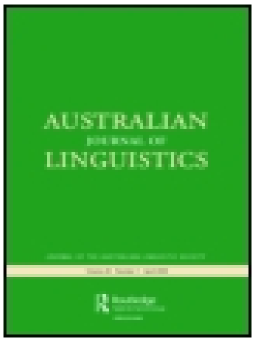

Australian Journal of Linguistics

ISSN: 0726-8602 (Print) 1469-2996 (Online) Journal homepage: http://www.tandfonline.com/loi/cajl20

Assessing Annotated Corpora as Research Output

Nick Thieberger, Anna Margetts, Stephen Morey & Simon Musgrave

Australian Journal of Linguistics, 2015
http://dx.doi.org/10.1080/07268602.2016.1109428

Routledge
Taylor & Francis Group

Assessing Annotated Corpora as Research Output*

Nick Thieberger, Anna Margetts, Stephen Morey and Simon Musgrave

University of Melbourne, Monash University, La Trobe University and Monash University

(*Accepted 24 June 2015*)

The increasing importance of language documentation as a paradigm in linguistic research means that many linguists now spend substantial amounts of time preparing digital corpora of language data for long-term access. Benefits of this development include: (i) making analyses accountable to the primary material on which they are based; (ii) providing future researchers with a body of linguistic material to analyse in ways not foreseen by the original collector of the data; and, equally importantly, (iii) acknowledging the responsibility of the linguist to create records that can be accessed by the speakers of the language and by their descendants. Preparing such data collections requires substantial scholarly effort, and in order to make this approach sustainable, those who undertake it need to receive appropriate academic recognition of their effort in relevant institutional contexts. Such recognition is especially important for early-career scholars so that they can devote efforts to the compilation of annotated corpora and to making them accessible without damaging their careers in the long-term by impacting negatively on their publication record. Preliminary discussions between the Australian Linguistic Society (ALS) and the Australian Research Council (ARC) made it clear that the ARC accepts that curated corpora can legitimately be seen as research output, but that it is the responsibility of the ALS (and the scholarly community more generally) to establish conventions to accord scholarly credibility to such research products. This paper reports on the activities of the authors in exploring this issue on behalf of the ALS and it discusses issues in two areas: (a) what sort of process is appropriate in according acknowledgment and validation to curated corpora as research output; and (b) what are the appropriate criteria against which such validation should be judged? While

[*] We gratefully acknowledge the input received from colleagues at the Australian Linguistic Society conferences at which aspects of this proposal were discussed. Two anonymous reviewers provided very useful comments.

the discussion focuses on the Australian linguistic context, it is also more broadly applicable as we will present in this article.

Keywords: Primary Data Curation; Citation of Data; Valuing Collections

1. Introduction

Across disciplines, there has been a clear international movement to *recognize* primary data as a valuable scholarly output of the research process. Ball and Duke (2012) note the importance of publishing primary data for verification of scientific analyses and it has become commonplace in science disciplines to publish articles with references to datasets accessible via repositories.[1] Indeed, a number of journals will not publish articles unless the underlying data have been made publicly available.[2] Costello (2009) summarizes the many advantages of accessible data, scientific and social, as well as canvassing some of the reasons that make scholars reluctant to publish their data. One of these is the lack of recognition given to the activity of making data accessible. The solution advocated by many in the scientific community (Costello 2009) is to establish procedures for the formal publication of data which include some form of peer-review (Lawrence *et al.* 2011), so that data published in this format are treated on a par with traditional academic output.

Data accessibility is integral to developing the highest research standards in linguistics and other disciplines. Paired with advances in digital media, accessible corpora of annotated language data not only allow for verification of current analyses; they will, in time, provide answers to as yet unknown research questions, as well as providing a cultural record of value to the broader community.

However, the compiling of data corpora, their annotation and dissemination comes at a price—investing time in these activities rather than in traditionally valued forms of publication such as journals and books. For all researchers, but particularly those in their early and mid-career, this is a formidable price to pay when track-record is measured first and foremost in publications (see, e.g. Borgman 2008: 125). Thus there is at present a strong disincentive for researchers to continue to annotate, curate and archive their data beyond the minimum required for their current research focus because of the considerable investment of time and effort required. This means that our current practices around data treatment may jeopardize future access to these valuable cultural artefacts.

Increasingly, funding agencies require primary data to be preserved as part of the research process. This is evident, for example, in the discussion of managing and sharing research data in Beagrie *et al.* (2010), Corti *et al.* (2014) or the detailed report by the USA's National Science Foundation (NSF Task Force on Data Policies

[1] In this paper we use the more generalized term 'repository' to include archives, such as the specialist linguistics archives discussed below.

[2] See for example PLOSone: http://www.plosone.org/static/policies.

2011) and its recent public access plan.[3] Tenopir *et al.* (2011: 1), in a study of data sharing by scientists, found that 'barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves'. As already mentioned, the time required to prepare the data for publication, together with the lack of recognition of such data as examples of scholarly output, are major impediments to making primary data available (see Lawrence *et al.* (2011) and Kratz and Strasser (2015) for discussions of this in relation to the sciences).

There is increasing 'top-down' pressure from government, research institutions and funding agencies who are introducing requirements for research data to be placed in publicly-accessible repositories. At the same time many researchers themselves are moving to place value on collections of primary data by, for example, listing collections of primary records on their CVs along with traditional research outputs. However, without discipline-specific assessment on what constitutes an appropriately annotated dataset and without a process of formal recognition of such annotated corpora as research output, the 'top-down' process will fail in what it aims to achieve. If researchers do lip-service to an administrative requirement rather than strive for high quality to satisfy peer-review standards the outcome will not be transparency of the research process and safeguarding of unique data. The proposal presented here aims to establish processes which communicate between these 'top-down' and 'bottom-up' approaches in a meaningful way.[4]

In this paper we use the terms *collection* and *curated corpora* interchangeably to mean a particular type of dataset, namely annotated corpora of text data transcribed from recordings that are included as a part of the dataset. Corpus linguistics has a long tradition (cf. McEnery 2006), and some of the established practices are akin to those which are discussed here. However corpus linguistic methods have not typically been applied to the world's small languages and mostly are not easily applicable. There are several features which distinguish such corpus linguistic datasets from the collections discussed in the current proposal. Datasets in corpus linguistic research typically come from the world's larger languages whose grammar and lexicon have already been extensively analysed.[5] There are therefore typically many language experts (as researchers or research assistants), as well as grammatical and lexical databases and existing conventions which can be drawn on in the annotation of such datasets. By contrast, for the collections discussed in our proposal, the data come from small

[3] http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.

[4] Moves towards seeking recognition for publication of linguistic data began more than a decade ago. See, e.g. http://www.als.asn.au/newsletters/alsnews200205.html#Postgrads, http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation, http://www.als.asn.au/newsletters/alsnews201202.html, http://www.paradisec.org.au/blog/2012/11/counting-collections.

[5] For example, in the *Routledge Handbook of Corpus Linguistics* (O'Keeffe & McCarthy 2010) the word 'indigenous' is not an index item, nor are 'endangered', 'documentation' or 'DoBeS'. The only reference that deals with 'less-studied languages' that we have found in a search of the corpus literature is Ostler (2008), and, due to the paucity of traditionally conceived corpora in these languages, he also includes archival collections of primary records in his discussion.

languages which are under-documented and under-described and typically arise from fieldwork with the language community. There tend to be few resources on the languages available beyond those created by the researchers who compiled the corpora and resources. There is limited access to research assistants who are knowledgeable in the language, literate and computer literate, and the corpora have a significant heritage value in that they preserve unique cultural and linguistic knowledge of a speech community.

In this paper we describe a proposal for establishing a peer-review process for linguistic corpora which is akin to that of traditional publications. We discuss how a peer-review process may be applied to the publication of linguistic data in the form of annotated text corpora, specifically addressing the questions of: (a) what sort of process is appropriate in according recognition to annotated text corpora as research products; and (b) what are the appropriate criteria by which such corpora should be judged? This process will give prestige and career-building recognition to the creators of well-annotated linguistic corpora which are available as resources for other users. This incentive will encourage researchers to engage in such activities to the overall benefit of the discipline, establishing that corpora of annotated texts are not merely epiphenomenal to the research process but are in fact central.

The following section briefly discusses the cross-disciplinary context of this proposal. Section 3 reviews related proposals to award scholarly recognition to annotated corpora that have been advanced. In Section 4 we present our proposal for the process and criteria including a discussion of what constitutes a 'well-formed' collection in the linguistic context.

2. Context

The discussions about data publication have so far been located primarily in the natural sciences, but the issues are cross-disciplinary. The Australian National Data Service, for example, states:

> Better data—better described, more connected, more integrated and organised, more accessible, more easily used for new purposes—allows new questions to be investigated, larger issues to be investigated, and data landscapes to be explored.[6]

The issues have also been raised explicitly in the humanities and the social sciences. In Australia, the *National Research Investment Plan* notes that what they call the 'human domain' (which is made up of disciplines in the humanities)

> Includes the capability to make old and new data discoverable and reusable and to extract greater value from existing corpora that are as varied as statistical data, manuscripts, documents, artefacts and audiovisual recordings. This domain will enable discovery and use of previously inaccessible information, stimulating

[6] http://www.ands.org.au/about-ands.html (25 February 2015).

> connections and synergies and catalysing innovative research. (Australian Research Committee 2012: 62)

In Europe, the Digital Research Infrastructure for the Arts and Humanities (DARIAH) notes that:

> Long-term viability and access to data and resources is one of the most crucial requirements for research within the digital arts and humanities. Robust policies for collection management and preservation of data will ensure that resources are available for future use by researchers, and will allow discovery and sharing through the DARIAH infrastructure. All users of DARIAH should have an interest in the long-term maintenance of the data they create and use. (DARIAH 2010)

Similarly, the European Science Foundation (2011) notes the increasing importance of databases, tools and online services in the humanities, and further that the lack of agreement on standards for metadata and other means of interlinking research material is an impediment to creating richer open data sources. The research data created by humanities scholars are typically of interest and more directly comprehensible to the broader population in ways that much science, technology, engineering and mathematics (STEM) primary data are not. Moreover data from endangered languages must be presumed to be of special interest to smaller language communities and potentially to other researchers in the humanities. There is a great need for long-term institutional preservation and access as recognized by the European Science Foundation (2011: 9) who note that a new academic recognition system must begin to recognize the scholarly value of electronic editions (which we take to include annotated primary texts) and to review them in highly ranked journals; and to evaluate them as research contributions (20).

A recent volume of the *Digital Libraries Magazine*[7] (*D-Lib*) focuses on the importance of publishing primary data. Notable in particular is Callaghan's (2015) work on citation of primary data in the Earth sciences which includes a proposed method for reviewing such collections and a preliminary assessment of the current state of some (we return to this in Section 4 below).

Recognizing annotated texts as scholarly output is in fact already standard practice in other areas in the humanities. For example, in philology and the study of historical manuscripts, the annotation of primary data, as represented by the critical edition, has always been treated as a standard and indeed highly-valued research output—see for example the Tertullian Project[8] with editions dating back to the fifteenth century. Digital critical editions include the Digital Homer project,[9] or the many research projects associated with the Arthur Schnitzler project.[10]

[7] *D-Lib Magazine* 21 (1/2) DOI: 10.1045/january2015-contents.

[8] http://www.tertullian.org/rpearse/eusebius/works.htm.

[9] http://wiki.digitalclassicist.org/Digital_Critical_Edition_Of_Homer.

[10] http://www.arthur-schnitzler.org.

In linguistics there is broad consensus on the value and importance of making data from under-documented and often endangered languages available for scholarly research and of ensuring such data are accessible to future generations of the language communities and researchers from linguistics and other fields alike. This can be seen in the recent development of archives to house such material.[11] This approach is rooted in our growing awareness of the fragility of what UNESCO terms *intangible cultural heritage* (UNESCO 2003). In linguistics, this concern relates specifically to the problem of language endangerment which has preoccupied many linguists over the past two decades (Crystal 2000; Evans 2009; Grenoble & Whaley 1998 inter alia). This concern has seen the emergence of the field of *language documentation* (Himmelmann 1998), that is, the collection as a matter of urgency of language data, and then preserving such data and making them accessible to multiple audiences.

In 2010 the Linguistic Society of America adopted a resolution which supports the recognition of linguistic corpora as research outcomes and the development of frameworks which will allow their assessment relative to more traditional publications:

> […] the products of language documentation and work supporting linguistic vitality are of significant importance to the preservation of linguistic diversity, are fundamental and permanent contributions to the foundation of linguistics, and are intellectual achievements which require sophisticated analytical skills, deep theoretical knowledge, and broad linguistic expertise;
>
> Therefore the Linguistic Society of America supports the recognition of these materials as scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty. It supports the development of appropriate means of review of such works so that their functionality, import, and scope can be assessed relative to other language resources and to more traditional publications.[12]

Several major initiatives in Europe and the US have provided grants specifically for work on endangered languages, including the programme Dokumentation Bedrohter Sprachen (DoBeS),[13] the Hans Rausing Endangered Languages Project,[14] and the joint National Science Foundation and National Endowment for the Humanities programme Documenting Endangered Languages.[15] All of these funding bodies have made it again a key part of their activities that records made in the course of funded research should be preserved and that data management plans be in place to ensure records are well described and structured for long-term access.

[11] There are a number of such archives, represented by the umbrella organization the Digital Endangered Languages and Musics Archives Network (DELAMAN.org).

[12] http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation (viewed 10 November 2014).

[13] http://dobes.mpi.nl.

[14] http://www.hrelp.org.

[15] http://www.nsf.gov/news/news_summ.jsp?cntn_id=110719&org=NSF&from=news, initially from 2004, and as a permanent programme from 2007.

3. 'Reviewing' Annotated Corpora

Although there is now a consensus on the importance of protecting linguistic data and making such data available, there is somewhat less consensus on what part of a corpus of linguistic data should be made accessible and how. While some researchers call mainly or exclusively for well-annotated texts with full and consistent morphological analysis and interlinearization (e.g. Haspelmath & Michaelis 2014) others call particularly for the publication and curation of primary data, such as Hill in response to Haspelmath and Michaelis:

> It is the primary data that is most important in 'documentation' […] Proper historical linguistics cannot be done without extended textual passages; word lists and example sentences is not enough. […] for documentation projects too much effort is spent fussing over things like glossing and time alignment. […] In the interests of science it is getting the data out there that matters most. (Hill 2014)

Recent discussions around acknowledging linguistic data corpora as scholarly output—so far mostly held on blogs, at conferences and topical research workshops[16]—have included a range of suggestions in which similar terms, such as the notion of 'review', have been used in various ways and with the possibility of confusion. It is therefore necessary to define what we are and are not talking about in the current proposal. At least four different ways of acknowledging data corpora as research have been suggested within the ongoing discussion. We will call them here: (a) corpus review; (b) corpus overview article; (c) corpus journal; and (d) peer-review of a corpus. Our focus will be on the last but we will briefly summarize all four approaches here along with their differences and their merits.

The notion of a *corpus review* is similar to that of a book review, i.e. a summary, description and critical assessment of a substantial piece of academic research in an academic journal. Book reviews play an important role in the publicity for books and in gauging how they are received by the scholarly community. Reviews of corpora will work along similar lines. For such reviews, a journal invites experts in the field to write an assessment and review of a corpus to be published in the journal's review section. Reviews will typically include information such as a summary of the data and information available in the corpus, comments on how to navigate it, the reviewers' assessment of strengths and weaknesses of the corpus, and who they see as likely users of the corpus. Such a review can also include comparisons with other corpora, e.g. on related languages, or a discussion of how the corpus interacts with other resources available for the language, e.g. a published grammar or dictionary which draws on corpus data.

Corpus reviews will be published in journals specializing in particular language groups or geographic areas or those dealing specifically with areas like language

[16] E.g. at the workshop Potentials of Language Documentation: Methods, Analyses, Utilization held in Leipzig in 2011 (Seifart *et al.* 2012), by Margetts *et al.* (2012), as well as on blogs such as: http://www.paradisec.org.au/blog/2012/11/counting-collections.

documentation, or whose audiences are likely users of corpora, such as typologists or researchers investigating narrative structure. Such review articles are already an established practice in journals like the *International Journal of Corpus Linguistics*[17] or *Ride*[18] which is dedicated to reviews of digital editions and resources. The journal *Language Documentation & Conservation* has created a new section for this purpose and corpus reviews are planned for future editions.

A second way to provide academic credit to corpus compilers is what we will call a *corpus overview article*,[19] written by the corpus compilers about the corpus, published in a peer-reviewed journal. Such articles will include the same kind of information as corpus reviews discussed above, e.g. a summary of the content and organization of the corpus, how to navigate it, comparison to other corpora, or discussion of how it interacts with other published language resources. Such articles provide the corpus compilers with a piece of traditional research output in the form of a peer-reviewed article which can be set as a standard reference when working with and citing corpus data, similar to the referencing conventions of the CHILDES and other Talkbank databases.[20] A current example of such an overview article is Salffner (2015) on the Ikaan language corpus.

Corpus reviews and corpus overview articles can play an important role in providing guidance to potential users in how to navigate and use corpora. They would provide a form of peer appraisal of the corpus (through the review or article itself, their citation index, and through increasing the number of hits to the corpus). Such information can then, in principle, be cited to demonstrate the compilers' track-record for funding applications, promotions, etc. However, overall, we see the benefit of such publications primarily in providing feedback to potential corpus users. The direct, practical benefit to the corpus compiler is relatively limited and this does not seem an effective solution to the issues raised above concerning data accessibility and career-building.

Within the aims of language documentation, it is quite possible that the person who assembles a corpus might not write papers or books about all the data they collected. As a concrete example, if a student was working on a documentation project for a PhD, they might collect data on several languages in one geographic region, assemble the data from the various languages into corpora, but then put their future scholarly efforts into describing and analysing one language only. In contrast, it is unlikely that a corpus linguist would assemble a corpus, publish an overview of it as an article, but then never produce papers or chapters based on the analysis of the data. The value to the broader research community in the language documentation is clear, but we argue that the academic rewards are not yet in place to support what is currently an altruistic practice of data preservation.

[17] https://benjamins.com/#catalog/journals/ijcl.

[18] http://ride.i-d-e.de/.

[19] Lawrence *et al.* (2011: 18ff) call this 'publication by proxy'.

[20] We are grateful to an anonymous reviewer for pointing out that articles of this type are normal practice in the field of corpus linguistics.

3.1. Corpus Journal (≈ Corpus as Article)

A third proposal has been made recently by Haspelmath and Michaelis (2014).[21] What they describe as a *corpus journal* is different in nature from the two approaches discussed above in that they propose that actual corpora be published, rather than the reviews and descriptions of such corpora. In this way their proposal goes a step further towards meeting the goals of both data accessibly and career-building.

> By publishing an annotated corpus, in a book-series-like or journal-like corpus-publication outlet […], the author gets credit, just like she gets credit for any other (book or) journal publication […]. Moreover, once a corpus is published, it is much more likely that it will actually be used […]. (Haspelmath & Michaelis 2014)

They maintain that

> Only publication in a selective peer-reviewed journal or series can create the kind of prestige that is necessary for career-building […].

In addition, Haspelmath and Michaelis suggest that creating prestige for this type of scholarly output will provide a powerful incentive for researchers to continue to process text data and to produce high-quality annotations. With the prospect of publishing the actual data, enhancing a corpus will become a viable time investment for the compilers rather than standing in direct competition to traditional publication efforts.

The peer-review process would include consistency checks of annotations and translations. In terms of the expected annotation status of corpora to be published, Haspelmath and Michaelis suggest the following as obligatory components:

(i) the primary text
(ii) an analysed version with morpheme breaks (if the language is morphologically complex)
(iii) glosses/annotations for grammatical categories and/or parts of speech
(iv) translation into English
(v) metadata

Corpora could optionally include further features, such as,

(vi) translations into languages that users may find convenient for various reasons, e.g. lexifiers for creole languages, major national languages (Russian, Chinese, Spanish etc.)
(vii) sound and video files
(viii) time-alignment
(ix) other annotation tiers (intonational information, reference tracking, IPA, etc.)

[21] Their proposal is specific to linguistics; but Lawrence *et al.* (2011: 21–23) discuss such possibilities under the label ‘overlay data publication’.

Such publications would result in electronically accessible, well-annotated corpora which are easily citable and which would be attractive resources to other researchers, such as typologists, but also to the language community.

There are many practical questions still to be addressed, such as what is considered an appropriate size of a corpus to be published in this way. For example, if the publication were to receive the acknowledgement equivalent to a single journal article it will probably be appropriate to submit different parts of a larger corpus as separate publications. Alternatively, and this is raised by Haspelmath and Michaelis, differences in corpus size could be addressed by having both a corpus journal and a book-series-like publication outlet. Depending on size, a corpus would then be recognized as a peer-reviewed article or as a monograph. It is also debatable whether translation into English should be mandatory or whether it could be replaced by other languages of wider communication. Another question is whether it is appropriate, as suggested by Haspelmath and Michaelis, to allow the publication of annotated transcripts without linked media for those languages or those texts where audio/video media are in principle available. It has long been acknowledged that transcripts are a subjective, selective and theoretically informed representation of linguistic data and constitute a process of analysis rather than a neutral form of representation (apart from Ochs' seminal 1979 paper see, e.g. Bucholtz (2007) and Margetts (2009)). Since, technologically, we are now able to provide recordings of the actual speech event via linked media, we maintain that published corpora should be required to do so where possible, that is, in Himmelmann's (2012) terms, providing raw and primary data along with structural data.

As Haspelmath and Michaelis point out, what they propose is different in several ways from making corpora accessible through repositories—an approach which we address in more detail below. Some of these differences can be seen as strengths of their proposal, others we see more critically and cautiously. An advantage a corpus journal has over repository-based corpora is that annotated texts published in a journal are more easily accessed and viewed. Despite recent efforts, most linguistics repositories have not reached a truly user-friendly interface and they are commonly difficult to navigate. Also, repositories typically include texts at different states of annotation along with completely unannotated data. This can make it difficult to find those texts which are sufficiently annotated for interested users to work with. By contrast, a corpus published in a corpus journal would be composed of texts with a consistent level of annotation available at a single location.

On the other hand, corpora published in this way do not constitute 'live' data; rather they represent a snapshot of a level of analysis current at the time of publication, and which cannot be updated. This is of course the case for any published analysis but in the case of corpus data we see this as more problematic. Changes in the ongoing analysis of text data must be expected to be the norm rather than the exception, and so a published (fixed) corpus, may soon be out of date.[22] In our view, greater benefit will come from an online corpus which can be incrementally improved over time as the analysis develops.

Furthermore—and this is one of the main reasons why we view the Haspelmath and Michaelis proposal with reservations—we agree that well-annotated data need to be preserved and to receive career-building acknowledgement but there is also a place for recognition of the curation of raw, little- or un-annotated primary data provided there is accompanying well-structured metadata.

The fourth proposal that has been made for providing academic recognition for corpus compiling and annotation is what we have called the *peer-review of a corpus* and which we conceive as being parallel to the peer-reviewing process of traditional research outputs such as books or journal articles. It is in many aspects akin to the proposal made by Haspelmath and Michaelis with the crucial difference that we conceive of 'publication' as taking place within the repository, therefore not only providing acknowledgment for the compiling of corpora but also their archival curation.

The processes involved in peer-review and publication of a corpus in this sense will be our focus in the remainder of this article. We present a proposal of how such a review process could work in practice, including establishing criteria to assess a corpus, and developing conventions to accord scholarly recognition similar to peer-review for conventional types of scholarly output.

4. The Proposal

In this section we describe our proposal for the assessment criteria to be applied in the review process. The proposal is based on the Australian context, and assumes that the ALS is the appropriate body to take responsibility for assessing data corpora. But we believe it is applicable and adaptable to the research context in other countries. The work of reviewing corpora outlined here will be part of normal academic 'service', in the same way as are grant and article reviewing currently.

4.1. The Review Process

We propose a process which would closely parallel the peer-review process applied to traditional publications and that should involve the following steps:

(1) Once a corpus has been uploaded onto the repository website, and has met the requirements of that repository, it is submitted by its creators to the ALS with a request for review of all, or a specified part, of the corpus.

[22] Of course published articles face similar challenges. For example, whilst scholars presumably submit articles on a particular grammatical feature when they believe that the analysis is complete, further research may require the findings of earlier articles to be revisited. A key difference here is that in the case of corpora we would not expect researchers to refrain from publishing them until they felt all aspects of the analysis of a given language were complete. In this sense corpora can be expected to be more open to change than some more traditional academic publications.

(2) Refereeing by an anonymous panel representing the ALS.
(3) Report by the panel to the creators, possibly with suggestions for improvements. The report could recommend one of the following:
  (a) accepting the submission as is or pending minor revisions;
  (b) resubmission after major changes or additions to the corpus;
  (c) rejecting the submission.
(4) Response by the creators to the panel, possibly including resubmission of the corpus after revisions based on the panel's recommendations.
(5) Announcement of the successful review of the corpus under the auspices of the ALS, for example in the ALS newsletter or website, or the *Australian Journal of Linguistics*. Publication of the announcement would constitute recognition of the corpus (or part thereof) as a research product.

Lest it be thought that this is too onerous a process which will not be implemented due to its complexity, we refer the reader to an already functioning review process in place for digital repositories, the European Data Seal of Approval.[23] The review places a value on the work of a repository, as recognized by an independent process. This value to the scholarly community will be the motivation for scholars to participate in both submitting a collection for review, and in acting on the review panel.

4.2. Assessment Criteria

In the same way that a range of criteria are applied when assessing traditionally published work, the following criteria can be used in assessing a corpus. We consider the criteria under three broad headings, namely *accessibility*, *quality* and *quantity*.

4.2.1. Accessibility criteria

4.2.1.1. Repository, long-term curation, accessibility of corpus. Accessibility is a fundamental criterion. A *sine qua non* for our purposes is that a corpus must be housed in a repository before it can be considered for evaluation. The repository must have a commitment to provide long-term curation and access to the corpus, which includes creating a persistent identifier and a citation form for items within the corpus. The repository should provide access to metadata and a clear means for accessing primary data with clearly stated access conditions that may include restrictions. Taken together, these conditions are similar in effect to the standard suggested by Callaghan (2015) that a data source should have a single web page or easily accessible file which provides basic information about the data. A corpus that is not publicly available with long-term access would not be considered for evaluation. Lawrence *et al.* (2011) have suggested that the kind of publication we are considering here (termed

[23] http://datasealofapproval.org/en/.

Publish ‘with a capital P’ by them), should be defined as to ‘make data as permanently available as possible on the Internet’.

Ideally, after a corpus has received recognition as scholarly output, it will be identified as such within the repository. Furthermore, while this model will allow for changes and improvements to the ‘published’ corpus to be made over time, the version which has received recognition would need to remain accessible in some form, just as earlier editions of books remain on library shelves. This is essential so that future users can access the ‘published’ corpus in exactly the same form as it was assessed.

4.2.1.2. Data format, non-proprietory. The files in the corpus have to be in formats that are non-proprietary, i.e. they need to be accessible to any user without the need to purchase specific software. Any structured elements relating to the software need to be documented (e.g. Toolbox files may require additional .typ and .lng files in order to be readable; or Elan files may have named tiers that need to be explained). This condition provides some assurance for continuing access to data but not all non-proprietary formats are formats which are likely to remain portable in the sense of Bird and Simons (2003). Portability implies the use of standardized or well-defined file formats and the review process should recognize the greater value of resources which are in portable formats.

4.2.1.3. Access restrictions to parts of a corpus. Only those contents of a corpus that are available to other users can be considered in the review. Corpora may include materials to which access is restricted, e.g. to a specific group of the language community where data may become publicly accessible only after a certain time period. Such data will not be assessed in the panel’s review and will not count towards the quantity assessment. However, if such data are accompanied with good metadata descriptions and clear information about when, to whom, and under what circumstances access will be granted then such protected data may be considered as increasing the quality of the corpus as discussed below.

Some linguistics repositories require users to register in some form, as for example the Endangered Languages Archive (ELAR) which has a registration process.[24] Provided that such registration processes and codes of conduct are open to all potential users free of charge, and do not present insurmountable barriers especially to members of language communities, then the fact of having to register would not be regarded as an impediment to open access. Such processes of registration are certainly far less onerous for language communities than having to travel to a library to inspect a printed book.

The aspects of accessibility discussed here closely parallel those treated in a recent paper which reviews some data sources in atmospheric science. Callaghan (2015) identifies four questions around accessibility which she treats as purely editorial,

[24] http://www.elar-archive.org/using-elar/registering.php.

that is, failing to meet any of these standards would mean rejecting the data source immediately without further review:[25]

1. Does the dataset have a permanent identifier?
2. Does it have a landing page (or README file or similar) with additional information/metadata, which allows potential users to determine that this is indeed the dataset they are looking for?
3. Is it in an accredited/trusted repository?
4. Is the dataset accessible? If not, are the terms and conditions for access clearly defined?

4.2.2 Quality criteria

Assessment of the quality of a corpus will be based on a number of criteria, including: (a) the nature and amount of contextual and background information; (b) structure of the corpus; (c) quality of metadata; (d) nature of linguistic annotations to the primary data; and (e) structural linking between raw data and their annotations.

4.2.2.1. Contextual background information and corpus structure. Contextual information about the corpus includes background information on how the corpus came into being, who prepared it and who funded it, what projects were part of the research.[26] It should also include an overview of what type of material the corpus contains and by what principles it is structured (e.g. by text type, by dialect, speaker groups, geographic location, etc.). This category would also include any materials that would be necessary for or would aid working with the data, such as lists of abbreviations, information on the orthography and its relation to the sound system, as well as resources such as a grammar sketch and word lists or dictionaries.[27] Such information will be vital for third parties to navigate and use the corpus. Finally, contextual information could also include ethnographic and geographic information about the speaker community.

[25] It is worth noting that in Callaghan's empirical test, these conditions turned out not to be trivial. She selected seven data sources, all of which passed test 1 because the tool used for selection was based on Digital Object Identifiers (DOIs), which are unique identifiers and remain fixed for the lifetime of the object that they refer to. However, two of these seven failed at least one of the other tests, and were rejected at this editorial stage. Another source of data only would have progressed to review on a generous interpretation of test 2 as it had a README file in raw .xml without stylesheet instructions, and yet another had confusing information about access. In other words, these editorial tests applied rigorously would have ruled out more than half of Callaghan's (admittedly small) sample.

[26] See for example the descriptions of CHILDES corpora at http://childes.psy.cmu.edu/manuals/.

[27] We envisage that in the future online grammars and well-structured lexical databases themselves will be counted as research outputs in the same way as we discuss here for text collections. While they are beyond the focus of the current discussion their accessibility along with the text data would be counted as raising the quality assessment of a text collection.

4.2.2.2. Descriptive metadata. The materials should be sufficiently described, that is, have descriptive *metadata*, to identify what the contents are, e.g. texts, media, vocabulary, information about speakers.[28] Further metadata such as content keywords, notes on the recording situation, time and location, information about any other material recorded during the same recording session, etc. would also be considered. Metadata will also be assessed in terms of whether they are easily accessible and searchable.

4.2.2.3. Raw, primary and structural data and their linking. The core material to be reviewed will be a combination of raw data (recordings) and primary data (transcriptions) combined with structural data (annotations). Recordings should be transcribed and translated into language(s) of wider communication and provided with time-alignment to create text–media linking between the recordings and the transcriptions. A substantial part of the corpus will be expected to be annotated with morphological breakdown and interlinear glosses.

Additional types of annotations will be considered to raise the quality rating of the corpus. Examples of such annotations include marking parts of speech, intonation contours, pause length, speech-accompanying gestures, tagging of syntactic constructions, referentiality features, detailed presentation of overlapping speech by different speakers, etc. Ideally, the corpus would also include appendices, or sketch grammars, explaining the theoretical basis for the analysis, for example criteria for parts of speech, overview of the syntactic constructions marked, and so on.

Additional raw data or data that are, for example, transcribed and translated but not further annotated may supplement the core corpus and will also increase the quality ranking. Un- and little annotated recordings can be of high value if they accompany transcribed and translated and interlinearized media in the same language which provides a model to potential users for further annotation and analysis.

Transcribed language materials without their original recordings can, in some cases, be considered, but typically only in circumstances where the raw data cannot reasonably be expected to be included, e.g. in the case of analysis of written sources, or of historical data, where there were no recordings or the recordings have been lost.

Corpora of transcribed and translated recordings without any further annotations relating to morphological breakdown and interlinear glossing can be submitted but will be considered to be of comparatively low quality ranking given that it will be considerably more difficult for third parties to work with such corpora and given that they represent a lesser degree of linguistic analysis and annotation.

Corpora which deviate from the described format of a core of raw data with time-aligned transcription, translation and basic morphological annotation and which provide only some of these features can in some cases be considered but would again be rated lower in terms of their quality criteria.

[28] Speakers may be anonymized and in this case information about naming procedures should be provided.

A corpus consisting exclusively of raw untranscribed data will not be considered eligible for review. Simply amassing many hours of recordings is not considered to be scholarly output.[29]

4.2.2.4. Content. The content of the corpus should ideally represent a range of speakers across age, gender and other relevant social variables, and a good range of text types (narratives, procedural, hortatory texts, written, songs, etc.) However, non-random specializations are expected and valued especially if they are paired with specialist annotations of the data. Content will therefore be assessed relative to the focus and expertise of the projects and researches who compiled the corpus, and also to opportunity. A musicologist will record more musical performances than will a project focussing on child-directed speech, or investigating syntactic phenomena. In other cases a single researcher may document a whole group of languages never previously recorded, and as a result of considerable diversity in the data there will be less depth to the analysis and level of annotation. Similarly, researchers working with highly endangered or moribund languages may have limited opportunities to record a diverse range of speakers or of discourse types.

4.2.2.5. Amount of data. While the quality criteria listed above are the most important metrics, they need to be applied together with an assessment of the quantity of material and the effort that has gone into producing the corpus. All other things being equal, a greater amount of analysis should be recognized by the review process as having greater academic value. Quantity can be evaluated in terms of the amount of text, the number of items, and the diversity among those items.

4.2.2.6. Combined scale of criteria. In the end there will be a complex interaction of the criteria established above that will be described in a review. As laid out above, the default submission to the review panel will consist minimally of recordings, plus their transcriptions, translations and annotations. Corpora may deviate from this standard and provide less or additional annotations and this will affect the quality ranking downwards or upwards.

A corpus must reach a minimum threshold in its ranking in order to be considered published with endorsement from the review committee.

Finally, (as discussed earlier in Section 1), we need to distinguish (a) language documentation corpora for which the linguist is the only researcher with knowledge of the language from (b) corpora for languages for which there are many research assistants available.

[29] Much legacy data fall into this category and are doubtless extremely significant from a heritage perspective, and as the basis for future scholarly work. Thus legacy data merit archiving and curation, but are not considered as a collection eligible for review.

4.2.3. Version control and editions

It will be necessary to develop a system for allowing improved versions of a corpus to be acknowledged, much as we currently allow for new editions of books that improve and adapt over time. Updated and revised corpora re-submitted for review will be assessed for the degree of difference they display from an earlier assessed version. Just as with books, however, it is also necessary to maintain an accessible version of the earlier edition.

This is problematic, because the corpora we are talking about are by their nature works in progress. It is relatively easy to update, correct and enlarge an annotated corpus or a dictionary, and to add new recordings, and this is a very positive thing for our field. Online publication means quick and easy access is provided to much larger numbers of scholars, community members and others, but it also means that we will inevitably be called upon to assess a work that is not in its final form. In this sense we may have to adopt the methodology of the hard sciences in treating the corpora we are asked to assess as a step along the way of building a larger dataset.[30] An updated version would count as an additional output if it makes a significant contribution beyond that of the original version.

Since corpora are works in progress, a researcher may decide to submit just one part for review. This may occur where a scholar is working on multiple varieties of related languages, but only some are judged sufficiently well analysed and understood to be reviewed. It may occur where some text genres are better understood than others, or have been more deeply analysed. It may also be that the review process itself suggests that only part of the corpus submitted for review was of sufficient quality that it should be recognized as scholarly output.

We therefore need mechanisms for splitting an archived corpus for publication purposes. This, we suggest, can and should be handled at the level of archival curation with the organization of material in a repository being sufficiently structured for sub-groupings to be identified within a deposit. For example in a DoBeS corpus, the researcher can decide on a structure of 'nodes' and this could be used to define the part of a corpus which was to be published. This approach would also have consequences for citation of the different parts, but again this is an issue which can be handled as part of archiving.

4.2.4. Some examples

We will now illustrate our approach by comparing two hypothetical corpora (which are however based on actually existing datasets). Corpus 1 is a small collection of materials from an under-researched language. The collection is stored in a recognized repository and meets basic accessibility criteria (as described above). However, the

[30] For example see the Long Term Ecological Research Network Data Access Policy http://www.lternet.edu/policies/data-access, or the NSF report, Today's Data, Tomorrow's Discoveries, http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.

corpus is limited in various ways. All of the data were collected from a single speaker; the corpus therefore consists entirely of monologic narratives and it is quite small. Each narrative is accompanied by a recording; the transcriptions include timings which align the text sections with these recordings. All of the texts are fully transcribed and annotated with interlinear glossing and free translation. The annotations were all prepared by a single linguist (working with the speaker who was recorded) and the level of consistency is high. No background information about the language and its speakers is provided; nor is there a lexicon or a sketch grammar.

Corpus 2 also contains materials from an under-researched language, but it is at once much larger than Corpus 1 and also rather less organized. This corpus is held in a recognized repository with a permanent identifier. However, the identifier resolves to a catalogue page which only lists the resources and which does not provide the user with easy access to relevant data. There is an introduction to the corpus, but it is listed in among the other files and it is not immediately obvious that it is present. The corpus contains recordings of many speakers in a variety of speech situations: monologues, dialogues and multi-party events of both formal and informal character. In addition to these more or less spontaneous materials, Corpus 2 also contains a group of recordings of elicited speech intended to illustrate the phonetic contrasts present in the language. Along with the raw data, Corpus 2 includes transcription and annotation for a portion of the materials, approximately 30% of the total recordings. Annotations were added by several linguists and the level of detail varies from one text to another. Also, some points are not annotated consistently across texts (although the scholars responsible note that this is the case in their introductory materials). Only a small proportion of the transcribed texts have time-codes included. There are also ethnographic notes about the speech community and notes on grammatical points of interest. These latter do not claim to be a sketch grammar, but they do improve the usability of the corpus.

We anticipate that the result of peer-review of Corpus 1 would be a recommendation to accept it as a publication as it stands. The review panel may suggest that providing more introductory material would enhance the value of the resources, but we suggest that this would not be an impediment to recognition of this corpus. The data are useable in the form in which they are presented and they make a contribution to research. On the other hand, we suggest that the review of Corpus 2 would give a 'revise-and-resubmit' verdict. The value of at least some of the resources is clear, but accessibility could be considerably improved by signalling more clearly where the essential introductory information was located. We think that the review panel might also suggest that either the level of consistency in annotations should be improved, or alternatively, that the scholars submitting the corpus for review should consider excluding some portions of the material from the version to be published in order to lessen this problem.

In both cases, the important judgment which has to be made by the panel is the same as the judgment made in reviewing the currently accepted types of publication: does this work make a contribution to the field?

4.2.5. Equivalence to traditional scholarly outputs

In terms of the level and nature of scholarly recognition awarded to peer-reviewed corpora of different sizes, there are in principle two approaches. A corpus could be deemed *equivalent* to a specific type of traditional scholarly output, and, e.g. according to its size and quality, be considered equivalent to either a journal article or a monograph. Alternatively, one could award professional recognition without comparing to traditional output and instead establish a new category of linguistic scholarly output, somewhat akin to 'non-traditional research outputs' in Australia as established in some disciplines, such as performing and visual arts.

Both approaches have their merits and their problems. We would like to suggest that this issue is less important in immediate considerations and may well prove to be a distraction from the pressing issue, which is to ensure that recognition is given at all. The practical implementation of such recognition will need to be discussed in direct consultation with the relevant agencies.

5. Conclusion

There are a number of issues that remain for future discussion. The focus and scope of the present discussion is to some extent determined by the authors' discipline and research background. The discussion is grounded specifically in the context of the description, analysis and documentation of endangered languages. This means our focus has been on text corpora documenting un- or under-described endangered languages which are commonly minority languages. However, we feel that the discussion can and probably should be extended to other types of corpora in the humanities and social sciences, including databases from psycholinguistic research (e.g. on first or second language acquisition), conversational analysis, oral histories and other areas, and we anticipate discussions with colleagues working in these fields.

We have well established systems for recognition of publications as scholarly output —including libraries and publishers—which have taken some 500 years to develop, however no comparable system yet exists for primary or raw data. Borgman (2007: 225) notes that 'only a few fields have succeeded in establishing infrastructures for their data, and most are still fledgling efforts'. In linguistics we are clearly ahead of the game, having established repositories and metadata schemas and developed some consensus about the need to preserve language recordings. We can now build on that consensus to recognize this priceless research material and to work to persuade the relevant authorities that creating annotated primary datasets is in itself a scholarly activity. There will be a need for advocacy both *up* to government research councils, universities and other academic institutions to acknowledge new kinds of scholarly output and also *among* peers to improve data creation processes and deposit in appropriate repositories.

ORCID

Nick Thieberger http://orcid.org/0000-0001-8797-1018
Stephen Morey http://orcid.org/0000-0001-6121-684X
Simon Musgrave http://orcid.org/0000-0003-3237-9943

References

Australian Research Committee 2012 *2012 National Research Investment Plan* Canberra: Department of Industry, Innovation, Science, Research and Tertiary Education. Available at: http://industry.gov.au/research/Documents/NationalResearchInvestmentPlan.pdf

Ball A & M Duke 2012 *How to Cite Datasets and Link to Publications* Edinburgh: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/how-guides#sthash.FmNPiXsU.dpuf

Beagrie N, BF Lavoie & M Woollard 2010. *Keeping Research Data Safe 2*. HEFCE. Available at: http://www.webarchive.org.uk/wayback/archive/20140-1405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Bird S & Simons G 2003 'Seven dimensions of portability for language documentation and description' *Language* 79: 557–582.

Borgman CL 2007 *Scholarship in the Digital Age* Cambridge, MA: MIT press.

Borgman CL 2008 'Data, disciplines, and scholarly publishing' *Learned Publishing* 21(1): 29–38. Available at: http://doi.org/10.1087/095315108X254476

Bucholtz M 2007 'Variation in transcription' *Discourse Studies* 9(6): 784–808. http://doi.org/10.1177/1461445607082580

Callaghan S 2015 'Data without peer: examples of data peer review in the earth sciences' *D-Lib Magazine* 21(1/2). http://doi.org/10.1045/january2015-callaghan

Corti L, V Van den Eynden, L Bishop , & M Woollard 2014 *Managing and Sharing Research Data: a guide to good practice* London: Sage.

Costello MJ 2009 'Motivating online publication of data' *BioScience* 59(5): 418–427. http://doi.org/10.1525/bio.2009.59.5.9

Crystal D 2000 *Language Death* Cambridge, UK, New York, NY: Cambridge University Press.

DARIAH 2010 Collection ingest, management and preservation policy. Digital Research Infrastructure for the Arts and Humanities. Available at: http://subugoe.github.io/website-DARIAH-EU/indexb808.pdf?option=com_docman&task=doc_download&gid=408&Itemid=200

European Science Foundation 2011 *Research Infrastructures in the Digital Humanities* (No. 42). Strasbourg: European Sceince Foundation. Available at: http://www.esf.org/fileadmin/Public_documents/Publications/spb42_RI_DigitalHumanities.pdf

Evans N 2009 *Dying Words: endangered languages and what they have to tell us* 1st edition Chichester, UK: Wiley-Blackwell.

Grenoble LA & LJ Whaley (eds) 1998 *Endangered Languages: language loss and community response* Cambridge, UK, New York: Cambridge University Press.

Haspelmath M & SM Michaelis 2014, March 11 *Annotated corpora of small languages as refereed publications: a vision*. Available at: http://dlc.hypotheses.org/691

Hill NW 2014, March 13 *Comment on "Annotated corpora of small languages as refereed publications: a vision."* Available at: http://dlc.hypotheses.org/691

Himmelmann N 1998 'Documentary and descriptive linguistics' *Linguistics* 36(1): 161–196.

Himmelmann NP 2012 'Linguistic data types and the interface between language documentation and description' *Language Documentation & Conservation* 6: 187–207.

Kratz JE, & C Strasser 2015 'Researcher perspectives on publication and peer review of data' *PLoS ONE* 10(2): e0117619. doi:10.1371/journal.pone.0117619

Lawrence B, C Jones, B Matthews, S Pepler & S Callaghan 2011 'Citation and peer review of data: moving towards formal data publication' *International Journal of Digital Curation* 6(2): 4–37. http://doi.org/10.2218/ijdc.v6i2.205

Margetts A 2009 'Data processing and its impact on linguistic analysis' *Language Documentation & Conservation* 3(1): 87–99.

Margetts A, S Morey, S Musgrave, A Schembri & N Thieberger 2012 'Assessing curated corpora as research output: issues of process and evaluation' Presented at the Annual Conference of the Australian Linguistic Society, University of Western Australia.

McEnery T 2006 *Corpus-Based Language Studies: an advanced resource book* Routledge Applied Linguistics. London, New York: Routledge.

NSF (Task Force on Data Policies) 2011 *Digital Research Data Sharing and Management* (No. NSB-11–79). Washington DC: National Science Foundation.

Ochs E 1979 'Transcription as theory' in E Ochs & BB Schieffelin (eds) *Developmental Pragmatics* New York: Academic Press. pp. 43–72.

O'Keeffe A & M McCarthy 2010 'Historical perspective: What are corpora and how have they evolved?' in A O'Keeffe and M McCarthy (eds) *The Routledge Handbook of Corpus Linguistics* London: Routledge. pp. 3–13.

Ostler N 2008 'Corpora of less studied languages' in A Lüdeling & M Kytö (eds) *Corpus Linguistics: an international handbook* Berlin, New York: Walter de Gruyter. pp. 457–483.

Salffner S 2015 'A road map to the Ikaan language documentation' *Language Documentation & Conservation* 9: 237–367.

Seifart F, G Haig, N Himmelmann, D Jung, A Margetts & P Trilsbeek 2012 *Potentials of Language Documentation: methods, analyses and utilization* Honolulu: University of Hawai'i Press. Available at: http://scholarspace.manoa.hawaii.edu/handle/10125/4540

Tenopir C, S Allard, K Douglass, AU Aydinoglu, L Wu, E Read, M Manoff & M Frame 2011 'Data sharing by scientists: practices and perceptions' *PLoS ONE* 6(6): e21101. http://doi.org/10.1371/journal.pone.0021101

UNESCO 2003 *Convention for the Safeguarding of the Intangible Cultural Heritage* UNESCO. Available at: http://www.unesco.org/culture/ich/en/convention/