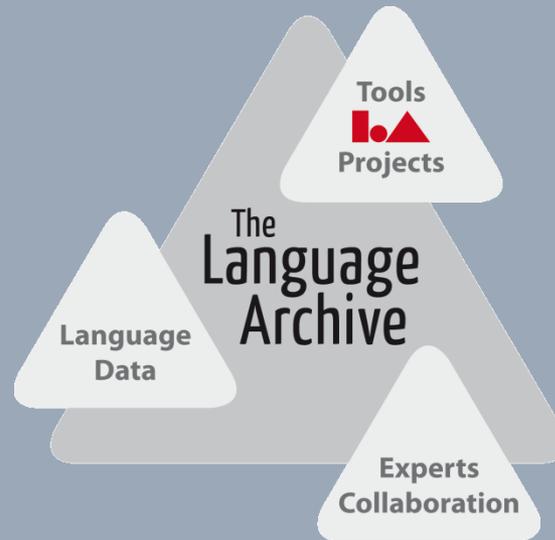


LREC
Workshop 18

2012-05-21
Istanbul



MAX-PLANCK-GESellschaft

Best practices in the design, creation and dissemination of speech corpora at The Language Archive

Sebastian Drude, Daan Broeder, Peter Wittenburg, Han Sloetjes

The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands



TOPICS

- 1. Introduction**
- 2. Corpus Design and Curation**
- 3. Annotation (including Transcription)**
- 4. Metadata & Data Management**
- 5. Data Preservation and Dissemination**
- 6. Legal and Ethical Issues**



TOPICS

1. Introduction
2. Corpus Design and Curation
3. Annotation (including Transcription)
4. Metadata & Data Management
5. Data Preservation and Dissemination
6. Legal and Ethical Issues



- The Language Archive: New unit at MPI-PL (NL)
- Technical aspects (less: content, corpus design)
- Long-term developments (>15y)
- TLA Data mostly from two research fields:
 - *Language acquisition* (CHAT / CHILDES)
 - Focus here: *Linguistic fieldwork & Language Documentation* (esp. DOBES)
- TLA: Archive, Lang. Archiving Software (“LAT”)
- Annotation tool ELAN



TOPICS

1. Introduction
2. **Corpus Design and Curation**
3. Annotation (including Transcription)
4. Metadata & Data Management
5. Data Preservation and Dissemination
6. Legal and Ethical Issues



- Static vs. growing corpora
- Static corpora at TLA:
 - *Corpus Spoken Dutch (CGN)* (COREX)
 - *Dutch Bilingualism Database (DBD)*
- Dynamic corpora at TLA (‘living’ archive) :
Documentation of Endangered Languages (DOBES)
- Different teams, goals, methods, standards
- Challenge: achieve (semantic) interoperability



(DOBES corpora, continued)

- Multi-purpose by design
- Often the major/only source of data for a lang.
- Also for other scientific disciplines:
 - anthropology, ethnomusicology, -botany, -history,...
- Basis for extensive description (grammar), and for typological studies
- Details of the analysis may improve over time
 - > evolving annotation



TOPICS

1. Introduction
2. Corpus Design and Curation
3. **Annotation (including Transcription)**
4. Metadata & Data Management
5. Data Preservation and Dissemination
6. Legal and Ethical Issues



- (Linguistic) primary data are direct representations or results of a *speech event*, for instance a written text or, in particular, an audio/video recording of a speech event
- (Linguistic) annotation of data is any systematic symbolic representation of properties of the (speech) event represented in the primary data or of earlier annotation
- ⇒ Transcription is (basic) annotation



- Technical data format for annotated speech in audio and video recordings: ELAN A.F. (XML)
 - > deeper analysis of sign language
 - > “paralinguistic” phenomena such as gestures
- Unlimited number of tiers (for any speaker)
- Hierarchical relations between tiers
- DOBES corpora: no agreed format or naming for the types of linguistic levels / annotation layers
- ⇒ Challenge for comparisons & interoperability



- Basic annotation: A transcription & at least one translation, optionally notes or comments
- Basic glossing: basic annotation & information on individual units (morphs / words), typically: an individual gloss (meaning or function), perhaps also categorical information (POS tag)
- Standardized by the “Leipzig Glossing Rules”
- Interlinear glosses often with Toolbox program (semi-automatic, lexicon string-matching)



ANNEX with ELAN annotation



MAX-PLANCK-GESELLSCHAFT

ANNEX manual ? embed settings

user: seba@mpi.nl logout

Text

Grid

Subtitle

Waveform

Timeline

Combined

Video display

min



Media information

min

Resource: 026_autobiogr-2.eaf
Media file: 026_autobiogr2.m4a

Elapsed time: 00:00:12:900

Selected chunk:

Begin time: 00:00:04:484

End time: 00:00:08:220

Text: kype i'atu ti a'yn , kype me

Mini Data Frame

min

(ite te) itemoajpu a'yt kype a'yn
itemoaipu a'yt kype a'yn kype i'atu ti
a'yn , kype me kype jatã net itã kitã
na'yt ne tazu'jyt tetam 'etu jatã pe a'yn
nanype a'ang a'yt i'atu a'yn kype mejka
itekozokotu a'yt kaminu'azan net
itekwhapu jatãtsu ika'akwahapejutu
mã 'jyt katu met akwawap wã a'yn
akwawap 'jyt jatã [mune] mã'jy[t]katu

Tier:

Font size:

Play selection

Clear selection

Timeline

00:08:500 00:00:09:000 00:00:09:500 00:00:10:000 00:00:10:500 00:00:11:000 00:00:11:500 00:00:12:000 00:00:12:500 00:00:13:000

SmngP@026

isso se chama a aldeia da formiguinha

Sort@026

jatã net itã kitã na'yt ne tazu'jyt tetam 'etu jatã pe a'yn

com@026

ref@026

4





- Abbreviations for abstract functional units vary
- ISOcat: Data category registry (concepts) hosted at and developed by TLA
- ELAN tiers and annotations can refer to ISOcat
- Reference to shared general concepts, or to concepts related with the RELcat registry
- Advanced glossings include additional other levels, e.g. intonation, syntactic structure, grammatical relations, DOBES GRAID annotation



ISOcat



MAX-PLANCK-GESellschaft



Welcome Guest Help

intransitive

- My Workspace
 - Public
 - Thematic Views
 - Metadata
 - Morphosyntax
 - Morphosyntax
 - Basics
 - Cases
 - FormRelated
 - MorphologicalFeatu
 - Operations
 - PartOfSpeech
 - RegisterDatingFrec
 - Semantic Content Rep
 - Syntax
 - Language Resource C
 - Lexicography
 - Language Codes
 - Terminology
 - Multilingual Informati
 - Lexical Resources
 - Lexical Semantics
 - Translation
 - Sign language
 - Audio
 - CLARIN-NL
 - GOLD

search result for 'intransitive'

#	Name	Version	Administration	Registration st.	Char	Type	Owned by
1322	Intransitive	1:0	private	private	👉	simple	Declerck, Thi
3080	AntiCausativeVoice	1:0	private	private	✓	simple	gold-user
3427	Processive	1:0	private	private	✓	simple	gold-user
3533	Transitivizer	1:0	private	private	✓	simple	gold-user
3457	Repetitive	1:0	private	private	✓	simple	gold-user
3247	ImpersonalPassiveVoice	1:0	private	private	✓	simple	gold-user
3276	Intransitivizer	1:0	private	private	✓	simple	gold-user
3548	Versive	1:0	private	private	✓	simple	gold-user
1225	absolute case	1:0	private	private	👉	simple	Francopoulo,
3003	light verb	1:0	private	private	✓	simple	Francopoulo,

Intransitive - 1:0

2. Description Section

Profile	Syntax
2.1 Language Section	
Language	English (en)
<i>2.1.1 Name Section</i>	
Name	Intransitive
Name Status	admitted name
<i>2.1.2 Definition Section</i>	
Definition	Refers to a verb that does not take a direct object; that is, to a verb that does not express an action which directly affects another person or thing.
Source	www.southwestern.edu/~carlg/Latin_Web/glossary.html
<i>2.1.3 Example Section</i>	



TOPICS

1. Introduction
2. Corpus Design and Curation
3. Annotation (including Transcription)
4. **Metadata & Data Management**
5. Data Preservation and Dissemination
6. Legal and Ethical Issues



- Speech data “archiving”: long-term availability
- IMDI, an XML-based metadata standard
- Metadata in separate XML files describe bundles of resources (multimedia & annotation files etc.)
- Persistent identifiers (handles) point to data f.
- A virtual hierarchy of “sessions” > a tree-like structure with sub-corpus IMDI files as nodes
- resources & metadata stored as separate files
- IMDI browser (standalone & online / web-based)



IMDI-Browser



MAX-PLANCK-GESELLSCHAFT

settings manual register user: seba@mpi.nl login logout

IMDI-Browser

- Aweti
 - From Language and Culture
 - Cultural Data
 - Linguistic Data
 - Dialogical
 - Elicitation
 - Monological
 - Biographies
 - 010_autobiogr
 - 025_autobiogr
 - 026_autobiogr
 - 026_autobiogr1.mp4
 - 026_autobiogr1.mpg
 - 026_autobiogr1.wav
 - 026_autobiogr2.mp4
 - 026_autobiogr2.mpg
 - 026_autobiogr2.wav
 - 026_autobiogr3.wav
 - 026_autobiogr-2.eaf
 - 026_autobiogr-2.txt
 - 043_autobiogr
 - 046-marriage
 - Descriptions
 - Explanations
 - Narratives-083
 - Narratives-157
 - Procedural
 - Word lists
- Bakairi, Katxuyana and Mawe project
- Bakola
- Baure
- Beaver Archive
- Cashinahua project
- Chaco languages
- Chintang and Puma Documentation Project
- Chipaya project

IMDI

ISLE Metadata Initiative

Session

Name 026_autobiogr
Title 026 tells the story of his life
Date 2002-06-20

Description

An autobiographical relatory given by 026, one of the older men of the village.
There is an audio and a video recording to this session. The main part of the session is on media files 026_autobiogr2.wav and 026_autobiogr2.mpg. Introductory and concluding remarks are on separate media files.

Location

Project Awetí

Keys

Path in BrowsableCorpus
BrowsableCorpus\MatFromLangCult\LinguisticData\NonElicited\Monological\Biographies
Path in Archive .\fr-Ing-clt\ling\monolog\biogr\
basic annotation Port. only

Content

Actors

Description

SD asks the consultant, 026, to tell his autobiography. 017 translates between SD and 026.



- LAMUS, AMS (online tools) upload and integrate sessions and manage the data in online archive
- TLA involved in European infrastructure project CLARIN (cooperation with DARIAH)
- Transfer to CLARIN CMDI component metadata
- ARBIL for creating and editing CMDI metadata
- Basis for a general virtual research environment for data management in CLARIN and beyond



TOPICS

1. Introduction
2. Corpus Design and Curation
3. Annotation (including Transcription)
4. Metadata & Data Management
5. **Data Preservation and Dissemination**
6. Legal and Ethical Issues



- For DOBES data, currently six copies are created automatically at three locations
- MPG 50 years bit-stream preservation guarantee
- Digital data centres: the more often the material is accessed and used, the better
- ⇒ Important goal: easy and useful access
- Tasks for technicians and researchers
- An important aspect of a fruitful relation between researcher and data centre is *trust*



TOPICS

1. Introduction
2. Corpus Design and Curation
3. Annotation (including Transcription)
4. Metadata & Data Management
5. Data Preservation and Dissemination
6. **Legal and Ethical Issues**



- Human subjects are recorded: intricate issues
 - Privacy of the human subjects
 - Extreme inequality wrt. resources and information
 - Impossibility of anonymizing the speakers
 - International setting of the projects
- Often access is determined individually
- Attempts at clear and fair conditions of use
- Code of conduct: no commercial use and other uses that are disrespectful to the culture
- Goal: open access as far as possible



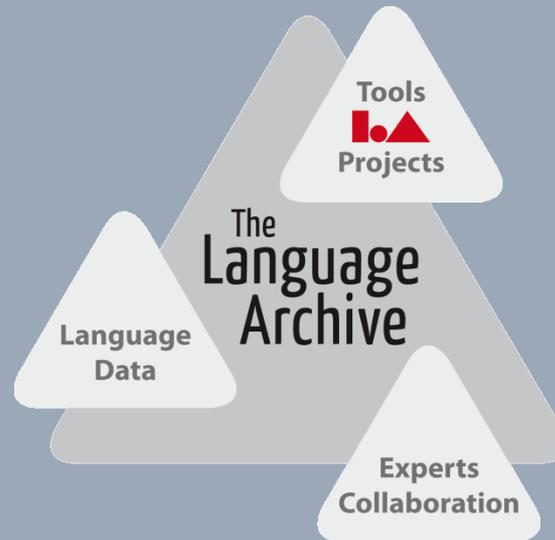
Four levels of access privileges:

1. Directly accessible via the internet
2. Users register and accept the Code of Conduct
3. Users apply to the responsible persons and make their usage intentions explicit
4. Completely closed material, except for the researcher and members of the speech communities

Management with AMS exploits the tree structure

LREC
Workshop 18

2012-05-21
Istanbul



MAX-PLANCK-GESellschaft

Best practices in the design, creation and dissemination of speech corpora at The Language Archive

Sebastian Drude, Daan Broeder, Peter Wittenburg, Han Sloetjes

The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands