Language Documentation and an Australian National Corpus

Simon Musgrave[a] and Sarah Cutfield[b]
[a]Monash University and [b]Australian Institute of Aboriginal and Torres Strait Islander Studies

1. Introduction

The data sets used in corpus linguistics and language documentation are often of different scales, and designed and used for different purposes. However, the over-arching goals of each subdiscipline are very similar:

"A corpus seeks to represent a language or some part of a language." (Biber, Conrad, & Reppen, 1998, p. 246)

"The aim of a language documentation ... is to provide a comprehensive record of the linguistic practices characteristic of a given speech community." (Himmelmann, 1998, p. 166)

We assume that the Australian National Corpus (AusNC) should have the goal of representing language in the Australian community. To this end, (and in addition to Australian Englishes) the AusNC should include records of traditional and contemporary Aboriginal and Torres Strait Islander[1] languages, community migrant languages, and sign languages (see Johnston, this volume). In order to meet these demands, the AusNC should be conceived of as multimodal and multilingual.[2] Designing a corpus along these lines presents technological, ethical, and logistical challenges. We suggest that recent experience in responding to these challenges in language documentation can inform the development of the AusNC.

In this paper, we begin by describing the similarities and differences between corpus linguistics and language documentation (Section 2). We identify some common technological and ethical challenges (Section 3) and suggest that for the purposes of developing the AusNC, a solution to some of these challenges may be to conceive of the AusNC as a set of distributed resources, rather than as a centralized holding (Section 4). We examine the consequences of such a model and recommend that a national audit of existing language materials is the crucial first step in the design of a multimodal, multilingual, and distributed AusNC (Section 5). Our conclusions are summarized in Section 6.

2. Corpora and Documentation

In this section, we consider some ways in which corpora and documentations differ. These include the type of data which is usually included in each sort of collection (Section 2.1), the ways in which that data may be manipulated (Section 2.2), the treatment of multilingual data (Section 2.3), and the different approaches to annotation taken in each case (Section 2.4).

[1] We use the terms 'Aboriginal' and 'Torres Strait Islander' to refer to the two distinct cultural groups, and 'Indigenous' as a hypernym to refer to both groups.

[2] This is an innovation in comparison with the British National Corpus and the American National Corpus, which are both English-monolingual and text-only.

2.1. Coverage – Types of Data

Both types of collection typically include data from a range of genres. Sources which discuss the design of a corpus give advice on achieving a balance between, for example, written text and transcribed oral texts, between formal and informal registers and so forth (Atkins, Clear, & Ostler 1992; Biber, 1993). For instance, the British National Corpus includes written and spoken language data (in the proportion 90% to 10%). Of the written material, 75% is informative and 25% is creative writing, and 60% of the material was published as books and 25% as periodicals. The spoken language data are divided into two parts, one a demographic survey and the other a collection of context-governed text. (Leech, 1992). Corpora are usually significantly larger than documentations. Linguists use corpora to describe linguistic phenomena on account of the frequency of their occurrence, and by means of their attested context(s) (McEnery, Xiao, & Tono, 2006).

A documentation also aims to cover a range of genres and registers (Himmelmann, 1998). However, collections of this type may include material which would normally be excluded from a corpus on the grounds that it is not sufficiently naturalistic (e.g., Evans & Sasse, 2007). Such material may include word lists, directly elicited phrases and sentences and speakers' judgements, research-focussed interviews, transcription checking sessions, and conversations about speakers' metalinguistic insights. All of these metalinguistic data are vital to developing a general linguistic description of the language, as a documentation is usually too small for all members of a paradigm or all senses of a lexeme to be reliably attested.

Although such sources are not amenable to many of the techniques of corpus linguistics, they are the only language materials which are available for a number of Australia's indigenous languages, and we therefore believe that a properly designed AusNC should be able to accommodate them.

2.2. Manipulation of Data

It is rather common for a documentation to include multiple versions of the same data. For example, the linguist(s) may include their first transcription of a piece of data, with the speakers' performance errors faithfully notated as well as a more polished version which excludes (or at least minimizes) such details. Neither of these versions may be considered as appropriate by the speakers or by their wider community when dissemination of material is planned. At that point, yet another version may come into existence which is carefully edited by the speech community and which may differ considerably from the original, more or less spontaneous, text. These different texts are all considered part of the documentation, and the decisions made by speakers in 'cleaning-up' the data often provide useful metalinguistic insights for researchers and the speech community alike (Mosel, 2008).[3]

Data collected in a corpus, on the other hand, is essentially static. Annotation, such as part of speech tagging for example, may be added to the basic data in order to make analysis possible, but the actual language data are not changed in this process. This relationship is made explicit where stand-off annotation is used (Ide & Suderman, 2007): The data are constant and many separate annotation layers can be added.

If the AusNC is to make use of the type of data discussed above, then a strategy for version control must be developed to manage ingestion of alternative versions of a single data set, and its related transcriptions. Apart from the benefits of being able to ingest multiple versions of language documentation data, this strategy would enable researchers to compare alternative analyses on a single data set, whether it originated in documentation or not. For example, discourse analysis transcriptions involve complex tagging of multiple simultaneous phenomena, and are often highly subjective. Having access to different transcriptions of the same data allows for comparison, and having primary data stored in an accessible location allows for direct reference to it and therefore greater transparency in our analyses.

---

[3] See Section 3.1 for discussion of the idea that any transcript can be regarded as an annotation of a recording when that exists.

2.3. Multilingual data

Documentations typically use more than one language. The primary data are in some language, the object language for that collection, while other parts of the collection will use some metalanguage or metalanguages. Technical information, such as grammatical analysis in the form of glossing, will often use English (or a semi-controlled vocabulary of it such as that advocated in the Leipzig Glossing Rules[4]) or some other widely used language. But other material, for example free translations of texts as given by language consultants, may use regional or national languages of wider communication. Annotations are in a language or languages of wider communication, such as the national language (Standard Australian English) and/or regional lingua francas such as Kriol. Metalinguistic data (as described in Section 2.1) often feature multilingual conversations.

The relationships between these different languages in a documentation are complex, and these relationships are complex both in the structures implied and in the levels of correspondence which are assumed. For example, interlinear glossing is a mode of presentation commonly used for language data in a documentation, and this format represents a complex data structure (Bow, Hughes, & Bird 2003). Where corresponding versions of material are collected in different languages, it may be unclear which sections should be taken to correspond. A free translation of a story may correspond to the object language version on a clause-by-clause basis, but it may equally only correspond at the level of large discourse units.

In comparison, corpora which include data from more than one language typically have simple relationships between the various languages. One case is that in which annotation is in a different language from that of the primary data. In such a case, it is essential for the purposes of corpus work that the annotations are explicitly linked to specific units of the data. Another possibility is that the corpus itself contains data from more than one language, in which case the paradigm is the parallel corpus where a unit of data in one language corresponds precisely to a unit in the other language or languages (see, for example, Steinberger et al., 2006).

2.4. Annotation

Several differences between corpora and documentations have been noted in previous sections, most of which we suggest can be viewed as differences in the approach to the issue of annotation. Documentations typically contain several layers of material, which might include interlinear glossing, free translation of different types, encyclopaedic information, ethnographic information, and what might be called 'meta-annotation' – for example, notes made by the transcriber about items needing to be checked, comments on the linguistic form, and the annotator's analytic comments which structure the annotations. Audio and video data may have time-aligned annotations, which can be generated in transcription software such as ELAN or Transcriber. Corpus material is typically less richly annotated. Tagging for various sorts of information which may be analytically interesting is possible, but it is normally task specific.

We suggest that these differences are related to the different purposes which annotation serves in each case. Documentation is conceived of as multi-purpose (Himmelmann, 1998). To this end, annotation in a documentation gives access to the data. The user who comes to the material with little background knowledge should be able to use the annotations to find information which interests them. Corpora also are ideally multi-purpose tools, but excessive annotation may impede access to the data for some users. Rich annotation of the data which is embedded in the files which make up the corpus can cause problems for a user coming to the data with new analytic purposes.

These two approaches can be reconciled when we conceive of data and annotation as separate entities, that is, when we use the concept of stand-off annotation (Ide & Suderman, 2007). We suggest that the use of a design based on stand-off annotation should be a crucial element of the AusNC, making possible the storage of data from a diverse range of sources in a way which makes that data

[4] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

maximally useable for as many people as possible. In addition, as we suggest in Section 3.1, a design which utilizes stand-off annotation will make multimedia data tractable within a corpus project.

3. Common Challenges

Technological and ethical challenges are common to both corpora and documentations. In this section we detail issues of: text and media (Section 3.1); metadata (Section 3.2); migration (Section 3.3); and ethics and access (Section 3.4).

3.1. Text and Media

While text is, and will remain, fundamental for most language research, audio-visual media (AV) is the means to the **real** data for spoken language. Documentation has exploited the fact that high quality AV is now affordable and easy.

Corpora have largely remained text-only. For example, the British National Corpus and the American National Corpus are both text-only. Although in each case some parts of the data are transcripts of spoken language, the original recordings are not accessible via the corpora.[5] Of course, text is highly searchable, which is essential in corpus linguistics.

We have already introduced the concept of stand-off annotation. The data to which such annotation relates need not be text data; what is essential is that the annotation is precisely linked to some section of primary data. The primary data itself might be text or it might be a section of an audio recording specified by time codes. On this basis, we suggest that AV data should be included in the AusNC. Where transcripts exist (treated as a form of annotation in themselves), the AV data provide the authoritative reference for researchers. At some time in the future, it is possible that tools will be developed for direct searching of at least audio streams.

3.2. Metadata

Metadata is data about data, for example, when a recording was made, who the speakers and the recorder are, which language(s) is being spoken etc. Detailed metadata is very important in documentation, and good standards have been (and still are being) developed. Examples include the Open Languages Archives Community (OLAC)[6] and ISLE MetaData Initiative (IMDI).[7] Corpus data which meet one of these standards make the data more accessible, and easily allow addition of data from other sources, such as data generated in language documentations. That is, we regard use of a well-accepted metadata standard as part of the common technical standard which should be established as part of the AusNC project.

3.3. Migration

A national corpus will ideally capture and represent language use over time in text, audio, and video formats. The average life of data formats is perhaps 5 years, and this presents a challenge when designing a corpus which aims to make the data available in the long term. Long-term storage implies the preservation of old technology and/or migration of data to new formats. Expertise in managing these issues is available domestically at archiving institutions such as Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[8] and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS).[9]

---

[5] In the case of the BNC, original sound recordings are deposited in the Sound Archive of the British Library.
[6] http://www.language-archives.org/
[7] http://www.mpi.nl/IMDI/
[8] http://www.paradisec.org.au/
[9] http://www.aiatsis.gov.au

3.4. Access and Ethical Issues

Corpus data have typically been taken from large speaker populations, and this practice means that individual speakers are very unlikely to be identifiable. This facilitates open access to the corpus data. Deidentification is also more easily achieved in a text-only corpus than a multimodal corpus which features recordings of speakers' voices, and their video image.

Documentations are different from traditional corpora with respect to speaker identification and access. Endangered languages with small speaker populations are the typical object languages of documentations. Speaker identification may be unavoidable in such cases, but, beyond that, may be desirable or necessary. In cases where the speaker (and their family or clan group) own or share the intellectual property in a particular story or song, the speaker may reasonably expect to be identified with their production of that story or song. Cutfield's experience in south-western Arnhem Land is that speakers universally expect to be identified with their language recordings (see also Thieberger & Musgrave, 2007).

Comprehensive coverage of genres and speakers often includes restricted or sensitive material. For example, in Aboriginal Australia, some cultural material is restricted to persons who have obtained a certain status (e.g., through initiations), or is restricted to men or women only. There are also traditional restrictions on naming persons who have died, or on viewing their image.[10] Documented material may also be regarded as sensitive for more personal reasons. Some personal histories include stories of settler violence, and of forced removals of part-European children from their Aboriginal families. Recordings of public meetings may also contain revelations of personal disputes. Such revelations are understood to be 'public' in the context of a community meeting, but the individuals involved may not agree for these to be made 'publicly' available to a wider audience.

These examples (and of course many others which we have not mentioned) raise the issues of consent and data distribution. Documentary linguists are generally diligent in discussing and recording access consent decisions with speakers. However, rapid development in technology influences what we may understand 'consent at a given point in time' to cover. 'Retrospective consent' refers to consent for existing materials being put to new purposes. For example, can we assume that the consent given by speakers in the 1980s for public access to their recordings will also apply to distribution of these recordings by means of the internet? This problem also applies into the future with 'prospective consent': What may be done with data in the future? (Thieberger & Musgrave, 2007)

An AusNC built along the lines which we suggest will be an innovative corpus project, as well as one of national importance. In such a project, it is imperative that we strive to include multimodal data, including such data from Australia's indigenous languages, even if challenging accommodations for a 'corpus' (as traditionally conceived) need to be made. To properly manage relative expectations of speaker identification, appropriate distribution of sensitive material, and consent-relevant-to-use, access-management conditions may need to be introduced to the corpus. Expertise on access and rights management is available at AIATSIS. Additionally, some material may need to be excluded from the corpus, where access restrictions are considered too limited for the purposes of the AusNC.

4. Distributed Resources

We suggest that there would be advantages to conceiving of the AusNC as a group of distributed resources rather than a centralized holding. There are several major documentation archives around the world (e.g., PARADISEC, AIATSIS, ELAR,[11] DoBeS[12]). Linguists usually archive their documentations at more than one of these locations. This practice results in some overlap in the collections of the respective archives, which is not viewed as a negative. Rather, it is considered an

[10] Many speakers are relaxing these restrictions in the context of language endangerment, to allow subsequent generations maximum access to language recordings.

[11] Endangered Languages ARchive <http://www.hrelp.org/archive/>

[12] Dokumentation Bedrohter Sprachen (Documentiation of Endangered Languages), Volkswagen Stiftung (Foundation) <http://www.mpi.nl/DOBES/archive_info/>

advantage to have multiple archived documentations, as this creates back-ups for each collection, and also creates multiple locations for local access (i.e., real-world access).

Such a model depends crucially on the establishment of common technical standards which must be satisfied by any data which is to be included in the AusNC. If this step can be accomplished, then it is straightforward to establish a 'virtual centrality' online for a distributed AusNC. Contributing archives will be linked by a set of network services based on a level of interoperability ensured by the technical standard. Access to any portion of data will be via the central services, but the location of the data can vary from item to item. New data which meets the technical standard can easily be ingested at any contributing archive. This last point implies that one part of the central network services will handle version control, ensuring that where data holdings are duplicated in more than one archive such data are always synchronized.

We suggest that although this model may be more difficult to implement initially, it would be easier to administer and maintain in the long term, with many institutions taking responsibility for data storage.

5. Consequences

Thus far, we have proposed a model for the AusNC which is multimodal, multilingual, and distributed. We have drawn on our experience in language documentation to highlight issues and possible solutions in the development of such a model. In this section we discuss the consequences of this model, and further refine the proposal. Specifically, we identify existing expertise (Section 5.1) and data sources (Section 5.2), and ask 'what might be overlooked?' (Section 5.3).

5.1. Existing Expertise

There exists in Australia significant expertise in digital archiving and data management, as well as in rights management. PARADISEC and AIATSIS are both world leaders in developing and managing digital AV collections. AIATSIS also has expertise in rights management, particularly as related to Indigenous cultural material.

Australian researchers who have been funded and trained by bodies such as HRELP[13] and DoBeS also have considerable expertise in the practices of collecting, managing, and distributing digital documentations of endangered languages.

There are also several Australian researchers with expertise in the theory and design of digital archives, collections, and corpora, as a result of their work with bodies such as PARADISEC and OLAC (Barwick & Thieberger, 2006; Bird & Simons, 2003).

5.2. Existing Data Sources

There are large holdings of language data already existing in Australia. These are varied in type, format, and coverage. As much of this material as possible should be incorporated into a national corpus. To this end, an audit is needed. An audit would need to review both data coverage (what is already covered?) and technical coverage (what can easily be brought up to a necessary standard?).

5.3. What Might be Missed?

Some language material is of great interest but would not necessarily be identified for inclusion in the AusNC, when the focus is firstly on an adequate coverage of Australian English and secondarily on data resulting from the documentation of endangered languages. We refer specifically to contemporary Indigenous languages (e.g., creoles and Aboriginal Englishes) and community (migrant) languages.

---

[13] Hans Rausing Endangered Languages Project, School of African and Oriental Studies, University of London <http://www.hrelp.org>

5.3.1. Contemporary Indigenous Languages

Creoles and Aboriginal Englishes are contact languages which are spoken in many areas of Australia as contemporary Indigenous codes and regional lingua francas. Examples include dialects of Kriol spoken across the Kimberley, Top End, and north-west Queensland, and Torres Strait Broken; and Aboriginal Englishes of south-western Western Australia, NSW, and Alice Springs (Sandefur 1979, 1986; Shnukal, 1994). These varieties typically have features in common with the traditional Indigenous languages of their regions (e.g., phonology, semantics), but they also have features attested in other contact languages around the world (e.g., morphosyntax).

Contemporary Indigenous languages are receiving increasing attention from linguists especially within the area of language contact (see references cited in the previous paragraph as well as McConvell & Meakins, 2005; Disbray & Simpson, 2005; Munro, 2005). Including these languages in the corpus has many advantages for analysis. For example, contact languages often change very rapidly, and can vary greatly in small geographic areas, on account of different substrate influences. Existing recordings of Australian contact languages vary across time and place, and thus allow researchers to develop insights into processes of change in contact languages (including decreolization) and substrate influences. Access to data with such temporal and geographical variety would be very difficult for a single researcher or research team to collect.

Many existing recordings of traditional Aboriginal and Torres Strait Islander languages also feature contact languages being spoken. For example, a parallel translation of a text in a traditional language may be offered in Kriol or Aboriginal English, or metalinguistic conversations between the researcher and language speakers may take place in a contact language. The metadata for these recordings may not accurately reflect the use of contact languages (i.e., the metadata may record only the name of the traditional language, the object language). This limitation presents an issue for the audit, as well as for researchers wishing to access data on contemporary Indigenous languages.

We propose that existing recordings of contemporary Indigenous languages be included in the AusNC and that a priority of the corpus be to seek to include new data on contemporary Indigenous languages. Additionally, we suggest that the AusNC ask all contributors to include in their metadata any instances of the use of contemporary Indigenous languages.

5.3.2. Community Languages

Community languages are the languages of migrant populations in Australia. There is already a substantial body of research on these languages (see Clyne, 2005), and a substantial amount of data already collected. However, unlike research on Aboriginal and Torres Strait Islander languages, there is no obvious central institution which encourages researchers in this field to archive their data for posterity, and to make it available for other researchers. Consequently, it is assumed that individual researchers have been responsible for maintaining their own collections. Identifying the numerous data on community languages presents an issue for the proposed audit.

The situation for community languages in Australia is also constantly shifting, with new migrant groups arriving (e.g., most recently from Somalia and Sudan), and intergenerational changes in language maintenance and use. Intergenerational change makes it all the more imperative to identify and make available older data on these languages, in order to accurately identify and describe language change phenomena in these communities. Further, this situation raises in rather an acute form the question of what constitutes a representative data set.

6. Conclusions and Recommendations

In this paper we have proposed that the AusNC should consist of a distributed set of multimodal and multilingual resources which meet common technical standards. The necessary expertise to design, develop, and maintain such a corpus is already largely available in Australia. We have identified some basic design principles for the corpus, namely: separating the data and its time-linked annotations; using sound protocols for version control across distributed storage; identifying and/or ingesting

metalinguistic data as part of the corpus; and, most importantly, developing metadata and technical standards in line with existing international best practice.

We also propose that an audit of existing language material held by Australian institutions and individuals is necessary. The audit would report to two questions: (a) What data exists and what coverage does it provide? (b) How much of this can easily be brought up to necessary technical standards?

7. Acknowledgements

The authors would like to thank the participants at the HCSNet Designing the Australian National Corpus Workshop for their questions and comments which assisted us in formulating our ideas, and to Michael Clyne for helpful advice. Remaining errors are our responsibility.

References

Atkins, Sue, Jeremy Clear, & Nicholas Ostler. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*, 1-16.

Barwick, Linda, & Nicholas Thieberger. (2006). Cybraries in paradise: New technologies and ethnographic repositories. In Cushla Kapitzke & Bertram C. Bruce (Eds.), *Libr@ries: Changing information space and practice* (pp. 133-149). Mahwah, NJ: Lawrence Erlbaum.

Biber, Douglas. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8,* 243-257.

Biber, Douglas, Susan Conrad, & Randi Reppen. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Bird, Steven, & Gary Simons. (2003). Seven dimensions of portability for language documentation and description. *Language, 79,* 557-582.

Bow, Catherine, Baden Hughes, & Steven Bird. (2003). Towards a general model of interlinear text. *Proceedings of EMELD Workshop 2003: Digitizing & annotating texts & field recordings*. LSA Institute, Lansing MI, USA. Available from http://emeld.org/workshop/2003/proceeding03.html (Retrieved 31/07/09).

Clyne, Michael. (2005). *Australia's language potential.* Sydney: University of New South Wales Press.

Disbray, Samantha, & Jane Simpson. (2005). The expression of possession in Wumpurrarni English, Tennant Creek. *Monash University Linguistics Papers, 4* (2), 65-86.

Evans, Nicholas and Hans-Jürgen Sasse. 2007. [On-line reprint, slightly revised, of Evans & Sasse 2004]. Searching for meaning in the library of Babel: field semantics and problems of digital archiving. *Archives and Social Studies: A Journal of Interdisciplinary Research* Vol. 1.0:63-123. On-line publication available at http://socialstudies.cartagena.es/index.php?option=com_content&task=view&id=53&Itemid=42, Accessed 17/08/09.

Himmelmann, Nikolaus P. (1998) Documentary and descriptive Linguistics. *Linguistics, 36* (1), 161-195.

Ide, Nancy, & Keith Suderman. (2007). GrAF: A graph-based format for linguistic annotations. In B. Boguraev, N. Ide, A.Meyers, S.Nariyama, M.Stede, J.Wiebe et al. (Eds.), *The LAW*: *Proceedings of the Linguistic Annotation Workshop* (pp. 1-8). Stroudburg PA: Association for Computational Linguistics.

Leech, Geoffrey. (1992). 100 million words of English: The British National Corpus (BNC). *Language Research, 28,* 1-13.

McConvell, Patrick, & Felicity Meakins. (2005). Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics, 25* (1), 9-30.

McEnery, Tony, Richard Xiao, & Yukio Tono. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Mosel, Ulrike. (2008, October). Putting oral narratives into writing: Experiences from a language documentation project in Papua New Guinea. Paper presented at the Linguistics Program seminar, Monash University.

Munro, Jennifer M. (2005). *Substrate language influence in Kriol: The application of transfer constraints to language contact in Northern Australia.* Unpublished PhD thesis, University of New England, New South Wales.

Sandefur, John R. (1979). *An Australian Creole in the Northern Territory: A Description of Ngukurr-Bamyili dialects (part 1).* Work papers of SIL-AAB. Series B, Volume 3.

Sandefur, John R. (1986). *Kriol of North Australia: A language coming of Age*. Work papers of SIL-AAB: Series A, Volume 10.

Shnukal, Anna. (1994). Torres Strait Creole. In Nick Thieberger & William McGregor (Eds.*), Macquarie Aboriginal words* (pp. 374-398). Sydney: The Macquarie Library Pty Ltd.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş et al.(2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (pp. 2142-2147). Available from http://www.sdjt.si/bib/lrec06/ (Retrieved 14/08/09).

Thieberger, Nick, & Simon Musgrave. (2007). Documentary linguistics and ethical issues. In Peter K. Austin (Ed.), *Documentary and descriptive linguistics: Vol. 4* (pp. 26-37). London: School of Oriental and Asian Studies.

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Musgrave, Simon and Sarah Cutfield. 2009. Language Documentation and an Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 10-18. Somerville, MA: Cascadilla Proceedings Project.

or:

Musgrave, Simon and Sarah Cutfield. 2009. Language Documentation and an Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 10-18. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2282.