Assessing curated corpora as research output:
issues of process and evaluation

Anna Margetts (Monash)
Stephen Morey (Latrobe)
Simon Musgrave (Monash)
Adam Schembri (Latrobe)
Nick Thieberger (Uni Melbourne)

Background

- 2002 ALS Newsletter: Open letter from linguistic postgraduate students
  - importance of documentation of endangered languages
  - need for recognition of language documentation in the academic context
- 2010 LSA Resolution Recognizing the Scholarly Merit of Language Documentation
  - support for the recognition of language documentation as scholarly contributions

Background

- Preliminary discussions between ALS and ARC in 2011 showed that curated corpora could legitimately be seen as research output
- Responsibility of the ALS (or the scholarly community more generally) to establish conventions to accord scholarly credibility to such products

Background

- Motion at ALS AGM 2011 :
  - that the ALS write to the ARC noting that curated corpora of linguistic data and accompanying analysis should be counted as research outputs subject to certain criteria being met.
  - that the ALS Executive establish a sub-committee to detail what constitutes a collection for these purposes and how such collections can be evaluated.

The sub-committee

- Self nominated
- To be extended

Corpora as research output

- ARC are thinking of
  - ERA process
  - possibly track record for applications
- But acknowledgement of documentation corpora could extend to other contexts
  - research degrees
  - job applications
  - tenure/probation conditions
  - promotion
  - sabbatical project

Parallels to other forms of output

- Critical editions
  - classic work with extensive commentary
  - counts as a book!
- Non-traditional eligible research output types
  - Original creative works
  - Live performance of creative works
  - Recorded created works
  - …
- Where do corpora fit (e.g. in ERA)?

Current focus

- Language documentation corpora
  - un- or under described languages
  - often minority languages
  - often endangered
- Discussion can be extended to
  - other type of corpora in the humanities
    - child language
    - CA
    - Oral history
    - …

Linguistic Corpora

- Web of Knowledge Data citation index (Thomson Reuters)

REPOSITORIES BY DISCIPLINE	**COVERAGE**
– Life Sciences	48%
– Social Sciences	20%
– Physical Sciences	23%
– Arts & Humanities	**7%**
– Multidisciplinary	2%

Current focus

- Establishing criteria and procedures to assess a corpus
- Develop conventions to accord scholarly credibility
- Similar to peer-review for conventional academic outputs

“Reviewing” corpora

- Three different things to distinguish
  - Corpus review (≈ book review)
  - Corpus overview by corpus creators (= peer-reviewed article)
  - Peer-review of a corpus “for publication” (≈ peer-review of book/article for publication)

Corpus review

- Corpus review (≈ book review)
  - New section in *Language Documentation & Conservation*
    - Reviews planed but none published yet
  - Expert review of a corpus, published in a journal
    - summary of info in the corpus
    - info about what's good or not about it
    - provides some publicity for the corpus & corpus creators

Corpus overview article

- Overview article by corpus creators
  - Info about corpus content and organisation
  - Could be set as standard reference when citing the corpus
- Peer-reviewed article in suitable journal
  - e.g. LD&C, AJL, Oceanic Linguistics, …
  - with increasing citation index if corpus is used

Peer-review of a corpus

- Parallel to peer-review of a book or article for publication
- Reviewers assess corpus and award “seal of approval” for scholarly credibility
- Should be anonymous
- Needs clear criteria & procedures

- **This** is what we are talking about today

Corpus review committee of ALS

- ALS to establish a committee of reviewers
  - Membership can be included as service on members' CVs.
  - Any review is written by members of that committee and produced without their names
    - Committee needs to be large enough to make its fairly anonymous

Suggested Process

1. Submission of corpus by creators to ALS with request for review

2. Review by ALS panel

3. Report of panel to the creators
   - possibly with suggestions for improvements.

4. Response by the creators
   - including possibly revisions to the collection.

5. Publication of report summary & creators' response
   = recognition of the corpus as a research product
   - E.g. in ALS newsletter, AJL, or ALS website

Assessment Criteria

- Quality criteria
- Quantity criteria
- Each criterion to be assigned a value
  - e.g. between 1 and 5
- Decisions we (as a research community) make can/will drive behaviour of corpus creators

“Quality” criteria

1. Collection is housed in a repository
   - which has a commitment to long-term curation and access
   - which provides a citation form for items within the collection.
2. Contextual information
   - background to how the collection came into being and overview of what it contains)
3. Metadata
   - sufficiently described to allow corpus to be located
   - What it contains (e.g., texts, media, vocab, speaker info, perhaps content keywords)

“Quality” criteria

4. Accessibility:
   - are the files in the corpus in a format that is freely accessible?
   - do they depend on certain software?
     - this can render them difficult or impossible to use
   - does the software require additional meta-files?
     - E.g. Toolbox files need .typ and .lng files to be read

“Quality” criteria

5. Annotations
   - transcription
   - text-audio linkage
   - translation into language(s) of wider communication
   - interlinear glosses
6. Content
   - good range of speakers of different ages and genders
     - relative to opportunity
   - good range of text types (narratives, songs, procedural, hortatory, written, etc .
     - as per Himmelmann (2006:21) — intro to Gippert et al.
     - relative to opportunity

“Quality” criteria

7. Supporting documentation
   - list of abbreviations used
   - information on orthography and its relation to sound system
   - grammatical information sufficient to allow further analysis
   - lexicon included (quality rated from short wordlist to detailed dictionary

“Quantity” criteria

- Hours/units of primary data?
- Proportion of data which is annotated etc is a quality issue, not a quantity issue
- Perhaps the quality rating may end up being independent of quantity?

Combined scale of criteria

- Scale of 1-5 for each criterion which add up to a total score
- Star system assigned to collection, e.g.
  - 5-star collection =
    - quantity: many hours of annotated data
    - high content quality
    - extensively annotated
    - sufficient descriptive metadata
    - …

Example

- 4-star collection
  - small collection
    - well annotated
    - good supporting documentation
    - …
  - larger collection
    - less annotated
    - less supporting documentation
    - …

Content relative to opportunity

- Moribund languages
  - Quality criterion 6: Content **relative to opportunity**
    - good range of speakers, of different ages and genders
    - good range of text types
  - A corpus of a moribund language may score higher than an equivalent corpus of a non-moribund language
    - lack of opportunity to record range of data & speaker
    - corpus of non-moribund language COULD & SHOULD have been planned with more variety & representativeness

Version control / editions

- What's a "new" corpus / new version?
  (≈ 2nd edition, substantially revised)
- A researcher may submit a corpus for review and over the years add or annotate more data
- We need criteria to deal with assessing new or updated parts of a corpus

Possible consequences

- Such a process may mean corpus creators would choose to submit only part of a corpus for review
  - quality valued more highly than quantity?
- Implications:
  - archives might have to split collections (could this be handled at metadata level?)
  - different parts of a collection would have different citation styles

Is this desirable?

- The two need not be incompatible
- But important to recognise that our community response to proposals such as the current one may affect what we do

Cultural Change

- We recognise that these are ambitious aims
- It has taken 500 years to establish criteria for determining scholarly value of published work
- So it is early days for figuring out how to assess new types of scholarly output

Cultural Change

- Need for advocacy
  - *up* to the ARC, universities and other gate-keepers to acknowledge new kinds of scholarly output
  - among peers to improve data creation processes and deposit in appropriate repositories

Your feedback

- Join the discussion list
- Slides, handout and a summary on the PARADISEC blog (http://paradisec.org.au/blog)
- Please add comments