LD Tools and Methods Summit Report

This document provides an overview of the main points arising from discussion at the Language Documentation Tools and Methods Summit (http://bit.ly/LDsummit2016) held at the University of Melbourne on 1-3 June 2016 and convened by Nick Thieberger and Simon Musgrave for the Centre of Excellence for the Dynamics of Language, funded by the Australian Research Council. Invited participants were asked to consider key issues that were pre-circulated and then prepare discussion points for the meeting. Each theme leader took notes and they are summarised below, with links to the original notes also provided below. There is necessarily some overlap between the reports on group discussions.

All presentations at the Summit can be seen on the Centre of Excellence for the Dynamics of Language website. A list of useful links is provided at the conference website: http://bit.ly/LDsummit2016. Images taken during the Summit can be seen here: https://flic.kr/s/aHskEFhbMm

This document should be cited as: Thieberger, Nick. 2016. Language Documentation Tools and Methods Summit Report. http://bit.ly/LDTAMSReport

Introduction

This conference brought together people interested in building better tools and methods to document small languages. While primarily aimed at tool developers and at the linguists using the tools, a major outcome of our work will be increased access for speakers to better records of more languages than is currently the case.

This conference will help to set the agenda for collaboration on standards and tool development and provide the CoEDL with direction for investment of funds. There were similar workshops/conferences run by emeld.org in the USA a decade ago, and a Digital Tools Summit in the Humanities was also run at Virginia in 2005.

We want to explore innovative tools and methods and identify current problems for fieldwork, recording, transcription, analysis, archiving and accessibility of language material.

Four themes were identified by the Summit organisers as set out below and groups were assigned to each theme. Each group had discussions in advance of the Summit and then met to write up their responses and additions to these discussion points. Full discussion documents are linked to below and this document summarises the main points. A list of members of each theme is given in the Participant list at the end of the document.

Theme 1: Archiving, Discovery, (re)Use (theme leader: Linda Barwick). Full discussion is at http://bit.ly/LDSummitTheme1

- How can we build on the foundation provided by OLAC to maximise discoverability of existing material?
- How can archived language resources be made more useful in terms of citation, persistent identification, ease of access, and the development of 'landing pages' that describe the collections?
- How can we extend the number of archives and the reach of archives to include more records, especially at-risk legacy records?
- Archiving software, visualisation of language collections

Theme 2: Workflows, Interoperability (theme leader: Sasha Arkhipov) . Full discussion is at http://bit.ly/LDSummitTheme2

- What is the range of workflows (from recording through to the archive) that are used in LD projects and how can they be improved?

Workflow blockages: how much is the lack of interoperability of our tools preventing the development of well constructed corpora? (problems: assigning metadata to items; knowing what has been transcribed, annotated, interlinearised; moving from complex multi-tier transcription to interlinearisation and losing part of the transcription, etc).

- Interoperability and the outputs of LD (standards for all kinds of material created by LD)
- Standard formats for complex annotation/IGT
- Metadata entry tools to help organise collections and prepare them for archiving

Theme 3: Data Enrichment (theme leader: Caroline Jones) . Full discussion is at http://bit.ly/LDSummitTheme3

- Recording and transcribing/annotating recordings (HTK, e.g. MAUS – forced alignment).
- Eventual automatic transcription
- Distributed annotation (including crowdsourcing):

online systems for annotating page images of notes (archival manuscripts: handwriting recognition)

annotating dynamic media

interlinearising annotations

- What emerging tools or methods can we look to and invest in?
- Increasing scope of recordings (e.g., Aikuma)
- Delivery of language records for speakers (phone apps, HTML5 services from archives)
- Dictionary creation and presentation systems (online, app-based)

Theme 4: Corpora, Scale (theme leader: Steven Bird). Full discussion is at http://bit.ly/LDSummitTheme4

• Corpus development for small languages: what standards should we be adopting or developing for corpora of small languages that may be different to those in use for large languages?

What frameworks are there that small textual/media corpora can be placed into for general use (e.g., developing EOPAS.org.au)

Interfaces, models and technologies for mobile language apps (scaling up recording and delivery)

Background reading

Emily M. Bender & Jeff Good. 2010. A Grand Challenge for Linguistics: Scaling Up and Integrating Models http://www.paradisec.org.au/BenderGoodScalingUp.pdf

Recommendations extracted from the theme group reports

1. Can the Center fund individuals who have created solutions and get-arounds to certain problems to create youtube videos and/or documentation on these?
2. Run a focussed workshop on data management/ metadata entry with participants working with several existing tools in advance to then provide a set of requirements
3. Possibility of asking CoEDL to invest in software that does do collaborative work (for each step of the workflow, transcription, annotation, analysis)
4. Can CoEDL fund research into the location of analog fieldwork media that needs to be digitised and archived?
5. Select 5 or 6 corpora representing sufficiently diverse languages. Of these corpora, 5 hours (out of roughly 50) should be well transcribed
    - take the 2000 most frequently used words and find the corresponding audio segments
    - if possible record respeak by multiple speakers of (a part of) these words
    - hand the data over to the CL/Speech recognition specialists (and actively collaborate with them in the development/improvement of the algorithms)
    - run the recognisers on (the rest of) the data and evaluate the results. The evaluation might involve creating a good transcription of previously untranscribed recordings with the "automatic" transcription as a starting point
6. Transcription acceleration. A working group has been established to look into methods for automated speech recognition and text/media alignment.
7. CoEDL could support the development of tools that are easier to learn, perhaps simpler versions of existing tools (as it has already done with Simple Elan)
8. There is a need for a directory of tools and methods, or for more publicity of existing lists (see ‘Pooling of documentation tools info’ below)

Table of contents

Summary of each group's discussion

Theme 1: Archiving, Discovery, (re)Use – Summary of outcomes of discussion

The group worked through the Theme 1 topics first and then returned to the "Big questions" posed to all four thematic groups. The discussions enabled participants to share an overview of current issues in the area and to identify some areas for targeted action. It was agreed that the most pressing current priority was to facilitate pre-ingestion collection processing, either through adapting and expanding the current SayMore tool, or by development of sets of metadata entry tools ranging from spreadsheets to apps. It was recognised that some development (e.g. FAIMS) is going on in other disciplines that needed to be investigated for potential collaboration.

Pre-ingestion

- Comprehensiveness of coverage:
    - finding and archiving unarchived collections;
    - Identifying languages that need attention;
    - OLAC-izing language resources in non-language archives (institutional repositories, etc).
- Collection processing:
    - tools to facilitate collection organisation and submission to archive
- Best practice information to guide potential depositors

Archive management

- Harmonizing platforms and structures where possible between archives
- Version control - helping depositors to improve their collections over time

Dissemination

- Maximise discoverability
    - Better landing pages that give a broad view of the collection's contents
    - Multiple channels for dissemination of metadata
- Better user experience:
    - Data visualisation
    - Data presentation (e.g. IGT with media)
    - Mobile dissemination platforms
    - Facilitate data enrichment by users
    - Finer grained discovery tools for heritage language learners
- Better understanding of archival collection/corpus (are they the same?)
    - Feeding typologists' demand for online searchable annotated corpora
    - Feeding speech technologists' demand for more diverse data
    - Facilitating post-archive indexing/transcription of records (crowdsourcing)

Designated community actions: Effort coordination and cooperation

- DELAMAN development and publishing of how-to resources (archiving for dummies)
- Improving cross repository search experience
- Outreach at conferences

Recommendations:

Run a focussed workshop on data management/ metadata entry with participants working with several existing tools in advance to then provide a set of requirements

1. How can we build on the foundation provided by OLAC to maximise discoverability of existing material?

1.1. Outreach to existing repositories to code their language content and expose that to OLAC?

Action: Prepare a roadshow to relevant conferences (librarians, museum curators).

1.2. How do we extend awareness of the contents of our collections so that speakers can find that their relatives' voices/images are recorded and can be accessed? (social network media is a good idea, so how to make it work?)

1.3. One page description of what would be required to assign ISO-639-3 codes to objects - disseminate via language blogs etc as well as making available on websites

1.3.1. Insert code in the catalog if there is a place for it to go

1.3.2. Crúbadán adding ISO 639 detection capability to web crawlers

1.3.3. Influence DOI to add language specific metadata (at least allow the use of 639-3 codes)

1.3.4. Wiki-athon type data collection/registration activity at conferences

1.3.5. Building a register of sites that includes ISO-639-3? (cf. ISLRN - not a good model) or subset of http://www.re3data.org/ registry of repositories

1.3.6. Add a record to an existing catalog (e.g. http://catalog.paradisec.org.au/collections/External/items/Arosi)

1.3.7. If a resource is located on a transient website then locate that website in the Internet archive and use that URL to the register. If it is not there than suggest it to the Internet Archive.

1.3.8. DELAMAN? to build database to feed LR records to OLAC (how could this be maintained?)

1.3.8.1. Needs a social media etc campaign to get material entered into the register/database

1.4. revisiting the OLAC categories and enriching them; standardizing as much as possible is the easiest way to get data from the hands of the documenters via the archive to OLAC

2. How can we extend the number of archives and the reach of archives to include more records, especially at-risk legacy records?

2.1. Funding sources coming online for this sort of work (e.g. ELDP's 'Legacy materials' grants) so we should be proactive in locating and planning applications

2.2. Plan outreach to other disciplinary communities collecting language data (anthropology, musicology, ecology etc)

Actions: feed information to relevant email lists; PARADISEC or other blog; share on social media

2.3. The question of the purpose(s) of archiving/corpora should remain central (and specifically addressed, not in general terms such as "to keep traces/memories of a dying language etc.")

2.4. Extending the number of archives: Is there/what is an effective, cost-efficient way to have mirrored archives? Is that a way of having institutional repositories be useful alongside DELAMAN level archives? Without a consistent set of standards and tools for integrating the "pieces" of data (i.e., ELAN transcription with a video), how useful are repositories at the different levels (university, local museum, etc) especially when different repositories have different goals/themes. (And how can we get non-linguists/non-language interested people to a) use the language materials in meaningful ways and b) to consider best practices in language documentation when they are doing research in other disciplines?)

2.5. A number of programs at NSF actually fund a single repository for the discipline – is it too late for that? Would that even be the way to go (at least on national levels)? (Seems like the question is setting the stage for more archives, not fewer.)

3. How can archived language resources be made more useful in terms of citation, persistent identification, ease of access, and the development of 'landing pages' that describe the collections?

3.1. Use wikipedia pages as landing pages integrate with glottolog

3.2. Encourage linguists to create landing pages to contextualise their collections

3.2.1. Finding aids and developing other compilations that make it one-stop shopping is essential.

3.2.2. Use field report sections within existing journals (LD&C, LSA etc, or data journals, cf http://www.ands.org.au/guides/data-journals) as incentive to linguists to write the finding aid for their collection, then link to relevant web pages and archives

3.2.3. Get body like https://rd-alliance.org/ Research Data Alliance to lobby for recognition of such curated archival datasets

3.3. Data processing: the metadata and cataloging, transcription, translation, glossing and linguistic analysis, annotation – this is a huge hurdle and where it seems the least likely to have the bottlenecks removed by automated

processing or annotation (see the discussion of transcription acceleration below). Potential of citizen science initiatives (such as found in other disciplines) for power of processing the data?

4. How to locate existing material and accession it into archives?
   - 4.1. Legacy collections in the hands of linguists are an impending crisis. How can we best use existing resources to deal with legacy collections?
   - 4.2. This might be an area where some citizen science activity could be helpful, with a wiki-like interface (multilingual?)…there are people who have for example, heard that their great-grandmother worked with someone on a dictionary. See the 'Lost and Found' survey
   - 4.3. Do we want to consider something like an en masse summer school where groups of people are trained in what to do and then spend an internship-like time period ingesting, digitizing, etc. to bring legacy collections in? How many DELAMAN archives are actively involved in training by doing for their students…and how much integration there is with I-schools, where professionals are trained in best practices in all these areas. Language archives and professionals could benefit from strengthening these relationships.
   - 4.4. Identify languages for which fieldwork has been done in the past generation or two but for which no records are identified by an OLAC search; contact the linguist and apply for funding to digitise any tapes they hold. Research project: look at olac/glottolog where there are no resources but we know fieldwork has been done; then follow up with linguist and offer to help; also target any areas where OLAC/glottolog mismatch. (this project is underway in 2016 in the CoEDL, as a result of collaboration between Robert Forkel and Nick Thieberger arising from this Summit)

5. Software and other facilitating resources for archiving
   - 5.1. Is there a way to take advantage of Islandora implementation by 3 DELAMAN archives to develop a common core to facilitate future convergence
     - 5.1.1. Share OLAC feed configuration from Islandora so other archives can use it
   - 5.2. Long term aim of providing off-the-shelf archiving solution for language archives (issues for longterm maintenance)
   - 5.3. Mainstream library cataloguing infrastructure lacks two main things: ISO 639-3 and language resource type vocabulary (OLAC only has 3 now; needs to develop a longer set)
   - 5.4. Need to add to mimetype registry - what mimetypes do we care about e.g. X-EAF
   - 5.5. Data Management Plans at NSF are part of the merit review process. Create a DELAMAN data management plan template and publish on DMPtool - https://dmptool.org

5.6. Need to have introductory materials explaining speakers' and recorders' rights/permissions to help organisations to decide whether to contribute

6. Visualisation of language collections

6.1. Ideally an archive would present data in relevant forms

6.1.1. .eaf shown with media

6.1.2. .flextext presented as IGT

6.1.3. What corpus exploration systems are there that could work together with a repository?

7. Discoverability of untranscribed / unannotated audio

7.1. The purpose of an archive/ a corpus: is it possible to consider that recordings accompanied by their metadata are sufficient for a linguist who has not taken part in the recording (and possibly accesses the data after the last speakers have disappeared) to conduct an analysis of the language that may lead to a grammatical sketch/ a full grammar ?

8. Sustainability of tools as an issue for archives/discovery/reuse

8.1. Sustainability is a major concern. There are many database projects (and websites!!), but who sustains them when the funding is gone, when the software is obsolete, and when the programmer can't work for free. Most archives (as I understand it) can't ingest software, so how are these kinds of resources preserved and accessible for the future? Many researchers seem not to have answers for this.

8.2. Learning from the past - a typology of defunct online projects

8.3. What resources do archives need the most investment in/for/toward?

9. How can we encourage more use of proper archives rather than websites and other transient locations?

9.1. How to encourage use of language documentation corpora for linguistic research

9.2. Archiving infrastructure solutions/recommendations for small to medium archives -- minimum hardware requirements (server, storage) plus software platform (~LAT platform or similar)

9.3. interoperability with social media platforms as a key component of discovery Cf also Bender & Good IVd – Data sharing

9.4. Solve the login barrier

9.4.1. Clickable agreements too hard for machines

9.4.2. Ease of access to language resources should also include automated access. A major obstacle in this respect are no or non-open licenses, and data behind logins, see https://blog.ldodds.com/2015/11/25/how-can-open-data-publishers-monitor-usage/ for a description of this scenario, its impact and alternatives.

9.4.2.1. Ideally access is as open as possible, but the risk is that deposits will not be made if access is open to the world, so there always

needs to be the option to ask for user credentials for some items in a collection.

9.4.3. Future convergence of well run websites and archives? Need institution to anchor permissions/legal regimes

10. Engaging the user community

10.1. mobile technologies seem to have a lot of utility for people in the developing world. It seems like it's a good area for investment by archives and by those developing other tech tools.

10.2. there are many opportunities that field/language documentation products present for use by others, but the lack of integration of the data in its many formats seems like a problem. Sure, there are 100 beautiful videos, each with transcription and translation to accompany it, but are those files easily "corpora-ble"? And then where is that corpora stored? Databrary (video data library for developmental science) is a great resource because of all the ways data can have value added to it by other researchers. Is language documentation missing the boat when we we used to be state of the art?

10.3. Data Management Plans at NSF are part of the merit review process. Reviewers and panelists help the discipline the more rigorously they review these. But finding easy ways to teach old dogs new tricks (and new dogs new tricks – when they come from programs that don't train them in these tools is essential). Not everyone can go to CoLang or get an EDLP grant with the included training. Should NSF's DEL program do a new PIs meeting? The vast differences in the kinds of knowledge people have is challenging for evaluating the probability of their success on a project.

11. IP/Licenses/open access

11.1. CoLang 2016 Susan Kung presentation on IP rights http://www.alaska.edu/colang2016/courses/workshop04/

11.2. breaking through the login barrier for archive access

11.3. Integrating ethics and permissions requirements/prompts into data collection tools and then harvesting that as part of the archive submission package

11.4. There is an issue with international copyright protection of “works” such as music, narratives, visual artworks etc. under the Berne convention 1886 http://www.wipo.int/treaties/en/ip/berne/ and the Beijing treaty on audiovisual performances http://www.wipo.int/treaties/en/ip/beijing/summary_beijing.html. Specifically we can’t infringe the rights of the performers/artists by publishing their work online without informed consent. Keeping these works behind a login means that they remain “grey [gray]” and don’t enter into copyright, thus protecting the rights of the creators, and also protects our institutions from potential legal action. http://apraamcos.com.au/about-us/faqs/general-faqs/

11.5. There is quite a difference between classic 'open data' types of data (census data, government statistics etc) and recorded human performance that needs to be handled sensitively.

11.6. Data behind logins is just a reality that we'll have to deal with when it comes to endangered languages collections. Even ELAR with its open access policy for all data collected by new grantees has a registration/login requirement also for their open data. This problem is not unique to our domain, lots of (research) data is available only after authentication/authorisation and people are working on solutions to make it possible to work with these materials in an automated way. This is all rather complex matter though, but hopefully we will be able to take advantage of developments done elsewhere, e.g. within EUDAT. See also https://rd-alliance.org/groups/federated-identity-management.html (which is not just about identity, contrary to what the name suggests).

11.7. MPI will be implementing a feature for our new repository solution that will allow users to download whole collections or subsets thereof as one (or a number of) zip files, rather than each file separately. This should also make it easier to work with larger sets of files, even though one has to log in once via the web interface to download them.

Archiving responses to the big Questions

1. What are the most pressing technological needs in making better records of the world's small languages?

1.1. Finding analog field-recordings that are not in a repository

1.2. Collection processing: easier ways of entering / recording metadata (solving the metadata bottleneck) plus turning a collection of files into a useable corpus (data management, including metadata) (e.g., SayMore)

1.3. Facilitating interaction with archival records (as depositor, as user)

2. What efforts are being made by current researchers to address these needs?

2.1. Excel metadata template and upload (PARADISEC) http://www.paradisec.org.au/PDSC_minimal_metadata.xls

2.2. Archive user commenting (built in to PARADISEC collection/items for logged-in users)

2.3. SayMore/Exsite9/CMDIeditor etc

3. What obstacles to more efficient work practices could be overcome by a targeted effort of programming over the next few years?

3.1. sharing of modular resources among developers

3.2. Create metadata interchange format/tool for DELAMAN archives

4. What emerging tools or methods can we look to and invest in?

4.1. Upscaling SayMore (invest in getting SayMore to cross-platform, import spreadsheets, provide collection level archive submission information package)

5. How can new tech developers be attracted to the field?

5.1. create more open projects

5.2. Without singling out anyone, in the past there have been large infrastructure projects that have developed guidelines and frameworks, sometimes getting to the level of being functioning systems, but ending up without content or users.

6. Why are there so few digital repositories for all the material being created by documentation projects?

6.1. Are there really so few?

6.2. If so, why is this a problem?

6.2.1. Too much work for the few existing repositories.

6.2.2. Demonstrates the priorities of relevant agencies and researchers that they have not established repositories.

7. Can we identify the successful systems/ tools we use and why they are successful? (e.g. OLAC; Toolbox; ELAN)

7.1. SayMore -- an example of a collaborative, distance project is NEH-funded Seminole Nation Language Project headed by Jack Martin of College of William and Mary http://www.alaska.edu/colang2016/models/ and http://muskogee.blogs.wm.edu/

Theme 2: Workflows, Interoperability – Summary of outcomes of discussion

Workflows and Interoperability

- The focus of this group is the range of workflows (from recording through to the archive) that are used in LD projects and how they can they be improved.
- Workflow blockages: how much is the lack of interoperability of our tools preventing the development of well constructed corpora? (problems: assigning metadata to items; knowing what has been transcribed, annotated, interlinearised; moving from complex multi-tier transcription to interlinearisation and losing part of the transcription, etc).
- Interoperability and the outputs of LD (standards for all kinds of material created by LD), e.g. standard formats for complex annotation/IGT
- Metadata entry tools to help organise collections and prepare them for archiving
- Sustainability -- of data? of tools?

Workflows & blockages

Where do LD workflows start and end?

Many LD workflows are concerned with getting new data through a chain of enrichment to ultimately publishing (analysed) data in an archive. However, the steps that come before and after should be paid attention:

(before)

- Locating existing archives/collections on the target languages/regions
- Learning about current best practices in tools and methods
- Familiarizing yourself with the suggested tools and methods

(after)

- Updating archived resources
- Mobilizing archived resources (involving research communities, speaker communities and interested public)

Some common workflows: "Contemporary fieldwork"
Record -- transcribe -- analyse -- provide metadata -- publish

Common bottlenecks:

- required time and efforts amount increases greatly from record to transcribe to analyse, so only a fraction of collected data gets full analysis;
- thorough metadata descriptions take long, especially with many sessions (texts, recordings) to describe
- getting all files converted into right format and named appropriately

Some common workflows: "Legacy documentation"
Find extant recordings/publications -- digitise -- convert to common format -- align -- analyse -- provide metadata -- publish

Common bottlenecks:

- locating data and getting access
- digitisation and physical quality issues (handwritten notes, old tapes and wax cylinders, OCR problems with specific typographic/transcription issues)
- making data from different researchers, linguistic traditions, times, transcription systems... converge in a single common format is a pain

Some common workflows: "Assembling a corpus"
(assemble data from different sources, not necessarily too old, into a single collection/corpus):
Find extant recordings/annotations -- convert to common format -- align -- analyse -- provide metadata -- publish

Common bottlenecks:

- same as above, with more accent on conversions between software formats

It seems that there are by now acceptable solutions for several of the common problems. No single tool is perfect, however. One natural approach would be to get a selection of tools which go together well, e.g.: SayMore+ELAN+FLEx+Praat+R. However, SayMore and FLEx are not supplied for Macs, and ELAN-FLEx conversions cannot (yet) allow a complete roundtrip.
Non-linearity of workflows
Although software is often designed with a linear workflow in mind (Stage 1 > Stage 2 > … > Stage N), in reality most often the users need to go back and forth between various stages.

Some parts of the workflow can run in a smooth and linear fashion, however the nature of linguistic analysis is such that one often needs to review the very basic stages (e.g. revising the initial transcription and/or segmentation). The tools should allow such reanalysis of the basic things to be as non-destructive as possible for the later stages/tiers of analysis.

This makes roundtrips between tools/formats very much desirable. An ideal case is thus an interchange format from which various tools could use and update only a part without breaking the rest.

On the other hand, archiving needs not be only done when the analysis is finalised -- because there is no such thing as a really final analysis (see Archiving below).

Specific bottlenecks: Segmentation

Segmentation into intonation units, words, phones. Can this step be automated to boost productivity?

Suggestions:

Segmentation into intonation units

- ANALOR http://www.lattice.cnrs.fr/analor?lang=fr

Forced alignment:

- MAUS & WebMAUS: http://www.bas.uni-muenchen.de/Bas/BasMAUS.html , https://clarin.phonetik.uni-muenchen.de/BASWebServices/index.html#/services

  Force aligner trained on several languages, can also run on pooled phonemic models for arbitrary language using SAMPA transcriptions
- eSpeak/Praat: Praat (http://www.fon.hum.uva.nl/praat/) has a built-in aligner using (an older version of) eSpeak (http://espeak.sourceforge.net/), run in TextGrid editor under Interval menu; adding new language models is in principle possible
- ProsodyLab Aligner (http://prosodylab.org/tools/aligner/): a trainable aligner in Python; does not need time-aligned data for training.
- See GORILLA project (http://gorilla.linguistlist.org/software/), in particular the ELAN2split and a collection of eSpeak language models

Specific bottlenecks: Transcription & Translation

Bottlenecks:

Speaker involvement to boost productivity:

- Tools need to be super easy
- Automatic syncing/backup via web to avoid data loss

Suggestions:

Use BOLD-type oral annotation:

Let's not think of traditional writing-based docu methods and BOLD-type docu methods as separate paradigms. These methods/tools should complement each other in the same workflow.

Recommendation:
Researcher working with speakers: use SayMore
Speakers working independently: use Aikuma-NG

- SayMore (http://saymore.palaso.org/):
    - takes a pre-existing audio file as input, allows segmentation, oral slow rendition as annotation of original file, and transcription
    - Output Elan and Flex compatible
- Aikuma-NG (https://github.com/aikuma/aikuma-ng, http://www.aikuma.org/)
    - A Chrome app based on Aikuma which was an Android app
    - Can take a pre-existing (ideally segmented) audio file as input and add slow rendition & oral translation as annotation
    - Use Bulk import in Beta version

Crowdsourcing of transcription and translation:

Writing-based tools for use by speakers

- Phonemica.net (http://phonemica.net/): crowd-transcribed/translated storytelling initially in Sinitic languages

Speakers are interested in

- games
- proficiency and literacy in language of wider communication

Use this for crowdsourcing

- Duolingo-style tasks
    - Hear segment and write it down
    - Hear segment and translate
- Game-format of transcription & translation tasks
- Task for school curriculum: translate target language segment into language of wider communication

Specific bottlenecks: Interlinearisation/glossing

FLEx as a glossing solution

Glossing in FLEx (http://fieldworks.sil.org/) is quite a common practice. Features to know of include:

- One can either do manual glossing (breaking words and selecting morphemes among homographs) or use one of the built-in parsers.

- There are two different parsers in FLEx, one Item-and-Arrangement oriented (XAmple, default), the other phonological rule-based (Hermit Crab).
- FLEx now allows collaborative workflows with Send/Receive synchronisation of whole projects (including texts and dictionary). However, people should avoid working on the same text in parallel (merging becomes problematic), tasks should be kept distinct.
- Export to ELAN via flextext XML format. FLEx currently imports, stores and re-exports media file references, time alignment and speaker attributes from/to ELAN. So this is currently possible:
    - segmentation/transcription/translation in ELAN ⇒
    - glossing in FLEx ⇒
    - re-import into ELAN to have a time-aligned interlinear text

Some stoppers for using FLEx and suggested solutions or workarounds:

- No Mac version

(Windows and Linux versions available)

Suggestions: Mac version is not expected in the near future. A possible workaround is installing Windows on Mac. Running Windows/FLEx in a virtual machine can also be an option although slower.

- Unable to import interlinear glosses

Import currently works for sentence-level tiers only, i.e. transcription, translation and comments; even if interlinear is converted into flextext XML, the word- and morpheme-level tiers are not imported. This is a serious showstopper e.g. for people who have already a substantial amount glossed in Toolbox.

Action: FLEx team will address this issue to implement interlinear import from flextext.
Getting interlinear into flextext is a much simpler task.
The logic is to map morph-gloss pairs to the lexicon, creating new morphs (lexicon entries) as needed. In case of homonymy and identical gloss the first match is substituted.

- No support for custom tiers

(e.g.: different kinds of transcriptions; nominal classes; semantic features; interactional analysis events...)

Workaround (currently available): adding “fake” writing systems allows to create additional tiers (e.g. adding Alutor-FON alongside Alutor-Latin and Alutor-Cyrillic allows to have three distinct transcription lines for texts in Alutor language). But this is an abuse of the writing system concept, as the goal is to deal with conceptually different transcriptions which can be written with the same writing system.

Action: FLEx team will address this issue to add support custom tiers at sentence ("phrase") level. However, word- and morpheme-level custom tiers would demand serious changes in the FLEx data model.

Glossing outside FLEx

FLEx parsers are only available inside FLEx. Other parsing options include:

- ELAN CorpA (http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php): a special edition of ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) with lexical items extraction and lexicon-based parsing (similar to Toolbox). Also has the ability to create groups of annotations and links between annotations. Current version is 4.8.0, which is slightly behind the regular ELAN version (4.9.4), but intended to keep up at a reasonable pace.
- UniParser (http://languedoc.philol.msu.ru:8082/fieldling/en/uniparser/) [Russian webpage, English forthcoming]: a user-configurable parser
    - Takes morphological description in a YAML-like format (stems, paradigms,...) and a list of wordforms (or a text), outputs an XML list of wordforms with analyses
    - Word-and-Paradigm approach: can but needs not provide morpheme breakdown and glosses; the minimal output is a list of grammatical tags for each wordform
    - No disambiguation: outputs all possible analyses
    - Regex to specify various types of constraints
    - Output suited for the EANC web corpora platform with complex searches, see examples at http://web-corpora.net/

Specific bottlenecks: Metadata

Major workflow bottleneck: Getting metadata into the format accepted/required by archive. In some cases this prevents archiving of text data

Q: Are there Tools/scripts/formats that help bridge the metadata notation practices of the researcher with the requirements of the archives?

Suggestions:

- Refer to metadata workgroup emerging from Theme 1 workgroup
- CMDI Maker (http://cmdi-maker.uni-koeln.de/): a browser app (can only create IMDI/CMDI files, not edit existing files or manage them)
- SayMore (http://saymore.palaso.org/): Manages some (fixed) metadata descriptions for recording sessions, renames files to keep file bundles consistent. Windows-only.
- ? As an interim solution, spreadsheets with well-defined templates/webforms, e.g. using SmartSheet (https://www.smartsheet.com/)
- ? A web-app to edit metadata files (CMDI) based on a generic XML editor, e.g. jEdi jQuery editor (https://github.com/UNC-Libraries/jquery.xmleditor)
- On a long term, INEL project (Hamburg) will eventually develop a metadata manager

Specific bottlenecks: Archiving

Should/can data centers and archives play a role in / give support during the corpus creation (e.g. as backup location for non-annotated data) or just curate the "final" corpus?

Should archiving be the last step in a workflow and just apply to a well annotated "finished" product? Rather archiving should be a repeated intermediate step between workflow steps, ie.

- Archive recordings with basic metadata
- Archive transcription, translation, interlinearisation of archived recordings as they are created
- Easy updating of archived resources

Pooling of documentation tools info

Linguists doing fieldwork are often not aware of available tools and solutions, even of not-very-recent features of familiar tools (e.g. Linux version of FLEx, alignment in Praat, sound playback in Toolbox...).
Need place to share knowledge, scripts, how-to tutorials, descriptions of get-arounds, read-mes with functionality as (one of) entry points: what are available tools for transcribing, annotation, interlinearisation, metadata creation etc.

We need go-to points for people at different levels.
Non-programming-savvy linguists without institutional technical support need a different first-port-of-call than programmers and programming-savvy linguists

Suggestions:

- Wikipedia page on tools for LD. Please contribute to this site. https://en.wikipedia.org/wiki/Language_documentation_tools_and_methods
- RichardLitt's GitHub repository (https://github.com/RichardLitt/endangered-languages): a list of open-source code that would be useful for documenting, conserving, developing, preserving, or working with endangered languages
- https://github.com/HughP/LinguisticsLibrarianList
- Reviews like those on the CAQDAS site would be good:
- http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/analysingvisual/index.htm
- LinguistList maintained Wiki ?
- How-to tutorials on Youtube

Can the Center fund individuals who have created solutions and get-arounds to certain problems to create youtube videos and/or documentation on these?

File transfer

Big file transfer from depositor to the archive. Alternatives to sending a hard drive? Suggestion, use OwnCloud (https://owncloud.org/)

Specific bottlenecks: Interoperability

Complex tools vs. simple apps

The workgroup's conclusion was that neither complex tools nor simple one-task all-user-adapted apps will be a sufficient solution without each other.
A way to lessen the programming burden on the way to having both types of applications could be simplified or cut-down apps "extracted" from bigger apps, cf. ELAN and Simple-ELAN.

Standard formats

TEI as a standard format for interlinear (IGT)

Evidently, TEI is seen as a possible framing solution. The advantages are quite obvious: an open and evolving standard, XML-based, comes with many tools and a large and active community.

What are the possible limitations that TEI could bring in? Are we safe to rely on it?

Transcribing natural spoken discourse calls for prosody-based and shallow-structured segments. The relation of what was actually said to what can be grammatically analysed is more distant. This is why distinguishing between transcription-based tiers and 'normalised' text is preferred.

Can we perhaps regard the normalised transcription with related morpheme glossing as 'written-type' IGT embedded into 'spoken-type' annotation?

To make things more complicated, there can be a mix of different media types serving as annotation/annotated data:

- primary (annotated) data: sound, video, text (string), manuscript/image (page + 2D selection as a unit)
- annotation: text, sound (oral transcription/"careful speech", oral translation, ...), video (sign language translation)

Standoff vs inline annotation

Should there be a tier with minimum intervening markup? Should it be the main (reference) tier or a(n optional) dependent tier?

Big questions and discussion

How can new tech developers be attracted to the field?

Problem (among others)

Tech developers face the same issues as documentary linguists: their main outputs (e.g. great software solutions to linguistics problems/needs) do not count as research output and scientific work

E.g. presentations / articles submitted to conferences / journals e.g. on user study of a particular resource that was created are rejected and not considered relevant enough for the wider community.

Suggestion

The journal Language Documentation & Conservation publishes technical descriptions of technical issues for a technical audience.

- Creates peer-reviewed research output for developers
- Creates searchable (findable!) platform for technical information

General problem: Valuation of various kinds of resources created for and during the language documentation: software, archives/corpora, databases etc.

What emerging tools or methods can we look to and invest in?

What opportunities do emerging technologies enable? / What is the low-hanging fruit (ie technologies that could be deployed now)? / What new tech could be useful?

If some of the functionality now delivered on the desktop by ELAN etc. could be moved to the web (which is feasible, I think), there would be a very big potential in "crowd-sourcing" transcripts and annotations. (Semi-)automatic support in these tasks (segmentation via silence detection, speaker diarization, ASR, automatic token annotation etc.) is also within reach, but still (too) hard to use productively for projects without dedicated technology support. Some prototype tools to invest in or be inspired by:

- Transcriber.js [http://modyco.inist.fr/transcriberjs/doku.php]
- WebAnno [https://webanno.github.io/webanno/]
- Camomile [http://www.chistera.eu/projects/camomile]

TEI as a standard format for interlinear (IGT): What are the possible limitations that TEI could bring in? Are we safe to rely on it?

It is certainly not safe to rely on TEI as a whole. The guidelines have way too many degrees of freedom so that any two independent endeavours to represent data in TEI are very likely to end up with solutions that may be semantically compatible but are structurally very different. That would not really be standardisation. You would therefore need to define a TEI-based IGT format, which can be a lot of work, but worthwhile. For "spoken language" (i.e. outside language documentation proper) it has paid off to approach the question in a bottom-up manner, by analysing existing tool formats, corpora and annotation/transcription guidelines and see what they have in common, where they differ, how they can be transformed to a TEI representation etc.

It may be a good idea to start an IGT TEIsation effort from existing language documentation unification efforts (maybe EOPAS is a candidate?), try to marry this to the ISO/TEI proposal

for spoken language transcription and only then add/modify what is missing/not adequate. (I'll bring some examples to illustrate this).
The most obvious limitation at the present moment is that there is no dedicated tool that reads, writes and/or processes TEI (for spoken language, for language documentation) "natively". Maybe TEI should therefore be viewed, for the time being, as something that comes in towards the end of the workflow, not as a format in which data are created and annotated (ELAN etc. can continue to do this), but as a format that is well-suited for storing completed corpora in a repository etc.

Recommendations for / standardization of media (especially video) formats / encoding

We have recently concluded a painful, year-long experimentation phase with different media formats, with the aim of defining sustainable standards for our archive. The results in brief are:

- PCM-WAV with 48kHz/16 bit (or 96kHz/24 bit) is fine for almost all purposes, we have no problems with audio
- Single image formats (like MPEG2000) would be ideal for archiving purposes, but require prohibitive amounts of disk space which we cannot afford.
- MPEG-4 (with some standard encoding parameters for image size, frame rate etc.) is the best solution for archiving data, and also (in a smaller version) for web delivery. MPEG-4 should also work for annotation tools (ELAN, EXMARaLDA), but some problems remain. The most reliable format across tools and operating systems seems to be (the totally ancient) MPEG-1 (again with some standard encoding parameters).
- We will therefore use MPEG-4 (big) as our archive format and derive an MPEG-4 web format (small) and an MPEG-1 tool format from that automatically. We will give up, once and for all, on all other (non-MPEG) video formats and containers (i.e. no more MOV, AVI, WMV etc.).
- We will discourage people from using freeware video converters/editors (like Super, Format Factory, Freemake Video Converter etc.), since these produce bad data. We will instead recommend EDIUS Pro (Windows) and/or Sorenson Squeeze as two commercial programs (with affordable licence fees) which we found reliably produce high quality results.

Theme 3: Data Enrichment – Summary of outcomes of discussion

The recommendation resulting from this group's discussion aims at bringing us closer to an answer to two questions:

1. how to speed up the transcription process / can computational techniques (speech recognition) be applied to language data of under-resourced languages in order to speed up the transcription process (and in order to get data annotated that otherwise would not be annotated)
2. if so, how much (annotated, clean) data is required in order to train models for a language such that the recogniser can successfully be applied to the rest of the data

of that language (criteria for the level of success to be defined). During the summit no one seemed to be able to give an estimation of the required minimal size.

This recommended project could consist of the following actions:
Select 5 or 6 corpora representing sufficiently diverse languages. Of these corpora, 5 hours (out of roughly 50) should be well transcribed

- take the 2000 most frequently used words and find the corresponding audio segments
- if possible record respeak by multiple speakers of (a part of) these words
- hand the data over to the CL/Speech recognition specialists (and actively collaborate with them in the development/improvement of the algorithms)
- run the recognisers on (the rest of) the data and evaluate the results. The evaluation might involve creating a good transcription of previously untranscribed recordings with the "automatic" transcription as a starting point

The last two steps might need several iterations.

Recording and transcribing/annotating recordings (HTK, e.g. MAUS – forced alignment).

1. Standardization/recommendations in the area of digital media formats, especially video. Audio is settled, video still problematic and depends very much on intended use. Useful remarks on this in Theme 2
2. Extract metadata from media header
3. Recommendations for quality also important in crowdsourcing situations e.g. to make sure that audio recordings are not only in mp3
4. Recommendations depend on the purpose of the material e.g. for gesture or eye gaze analysis (on the basis of video) the demands on resolution/quality will differ from those for other purposes
5. speaker recognition & diarization and other kinds of automatic audio analysis and annotation will be easier if all speakers involved would be recorded with separate, direct microphones (such as lavalier mics)
6. Need a researcher-friendly presentation of recommendations on a website. It could list the implications of a choice for a particular format in terms of which conversion and transcoding tools can be used
7. Ask all workshop participants to contribute 300 words on the tool they use or develop for the CoE website. Any site with recommendations, pros and cons, reviews etc. has to be updated regularly
8. Discovery in and transcription of data might be accelerated by using whatever technology is available for automatic detection. This could e.g. run in a workspace on a server. Ideas for detection tasks and/or technologies:
    1. use face recognition to find people

2. detect how many people are in a recording, find chunks where two people talk
3. apply silence recognition
4. find scene changes, background noises, chunks with any speech or with child speech etc
5. in other words: automatically create a protocol and for some tasks create an annotation file or metadata file

9. Application of forced alignment:

1. integration of forced alignment in ELAN would be useful. Possible use case: convert a symbolic subdivision tier into a time aligned tier and force align the words
2. is it possible to have stock grapheme-to-phoneme conversion for groups of languages, maybe cover the whole of IPA?
3. automatic transcription will be very hard for many languages, probably out of reach
4. what about training up for every language in the world?

How then would ELAN work with multiple audio files?
CorpAfroAs – always record speakers with one recorder+lavalier per speaker, and batteries when they are moving around (recording cooking sessions for instance). One problem is then the monitoring of the various devices: whether the REC or mic is still on, whether the batteries are ok, whether the SD card is not full yet, etc. Perhaps a central monitoring station (wireless) for separate devices would be a good technological improvement ? As far as reverb is concerned, Alexis Michaud (CNRS Lacito) has very good advice on setting up anechoic chambers in remote places, using local materials.
There are standard tools today that deal with cleaning up audio files and even removing reverb, e.g. iZotope's RX 5 (https://www.izotope.com/en/products/repair-and-edit/rx.html) or Zynaptics Unveil (http://www.zynaptiq.com/unveil/).

Eventual automatic transcription. How close are we? What will this take?

1. Automatic segmentation & speaker recognition on the basis of audio would be very useful (if sufficiently accurate), especially for projects in which not all data can be transcribed
2. What (good, pref. open) software in this respect is available?
   2.1. one open source audio tool suite that includes support for speaker recognition: https://github.com/tyiannak/pyAudioAnalysis (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144610)
   2.2. In Cologne the open-source speech recognition engine KALDI (http://kaldi-asr.org/doc/about.html) was tested on a collection of language documentation corpora in order to perform IPA/SAMPA based search on untranscribed audio with very limited success

2.3. Some papers about multilingual or even universal speech recognition:
   2.1. http://www.cs.cmu.edu/~tanja/Papers/ICMI02-Wang.pdf
   2.2. http://www.atlantis-press.com/php/paper-details.php?from=author+index&id=25846940&querystr=

In small and medium size documentation projects but certainly in those projects where much more data is collected than can be transcribed. If there is no existing software that is good enough and that preferably is open, then this might be technology where the field could invest in.
The AUVIS project (Cologne) tried to train the open-source speech recognition engine KALDI (http://kaldi-asr.org/doc/about.html) on a collection of language documentation corpora from the (DoBeS NTVR project) in order to perform IPA/SAMPA based search on untranscribed audio with very limited success.

We need to talk to people who do ASR full-time so we can benefit from the decades of literature figuring this out

We have no idea how much audio data is required for training, it also depends on recording quality. Find out many hours are required? Maybe ask the ASR experts what their challenges are going to be so we can include relevant datasets.

ASR may need a two-step process.

A. Transcription
2 hours++ transcribed in ELAN, time-aligned at utterance level / prosodic phrase. 5 hours would better because we could rotate the training and testing set.

1. Plus we need a text file with all tokens, with the phonemic transcription tokenized as a pronunciation dictionary for the HTK.

It needs a speech community you can go back to. Maybe look at a user interface that a community member can contribute via e.g. 'word spotting' type game.
What would be good datasets to test out the ASR?
It would make sense to create forced aligners for languages / language types for which we have sufficient data that we can use it later on.

B. ASR
Then get speakers to respeak and get more talker variability etc and retrain the tools (Forced alignment on the new corpus). Basically time align the new recordings and retrain the tools. Then start experimenting with normal speech recognition on a ?10 hour corpus. Then cycling and recycling through the data. Recoding dictionary for alternate pronunciations etc to improve the recognition.

How much interlinearising do we need to do? At what point is this unnecessary because of search capabilities? At what point is it more financial and practical to hire someone to write a parser? How to get a parser that can update its decisions? ELAN Corpa does this.

Crowdsourcable parts are the respeaking, repeating slowly. Cf. Aikuma. There are patterns in our elicitation sessions which could be analysed and converted into a form for crowdsourcing apps. E.g. number of reps we need on average. We need good metadata on the crowdsourced contribution.

Could there be ways of using ELAN and Flex as a group. Some Elan users have a method of assigning tiers to different people and then merging them. 10 years ago there was a collaborative ELAN peer-to-peer network which came to the point where it was a demo. General agreement that an application should allow for group work and not necessarily be based on one user-one computer-one project scenario. There is a potential problem of version control in his collaborative work, so we want a project management tool to manage crowdsourcing and version control. Not possible to retrofit some older programs into collaborative model. Possibility of asking CoEDL to invest in software that does do collaborative work.

1. Online systems for annotating page images of notes (archival manuscripts: handwriting recognition). Which systems do we know of or have we used? What are their strengths and limitations?
2. Maybe https://transkribus.eu and http://transcriptorium.eu can be sources of inspiration.
3. Online systems for annotating dynamic media, any good models?
4. Online systems for interlinear annotations, any good models?

Theme 4: Corpora – Scale, Summary of outcomes of discussion

Session 1: App Proposals

This group addressed the issue of scale by looking at ways in which app development could contribute to greater structured output from recording, surveys and games.

- Recruiting participants – Steven Bird (incl 2-page template for specifying apps)
- Household survey app – Nick Evans, John Mansfield, Steven Bird
- DirKap (directional app using iwatch) – Damir Cavar, Stefan Schnell
- Game ideas: Ngaanyatjarra running, Love, Word – Ben Foley and Jane Simpson
- Uremu (song learning app) – Dan Kaufman, Mat Bettinson
- Cooking App – Amina Mettouchi
- Misc resource: Web Apps terminology – Steven Bird and Mat Bettinson

Next steps: participants finalise their app proposals (recommended to do something close to the 2-page template); collectively identify common components, e.g.:

- Language chooser (generic)

- Participant metadata (generic)
- Elan export (generic)
- File list (generic)
- Storage API for IGT + audio + lexicon – Flextext, Elan (generic)
- Voice recorder + checklist + camera + GPS (household survey, recruiting participants)
- Voice recorder + slide sequencer (cooking app, Uremu)

Session 2: Corpus issues

Jane explained CoEDL's expectations for 10 corpora – looking for a road-map. Steven pointed out the difficulty of addressing this effectively in the short time available, given the broad range of complex issues involved. We captured the following notes from our session:

Things to bear in mind:

- The trust that exists between speakers and linguists
- The reason for collecting the corpus (e.g. answering linguistic questions, justifying a descriptive observation relative to the base data, presenting data to speakers/world)
- The idea that the Internet is mediating the use of the information

Approaches to the "corpus chaos"

- Create a new standard and supporting infrastructure;
- Do something exemplary for others to learn from (consensus choice);
- Use NLTK corpus readers plus programs;
- Map legacy data to standardized fields (when possible) to facilitate searching over different formats within a single corpus (the kratylos solution);
- Allowing end users to add annotation layers rather than forking a corpus for specialized needs (cf "standoff annotation");

Recommendations:

- Use versioning
- Distinguish private vs public repositories
- Be explicit about the relation of corpora to the sub-corpora derived from them, and having ways of synchronising them
- Recognise that corpus creation is a process, and while we have ideas about the desired endpoint, it's also important to consider "daily hygiene" (a set of recommended best practices?)
- Simple access management e.g. a check box on an offline file which says ‘release for community’ and then uploads the file and makes it visible online

Corpus development for small languages: what standards should we be adopting or developing for corpora of small languages that may be different to those in use for large languages? “a corpus, what for?” = an archive may be just a recording that sits there waiting

to be used, but a corpus is a collection of data organized with a purpose. And there should be a purpose to at least part of the linguistic materials we are collecting. Otherwise we might (?) end up with big data - but useless data. Technically speaking, a tool that could automatically recognize and chunk long rushes of video into parts that could be used for gesture analysis for instance would be great !

What frameworks are there that small textual/media corpora can be placed into for general use (e.g., developing EOPAS.org.au)
Interfaces, models and technologies for mobile language apps (scaling up recording and delivery)
What are the demands for structuring a corpus for archiving it? And what are the demands for structuring a corpus for searching it? Are these demands compatible?
For example: for archiving, we may want to have all the files associated with a single recording session bundled together (sound, tier-and-time-aligned transcripts, derived files e.g. pos-tagged files, community-friendly files. For quantitative searching, we may want to have all the files of a single type stored together (a folder of transcripts, a folder of pos-tagged files.. or even as Mosel does, a file of texts of a particular genre). The files associated with the same session would then be linked via their names (xxx.eaf, xxx.eaf. xxx.txt...), and a catalogue. [The alternative is a complex search that searches just files of type.txt say across all folders consisting of bundles of files linked to a session.]

Where do we store metadata and how do we avoid version control problems?

- Catalogue (csv) - this will be needed in any case to list derived files
- Headers in the file (.eaf, .txt, etc) - this is a good idea if files get separated

Other means - labelling files, labelling folders - obviously need systematicity and can be helpful as a fast way of eyeballing data, but shouldn't be the prime location of metadata.

Other metadata problems come with information that pertains to part of a file but not all of it. Two examples are **speaker** and **genre**.
Speaker: We are used to a tier structure labelled with speaker, which allows for searching by speaker, and quantifying data filtered by speaker.
Genre: This is another matter. Sometimes it characterises a whole file. But in one session there may be lots of different genres (and of course the commentary in between characteristic genres). A genre can span speaker-labelled tiers (as in co-constructed narratives). So, if we want to get all the vocabulary used in a particular genre, how do we extract it?

List of participants in the LDTAM Summit 2016

Theme 1: Archiving, Discovery, (re)Use
*Linda Barwick
Colleen M Fitzgerald

Felix Rau
Gary F. Simons
Gary Holton
Jenny Green
Nick Thieberger
Paul Trilsbeek
Robert Forkel

Theme 2: Workflows, Interoperability
*Alexandre Arkhipov
Anna Margetts
Christian Chanard
Denny Moore
Jason Naylor
Julia Colleen Miller
Mark W Post
Thomas Schmidt

Theme 3: Data Enrichment
*Caroline Jones
Amina Mettouchi
Andrew Margetts
Erich Round
Han Sloetjes
Jan Strunk
Janet Wiles
Kellen Parker van Dam
Rachel Nordlinger
Simon Musgrave
Stephen Morey

Theme 4: Corpora, Scale
*Steven Bird
Ben Foley
Damir Cavar
Daniel Kaufman
Jane Simpson
John Mansfield
Mat Bettinson
Nick Evans
Stefan Schnell