Corpora and archiving in language documentation, description, and revitalisation

Peter K. Austin
SOAS, University of London
pa2@soas.ac.uk

Abstract

A great deal of work over the past 25 years on documentation, description, and revitalisation of minoritised languages, especially those categorised as endangered languages, has centered around corpus creation and archiving. In this paper, I discuss what can be involved in such work, both for newly collected materials as well as historically existing ones ('legacy corpora'), and critically examine some of the issues and challenges involved in such work. Examples are presented from a range of sources, including the author's work on value-adding in several legacy projects involving Australian Indigenous languages.

Keywords: corpus, archiving, annotation, metadata, Australian Aboriginal languages

1. Introduction[1]

The past 25 years has seen a major growth in interest, both from researchers and language communities, in languages and cultures around the world which are under social, economic, and political pressure from other languages and cultures that are perceived as more dominant and powerful. This is especially the case for endangered languages, whose linguistic ecology involves shifts in language status, functions, use, and acquisition, leading frequently to their dispreference as a regular means of communication and transmission of knowledge and culture (see Austin and Sallabank 2011, 2022; Grenoble 2011).

---

[1] This paper began as lecture notes for a seminar at the *FieldLing Summer School* in Paris, September 2021, and refined in lectures at the *LDSS Language Documentation Summer School* in Viterbo, July 2022. At Moreno Vergari's invitation I have revised and elaborated my lecture notes for publication. I am grateful to Lise Dobrin, David Nathan, and Julia Sallabank for earlier discussion of several of the topics covered, and for questions and feedback from seminar audiences that have improved the resulting paper. None of these people are responsible for any errors. I thank Andrew Garrett, Andy Cowell, and Jorge Labrada for information about the derivative corpora discussed in Section 3 below.

One response from researchers has been the development and application of principles and practices of language documentation, an approach to language study whose central goal is the creation of archiveable corpora illustrating language performances evincing language in use (Woodbury 2011, Austin 2016, Seifart et al. 2018). In the following sections, we discuss various understandings of language documentation, its application to new language research as well as existing (legacy) materials, and some of the challenges that this work presents for researchers and communities, as well as potential users of the corpora thus created.

2. What is a corpus (plural corpora)?

The traditional definition of a corpus is as follows (see also Crystal 1992: 85):[2]

> a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus Linguistics deals with the principles and practice of using corpora in language study. A computer corpus is a large body of machine-readable texts.

Note that there are several important aspects of this approach:

1. an emphasis on written text;
2. the research analysis methods are primarily quantitative;
3. the use of software tools to parse and tag the corpus;
4. a particular collection is justified by a research hypothesis (goals for the project).

For an introduction to the theory and practice of Corpus Linguistics see McEnnery and Hardie (2011).

The creation of a corpus in this approach typically involves collecting a set of texts (either already born digital or by scanning print publications), typically with a defined size (e.g. 10 million words), and content (e.g. Old English literature),. An example is

[2] https://fdocuments.net/document/corpus-linguistics-and-corpora-corpus-corpus-plural-corpora-a-collection.html?page=3 (accessed 2022-09-10)

the British National Corpus (constructed 1980-1990, 1 million words, multiple written genres in British English).[3] Often, there is an attempt to make the sample or collection representative with regards to the research hypothesis; there may also be an attempt to balance the corpus to represent various non-linguistic variables (e.g. x% novels, y% poetry, z% conversation), e.g. frTenTen, which is a balanced corpus of French on the web (current, 10 billion words, covering European, Canadian and African French).[4]

A different approach to corpus creation and analysis was introduced in about 1995 with the development of language documentation (also called documentary linguistics, see Himmelmann 1998, Austin and Grenoble 2007). Gippert, Himmelmann and Mosel (2006: v) define this new perspective as:

> concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties

They argue that the outcome of such an approach is an annotated and translated corpus of representative and multipurpose materials on a language or variety, deposited in a major archive such as TLA/Dobes,[5] or ELAR,[6] with an accompanying apparatus such as a grammatical sketch and cataloguing metadata. In contrast to traditional corpus linguistics approaches, the documentary corpus (what Himmelmann 2012 calls "primary data") should aim to be publicly accessible to a wide audience of users, including members of the speech community, for both ethical and accountability[7] reasons. Within the conceptualisation of the Dobes project, particular data structures and software tools (such as CIMDI,[8] ELAN[9]) were employed to analyse and access the corpus. A broader definition of a documentary corpus is given by Woodbury (2003, 2011) as "transparent records of a language", prepared in such a way that it would be accessible to philologists and others many years into the future (see also Woodbury 2014).

---

[3] See https://www.english-corpora.org/bnc/ (accessed 2021-09-0)

[4] See https://www.sketchengine.eu/frtenten-french-corpus/ (accessed 2021-09-01)

[5] https://archive.mpi.nl/tla/ (accessed 2021-09-01)

[6] https://archive.mpi.nl/tla/ (accessed 2021-09-01)

[7] more recently, there has been a growing emphasis on openness of access and citation of both data and analyses under the banner of 'reproducibility' (see Berez-Kroeker et al. 2018).

[8] https://www.clarin.eu/content/component-metadata (accessed 2021-09-01)

[9] https://archive.mpi.nl/tla/elan (accessed 2021-09-01)

Nathan and Austin (2004) propose that the corpus also needs to be accompanied by a rich conceptualisation of metadata of several types (see also Austin 2006: 93): *cataloguing* — title, speakers, collectors, time and place of recording, language name etc., *descriptive* — information about content, relationship to other resources etc., *structural* — what organisational devices and patterns exist in the document etc., *technical* — performance and preservation information, description of formats etc., and *administrative* — work log, responsibilities, access protocol statements etc. Austin (2013) additionally suggests that corpora should have associated with them meta-documentation, i.e. metadata at the project level setting out the project goals, corpus theory, data collection and analytical methods, stakeholders, ethics (including informed consent), and access and usage rights.

The process of creation of a corpus within documentary linguistics also differs from traditional corpus linguistics approaches, and may involve a range of data collection contexts and methods, each with its own strengths and weaknesses (Lüpke 2009):

1. *elicitation* (interviewing), that typically involves translation (Lsource→ Ltarget, Ltarget → Lsource, sometimes with a lingua franca intervening in both directions) and grammaticality and/or acceptability judgements;
2. *narratives* (telling stories, often monologues of folk stories);
3. *conversation* (involving two or more participants);
4. *experimentation* (using puzzles, games, and other tasks, such as video and image descriptions);
5. *participant observation* (spending time with speakers, observing language use and attempting to use the language oneself, see Dobrin and Schwartz 2016).

Metadata collection and management can be done manually (with pen and paper), though for ease of storage, searching, sharing, and restructuring, electronic representation is preferable. Researchers typically employ one or more of:

1. general office software, with which they create plain text or formatted documents (e.g. Word tables), spreadsheets (e.g. Excel), or databases (e.g. Access, MySQL);
2. dedicated metadata software, e.g. SayMore[10] or CIMDI Maker[11].

[10] https://software.sil.org/saymore/ (accessed 2021-09-01)

[11] https://cmdi-maker.uni-koeln.de/ (accessed 2021-09-01)

Individual archives may have preferences regarding metadata tools and formats, and researchers are advised to check when first designing a project and planning its corpus structure and management. In this context, having a consistent file-naming system and folder-naming system, and applying both rigorously, is highly important for corpus management. For file naming:

1. use only ASCII symbols without punctuation or spaces (hyphen and underscore can be used for character separation, if necessary). If dates are included, use the ISO 8601 standard of yyyymmdd or yyyy-mm-dd, e.g. 2021-09-01 for 1st September 2021;
2. names should contain one and only one period, which precedes the file extension, which is a sequence of three or four characters showing the file type, e.g. .docx, .txt, .wav, .jpg, .eaf;
3. names should be kept short for easy identification and management. Relevant metadata should be represented separately and not be incorporated in file names;
4. ease of name sorting should be taken into consideration, e.g. names such as 2021-09-01_FieldLing_corpus.pptx will be easily sorted sequentially according to date.

For folder management, a rigorously applied hierarchical structure should be established. According to individual researcher preferences, folders could reflect documentation sessions, languages, speakers, or data types. A reliable and strictly applied corpus backup strategy is also an important part of any project (see Austin 2006: 89).

3. Derivative corpora ('legacy materials')

It is rarely the case that first-hand research is carried out on languages or communities that have never been documented before, so typically there already exists material in some form, e.g. in missionary or traveller reports, government records, or from previous linguistic or anthropological researchers. With careful use, these legacy materials can provide valuable information to contemporary researchers and communities, and may assist language recovery or revitalisation (Austin 2017). In some cases (e.g. much of eastern Australia), there are no contemporary fluent speakers of a given variety, and legacy materials are the richest or only sources for description and revitalisation. Sometimes, it is also the case that *in situ* field research in communities is

not possible due to danger from terrestrial phenomena (e.g. earthquakes, floods), violence (e.g. civil war or criminal gangs), or from disease, including pandemics like Ebola and Covid-19.

There are a number of available documentary corpora where researchers have taken various approaches to enhancing (adding value) to existing legacy sources. Some examples are the following:

1. my work with hand-written fieldnote materials collected by Stephen Wurm in 1955-1957 on several sleeping languages from New South Wales and Queensland, Australia, using the *Linguists Toolbox* program to create lexicons, glossed and annotated texts, and structured metadata. Examples are Guwamu (see Austin 2006) and Malyangapa (Austin 2002). The work on these projects involves:
    a. typing up the original fieldnotes, adding structured metadata on sources (speaker, recorder, fieldnote location of sentences), abbreviation definitions, and tracking the date of last edit;
    b. creating a set of analysed sentences with the original phonetic notation, and adding a unique identifier (snum), phonemicization, morpheme glossing, morpheme and word level part-of-speech, free translation into English, notes, links to the lexicon (via unique lexnum identifiers), and links to the abbreviations and sources;
    c. creating a lexicon, with each entry comprising a lexnum (unique identifier), headword, gloss, definition, scientific name, scientific name source, optional picture, semantic relations (synonym, antonym, cf.), notes, cognates, and links to example sentences (using snum, from which the phonemic representation and free translation can be identified).
2. Miwok tape recordings from the 1960s by Sarah Ballard with Catherine Callaghan,[12] transcribed and annotated in ELAN by Andy Cowell into a new corpus deposited in the California Language Archive;[13]
3. Makah recordings by William Jacobsen[14] transcribed and annotated in ELAN by Jorge Emilio Rosés Labrada and Erin Hashimoto into a new corpus in the

[12] http://cla.berkeley.edu/collection/10086 (accessed 2021-09-01)

[13] http://dx.doi.org/doi:10.7297/X2251GC0 (accessed 2021-09-01)

[14] http://cla.berkeley.edu/collection/10028 (accessed 2021-09-01)

California Language Archive.[15] The goals of this project were software and linguistic training for Hashimoto, and provision of more accessible and user-friendly materials for community members in the Makah Language Program (MLP),[16] in agreement with MLP Language Specialist, Maria Parker Pascua. Labrada also wished to acoustically study glottalized resonants in Makah;

4. the ELAR deposit on Tonsawang (Sulawesi, Indonesia)[17] collected 2016-2018 by Tim Brickell (and partially transcribed in ELAN) was analysed by SOAS MA student Rebekah Hayes who extracted 1,408 examples, annotated them using Excel for morphological analysis, grammatical functions, word order, NP type, and case-marking in order to analyse verbal constructions and the distribution of voice markers. Her MA thesis is based on this analysed derived corpus.

4. Archiving

Henke and Berez-Kroeker (2016: 411) argue that:

> It is difficult to imagine a contemporary practice of language documentation that does not consider among its top priorities the digital preservation of endangered language materials. Nearly all handbooks on documentation contain chapters on it; conferences hold panels on it; funding agencies provide money for it; and even this special issue evinces the central role of archiving in endangered language work. In fact, archiving language data now stands as a regular and normal part of the field linguistics workflow.

An archive is a trusted repository with a collection policy and a commitment to appraise the value of materials it receives as a potential deposit, to preserve selected items, to make known their existence, and to enable access to them (or their content, via a catalogue). Archives typically have an online catalogue that presents metadata about their collections, often in a standardized format; some have finding aids to assist users with accessing collections, and all will have access management protocols defining who can use the materials and how they may be used (but see 5.4 below). Many research funders now require that projects have a data management plan and archive their

[15] http://dx.doi.org/doi:10.7297/X2ZW1J3J (accessed 2021-09-01)

[16] https://makahmuseum.com/departments/makah-language-program/ (accessed 2022-10-10)

[17] https://www.elararchive.org/uncategorized/SO_27fb7171-6818-4e10-9d8c-d09554fa43c5/ (accessed 2021-09-01)

materials in a recognised repository.[18]

It is important to emphasise that placing materials on a website is not archiving:

1. websites are volatile and rarely have institutional support like an archive does;
2. files on websites can become obsolete and no longer be accessible; archives typically plan for 'forward migration' of file formats;
3. access to websites cannot typically be controlled to the degree that archives allow (e.g. restricting access by user type or content of the materials);
4. publication on websites often does not involve curation or editorial control. Archives have collection policies, make judgements about selection of deposits, and curate them, ensuring at least some level of quality control.

Archives can be classified according to the types of material they contain:

1. *physical* (analogue) – contain paper records, tape recordings, physical objects, e.g. Smithsonian Institution,[19] British Library (BL),[20] Bibliothèque nationale de France (BNF)[21]
2. *digital* – contain electronic files only: audio-visual, text, still images, maps, e.g. ELAR, TLA, AILLA,[22] Pangloss[23] (see DELAMAN[24] for a list)
3. *mixed* – contain both analogue and digital materials, e.g. AIATSIS,[25] CLA,[26] ANLA[27]

They may also be categorised according to their scope of coverage:

1. *international* – world-wide or multi-country coverage, e.g. ELAR, TLA, BL, BNdeF, AILLA, Pangloss
2. *national* – covering one country, e.g. AIATSIS

---

[18] there is a free online course about archiving at https://archivingforthefuture.teachable.com/; note that it does not cover how to use other people's collections or legacy materials (cf. Section 3).

[19] https://www.si.edu/ (accessed 2022-09-10)

[20] https://www.bl.uk/ (accessed 2022-09-10)

[21] https://www.bnf.fr/fr (accessed 2022-09-10)

[22] https://ailla.utexas.org/ (accessed 2022-09-10)

[23] https://pangloss.cnrs.fr/?lang=en (accessed 2022-09-10)

[24] https://www.delaman.org/ (accessed 2022-09-10)

[25] https://aiatsis.gov.au/ (accessed 2022-09-10)

[26] https://cla.berkeley.edu/ (accessed 2022-09-10)

[27] https://www.uaf.edu/anla/ (accessed 2022-09-10)

3. *regional* – covering an area in a country, e.g. CLA, ANLA
4. *local* – covering a town or community, e.g. local museums
5. *personal* – records of an individual or family

5. Language revitalisation and archived corpora

Language revitalisation is generally understood to mean efforts undertaken to increase the vitality of a language or variety by taking action to increase its domains of use and/or increase the number of users (often in the context of 'reversing language shift', Fishman 1991, Olko and Sallabank 2020). It tends to be primarily focussed on children, but may also include adult learners (so-called 'new speakers'). There are a large number of active language and cultural revitalisation programmes around the world; some of these are long-standing, e.g. Māori (New Zealand), Hawaiian, and Welsh, among many others. Language community members are often more interested in revitalisation than documentation, and there is a common assumption that revitalisation means formal language learning (school lessons, immersion).

Quite a number of communities have a desire to use archived corpora to support language learning and cultural recovery. Online corpora, such as those in the archives mentioned in Section 4, might seem at first sight to be great potential sources of instances of language use that could supplement the knowledge and use by contemporary speakers (where they exist) for the purposes of materials development, curriculum design, testing, and language learning and revitalisation. However, an exploration of existing online corpora reveals that there are often numerous problems with using such materials. We can identify at least the following issues concerning:

1. epistemology of archival content;
2. nature of materials in the corpus;
3. corpus rights and responsibilities;
4. accessing a corpus and using it.

5.1 Archival content

The materials to be found within an archived corpus do not exist in a vacuum but have their own socio-cultural context within which they were created and now exist, what Dobrin & Schwartz (2021) have called their "social lives". This is the broad context of corpus compilation reflecting the identity and history of individuals and groups

(researchers, consultants, community), relationships (Christensen 2018), types of interactions, and the assumptions and goals brought to the work by those involved (the stakeholders in the project). These are often implicit but need to be understood in order to make proper sense of the materials. It is commonly the case that they are rarely documented or made explicit by corpus creators as part of their meta-documentation, so socio-historical research on the corpus and its creation needs to be undertaken (see Austin 2017, and the case studies reported in the special issue of the journal *Language Documentation and Description* volume 21). Among the topics that can be fruitfully explored are:

1. the *biography* of the creator(s): their prior knowledge and/or study and/or exposure to the language and culture, the identity of their teachers/mentors/correspondents, how and when they learnt the language, how long they worked on the language and culture and at what point in their careers, how the work was funded and with what goals, whether there were previous studies of the language or the community that they could have had access to. There are also biographical aspects of the recorded consultants, including how and under what circumstances they learned and used the language, any prior experience with language teaching, and their motivations for engagement in the project. In addition, it is important to try to identify how and by whom knowledge representations were added to the recorded events (transcription, translation, contextualisation, metadata);
2. aspects of the *historical period* when the corpus was compiled: what the nature and impact of contact was between the corpus collectors and the recorded individuals and communities, prior inter-community relations and interactions (including colonialism and other forms of socio-cultural repression), the linguistic and cultural models known to the corpus compilers, and possibly influential descriptive categories and formats they may have drawn upon, e.g. traditional grammar based on Latin or Greek models, or structural, generative or functional linguistic and cultural approaches.

5.2 Corpus form, content, and interpretation issues

Archived corpora can often present challenges to users in terms of accessing and making sense of the content within them. Audio recordings may be poor quality, noisy, and difficult to hear and comprehend, especially for multiparty conversations. Similarly, video recordings may be poorly recorded or unwatchable, have poor audio, and only

partially incorporate (or miss entirely) contributors who are out of frame. If video or audio has been edited before deposit, it is possible that crucial contextualising information is omitted so as to focus on what the researchers think is important cultural and/or linguistic "data".

For textual sources, hand-written or typed text can be difficult to read or interpret, with crossing out, abbreviations, or other obscurities, all of which requires some degree of philological analysis (see Joby 2021 for an example, and Nathan et al. 2009). In digital text files, characters may be mismapped or omitted due to font problems, tabbed or spaced text may not align, and structured text may be uninterpretable if the structure definition is missing (e.g. Toolbox files without their associated .typ specifications). Where sources have been retranscribed, they should ideally link back to documents or image files on which they are based (so that interpretative steps can be retraced); a very nice example of this which includes a 'diplomatic edition' as well as an edited and cleaned-up version is the Dawes Manuscript[28] (see Nathan et al. 2009). Other problems with text sources include over-distinguishing or under-distinguishing crucial contrasts, in phonology (voicing, aspiration, vowel quality or quantity, tone), morphology (incomplete or misinterpreted paradigms), or syntax (role of case, transitivity-altering constructions, variable word order, cross-clausal linkage such as switch-reference).

Implicitly structured materials, e.g. those using typography or page layout to distinguish analytical categories or kinds of information, can be made more useful by encoding the structure separately from the form, e.g. through Extensible Markup Language (XML) representations (for a case study, see Austin 2022), or using a database model. Unfortunately, structure is not always computable from typography and may need to be manually added (Austin 2022 reports issues with over-use in the legacy source of quotation marks (for multiple purposes and often redundantly), unclear scoping, and spelling errors in structural cues, such as part-of-speech labels).

Other potentially problematic issues include cryptic or incorrect glossing, because the corpus compiler(s) or translators could not understand the language consultant's accent or pronunciation, or because the semantics of the source language terms were misunderstood (the so-called "gavagai problem" of Quine 1969). This can be compounded where a lingua franca is involved and where the collector and/or consultant speak different varieties of it, or one or both have incomplete competence in

[28] www.williamdawes.org (accessed 2021-09-01).

it. We can also find within corpus materials changing interpretations over time (especially changes in transcription and/or translation), and researchers misrepresenting utterances because of what they think they heard rather than what is in the recording. It is thus important to establish a timeline and map particular materials to it. There can also be interventions by speakers due to analytical decisions they make, including "cleaning up" a recording or transcription for reasons of prescriptivism or purism.

Corpus materials may also contain dated content that uses expressions that are no longer considered to be acceptable (e.g. "primitive tribe"), or are deemed to be inappropriate, e.g. personal remarks about the ancestors of living persons. Some content may also be inappropriate for various audiences, e.g. taboo, sacred, violent, or sexually-explicit. A not uncommon issue for revitalisation is mismatches between forms and expression in a corpus and the usage of knowledge-holders other than the people recorded, especially where there is a temporal and/or geographical difference between the corpus and contemporary sources. This can lead to conflicts about what and who is "right", especially for shifting languages undergoing change.

Finally, the relevant sociolinguistic and cultural context for recorded instances of language use in a corpus may be missing due to decontextualisation in the collection process (e.g. recording monologic narratives by an isolated individual to get a "good recording", when the cultural context of narration is a group and multi-performer one). Recovering aspects of the ethnography of speaking (who says what to whom when and where) for revitalisation can consequently be difficult, especially where the focus in the corpus is on a limited set of topics, genres, and interaction types (e.g. monological narratives, interviews about grammatical topics) thereby placing particular restrictions on the usability of corpora for language learning. Austin and Sallabank (2018) discuss challenges of this type in some detail.

5.3 Corpus rights and responsibilities

Language documentation projects typically involve many stakeholders who may have different kinds of interests in the materials collected and the analyses created. Control, consultation, and decision-making are important to work through when deciding what kind of documentary material to include in any corpus and how it can be used. For legacy materials there may be possible mismatches between past situations and the present (see also O'Meara and Good 2010, Innes 2010):

1. current membership of a contemporary ‘community’ may not coincide with past membership;
2. people who provided legacy materials may no longer be viewed as rightful members of a given group and therefore their information may be deprecated;
3. agreements, if any, between the original collector and the community or particular individuals at the time of collection may be unclear, and such agreements may not have been documented explicitly. There may also be issues about the relationship between any such past agreements and arrangements that are currently being negotiated between contemporary researchers and other stakeholders, e.g. researchers being told not to distribute copies of legacy materials without permission of current Indigenous groups who self-identify as descendants of the recorded speakers.

It is important to clarify rights and responsibilities, before creating and using any corpus, but especially one involving legacy materials. This includes exploring questions such as the following: who holds what rights (hereditary ownership, copyright, performance rights)? Are the rights documented? How do we establish rights retroactively? What if the researcher is not sure about speaker or performer rights? How do we determine rights when there are multiple contributors and data comes from multiple media? What happens to ‘orphan works’ where the original stakeholders can no longer be identified (e.g. materials passed from one researcher to a later researcher, possibly without consultation with the original community)? When analysing corpus materials, including using them for revitalisation, it is important to clearly document the various contributions to the work, including those of the original creators, research assistants, linguist-editors, archivists, other researchers, and current community members. In the case of derivative corpora, access rights and the relationships to the orginal legacy materials need to be decided and clearly documented.

5.4 Accessing the corpus and using it.

Archival corpora, especially those available online, can present a range of challenges for users who wish to access them, especially in the context of language revitalisation. Wasson et al (2016: 669) give a frank assessment of user experiences based on interviews with a range of archivists, identifying “a rich list of problems that might be

encountered by users of language archives". The most significant of these can be summarised as follows:

1. lack of contextual information at the deposit level, or in the metadata;
2. inadequate search/browse functions;
3. problems with the interface/information display, especially on mobile devices;
4. users may be frustrated when they do not have access to collections and need to make a request to the original depositor, who it may then be difficult for the archivist to locate;
5. there may be technology issues with the corpus, e.g. outdated file formats, broken scripts, Flash/Java problems, and so on;
6. the interface language(s) may not accessible to would-be users, e.g. while the AILLA archive contains materials collected in Brazil access to cataloguing metadata is provided in Spanish and English only.

In addition to these, online digital archives assume a high level of information technology and media competence, e.g. how to download and save files, as well as access to and knowledge of specialist software, such as ELAN, Praat, or Toolbox/FLEx. Finding a file of interest in an archive is usually only the first step to being able to display or interact with it in meaningful ways.

6. Conclusions

Creating and analysing corpora can be very rewarding, and can enable various exciting kinds of linguistic and cultural research. However, working with corpora can often involve dealing with complex issues and challenges about the form, content, context, and use of materials and analyses within and arising from them. Good corpus management principles and practices (e.g. file naming, folder structure, backup, choice of appropriate software tools) will ensure better outcomes, and make creation, analysis, and preservation processes easier. It is essential to build in archiving plans and get relevant advice from the conceptualisation and beginning of a project.

Maximising opportunities for use of a corpus requires thinking seriously about data entities, data types and relationships, and being explicit about them in the project design and application (e.g. in database design or XML tagging). There are essential roles for richly articulated metadata and meta-documentation that should be appreciated from the

initiation of a project. By creating good meta-documentation now we can reduce legacy corpus problems for future users, including researchers, communities, educators, and others.

There are potentially interesting and engaging opportunities for researchers and communities to add substantial value to corpus materials, and to create rich secondary corpora that may be very useful in the context of language and cultural revitalisation. This is especially true if the collaborators are able to work directly with historical sources (rather than reproductions of them that may contain introduced errors) and contemporary knowledge holders to elucidate the sources and the contexts surrounding their creation, analysis, and current status. Special attention needs to be paid to linguistic and cultural rights, recognising that this can be a complex ethical, historical, and political matter. Careful work with corpora that is aware of potential challenges can also be very rewarding for researchers and communities, especially where there is unique documentation of languages or varieties, or areas of knowledge, that are no longer available, and that can serve as important sources for language and cultural support and revitalisation.

References

AUSTIN, Peter K. 2002. "Developing Interactive Knowledgebases for Australian Aboriginal Languages — Malyangapa". Unpublished paper presented at Workshop on Australian Aboriginal Languages, University of Melbourne, March 2002. Online at http://emeld.org/workshop/2003/Malyangapa.pdf

AUSTIN, Peter K. 2006. "Data and language documentation". In GIPPERT, Jost, Nikolaus HIMMELMANN and Ulrike MOSEL (eds.) *Essentials of Language Documentation*, 87-112. Berlin: Mouton de Gruyter.

AUSTIN, Peter K. 2013. "Language documentation and meta-documentation". In JONES, Mari & Sarah OGILVIE (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, 3-15. Cambridge: Cambridge University Press.

AUSTIN, Peter K. 2016. "Language documentation 20 years on". In PÜTZ, Martin & Luna FILIPOVIĆ (eds.) *Endangered Languages and Languages in Danger: Issues of ecology, policy and human rights*, 147-170. Amsterdam: John Benjamins.

AUSTIN, Peter K. 2017. "Language documentation and legacy text materials". *Asian and African Languages and Linguistics* 11: 23-44.

AUSTIN, Peter K. 2020. “Language documentation and revitalisation”. In OLKO, Justyna & Julia SALLABANK (eds.) *Revitalizing endangered languages: a practical guide*, 199-219. Cambridge: Cambridge University Press.

AUSTIN, Peter K. 2022. “Making 2,180 pages more useful: the Diyari dictionary of Rev. J. G. Reuther”. To appear in Eda Dehermi & Christopher Moseley (eds.) *Endangered Languages in the 21st Century*. London: Routledge.

AUSTIN, Peter K. and Lenore GRENOBLE. 2007. “Current trends in language documentation”. *Language Documentation and Description* 4: 12-25.

AUSTIN, Peter K. and Julia SALLABANK. 2011. “Introduction”. In AUSTIN, Peter K. and Julia SALLABANK (eds.) *The Cambridge Handbook of Endangered Languages*, 1-24. Cambridge: Cambridge University Press.

AUSTIN, Peter K. and Julia SALLABANK. 2018. “Language documentation and language revitalisation: Some methodological considerations”. In HINTON, Leanne, Leena HUSS and Gerald ROCHE (eds.) *Handbook of Language Revitalisation*, 207-215. London: Routledge.

AUSTIN, Peter K. and Julia SALLABANK. 2022. “Endangered languages”. In WEI, Li, Zhu HUA and James SIMPSON (eds.) *Routledge Handbook of Applied Linguistics*. 2nd edition. London: Routledge.

BEREZ-KROEKER, Andrea L., Lauren GAWNE, Susan Smythe KUNG, Barbara F. KELLY, Tyler HESTON, Gary HOLTON, Peter PULSIFER, Dsvid I. BEAVER, Shobhana CHELLIAH, Stanley DUBINSKY, Richard P. MEIER, Nick THIEBERGER, Keren RICE, and Anthony C. WOODBURY. 2018. “Reproducible research in linguistics: A position statement on data citation and attribution in our field”. *Linguistics*, 56(1): 1-18. https://doi.org/10.1515/ling-2017-0032

CHRISTEN, Kim. 2018. “Relationships, not records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online”. In SAYERS, Jenetry (ed.) *Routledge Companion to Media Studies and Digital Humanities*, 403-412. London: Routledge.

CRYSTAL, David. 1992. *An Encyclopedic Dictionary of Language and Languages*. Oxford: Blackwell.

DOBRIN, Lise & Saul SCHWARTZ. 2021. “The social lives of linguistic legacy materials”. *Language Documentation and Description* 21: 1-36.

FISHMAN, Joshua A. 1991. *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Clevedon: Multilingual Matters.

GIPPERT, Jost, Nikolaus P. HIMMELMANN and Ulrike MOSEL (eds.) 2006. *Essentials of language documentation.* (Trends in Linguistics. Studies and Monographs, 178). Berlin: Mouton de Gruyter.

GRENOBLE, Lenore A. 2011. "Language ecology and endangerment". In AUSTIN, Peter K. and Julia SALLABANK (eds.) *The Cambridge Handbook of Endangered Languages*, 27-44. Cambridge: Cambridge University Press.

GRENOBLE, Lenore A. 2013 "Language revitalization." In BAYLEY, Robert, Richard CAMERON and Ceil LUCAS (eds.) *The Oxford Handbook of Sociolinguistics*, 792-811. Oxford: Oxford University Press.

HENKE, Ryan & Andrea L. BEREZ-KROEKER. 2016. "A Brief History of Archiving in Language Documentation, with an Annotated Bibliography". *Language Documentation and Conservation* 10: 411-457.

HIMMELMANN, Nikolaus P. 1998. "Documentary and descriptive linguistics". *Linguistics* 36, 161–195.

HIMMELMAN, Nikolaus P. 2012. "Linguistic data types and the interface between language documentation and description". *Language Documentation and Conservation* 6, 187-207.

INNES, Pamela. 2010. "Ethical problems in archival research: Beyond accessibility". *Language and Communication* 30(3): 198-203.

JOBY, Christopher. 2021. "Revisions to the Siraya lexicon based on the original Utrecht Manuscript: A case study in source data". *Historiographica Linguistica* 48(2/3): 177-204. https://doi.org/10.1075/hl.00084.job

LÜPKE, Friederike. 2009. "Data collection methods for field-based language documentation". *Language Documentation and Description* 6: 53-100.

McENNERY, Tony and Andrew HARDIE. 2011. *Corpus Linguistics: Method, Theory, and Practice*. Cambridge: Cambridge University Press.

NATHAN, David and Peter K. AUSTIN. 2004. "Reconceiving metadata: language documentation through thick and thin". *Language Documentation and Description* 2, 179-187.

NATHAN, David, Susannah RAYNER & Stuart BROWN (eds.) 2009. *William Dawes: notebooks on the Aboriginal language of Sydney: a facsimile version of the notebooks from 1790-1791 on the Sydney language written by William Dawes and others*. London: SOAS University of London.

OLKO, Justyna & Julia SALLABANK (eds.) 2020. *Revitalizing endangered languages: a practical guide*, 199-219. Cambridge: Cambridge University Press

O'MEARA, Carolyn and Jeff GOOD. 2010. "Ethical issues in legacy language resources". *Language and Communication* 30(3): 162-170.

QUINE, William. V. O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

SEIFART, Frank, Nicholas EVANS, Harald HAMMARSTROM and Stephen C. LEVINSON. 2018. "Language documentation twenty-five years on", *Language* 94(4): e324-e345.

WASSON, Christina, Gary HOLTON and Heather S. ROTH. 2016. "Bringing User-Centered Design to the Field of Language Archives". *Language Documentation and Conservation* 10: 641-681.

WOODBURY, Anthony C. 2003. "Defining documentary linguistics". *Language Documentation and Description* 1: 35-51.

WOODBURY, Anthony C. 2011. "Language documentation". In AUSTIN, Peter K. and Julia SALLABANK (eds.) *The Cambridge Handbook of Endangered Languages*, 159-186. Cambridge: Cambridge University Press.

WOODBURY, Anthony C. 2014. "Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire". *Language Documentation and Description* 12: 19-36. https://doi.org/10.25894/ldd161