <unclear>
191      <choice>
192          <w xml:id="w4">through</w>
193          <w xml:id="w4a">to</w>
194      </choice>
195  </unclear>
196  <w xml:id="w5">is</w>
197  <w xml:id="w6">your</w>
198  <w xml:id="w7">determination</w>
199  <!-- time information within a u element, see 5.2 -->
200  <anchor sync="#T1"/>
201  <w xml:id="w8">at</w>
202  <anchor sync="#T2"/>
203  <w xml:id="w9">all</w>

```

29

INSTITUT FÜR
DEUTSCHE SPRACHE

TEI/ISO STANDARD FOR SPOKEN LANGUAGE TRANSCRIPTION

- More than a proof of concept?
- Added value: Relation to TEI and to other ISO standards, relation to the world of written language corpora
- Macro and micro Structure
- Reference document + document grammars + examples
- Two established organisations behind it
- TEI Drop: tool for converting ELAN/EXMARaLDA/FOLKER/CHAT/Transcriber to TEI
- Basis for future developments?
 - AGD/DGD at IDS Mannheim
 - HZSK at University of Hamburg
 - GeWiss corpus at University of Leipzig
 - ANNIS platform at HU Berlin
 - French partners: ESLO corpus at Orléans, CLAPI database at Lyon
 - Potential partners: Czech National Corpus, Reference Corpus of Contemporary Portuguese, Speech Island Database in Austin/Texas

30

STANDARDS FOR TRANSCRIPTION AND ANNOTATION

- HIAT for use with EXMARaLDA (2004)
- cGAT: GAT for use with FOLKER/EXMARaLDA (under construction)
- FOLK guidelines for orthographic normalisation (under construction)
- STTS tagset for POS-tagging spoken language (Westpfahl/Schmidt 2013, u.c.)
- Project specific guidelines:
 - Disfluency Annotation (Schmidt/Hedeland 2012)
 - Phonetic Annotation (Lleo et al.)
 - Dialect data (SiN project, Schroeder et al.)
- ...

31

Annotations

Transcription	da	gehst	de	jetz	einfach	über	dem	bild
Normalisation	da	gehst	du	jetzt	einfach	über	dem	Bild
Lemmatisation	da	gehen	du	jetzt	einfach	über	d	Bild
POS	ADV	VFIN	PPER	ADV	ADJD	APPR	ART	NN

- Transcription: modified orthography / literary transcription / eye dialect
- Normalisation: standard orthography, semi-automatic(error rate ca. 20%, manual correction)
- Lemmatisation: automatique (TreeTagger, taux d'erreur ca. 2%)
- POS-Tagging: automatique (TreeTagger, taux d'erreur ca. 12%)

ORTHONORMAL

The screenshot shows the OrthoNormal 0.6 software interface. The main window displays a transcription of a speech act. The transcription text is as follows:

```

[25:52.98 25:54.53 LK] wann willen [will denn] die ne [eine] habilitation [Habilitation] schreiben
[25:52.98 25:54.17 WV] ja n [t] des [das] hab [habe] ich nämlich zu ihr
[25:54.53 25:58.19 VW] hab [hab] ich nämlich zu ihr auch gesagt na ja mit familie [Familie] und einem
[25:56.36 25:57.98 AM] deswegen keine kinder [familie]
[25:58.19 25:58.58 LK] warum
[25:58.58 26:00.18] (f. o)
[26:00.18 26:01.50 WV] weswegen was keine
[26:00.18 26:01.85 LK] wie viel [Viele] kinner [K]
[26:01.38 26:01.85 LP] zwei
[26:01.65 26:03.36 AM] keine kinder [Kinder] kr

```

Below the transcription, a red box highlights the word "familie". The software interface also includes a morphological analysis table on the right side.

Wort	Normal	Lemma	POS	p(POS)
anders	anders	anders	ADV	1.0
andersch	anders	anders	ADV	1.0
angeluckt	angeguckt	angucken	VVPP	1.0
angrufen	angrufen	angrufen	VVINF	1.0
angst	Angst	Angst	NN	1.0
anita	Anita	Anita	NE	1.0
ankucken	angucken	angucken	VVINF	0.994905
anrufen	anrufen	anrufen	VVINF	0.999721
ans	an d	an d	APPR ART	0.926526 ...
anstehn	anstehen	anstehen	VVFIN	0.999751
anstellen	anstellen	anstellen	VVIZU	1.0
arbeiten	arbeiten	arbeiten	VVFIN	0.537892
arbeiten	arbeiten	arbeiten	VVINF	0.740092
argument	Argument	Argument	NN	1.0
armen	arm	arm	ADJA	1.0
armer	Armer	Armer	NN	0.966827
au	auch	auch	ADV	1.0
auch	auch	auch	ADV	1.0

Modus: Normalisieren Tagging XML
 Automatisches Weiterlesen

STANDARDS FOR MEDIA DATA

- Audio:
 - Uncompressed PCM, WAV
 - Sampling rate: 48 kHz
- Video:
 - Single images (ideally...) for archiving (MJPEG2000) ➔ disk space!
 - MPEG-4 as „master format“ (encoding and container), 25fps, 4 key frames per second, high resolution
 - transcode for specific usage scenario, e.g. WEBM with lower resolution, fewer key frames, for web delivery

34

INSTITUT FÜR
DEUTSCHE SPRACHE

STANDARDS FOR METADATA

- Catalogue metadata vs. Corpus design metadata vs. Organisational metadata
- Individual solutions (XML with partly controlled vocabulary)
 - IMDI metadata for Language Archive at MPI Nijmegen
 - CoMa data model in EXMARaLDA
 - MeMaSyCo for AGD
 - TEI Metadata header
 - ...
- Comparable structure
- Common framework: CMDI in CLARIN

35

INSTITUT FÜR
DEUTSCHE SPRACHE

METADATA: HAMATAC IN EXMARALDA CORPUS MANAGER

	Communication MT_270110_Shirin
Description (Communication)	
Are the participants acquainted?	No
Communication type	map task
Project name	Maptask
Description (Location)	
City	Hamburg
Country	Germany
PeriodStart	27.01.2010 00:00:00
Description (Location)	
Precision	time not exact
Languages	
Communication (Language)	
LanguageCode	deu
Description (Transcription)	
Alignment status	fully aligned
Annotation type: disfluency	manual annotation of disfluency phenomena
Annotation type: pho	manual annotation of phonetic phenomena
Annotator: c, sup-pos	Fideniz Ercan
Annotator: disfluency, pho	Yael Dilger
Annotator: pos	TreeTagger
Segmentation algorithm	HIAT
Transcriber	Kim Chi Hamze
Transcription checker	Secil Yusun
Transcription convention	orthographic transcription/simplified HIAT
Transcription name	MT_270110_Shirin
Transcription status	fully transcribed

36

METADATA: HAMATAC IN EXMARALDA CORPUS MANAGER

Speaker: Dav (David, Sex: male)

Description (Speaker)
Function subject
2 Locations
Birth (Location)
Country Germany
PeriodStart 01.01.1980 00:00:00
Description (Location)
Precision month and day not exact
Residence (Location)
Country Germany
PeriodStart 01.01.1980 00:00:00
Description (Location)
Precision month and day not exact

4 Languages

L1 (Language)
LanguageCode deu
Description (Language)
Age of acquisition 0
Area of acquisition Hamburg
Usage exclusively
L2 (Language)
LanguageCode eng
Description (Language)
Age of acquisition 6
Usage rarely
L1 (Language)
LanguageCode fra
Description (Language)
Usage rarely
L2 (Language)
LanguageCode spa
Description (Language)
Age of acquisition 18
Usage rarely

37

METADATA: GENERAL STRUCTURE

```

classDiagram
    class Corpus {
        Recording
        Transcription
        Attached File
    }
    class Recording
    class Transcription
    class Attached File
    class Communication
    class Speaker

    Corpus "1" -- "n" Recording : 
    Corpus "1" -- "n" Transcription : 
    Corpus "1" -- "n" Attached File : 
    Recording "n" -- "1" Communication : 
    Transcription "n" -- "1" Communication : 
    Attached File "n" -- "1" Communication : 
    Communication "n" -- "m" Speaker :
  
```

- Communication = Speech Event = Session
- Speaker = Person = Actor
- What is hidden in „Attached file“?

38

INSTITUT FÜR
DEUTSCHE SPRACHE

CMDI VERSION

```

922 ▾      <HZSKSpeaker ComponentId="clarin.eu:cri:c_1345180279130">
923          <Sngle>Dim</Sngle>
924          <Function>subject</Function>
925          <Sex>male</Sex>
926          <BirthDate>1985</BirthDate>
927          <BirthCountry/>
928 ▾          <ActorLanguages>
929          <ActorLanguage>
930              <MotherTongue>true</MotherTongue>
931          <Language>
932              <LanguageName>Russian</LanguageName>
933 ▾              <ISO639 ComponentId="clarin.eu:cri:c_1271859438110">
934                  <iso-639-3-code>rus</iso-639-3-code>
935                  </ISO639>
936          </Language>
937      </ActorLanguage>

```

39

INSTITUT FÜR
DEUTSCHE SPRACHE

GOOD PRACTICES

- Researchers and funding agencies must know about standards, about recommended technologies, about solutions to be avoided
- Practices in corpus construction, exploitation, dissemination must be documented and discussed

DFG Roundtables on handling language corpora with data experts and leading researchers ➔ Recommendations for researchers submitting a proposal / evaluating a proposal

- Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora
- Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora

Published via DFG website, English translations under way, to be published via CLARIN

40

IDS INSTITUT FÜR
DEUTSCHE SPRACHE

GOOD PRACTICES

- Ruhi, S. / Haugh, M. / Schmidt, T. & K. Wörner (eds.) (2014):
Best Practices for Spoken Language Corpora in Linguistic Research.
Newcastle: Cambridge Scholars Publishing.
- Kirk, John, & Andersen, Gisle (eds.) (2015):
Compilation and Annotation of Spoken Corpora: Towards Best Practice.
(Special issue of the International Journal of Corpus Linguistics)
- Description of corpora and corpus compilation workflows
- Methodological issues in corpus design and use
- Technological issues (tools and formats)
- Organisational issues (centres, archives, infrastructures)

41

IDS INSTITUT FÜR
DEUTSCHE SPRACHE

GOOD PRACTICE IN 20 SECONDS

- **Always obtain informed consent**
- **Collect and check metadata immediately**
- **Use uncompressed audio formats**
- **Use ELAN, EXMARaLDA, FOLKER, Praat or Transcriber for transcription / annotation**
- **Backup and version control for your data**
- **Keep in touch with a centre for publication / long-term archiving**

SUMMARY: STANDARDS IN SPOKEN CORPORA

- Good interoperability between major tools for transcription and annotation
- ISO/TEI as a candidate for a „real“ standard
- Industry standards for audio / video
- Efforts for metadata standardisation under way, still some way to go
- Best practices starting to get documented

43

OUTLOOK: (MORE) COMMON GROUND?

- ✓ Standards as a way of expressing commonalities between different solutions
- ✓ Standards as way of making data reusable / sustainable

Standards as a way of making technology development more efficient / more effective / relevant to a larger group of users?

- Existing tools should be able to read and write standard formats
- New tools might be based on established standards

44