

EMELD 2006
Tools & Standards: The State of the Art

Preparatory Notes for Group 2

Transcription and annotation of primary data
transcription, time alignment, creating IGT
Elan, TasX, IGT Editor, etc.

Members: Scott Farrar, Naomi Fox, Dafydd Gibbon, Reinhard Hiss, Jermay Jiancuo, Trevor Johnston, Alexander Nakhimovsky, Robert Neumann, Alexis Palmer, Ann Sawyer, Nick Thieberger, John Thomson, Imelda Udoh, Rhea

Assignment: Sessions

- Sessions: The working groups will be asked
 - (1) to critique existing tools and standards,
 - (2) to identify gaps in the toolset, envisioning tools and functions which don't yet exist, and
 - (3) to consider larger issues having to do with the development of digital tools for linguistics, e.g., interoperability of tools, duplication of functionalities, needs of different user groups.
- There will be three working group sessions during the conference; and we will ask the working groups to devote one session to each of the three tasks above [\[JG's Paper\]](#).

Outcomes

- Comments on the tools that are, or should be, listed in the E-MELD software database (include non-custom tools, interoperability, resource migration, tool acceptability):
<http://emeld.org/school/toolroom/software/index.cfm>
- A description of needed tools and standards. In Session 2, the workgroups will be asked to envision desirable tools and functionalities that do not yet exist, e.g., automatic transcription of audio and video, automatic annotation of a text based on previously annotated texts.
- Discussion of the general situation in linguistics with regard to digital tools and standards, including comments on some or all of the issues raised in Good's paper: Creator – Archivist - User
- And, as a final activity, we would like you to review the handout on E-MELD outcomes which was distributed in the first session and indicate which you consider most important to maintain and/or pursue further. **What's this?**

Group 2 specifics & link suggestions

- Transcription and annotation of primary data
 - Transcription [\[Tools\]](#)
 - Where do the labels/categories come from? [\[OLAC\]](#), [\[GOLD\]](#), [\[IPA\]](#)
 - time alignment
 - Where do serial and parallel structures come from? [\[LDC\]](#), [\[AG\]](#)
 - creating IGT
 - Elan [\[ELAN\]](#), [\[TASXforce\]](#)
 - IGT Editor
 - (Interlinear) Script Annotation Manager [\[SAM\]](#)
 - Interlinear Glossed Text
 - [\[EMELD-IGT\]](#)
- Standards: EAGLES Handbooks [\[1997\]](#), [\[2000\]](#)

Relevant papers for Group 2

(offline links, see E-MELD site: [\[EMELD 2006 papers\]](#))

- **Session II: Documentation and Annotation**

- Andrea Berez, Gary Holton (Wayne State University, University of Alaska, Fairbanks): Designing community-tech workflows: A field linguist's guide to putting good practice language technology into the hands of speakers [\[Berez\]](#)
- Moses Ekpenyong, Nnamso Umoh, Mfon Udoinyang, Golden Ibiang, Eno-Abasi Urua, Dafydd Gibbon (University of Uyo, Universität Bielefeld): Infrastructure to Empowerment: An OSWA+GIS Model for Documenting Local Languages [\[Ekpenyong\]](#)
- Dafydd Gibbon (Universität Bielefeld): Fieldwork and computing: PDA applications [\[Gibbon\]](#)
- Thorsten Trippel (Universität Bielefeld): The missing links in documentary linguistics: An approach to bridging the gap between annotation tools [\[Trippel\]](#)
- Chris Hellmuth, Tom Myers, Alexander Nakhimovsky (Colgate University): Linguist's Toolbox and XML Technologies [\[Hellmuth\]](#)

- **Session IV: Databases and Corpora**

- Trevor Johnston, Onno Crasborn (Macquarie University, Radboud University Nijmegen) [\[Johnston\]](#)

1: Existing tools & standards

- Brainstorming round table on actual tools used
 - List of tools known, actually used, rumoured to exist
- Discussion of terminology
 - Standards, best practices
 - Metadata, annotation (markup, transcription, time-alignment/labelling)
- Tool classification:
 - Input / Output / Processing methods
 - E-MELD, EMELD-*pro*
 - Presentation of tools: tree, table, wiki, ...
- Workflow: data flow between tools, interoperability

1: Existing tools & standards

- Terminology (1):
 - Standards, “best practice”: a profusion of private definitions:
 - ISO, DIN, ...
 - Widely used *de facto* standards (despite possible disadvantages):
 - PC, MS-Windows, MS-Office, ...
 - ELAN, Praat, Transcriber, Toolbox, ...
 - WAV, MP3, ...
 - Consensual “best practice”
 - Protocols, interfaces, specifications

1: Existing tools & standards

- Terminology (2):
 - Metadata
 - Annotation:
 - Markup:
 - Content: e.g. assignment of categories to transcriptions and texts
 - Structure: document constituency, networks
 - Rendering: layout, font, speech output, ...
 - Transcription: symbolic representation of speech
 - Time-alignment (labelling): assignment of time-stamps to transcriptions

1: Text & character handling, etc.

- Input methods: keyboarding tools, import, OCR, dedicated hardware, speech/dictation, ASR, forced alignment (trained vs. generic)
- Output methods: font editors, fonts, export
- Workflow specifications: which tools interoperate with which other tools?
- Domain handling:
 - Text
 - Speech: phonetics, phonology, prosody
 - Sign languages
 - Conversational gesture
 - Artefacts
 - ...

1: Tool classification - EMELD

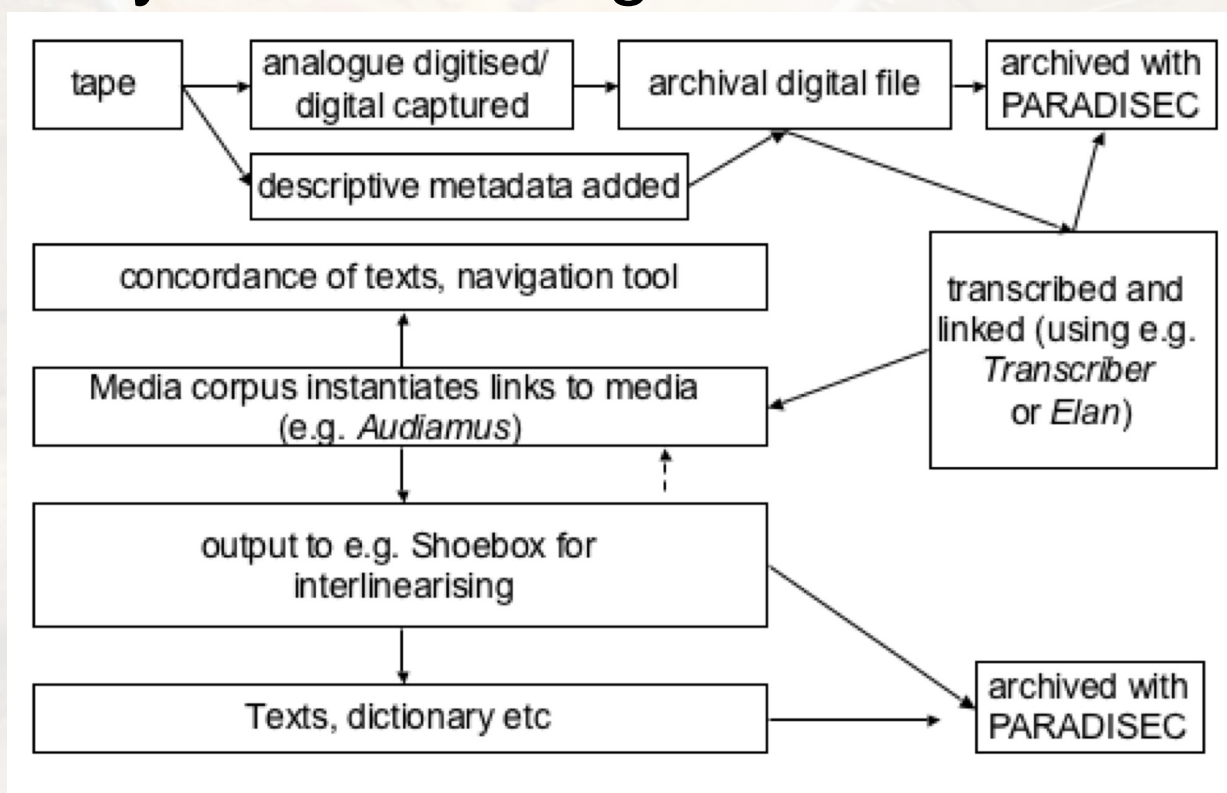
- Audio Editing/Conversion (7)
- Comparative Linguistics (1)
- Concordance (5)
- Corpora (8)
- Lexicon Management (16)
- Taggers (15)
- Text Editor (21)
- Transcription (14)
- Video Editing/Conversion (5)
- Video/Audio Alignment (7)

1: Tool classification – EMELD*pro*!

- **Annotation:** ANVIL, CLAN, CLAware, ELAN, EXMARALDA, HIAT-DOS, Praat, SignStream, SIL speech, analyser, Snack, SyncWriter, TASX annotator, Transana, Transcriber, WaveSurfer
- **Multifunctional Lexicon/Interlinearisation DBMS:** Access, Excel, Fieldworks Language EXplorer (FLEX), FileMaker, ILEX, Lexique Pro, Lingualinks, OpenOffice, Toolbox, Word, WordSurv
- **Interlinear & structured format editors & converters:** ITE, ITP, SAM (Script Annotation Manager), XMLspy; ECONV, Bielefeld Scripts
- **Output tools:** MS KB layout & creator, Fontlab, Graphite, Keyman, Typetool, Ukelele, UniScribe
- **Taggers & morphological analysers:** AMPLE, Brill, MaxEnt, OpenNLP, PC-KIMMO
- **Toolkits:** AGTK, BLARK (Basic Language Resource Kit), NLTK

1: Tool environment - workflow

- Proposal by Nick Thieberger:



- See revised version by David Nash:
 - <http://www.anu.edu.au/linguistics/nash/fm/flow.html>

1: Towards a critique model - ELAN

Tool	Transcription / text handling	Alignment handling	Interoperability: import / export	Interoperability: platform, user interface	Other comments
ELAN	Time-consuming cut-paste of existing transcriptions (better: file hacking outside the tool?)	Audio, Audio + Video, multi-channel, Flexible playback (WMP, QT, JMF), slow replay	XML, Unicode, but import/export unclear – EASY TO LOSE MEDIA FILE!	PC: Windows, Linux; Mac	Manual not very good. Hard to learn without training. But still: tool of choice for many annotation tasks. Good response to community input
ELAN <i>pro</i> !	Automatic dictionary creation, extended search functionality / concordancing. Integration with GOLD	“Switch to Praat” facility. Markup pins / areas on video display	Ready-made XSL stylesheets & templates for Naomi's applications	ELAN player to put on CD for use in language community	More example sets. Great improvement of task-oriented tutorials needed. Video annotation for web streaming. Hard copy production.

2: Gaps, lots of - Dreams, too many

- **Input methods:**

- keyboarding tools, OCR, import interoperability
- dedicated hardware, speech/dictation, ASR, forced alignment (trained vs. generic)

- **Output methods:**

- font editors, fonts, export interoperability
- Unicode character handling problems (Praat, Transcriber), base plane of Unicode (cf. 8/16/32 bit codes) & import/export, sorting, normalisation, rendering

- **Text handling:**

- General format conversion
- Corpus linguistics: automatic distributional analysis
- Machine learning: grammar induction, lexicon induction

3: Recommendations - tool classification

- Adopt an ontology for tool classification, including:
 - Input methods: including keyboarding tools, import, OCR, dedicated hardware, speech/dictation, ASR, forced alignment (trained vs. generic), touchscreens (e.g. for character tables)
 - Output methods: specify formal character model (e.g. Hughes, Trippel, Gibbon on character semantics) font editors, fonts, export
 - Data processing: corpus linguistic tools, e.g. taggers
 - Workflow specifications needed:
 - which tools interoperate – HowTo, FAQ, Wizard, ...?
 - collect workflows/tool inventories from existing projects

3: Recommendations - tool information

- Other presentation styles: tree, table, wiki, ...
- Other repositories
 - [DFKI software registry]
 - [Nadine Borchardt's links]
 - [Stanford links]

3: Recommendations - tool quality

- Evaluation
 - Specification of evaluation dimensions & procedures
 - ISO; EAGLES handbooks (corpora, evaluation)
 - consumer/user evaluation
 - review system for an online journal?
 - wiki
- The bottom line – what the beginner wants::
 - endorsement/deprecation:
 - ratings by experts wrt well-defined criteria (as in consumer magazines)
- Steven Krauwer's BLARK specs
 - EZ-EMELD?

Recommendation: Learn from others!

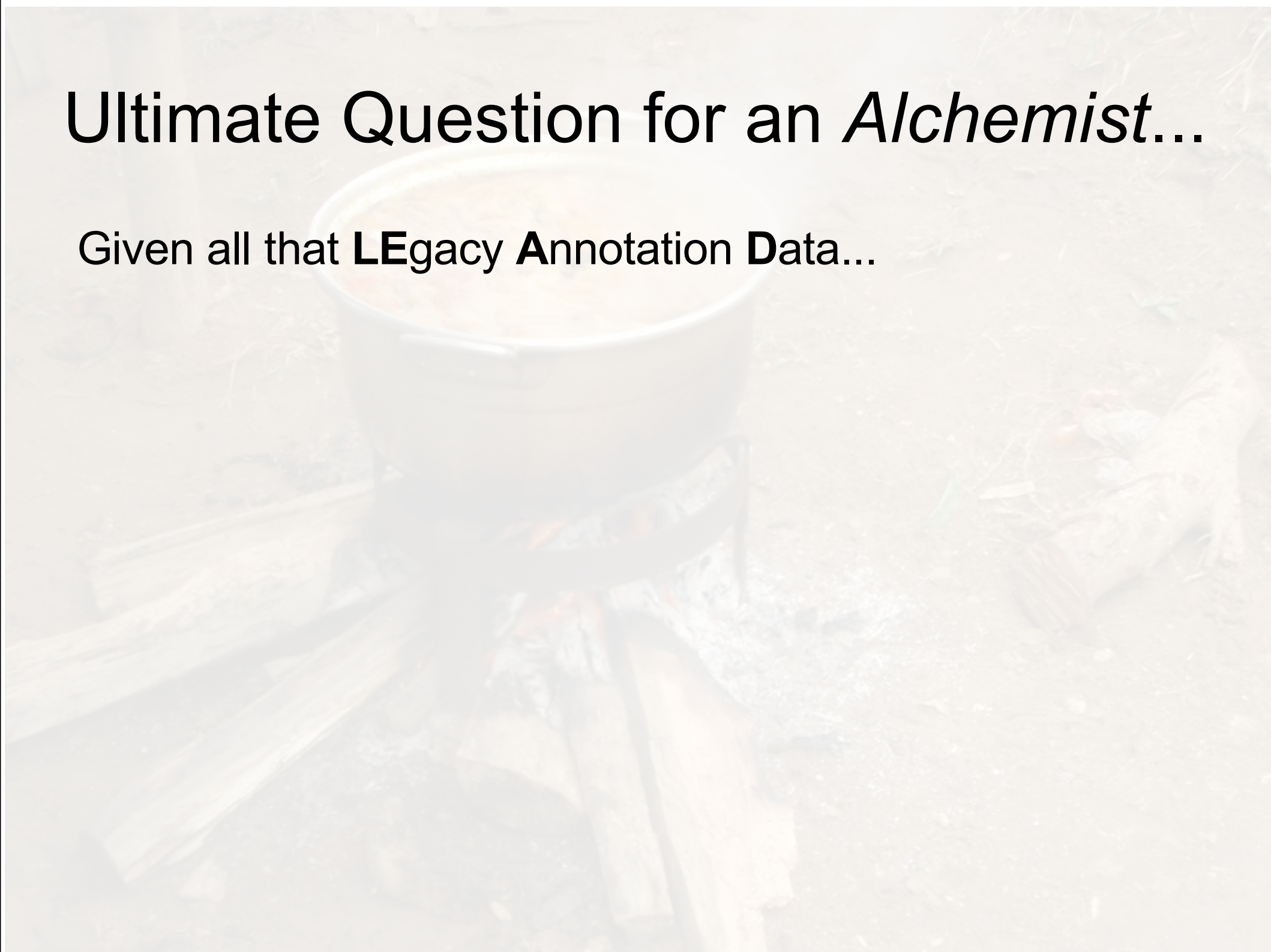
- NLP (Natural Language Processing):
 - Text normalisation, statistical analysis (corpus linguistics), information retrieval, lexicon induction, ...
- Speech Technology (ASR, TTS, ...):
 - Speech corpus creation: pre-recording, recording and post-recording procedures/standards
 - Automatic & semi-automatic segmentation and time-alignment; forced alignment of transcriptions and signals
 - EAGLES projects in 1990s: Evaluation of resource quality
- Computer Science:
 - Software engineering, standard development procedures
 - Machine learning, data mining, database views

More Recommendations...

- Publicise best practice:
 - Integration of GOLD support into more tools
 - ELAN
 - Taggers
 - ...
- E-MELD as a clearing-house?
 - Information exchange on tools, formats, modelling conventions for XML (data structures)...
 - “People should be less nervous about publishing stuff that is imperfect.”

Ultimate Question for an *Alchemist*...

Given all that **LE**gacy **A**nnotation **D**ata...



Ultimate Question for an *Alchemist*...

Given all that **LEgacy Annotation Data**...

How do we transform LEAD into GOLD ?

