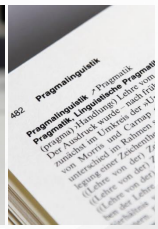
INSTITUT FÜR DEUTSCHE SPRACHE



Thomas Schmidt, Programmbereich „Mündliche Korpora“

STANDARDS IN SPOKEN CORPORA

Mitglied der Leibniz-Gemeinschaft

INSTITUT FÜR DEUTSCHE SPRACHE

OUTLINE

(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?

INSTITUT FÜR DEUTSCHE SPRACHE

SFB 538

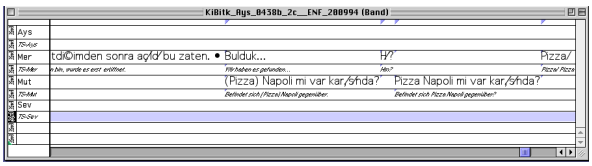
Research Centre on Multilingualism 1999-2011
Over 20 projects organised into four groups
- E: Multilingual Acquisition
- K: Multilingual Communication
- H: Historical Multilingualism
- T: Transfer

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

INSTITUT FÜR DEUTSCHE SPRACHE

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

INSTITUT FÜR DEUTSCHE SPRACHE

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

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

hzsk hamburger zentrum für sprachkorpora

Korpora des Sonderforschungsbereichs 538 Mehrsprachigkeit

Die folgende Tabelle gibt einen Überblick über alle Korpora, die zwischen 1999 und 2011 am Sonderforschungsbereich 538 Mehrsprachigkeit erstellt wurden und nun vom Hamburger Zentrum für Sprachkorpora (HZSK) gepflegt werden. Klicken Sie auf Mehr Information, um detaillierte Angaben zur Größe eines Korpus, zu den Zugangbedingungen etc. zu erhalten.
Der Großteil der Korpora sind EXMARaLDA-Korpora. Um mehr über EXMARaLDA-Korpora zu erfahren, empfehlen wir einen der folgenden Schritte:

- lesen Sie das Dokument How to use an EXMARaLDA corpus
- sehen Sie sich das EXMARaLDA Demokorpus an, das ohne Passwortschutz zugänglich ist
- sehen Sie sich das Hamburg Map Task Corpus an, das exemplarisch die Möglichkeiten eines mehrebenenannotierten EXMARaLDA-Korpus veranschaulicht
- besuchen Sie die EXMARaLDA-Website

Korpusname Projekt / Dateninhaber Ort	Kurzbeschreibung	Schlüsselwörter	Sprache(n)
Hamburg Adult Bilingual Language (HABLA) Mehr Information			
LL1 / Tanja Kupisch spoken/audio/exmaralda	Audioaufnahmen (semispontane Interviews) mit deutsch/italienisch und deutsch/französisch zweisprachigen Sprechern, etwa zwischen 15-55 Jahren. Die simultanen Zweisprachigen mit Deutsch und Französisch/Italienisch als L1 wurden für jede Sprache einmal aufgenommen. Die konsekutiven Zweisprachigen mit Deutsch als L1 und Französisch/Italienisch als L2, oder Französisch/Italienisch als L1 und Deutsch als L2, haben alle einen Erwerbsanfang von 11 bis 38 Jahren und wurden in ihrer L2 aufgenommen.	L2-Daten Querschnittsdaten Sprachenerosion Zweisprachigkeit bei Erwachsenen konsekutive Zweisprachigkeit simultane Zweisprachigkeit	German, Standard (deu), French (fra), Italian (ita),
Hamburg Corpus of Polish in Germany (HamCoPoliG) Mehr Information			
H8 / Bernhard Brehmer spoken/audio/exmaralda	Semispontane Daten (3 Themen) und Erzählung einer Bildergeschichte (aus: 'Vater und Sohn') von bilingualen (Polnisch und Deutsch) und monolingualen (Polnisch) Erwachsenen (16-46 Jahre).	L1-Daten Querschnittsdaten Sprachenerosion Zweisprachigkeit bei Erwachsenen Zweitsprachenerwerb bei Erwachsenen kindlicher Zweitsprachenerwerb konsekutive Zweisprachigkeit simultane Zweisprachigkeit	Polish (pol),
Hamburg Corpus of Argentinean Spanish (HaCASpa) Mehr Information			
H9 / Christoph Gabriel spoken/audio/exmaralda	Audio- und Videoaufnahmen von spontansprachlichen und laborphonologischen Daten von erwachsenen Sprechern des Porteño-Spanischen (Alter 18-69) aus zwei geographischen Gebieten Argentiniens (mittels u.a. Lesetext, Nacherzählung, gelesene Frage-Antwort-Paare, Intonationsfragebogen, freie Interviews und MapTask); insgesamt 7 Teilexperimente. Das in den Experimenten als Stimuli verwendete Material ist in den Transkriptionen als Annotation festgehalten.	Kontaktvarietät Querschnittsdaten Sprachkontakt regionale Varietät	Spanish (spa),
Dolmetschen im Krankenhaus (DiK) Mehr Information			
K2 / Kristin Bührig, Bernd Meyer spoken/audio/exmaralda	Audioaufnahmen verschiedener Arten von Arzt-Patienten-Kommunikation im Krankenhaus. Monolinguale Gespräche auf Deutsch, Portugiesisch und Türkisch im jeweiligen Land aufgenommen, sowie gedolmetschte Gespräche (Deutsch-Türkisch, Deutsch-Portugiesisch und Deutsch-Portugiesisch/Spanisch) in Deutschland aufgenommen, etwa 15-20 Aufnahmen pro Kategorie. Die dolmetschenden Personen sind zweisprachige Pflegekräfte oder Familienangehörige der Patienten, alle leben in Deutschland, verfügen jedoch über unterschiedliche Kompetenzen im Deutschen.	Arzt-Patienten-Kommunikation Kommunikation in Institutionen Konsekutivdolmetschen Laiendolmetschen gedolmetschte Kommunikation	German, Standard (deu), Portuguese (por), Turkish (tur),

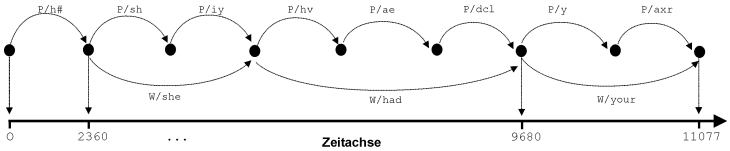
8

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“

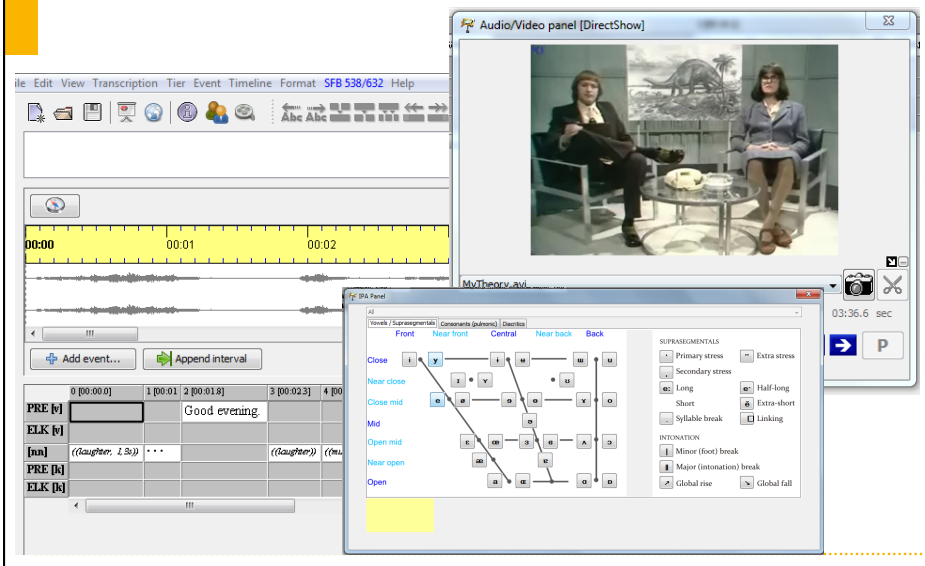


- Data formats: Open standards: Unicode and XML
- Tools for working with these formats ➔ Partitur-Editor, Corpus Manager, EXAKT
- Guidelines for working with these formats ➔ HIAT transcription conventions, specific annotation guidelines, …

9

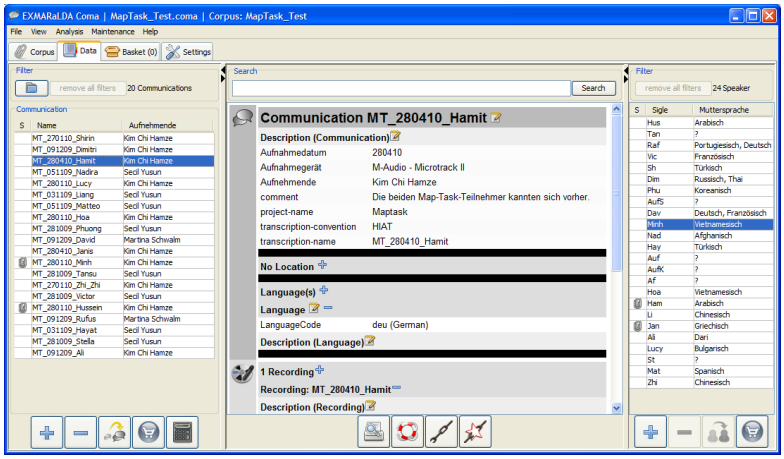
INSTITUT FÜR DEUTSCHE SPRACHE

PARTITUR-EDITOR



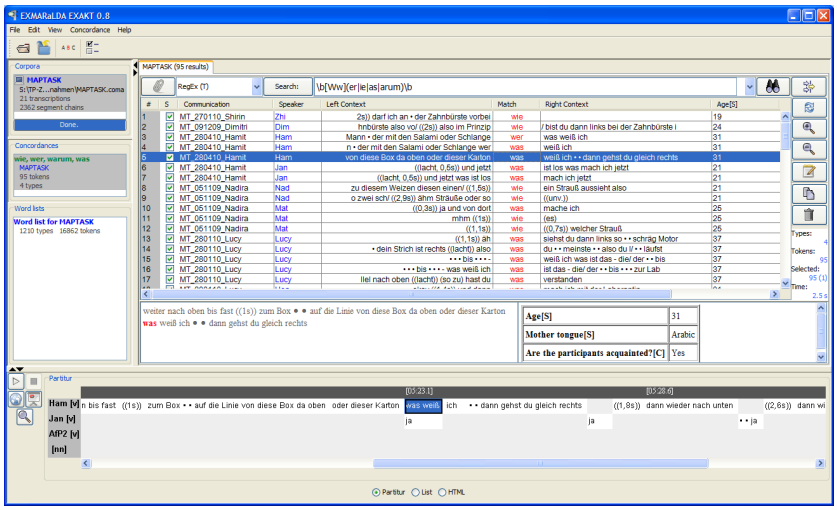
INSTITUT FÜR DEUTSCHE SPRACHE

COMA



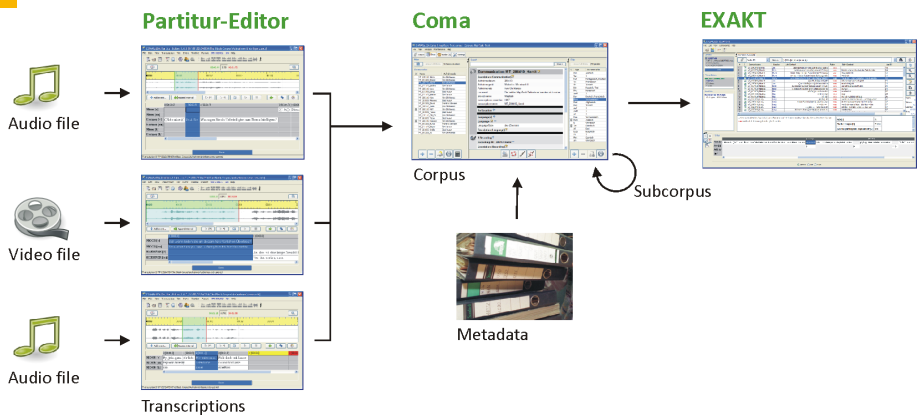
INSTITUT FÜR DEUTSCHE SPRACHE

EXAKT



INSTITUT FÜR DEUTSCHE SPRACHE

CORPUS WORKFLOW



INSTITUT FÜR DEUTSCHE SPRACHE

EXMARaLDA AND STANDARDISATION

Data types (what to standardise?)

- Recordings: Audio/Video
- Transcriptions: Transcription proper/Annotations
  - Macro structure: How to represent labels and time information? ➔ tool formats
  - Micro structure: What to put on the labels? ➔ transcription conventions, annotation guidelines
- Metadata about a corpus, about recorded interactions / recordings / transcriptions, about speakers
- Relations between different data types

INSTITUT FÜR DEUTSCHE SPRACHE

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

INSTITUT FÜR DEUTSCHE SPRACHE

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'

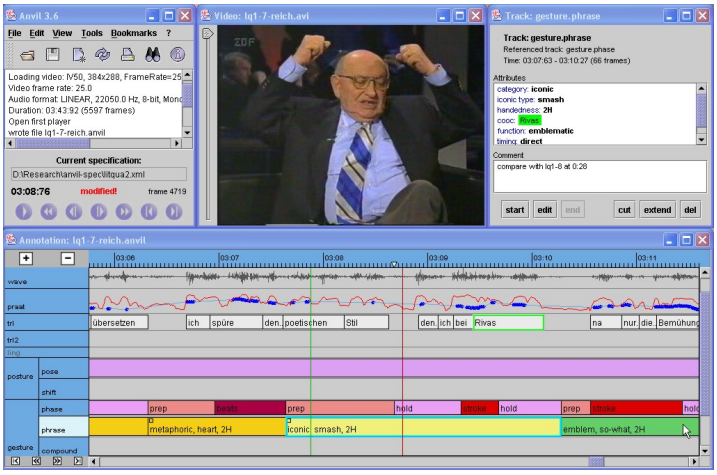
Goal: Interoperability between existing tools

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.

INSTITUT FÜR DEUTSCHE SPRACHE

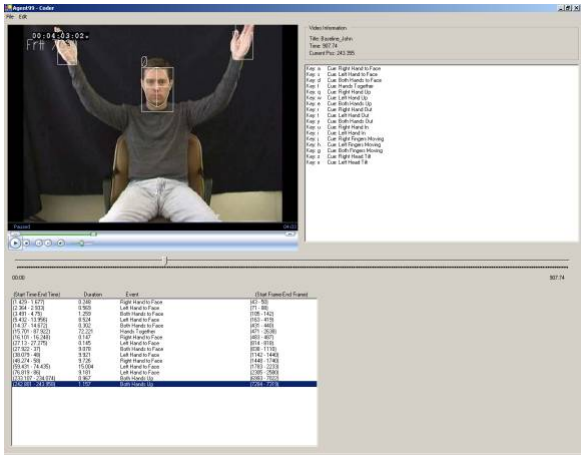
TOOLS (1): ANVIL



Developer: Michael Kipp, DFKI Saarbrücken

INSTITUT FÜR DEUTSCHE SPRACHE

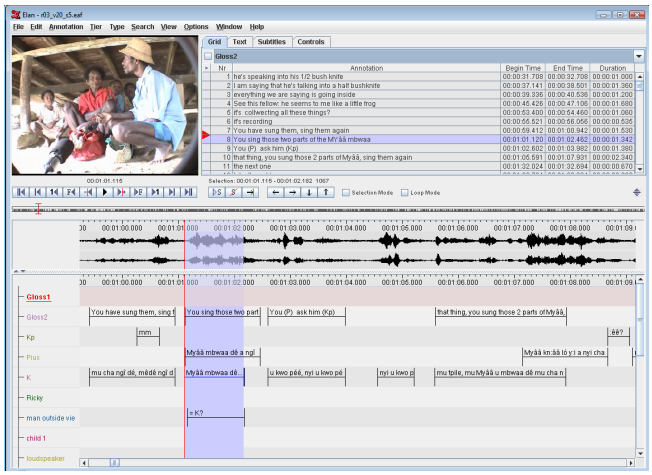
TOOLS (2): C-BAS



Developer: Kevin Moffit, University of Arizona

INSTITUT FÜR DEUTSCHE SPRACHE

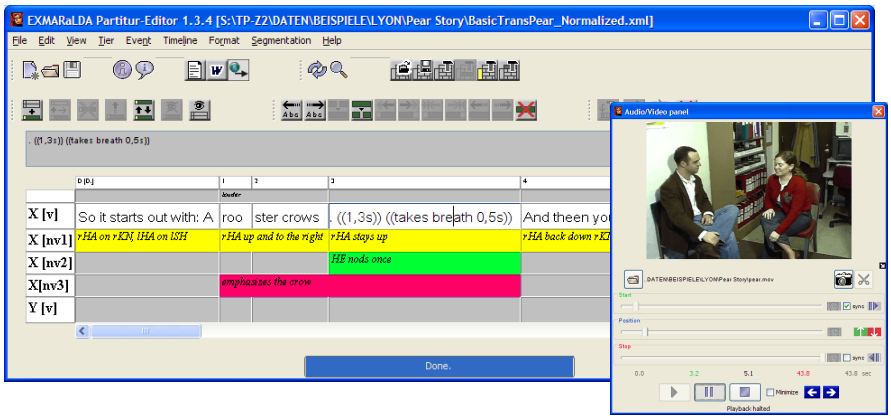
TOOLS (3): ELAN



Developer: Han Sloetjes, MPI Nijmegen

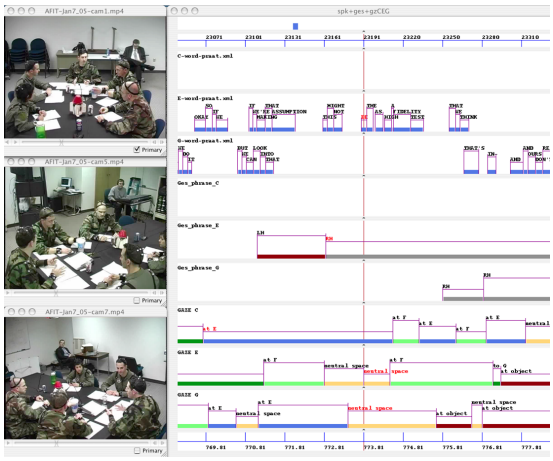
INSTITUT FÜR DEUTSCHE SPRACHE

TOOLS (4): EXMARALDA EDITOR



INSTITUT FÜR DEUTSCHE SPRACHE

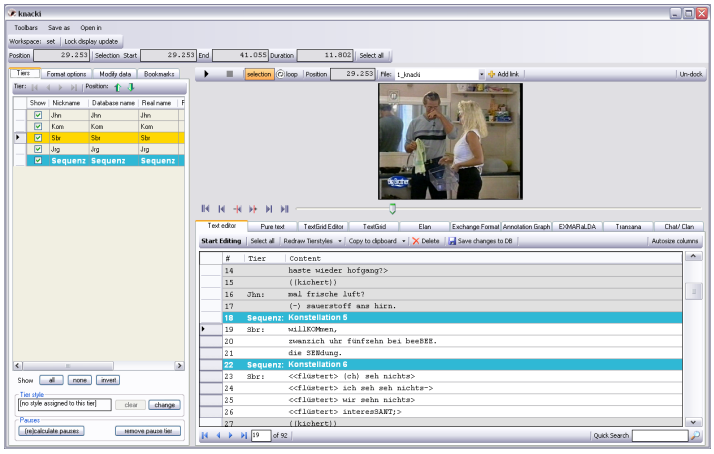
TOOLS (5): MACVISSTA



Developer: Travis Rose, Virginia Tech

INSTITUT FÜR DEUTSCHE SPRACHE

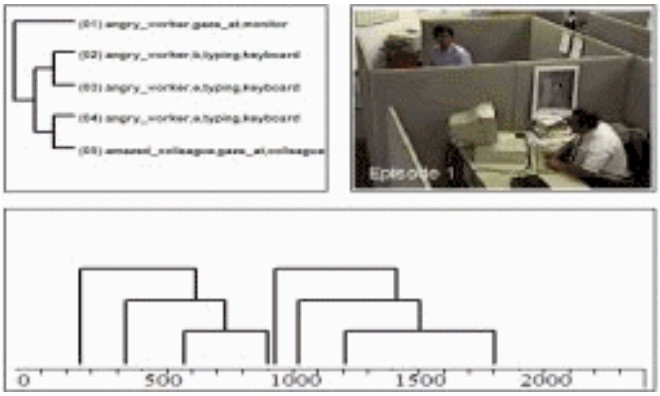
TOOLS (6): TRANSFORMER



Developer: Oliver Ehmer, University of Freiburg

INSTITUT FÜR DEUTSCHE SPRACHE

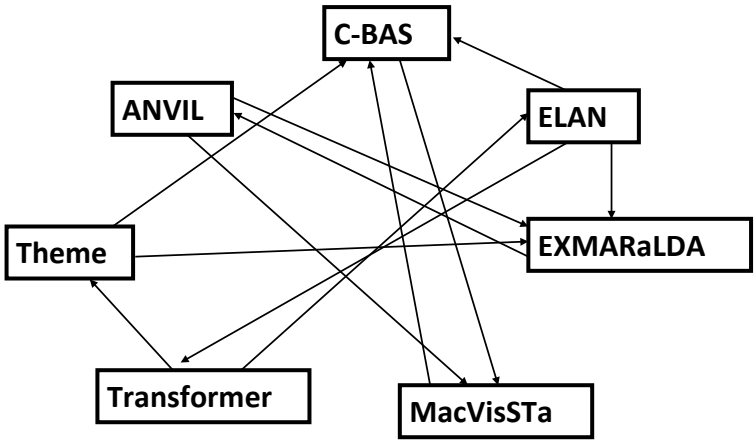
TOOLS (7): THEME



Developer: Magnus Magnusson, NOLDUS

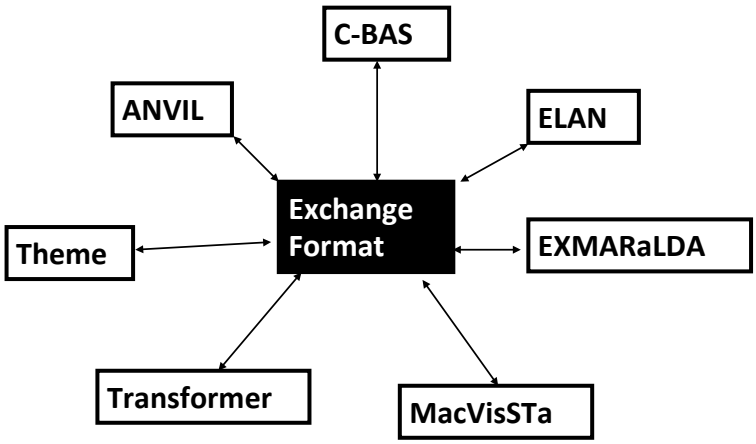
INSTITUT FÜR DEUTSCHE SPRACHE

INTEROPERABILITY



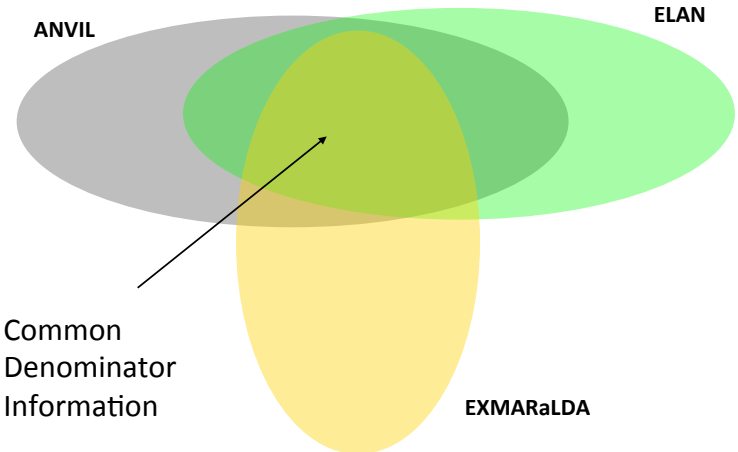
INSTITUT FÜR DEUTSCHE SPRACHE

INTEROPERABILITY



INSTITUT FÜR DEUTSCHE SPRACHE

DATA MODEL COMPARISON



INSTITUT FÜR DEUTSCHE SPRACHE

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

INSTITUT FÜR DEUTSCHE SPRACHE

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

```
<!-- ********************************** -->
<!-- The actual transcription, see 5 and 6 -->
<!-- ********************************** -->
<body>
   <!-- annotationGrp grouping u with dependent annotations, see 5.4 -->
   <annotationGrp xmlns="http://standoff.proposal" who="#SPK0" start="#T0" end="#T9"
      xml:id="ag1">
      <!-- utterance, see 5.2 -->
      <u xmlns="http://www.tei-c.org/ns/1.0" xml:id="u1">
         <!-- unit above the token and below the u level, see 6.6 -->
         <seg xml:id="seg0" type="utterance" subtype="declarative">
            <!-- (word) token, see 6.1 -->
            <w xml:id="w1">And</w>
            <w xml:id="w2">what</w>
            <w xml:id="w3">comes</w>
            <!-- uncertainty on the transcriber's part, see 6.5 -->
            <unclear>
               <choice>
                  <w xml:id="w4">through</w>
                  <w xml:id="w4a">to</w>
               </choice>
            </unclear>
            <w xml:id="w5">is</w>
            <w xml:id="w6">your</w>
            <w xml:id="w7">determination</w>
            <!-- time information within a u element, see 5.2 -->
            <anchor synch="#T1"/>
            <w xml:id="w8">at</w>
            <anchor synch="#T2"/>
            <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

INSTITUT FÜR DEUTSCHE SPRACHE

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

INSTITUT FÜR DEUTSCHE SPRACHE

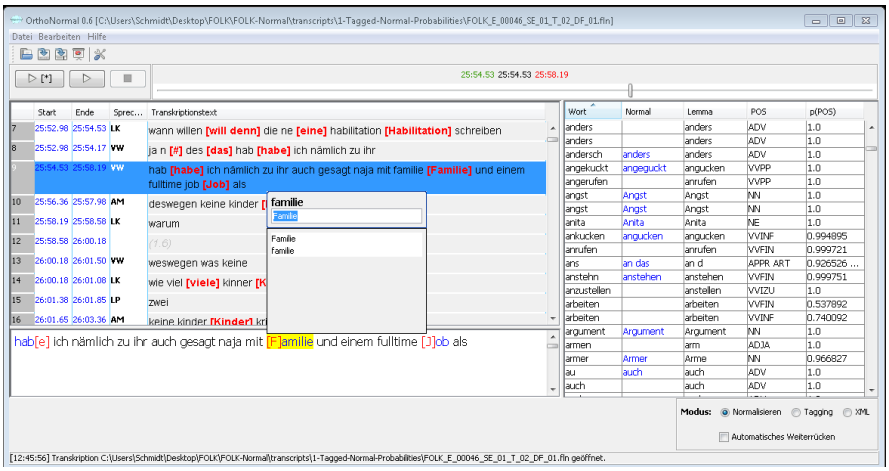
Annotations

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%)

INSTITUT FÜR DEUTSCHE SPRACHE

ORTHONORMAL



INSTITUT FÜR DEUTSCHE SPRACHE

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

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

Communication (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

INSTITUT FÜR DEUTSCHE SPRACHE

METADATA: HAMATAC IN EXMARALDA CORPUS MANAGER

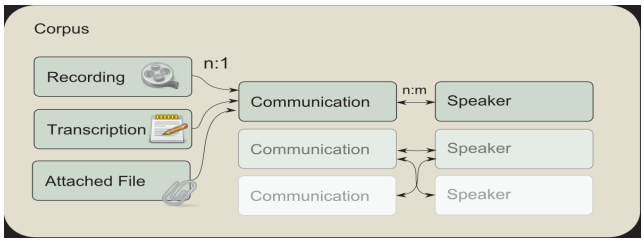
Speaker: Dav (David, Sex: male)	
Description (Speaker)	
Function	subject
2 Locations	
Birth (Location)	
Country	Germany
PeriodStart	01.01.1980 00:00:00
Description (Location)	
Precision	month and day not exact
Residence (Location)	
Country	Germany
PeriodStart	01.01.1980 00:00:00
Description (Location)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

INSTITUT FÜR DEUTSCHE SPRACHE

METADATA: GENERAL STRUCTURE



- 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:cr1:c_1345180279130">
923     <Sigle>Dim</Sigle>
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:cr1: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

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

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

INSTITUT FÜR DEUTSCHE SPRACHE

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

INSTITUT FÜR DEUTSCHE SPRACHE

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