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

BEST PRACTICES AT TLA

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

BEST PRACTICES AT TLA

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

Introduction

- 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

BEST PRACTICES AT TLA

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

Corpus Design and Curation

- 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

Corpus Design and Curation

(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

BEST PRACTICES AT TLA

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**

Annotation

- (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

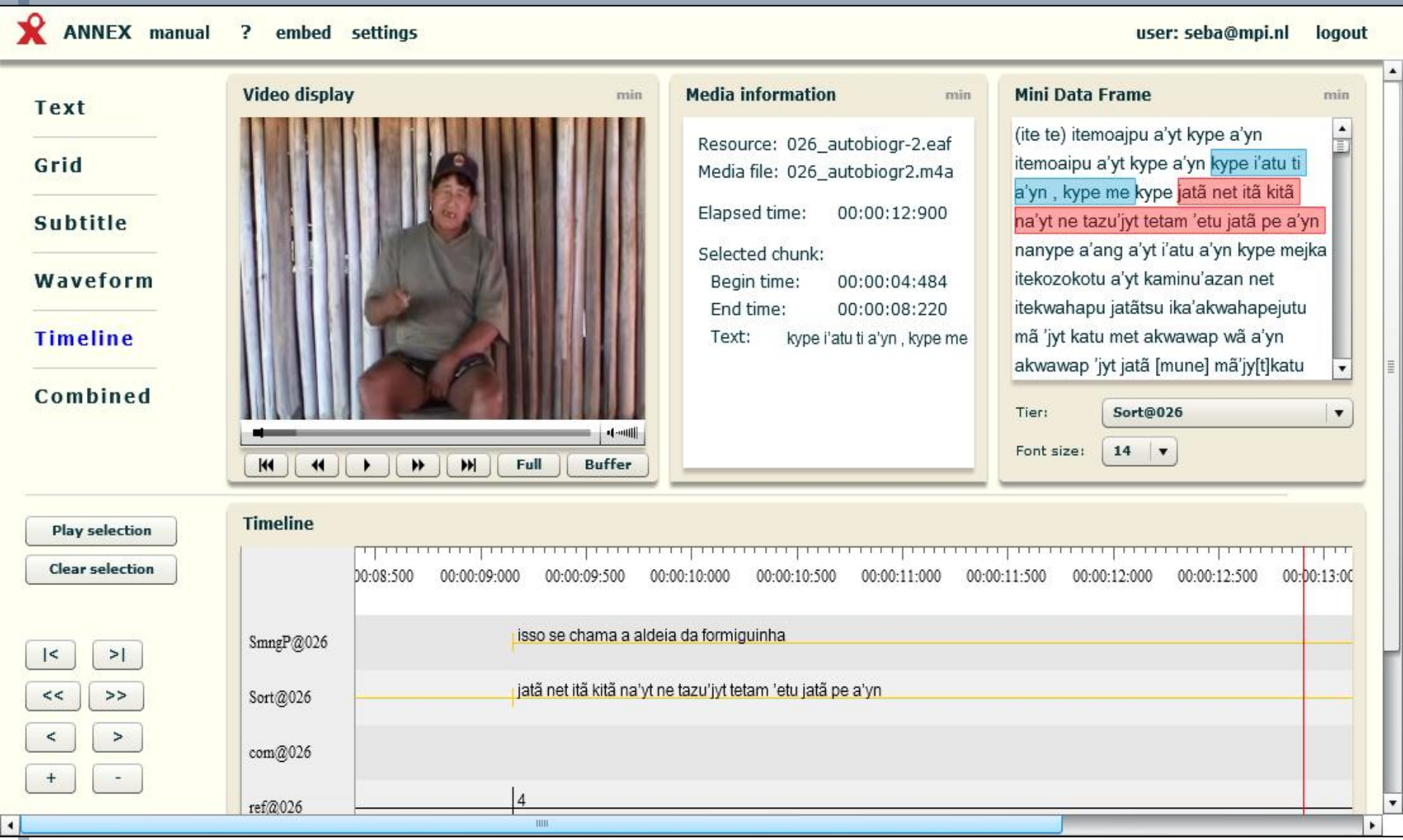
Annotation

- 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

Annotation

- *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)

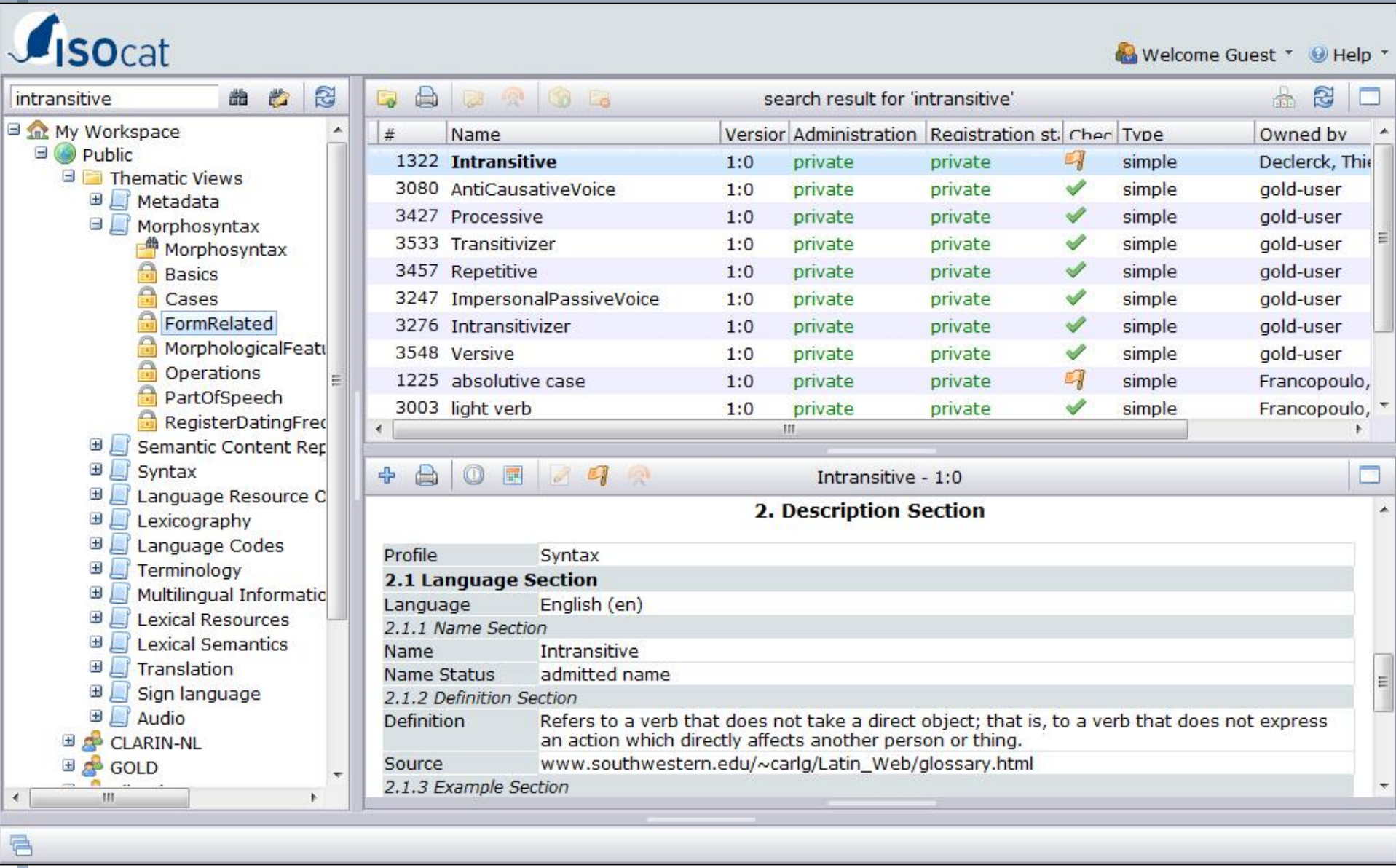
ANNEX with ELAN annotation



Annotation

- 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


ISOcat
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 Informatic
Lexical Resources
Lexical Semantics
Translation
Sign language
Audio
CLARIN-NL
GOLD
search result for 'intransitive'
Name Versior Administration Registration st. Chec 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 absolutive 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)
2.1.1 Name Section
Name Intransitive
Name Status admitted name
2.1.2 Definition Section
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
2.1.3 Example Section

BEST PRACTICES AT TLA

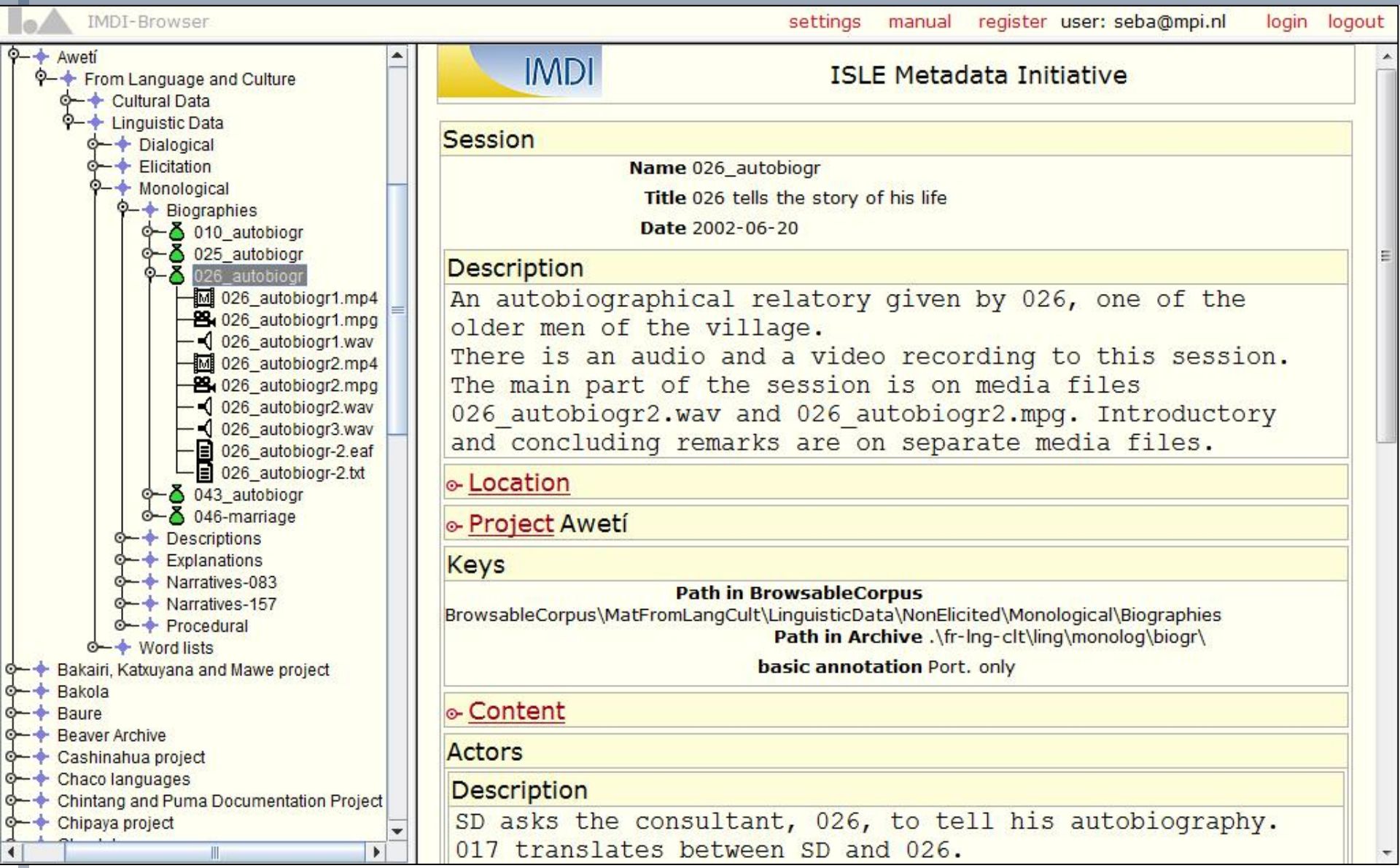
TOPICS

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**

Metadata & Data Management

- 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



Metadata & Data Management

- 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

BEST PRACTICES AT TLA

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**

Data Preservation and Dissemination

- 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*

BEST PRACTICES AT TLA

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**

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

Legal and Ethical Issues

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