Applied Linguistics 2016: 37/1: 88–99 © Oxford University Press 2016
doi:10.1093/applin/amv076 Advance Access published on 14 January 2016

Documentary Linguistics: Methodological Challenges and Innovatory Responses

NICK THIEBERGER

University of Melbourne
E-mail: thien@unimelb.edu.au

> Seeing the present situation, I think that, at the very least, it behooves us as scientists and as human beings to work responsibly both for the future of our science and for the future of our languages, not so much for reward according to the fashion of the day, but for the sake of posterity. What we need to do now stares us in the face. If we do not act, we should be cursed by future generations for Neronically fiddling while Rome burned. (Krauss 1992: 8)

Language documentation emphasizes the importance of fieldwork and of recording a broad range of cultural practices and the curation of records of a language in addition to more traditional grammatical description. As it has become clear that most of the world's languages will be under pressure from larger languages and risk being lost, a number of linguists have focused on the methods that can be employed to create better records and to support the ongoing use of the languages. This article provides an overview of the field and the advances that have been made in the recent past in providing better methods for documentation of the world's languages.

INTRODUCTION: LANGUAGE DOCUMENTATION

As it has become clear that many of the world's languages (those which I will call *small* languages) will be under pressure from larger languages and risk being lost, a number of linguists have focused on the methods that can be employed to create better records and to support the ongoing use of the languages. The foundation reference in this field is Himmelmann (1998), who set out the criteria for 'language documentation' (LD) to distinguish the new field from the earlier model of 'language description'. The main distinction is that *description* was very much focused on academic work like grammars and scholarly articles, while *documentation* acknowledges the importance of these while also arguing for the creation and valuing of primary records and of working collaboratively with speakers to encourage their involvement in the process of recording their own languages and of recording a broad range of cultural practices. My main focus in this article will be on this novel emphasis on records of small and endangered languages being treated as data that can then be used for further research, and that provide a corpus that gives access to language phenomena not identified by the original recorder. This can be seen as a convergence of field-based linguistics with traditional applied linguistics methods that include corpus creation and exploration. The emergence of LD serendipitously coincides with and benefits from new technological methods that permit and contribute to new ways of recording, storing, and interacting with the language records. New technological tools and methods mean that more recording is possible than before, but, while technology permits more and better-quality recordings, it also introduces new risks in data loss through fragile digital storage. Hence, the need for training in the use of new methods, and the development of tools to take advantage of technological advances. Time-aligned transcriptions, virtually unknown a decade ago, are now a standard way of creating transcripts. For more information on how LD is characterized by leaders in the field, see the collection of papers in Gippert *et al.* (2006), Seifart *et al.* (2012), Austin and Sallabank (2011), or Woodbury (2011).

The theoretical turn of linguistics that began in the 1960s meant that linguists were focusing on abstract formal representations that required introspection but not fieldwork-based corpora, with a consequent decline in the number of fieldworkers recording original language records. At the same time, on a sociopolitical level, there was a rise of nationalist independence movements and increased recognition of the injustice suffered by indigenous people, including the threat to their languages. Further, a general awareness of the loss of biological diversity (cf. Lovejoy 1980) helped to provide metaphors of imminent threat that were understood by the broader population and were available to be adopted by language activists. Of the more than 7,000 small languages in the world, linguists have predicted that half will no longer be spoken within the next century, and some put the figure as high as 90 per cent (Krauss 1992: 7). Krauss is widely quoted as estimating the rate of loss of a language every two weeks, and he observes that, 'for scientific purposes, it is most urgent to document languages before they disappear... By documentation I mean grammar, lexicon, and corpus of texts. This is a tradition well proven in the history of linguistics. To this we can now add documentation on audio and videotape. There must also be a network of repositories and centers for safeguarding and using this documentation' (Krauss 1992: 8). The decades since Krauss's statement have seen an increase in attention paid to language endangerment issues, with many popular news items and a number of accessible books on the subject (e.g. Harrison 2007; Evans 2010).

While Campbell *et al.* (2013) question the rapidity of projected language attrition, suggesting that the rate of loss is more like one language every three months, linguists in general agree that there is an urgency to record as much as possible of the world's language diversity. They have been active in mobilizing public interest, via organizations like the Foundation for Endangered Languages,[1] The Endangered Language Fund,[2] The Endangered Languages Alliance,[3] and the Living Tongues Institute,[4] among others.

EMELD[5] was a major project in this field that promoted standards for the description of endangered language data.

Documentation can also be distinguished from earlier methods by the range of topics it encourages linguists to record. It emphasizes creating a lasting, multipurpose record of a culture (Himmelmann 2006: 1), because, as Evans puts it, 'undocumented languages contain too much information to be wasted on linguists alone' (Evans 2012: 183). It follows that fieldwork can focus on an almost infinite range of topics that are the everyday and historical reality of the people a linguist is recording. Thieberger (2012) is a guide to some of the kinds of specialist information that can be required of a linguist in the field, including such topics as astronomy, botany, ethnography, gastronomy, geography, mathematics, musicology, toponomy, kinship, audiovisual technology, and copyright and legal concerns. This breadth then suggests the need for interdisciplinary collaboration and teamwork in the field as discussed by Evans (2012), who recounts experiences of fieldwork with colleagues from various disciplines, including musicologists, biologists, and anthropologists. Others who have written on the topic of collaboration, both interdisciplinary and with members of the speech community, include Dwyer (2006), Glenn (2009), and Yamada (2007).

Another way to characterize the distinction between language description and documentation is to acknowledge the colonial nature of extracting information for the (almost exclusive) benefit of the descriptive linguist. In the old paradigm, the recordings (if any were made) belonged to the linguist who may have made them available to the speakers recorded, but generally did not. This is further evidenced by the lack of recognition of the importance of recordings: linguists provided no repositories for recordings they did make; they created no reward to recognize the work needed to curate collections of primary records; they were not particularly encouraged to make the records in the first place, nor trained in how best to do that work. Poor records held by individual researchers are not easily made accessible to the people they recorded or the source communities of speakers. It is taking some time to convince field linguists of the need to change this practice, as witness how little is being archived by linguists who are not grantees of the large funding agencies mentioned below.[6] Documentary linguists have been active in establishing language archives and in promoting the need to create archival records in the course of fieldwork, facilitating the curation of these records, their description, and thus their accessibility to others, not reliant on the grace and favour of the original researcher providing access on an ad-hoc basis to records they hold in their offices or attics. The transparency with which the language records are described and made available via documentary efforts can be seen as a post-colonial repatriation. Much more remains to be done to make such records available, especially in remote areas with little access to communication technologies, but the first step for the linguist is ensuring that their records are created and described sufficiently to allow them to be curated and discovered.

The tension between the academic research agenda and the desires of speakers nevertheless remains and requires constant reflection and negotiation.

A critique of some LD literature is that it is unrealistic in the amount of information that it suggests a researcher can collect. This is clearly borne out if we look at the comparative paucity of collections created over the past decades, indicating that more commentators are preaching than are practising. While funding bodies have required records to be archived, even their own grantees have not always complied,[7] and researchers outside of these funding regimes are, unfortunately, still recalcitrant in following the sorts of recommendations that have been clearly articulated by, for example, the EMELD[8] conferences in the USA, or the standards promoted by the DoBeS[9] funding programme beyond its own grantees via the tools it developed, and also the Summer Institute of Linguistics[10] and the many tools it provides for free to the research community.

The refocusing of academic linguistic interest entailed in the LD effort means, among other things, training linguists to produce good records of languages they work on. Training is offered within some university departments; for example, my own department at the University of Melbourne regularly runs courses that are available to all comers, and deals with audio and video recording methods, workflows for data management, and clinics on the current preferred tools. Similar courses are run in other places, and there is regular training provided by the summer schools of the biennial CoLang[11] in the USA and the 3L consortium[12] in Europe. There are new types of recording methods and equipment being developed all the time, so it is important for all concerned—linguists, speakers, and other researchers interested in recording oral traditions—to keep up to date with what is the current best way to do this work. The Resource Network for Linguistic Diversity[13] provides a mailing list and a set of webpages of resources for supporting LD work.

This kind of training often also provides outreach outside of academia to community groups focused on language recording and revitalization. The community response and publicity around endangered languages and the need to promote more recording of them has led to several new funding sources and university courses. The major funders of this work are, in Europe, the Arcadia-funded Endangered Languages Documentation Programme[14] at SOAS (University of London) and the, now concluded, Dokementation Bedrohter Sprachen[15] (DoBeS) in Germany. In the USA, the Documenting Endangered Languages[16] programme of the National Science Foundation and National Endowment for the Humanities has funded over 95 projects. A number of universities offer specialization in LD methods. The University of Hawaii at Mānoa[17] is notable in having set up a degree programme, the premier conference on these issues—the biennial International Conference on Language Documentation & Conservation,[18] and the journal *Language Documentation & Conservation (LD&C)*, which covers issues discussed here and has been published since 2007. Its open access contents[19] give a good indication of the range of issues that are of current concern in the field, with articles on topics including: the nature of endangered languages themselves, collaborative work, training speakers in recording and revitalization methods, archiving and presentation of legacy records for reuse, and reviews of software and equipment. In addition, *LD&C* has published eight books as special publications, all freely available online,[20] whose titles are listed as follows *(up to 2014)*:

1 Documenting and Revitalizing Austronesian Languages
2 Fieldwork and Linguistic Analysis in Indigenous Languages of the Americas
3 Potentials of Language Documentation: Methods, Analyses, and Utilization
4 Electronic Grammaticography
5 Melanesian Languages on the Edge of Asia: Challenges for the 21st Century
6 Microphone in the Mud
7 Language Endangerment and Preservation in South Asia
8 The Art and Practice of Grammar Writing

Most titles are self-explanatory, and deal with specific geographical regions or with methodology. Uniquely, the volume 'Microphone in the Mud' is a novel about fieldwork experiences in the Philippines.

TOO MANY LANGUAGES, TOO LITTLE TIME

Clearly, there is a need to scale-up the recording of the world's languages, devising better ways for linguists to do their work, and training speakers to create records themselves. There are all kinds of valuable records currently produced by speakers of small or endangered languages, including YouTube videos, tweets, and Facebook pages. However, they are not typically easily locatable, and they are at risk of being lost due to the ephemeral platforms on which they are stored. The role of the professional linguist is to create enduring records and to provide the repositories that will ensure their longevity, especially as the kinds of records created by a concerted fieldwork effort are likely to be richer, more detailed, and more representative of a range of speakers and genres than are the outputs of social media.

Exciting new methods are emerging for recording, transcribing, and analysing more than was previously possible. Aikuma,[21] for example, is a phone app for recording texts and respeaking them, allowing a speaker to record a story in their own language and then respeak it in a metropolitan language, essentially providing an oral transcription (Reimans 2010). Woodbury (2003: 45) suggests that our time as linguists is better spent not interlinearizing texts, but instead asking elders to slowly 'respeak' texts to a second recording so that anyone with training in hearing the language can make the transcription if they wish. The promise this offers is that many more hours of recordings can be made, and be interpreted by non-speakers of the language in future, all without the need for written transcripts. So, instead of the two or three hours of annotated recordings typically produced in the past, or the tens of hours produced by a serious documentation project (both of which are nevertheless extremely valuable), it should be possible to produce hundreds of hours of recordings, representing a range of speakers and of discourse types.

A further exciting development is the ability to force the alignment of textual transcripts and audio recordings at the level of phonemes, for example, using the MAUS system (Schiel *et al.* 2011), which returns a time-aligned version of a transcript. Recent work (Strunk *et al.* 2014) shows that this system can be used on small languages, those for which there is no previous training material. Thus, a known pairing of text and media based on the digital acoustic characteristics of the media file could then be used to infer the same relationship in untranscribed media, resulting in automatically generated (partial) transcripts of previously untranscribed media.

CORPORA OF THE WORLD'S LANGUAGES

LD also draws on the established methods of corpus linguistics (McEnery and Wilson 2001) to create small corpora on which further analysis can be based. This is an innovation in a field that has previously relied on the linguist to observe and analyse, but not to provide the data on which this analysis is based. In this way, evidence for analytical claims is reintroduced into linguistic fieldwork methodology. Further, as the development of collections of records is central to LD, the community of scholars has begun to engage with the need to provide a system of recognition of collections as scholarly output to encourage properly formed collections. While good intentions abound, the number of collections of records for small languages is not growing as quickly as may have been hoped in the late 1990s (as noted earlier). If we take seriously the importance of creating primary language records that can be used by other scholars and acknowledge that it takes some effort to create, analyse, describe, and curate these records so that others can make sense of them, then it is important to provide the creators of these collections with recognition, in the normal way provided to other scholarly output. This is an established mechanism in STEM disciplines, with some journals now requiring primary records to be publicly available before an article citing them can be published. Costello (2009) summarizes the many advantages of accessible data as well as canvassing some of the reasons that make scholars reluctant to publish their data. The solution advocated by many in the scientific community 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 is treated on a par with traditional academic output. The Australian Linguistic Society has established criteria for reviewing such collections (see Thieberger *et al.* 2016) which have had broad support among their membership. To encourage the use of such collections of primary records, it is useful to have a portal or 'landing page' that acts as a finding aid to the contents. This is provided, for example, as a webpage for the Beaver[22] language collection, or the guide to the

Ikaan collection published as a journal article (Salffner 2015), or, more simply, the guide to my collection of South Efate[23] material.

A speaker of a small language who knows about a collection of recordings in their home village or a researcher who discovers or creates a new collection of material in the focus language of their research can publish the fact of its existence in research articles, books, or notes, but that is not going to happen for a while, and, even then, it is not going to reach a very wide audience. How is the rest of the world going to find out that this new source of information in an otherwise largely unrecorded language exists? No one else has yet found these records because either they are in a personal collection or the institution it is in did not catalogue it as being *in* a particular language, and perhaps did not even note that it was *about* a particular language.

Nowadays, you could blog about this new discovery, but blogs typically do not have many readers, so that would not get the coverage you need. Google would find the language name in the blog and that is better than there being no public record of your discovery. But, if the language name is also a common word in another language (e.g. *Noone, Karen, Kola, Titan, Maria, Mono, Mum*—which are all language names), then your prized discovery will be buried deep in a web search.

To make such searches more successful, there is a standard code for each of the (currently) 7,865 languages (ISO 639-3).[24] This three-letter code is a way of uniquely identifying every language in the world, so the language *Kola*, mentioned above, has the code 'kvv', and *Maria* (in Papua New Guinea) has the code 'mds'. If you can find a way of associating your newly found collection with this standard code (e.g. by lodging records in an archive), then searches will be more targeted.

A broader view of what is available for any given language is provided by the Open Language Archives Community, whose service harvests metadata from all subscribing language archives and then presents that as a single page—http://www.language-archives.org/language/XXX—where the last three letters are the language code (just mentioned). This list is updated daily and gives one of the few overviews of what information is available for each of the world's languages, but is only as good as the collections it harvests from cooperating archives. The Endangered Languages Catalog,[25] a joint project of the University of Hawai'i and the LinguistList, provides links to relevant information for all of the world's endangered languages, that is, to about a third of the total of some 7,000 languages. Wikipedia also provides a good point of entry for small languages, but many of these pages need more information than is currently provided. Glottolog is a service that lists all languages and language varieties together with references to published material about them.[26]

Metadata in a standard form allows this kind of searching and, when used with a broader description of the records it describes, maximizes the possibility that the information will be found by anyone searching for it on the web.

Using standard metadata terms, such as those recommended by the global library community (Dublin Core)[27] or the Open Archives Initiative,[28] allows our small community of language archives to benefit from an established international infrastructure. Of course, any additional information, including transcripts of media, can also be searched, and this should be considered part of the extended metadata associated with, for example, media files.

REUSE

Documentation is focused on creating good records, and for linguists, this is usually the basis for the linguistic analysis of a language. Narratives recorded in the course of fieldwork will become part of the set of material (along with your own observation, elicitation, questionnaires, or stimulus tasks) that is used to describe the grammar of the language. Unfortunately, there are languages for which we have only grammatical analyses and no further information—that is, no collections of recordings, no texts, no dictionary—which are the kinds of resources that speakers typically want to find about their languages (Moore 2006). But, it is not just speakers who want to read texts in a small and otherwise little-described language. Linguists also can use this material to explore questions the original linguist did not ask or was not interested in. There are obviously many ways in which language records can be used, especially now that online multimedia is so accessible and widespread. Online video of people telling their own stories or performing music is now a simple matter for anyone with fairly accessible equipment (like a camera or mobile phone). The problem is that this valuable record may be lost when the website it is hosted on is closed down. To ensure that our grandchildren can still access these records, we have to create the records in ways that allow them to be reused. That is, they must exist in formats that allow them either to be used immediately or to be converted to a usable form without too much effort.

This is a requirement of all kinds of research outputs, not just linguistic data, as reworking data by hand is too time-consuming, and the 'amount of data produced far exceeds the capabilities of manual techniques for data management' (Borgman 2007: 6). This implies that we have tools that produce output in reusable forms and that linguists have training in what it means to create reusable data. So, for example, a dictionary that is created as a Microsoft Word document has very limited use beyond a printed dictionary, and a finderlist—an English index of the words in the dictionary—has to be created by hand once the main dictionary is finished. Any web version of the dictionary will also take a great deal of manual work to construct. If, on the other hand, the dictionary was created from a lexical database, using the sort of software that is designed to make dictionaries (e.g. like Fieldworks[29]), then a finderlist and web-based version of the dictionary will be available automatically from the same underlying data. Dictionaries created in lexicographic databases can be output in various formats (finderlists, topical dictionaries, learner's dictionaries), and as books, websites, or phone apps. Transcripts of recordings that are created properly (e.g. using the free software Elan[30]) can also be used as subtitles for videos for wider distribution, produced as stories linked to media, or as an iTunes installation, or a website.

Documentation can take many forms and, as I have emphasized here, there is a multiplier effect in building records properly with sufficient description to allow them to be retrieved and reused. Multiple possible output formats can be derived from the same primary records, for online access, or via portable devices and mobile phones, in addition to books of narratives or dictionaries as discussed in Thieberger and Berez (2012).

DIGITAL ARCHIVES

As will be apparent by now, the creation of fragile digital language records risks producing endangered records of endangered languages (cf. Bird and Simons 2003: 567). In response, a dedicated network of digital archives has been established to curate and provide access to this material, making up the Digital Endangered Languages and Musics Archives Network, an umbrella body whose webpages provide links to local archives. The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) is the archive I have worked on since 2003. It was initially focused on digitizing analogue field recordings, and the collection now holds samples of over 840 languages, in more than 5,100 hours of audio recordings. We are active in recommending standards for researchers to use, and in helping build archival collections so that the work of naming files and identifying their contents can be done in the course of their creation by researchers in the field rather than trying to retrofit the collection post-hoc.

These archives must be distinguished from websites. As noted earlier, putting images and media online in a website allows it to be seen by others, but websites are usually temporary and are not an alternative to long-term archiving. An archive will also catalogue its collection using standard terms that allow it to be searched, as discussed below when we talk about blogs. Digital archives allow incremental deposit, so, at PARADISEC, we encourage deposit of recordings from the field, providing a safe backup for the fieldworker and citable form of the records. As a recording is transcribed, the transcript is added to the collection, and as the transcript is improved, newer versions can be accessioned. As it is translated and annotated, these additions too can be made to the collection.

In places where the Internet is difficult or expensive to access or not available at all, language records need to be made available in other ways. Books of stories, dictionaries, and collections of images (plants, animals, fish, and so on) with local names and uses are all valuable ways to return information collected with a community. If there is electricity and there are computers, then a local installation of iTunes can be the source for creating playlists of traditional songs, or your favourite old people telling stories.

CONCLUSION

There is an urgency in creating the best possible records of the world's languages. For each language that has previously had no permanent records, the intervention of an outside researcher or the training of a local researcher can result in speakers of that language having a presence on the web, and having records their grandchildren can look to. For a linguist engaged in a period of fieldwork in a small community, it makes sense to make the best possible recordings. No one else has recorded the language before, and some time and effort have been required to get into the field and establish relationships with the speakers there. It is unlikely a similar effort will recur.

Twenty-five years after Krauss's (1992) challenge, it should now be a normal part of research practice to create primary language records in a form that we can reuse, that we can cite in our research, and that can be of use to speakers of the language. These records may be digital versions of older analogue manuscripts or they may be born-digital recordings of performances, but they will only become part of the broader research effort, or be useful to the speakers themselves, if they can be located and reused. LD offers an opportunity to speakers of these languages and to academic linguists to create better records that will enrich knowledge of the diversity of human experience and will provide the basis for supporting the ongoing use of small languages.

NOTES

[1] http://www.ogmios.org/.
[2] http://www.endangeredlanguagefund.org/.
[3] http://elalliance.org/.
[4] http://livingtongues.org.
[5] http://emeld.org/.
[6] This point is elaborated on the blogpost here: http://www.paradisec.org.au/blog/2011/06/5649/.
[7] See, for example, the blog item here: http://www.paradisec.org.au/blog/2011/06/5649/.
[8] http://emeld.org/.
[9] http://dobes.mpi.nl/dobesprogramme. Accessed 26 July 2015.
[10] http://sil.org.
[11] For example, http://www.alaska.edu/colang2016/.
[12] For example, http://www.hrelp.org/events/3L/.
[13] http://rnld.org.
[14] http://www.hrelp.org/languages. Accessed 26 July 2015.
[15] http://dobes.mpi.nl/dobesprogramme. Accessed 26 July 2015.
[16] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816. Accessed 26 July 2015.
[17] http://ling.hawaii.edu. Accessed 26 July 2015.
[18] http://icldc4.weebly.com. Accessed 26 July 2015.
[19] The contents of all volumes can be found here: http://nflrc.hawaii.edu/ldc/?cat=2. Accessed 26 July 2015.
[20] These volumes are available here: http://nflrc.hawaii.edu/ldc/?cat=3. Accessed 26 July 2015.
[21] http://lp20.org/aikuma/. Accessed 26 July 2015.
[22] http://dobes.mpi.nl/projects/beaver.

[23] http://languages-linguistics.unimelb.edu.au/thieberger/sefate.html.

[24] http://www-01.sil.org/iso639-3. Note that not every language is represented here, and there is some controversy about the way in which codes are assigned and what form they take (see Morey *et al*. 2013).

[25] http://www.endangeredlanguages.com/.

[26] http://glottolog.org/glottolog/language.

[27] http://dublincore.org.

[28] https://www.openarchives.org.

[29] http://fieldworks.sil.org.

[30] http://tla.mpi.nl/tools/tla-tools/elan/.

ACKNOWLEDGEMENTS

Thanks to Lauren Gawne, Simon Musgrave and two anonymous reviewers for feedback on an earlier version of this article. My work has been supported by ARC grants DP0984419 and FT140100214.

REFERENCES

Austin, P. and **J. Sallabank**. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Bird, S. and **G. Simons.** 2003. 'Seven dimensions of portability for language documentation and description,' *Language* 79: 557–82.

Borgman, C. L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press.

Campbell, L., N. H. Lee, E. Okura, S. Simpson, and **K. Ueki**. 2013. 'New knowledge: Findings from the catalogue of endangered languages (''ELCat'')'. Paper presented at the 3rd International Conference on Language Documentation & Conservation, available at http://scholarspace.manoa.hawaii.edu/handle/10125/26145.

Costello, M. J. 2009. 'Motivating online publication of data,' *BioScience* 59/5: 418–27, available at http://doi.org/10.1525/bio.2009.59.5.9.

Dwyer, A. 2006. 'Ethics and practicalities of cooperative fieldwork and analysis' in J. Gippert, N. Himmelmann, and U. Mosel (eds): *Essentials of Language Documentation*. Mouton de Gruyter, pp. 31–66.

Evans, N. 2010. *Dying Words: Endangered Languages and What They Have to Tell Us*. Wiley-Blackwell.

Evans, N. 2012. 'Anything Can Happen: The Verb Lexicon and Interdisciplinary Fieldwork,' in Nicholas Thieberger (ed.), *The Oxford* Handbook of Linguistic Fieldwork. Oxford University Press, pp. 183–208.

Gippert, J., N. P. Himmelmann, and **U. Mosel (eds)**. 2006. *Essentials of Language Documentation*. Mouton de Gruyter.

Glenn, A. 2009. 'Five dimensions of collaboration: Toward a critical theory of coordination and interoperability in language documentation,' *Language Documentation & Conservation* 3/2: 149–60.

Harrison, K. D. 2007. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. Oxford University Press.

Himmelmann, N. P. 1998. 'Documentary and descriptive linguistics,' *Linguistics* 36: 161–95.

Himmelmann, N. P. 2006. 'Language documentation: What is it and what is it good for?' in J. Gippert, N. Himmelmann, and U. Mosel (eds): *Essentials of Language Documentation*. Mouton de Gruyter, pp. 1–30.

Krauss, M. 1992. 'The world's languages in crisis,' *Language* 68/1: 4–10.

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

Lovejoy, T. E. 1980. 'Foreword,' in M.E. Soule and B.A. Wilcox (eds): *Conservation Biology: An Evolutionary-Ecological Approach*. Sinauer Associates, pp. ix–x.

McEnery, T. and **A. Wilson**. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press.

Moore, R. E. 2006. 'Disappearing, Inc.: Glimpsing the sublime in the politics of access to endangered languages,' *Language and Communication* 26: 296–315.

Morey, S., M. W. Post and **V. A. Friedman**. 2013. 'The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization'. Paper presented at the Conference Research, Records and Responsibility: Ten Years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures,' available at http://hdl.handle.net/2123/9838.

Reiman, D. W. 2010. 'Basic oral language documentation,' *Language Documentation & Conservation* 4: 254–68.

Salffner, S. 2015. 'A guide to the Ikaan language and culture documentation,' *Language Documentation & Conservation* 9: 237–67.

Schiel, F., C. Draxler, and **J. Harrington**. 2011. 'Phonemic segmentation and labelling using the MAUS technique,' Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research'. University of Pennsylvania.

Seifart, F., G. Haig, N. P. Himmelmann, D. Jung, A. Margetts, and **P. Trilsbeek (eds)** 2012. *Potentials of Language Documentation: Methods, Analyses, and Utilization*. LD&C Special Publication No. 3. University of Hawai'i Press, available at http://nflrc.hawaii.edu/ldc/?p=247.

Strunk, J., F. Schiel, and **F. Seifart**. 2014. 'Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS' in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 3940–7.

Thieberger, N. (ed.). 2012. *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press.

Thieberger, N. and **A. Berez**. 2012. 'Linguistic data management,' in N. Thieberger (ed.): *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press, pp. 90–118.

Thieberger, N., A. Margetts, S. Morey, S. Musgrave. (2016). 'Assessing annotated corpora as research output,' *Australian Journal of Linguistics* 36: 1–21.

Woodbury, A. C. 2003. 'Defining language documentation,' in P. K. Austin (ed.), *Language Documentation and Description*. SOAS, Vol. 1, pp. 35–51.

Woodbury, T. 2011. 'Language documentation,' in P.K. Austin and J. Sallabank (eds): *The Cambridge Handbook of Endangered Languages*. Cambridge University Press, pp. 159–86.

Yamada, R. -M. 2007. 'Collaborative linguistic fieldwork: Practical application of the empowerment model,' *Language Documentation & Conservation* 1/2: 257–82.