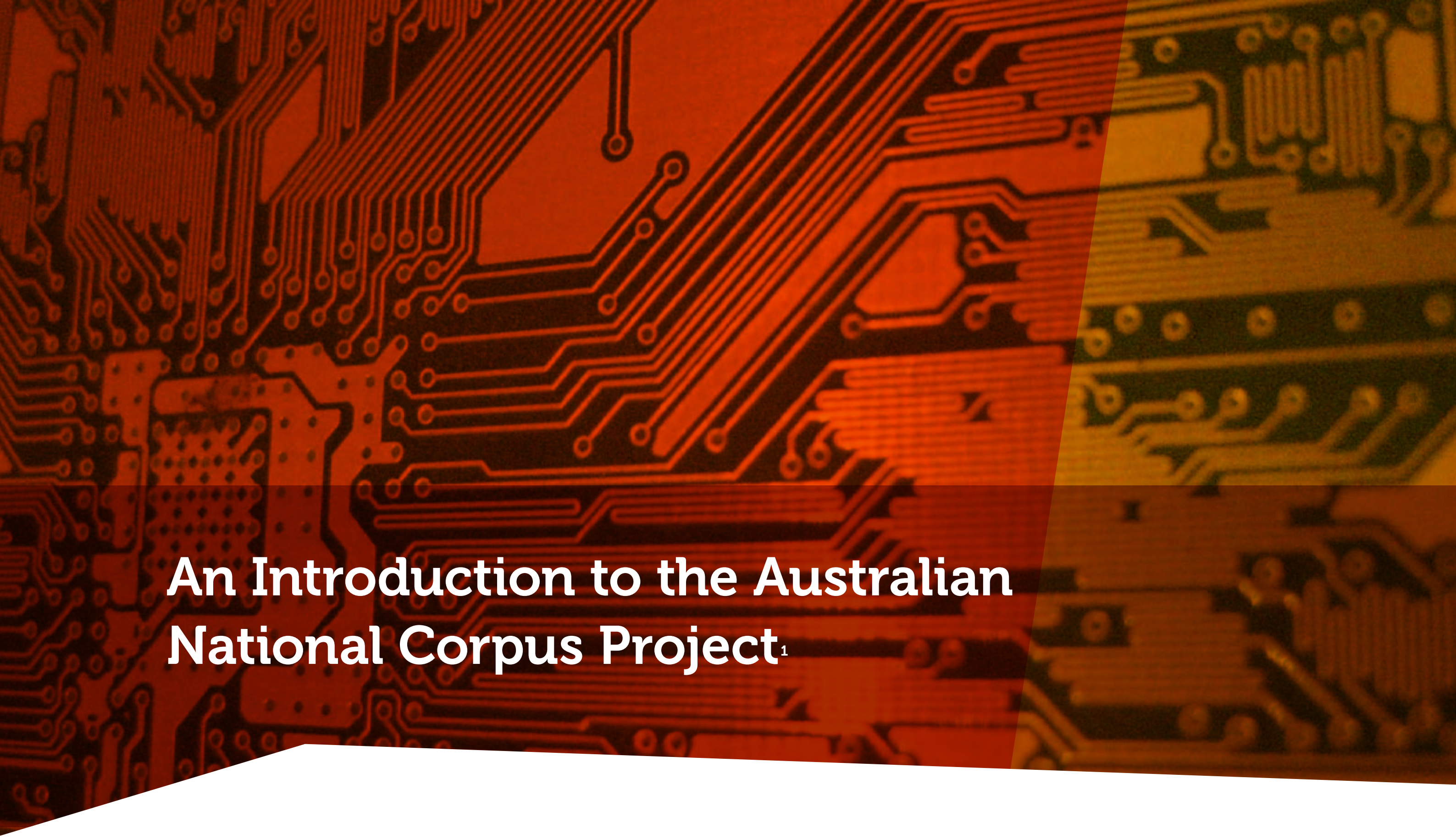


An Introduction to the Australian National Corpus Project[1]

» SIMON MUSGRAVE

In the last few years, a consensus has emerged from researchers in various disciplines that a vital piece of research infrastructure is lacking in Australia, namely, a substantial collection of computerised language data. A result of this consensus is an initiative aimed at the establishment of an Australian National Corpus. Australia's language data resources at present remain scattered and relatively inaccessible, and the Australian National Corpus Initiative constitutes a sustained effort on the part of linguists, applied linguists and language technologists to overcome this data inaccessibility by establishing a massive online database of spoken and written language in Australia, in all its forms and diversity (audio files, written texts, etc.). It represents a major expansion of Australia's e-research infrastructure in the humanities and social sciences. The Australian National Corpus is being developed (currently with support from the Australian National Data Service) as a linked set of multimodal and multilingual language resources that represents the Australian linguistic landscape, unified through common technical standards.[2]

The Australian National Corpus (AusNC) is therefore an ongoing project to collate and provide access to a wide range of samples of Australian language for use in academic research. Building on earlier work collecting corpora in different disciplines, AusNC will bring existing and newly collected samples together in one place and provide tools to help researchers annotate, analyse and work collaboratively on this data.

The corpus will contain collections of:

- published texts from many genres
- transcribed speech, often with aligned audio files
- visual records of interaction (video)
- electronic texts including email, blogs and social media.

AusNC aims to illustrate Australian English in all its variety: situational, social, generational, and ethnic; and to document languages other than English used in Australia, including Auslan, and the community languages of immigrants. The corpus will be built from existing collections contributed by researchers from many disciplines; these will be adapted as needed to allow them to be properly integrated into AusNC. The project welcomes suggestions and offers of already digitised collections of English and other languages used in Australia, to complement the main collection. All these different types of language data will then support a very wide range of researchers and their needs, including those of:

- linguistic researchers
- English language teachers, for school and adult education
- lexicographers and terminologists

- translators and interpreters
- speech and language pathologists
- natural language processing and language engineering
- other language-oriented research in the humanities and social sciences.

The last category mentioned above could include work in history, sociology, social psychology, cultural studies; indeed, any area which relates to Australian society and culture.

AUSTRALIA WAS A SITE OF GREAT LINGUISTIC DIVERSITY BEFORE EUROPEAN SETTLEMENT

Initial discussions concerning a possible AusNC have emphasised the diversity of research agendas which it might support and the corresponding diversity of content which might be desirable. These have particularly concentrated on three areas. Firstly, there is a consensus that an AusNC must have a carefully planned core component which is comparable to other large corpora, but questions remain about whether (or how much) technological change should influence this design. Secondly, there also consensus that an AusNC should represent language use in Australia beyond Australian English, which would make it significantly different from existing national corpora. Thirdly, if an AusNC is to accomplish the various goals mentioned here, it is clear that the design of the technical infrastructure will be of great importance.

The design of the AusNC has not yet been finalised, but there is little doubt that it will include a very substantial body of text data which can be utilised for comparison with sub-corpora of the British National Corpus (BNC)[3] or the American National Corpus.[4] Nevertheless, questions remain about the extent to which it is sensible to make comparability a high priority. In particular, the BNC was assembled around 1990, and therefore computer-based text types are scarcely represented in it. Any attempt to represent the use of the English language in Australia in the first decades of the twenty-first century obviously cannot afford to neglect such genres, and the AusNC initiative can be expected to include substantial amounts of such data. But should this be seen as an aspect of the corpus additional to those sections which provide comparability with earlier collections, or should some elements of comparability be sacrificed in order to make coverage of the newer genres more complete? Inevitably, such decisions will in the end depend also on resource allocation, but the decisions will have to be made relative to the expressed needs of various research communities.

The development of computer-mediated communication and the recognition of computer-based textual genres is one important change since the BNC was assembled. Another is the huge improvement in the possibilities for creating and disseminating high-quality recordings, both audio and video, of language in use.[5] Concurrent with these developments, and interdependent with them, has been an increasing focus on multimodal data as the basis for comprehensive language research. This change is in turn interdependent with the emergence of language documentation as a sub-field of linguistics.[6] A major corpus being designed now must take these developments into account, which means that the AusNC will include a substantial component of recordings of actual language use of various types. For such material, the actual multimodal recordings will be the basic data. This contrasts with the approach of the BNC, which includes approximately 10% of data from spoken language, but only transcripts are immediately accessible for analysis; the original sound recordings are part of the Sound Archive of the British Library, but are not treated as a part of the corpus itself. The proposed inclusion of audio(visual) recordings and computer-mediated communication in AusNC inevitably means that at least part of the language data held in the corpus will not be directly comparable with other major corpora, but does, on the other hand, raise extremely interesting new research possibilities.

AusNC has as one of its aims to represent language in Australia in total, that is, to go beyond only representing the use of (more or less) standard English in Australia. This aim is of considerable importance to many members of the research communities involved in the initiative, and can be considered a core objective. Australia was a site of great linguistic diversity before European settlement.[7] A small part of that diversity remains and the indigenous peoples of Australia also speak distinctive varieties of English (scarcely represented in written texts) and various contact varieties.[8] In addition, there has been a huge change to the language picture of Australia as a result of migration in the last half century.[9] Further, Australia is one of the few places in the world where a sign language has been documented in detail with extensive video collections available to represent Auslan in AusNC.[10] Ideally, all of this diversity will be represented in the AusNC.

THE AUSTRALIAN NATIONAL CORPUS IS INTENDED TO MEET A CRUCIAL NEED BY PROVIDING ACCESS TO LARGE AMOUNTS OF DATA ON LANGUAGE AND LANGUAGE USE IN AUSTRALIA

A crucial step in designing the AusNC is the creation and promulgation of a set of technical standards. These standards will have to specify the required formats of material which can be accepted into the corpus, the associated metadata which will be necessary to access it, the discovery and access systems to be used, and a storage architecture.[11] A further important aspect of the standards associated with the corpus is a framework for handling legal and ethical issues that come with managing large bodies of data. These questions have been addressed, initially in a workshop funded by the Australian Academy of the Humanities, and then with the help of legal counsel, with the result that AusNC has a legal framework that includes a contributor's licence, end-user agreement, and other legal protections.[12]

The AusNC project is based on a Statement of Common Purpose adopted by the Australian Linguistic Society and the Applied Linguistics Association of Australia in 2008. One part of the

You are here: Home

Welcome to the Australian National Corpus

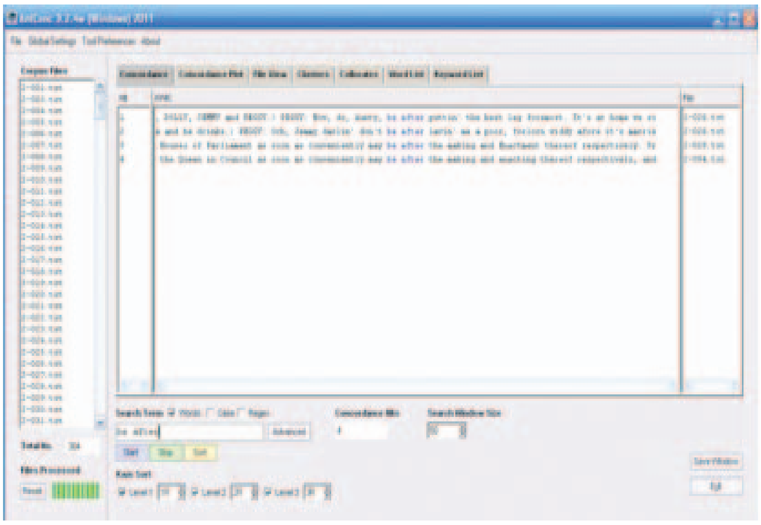
The Australian National Corpus (AusNC) is an ongoing project to collate and provide access to a wide range of samples of Australian language for use in academic research. Building on earlier work collecting corpora in different disciplines, AusNC will bring existing and newly collected samples together in one place and provide tools to help researchers annotate, analyse and collaborate on this data.

The corpus will contain collections of:

- published texts from many genres
- transcribed speech, often with aligned audio files
- visual records of interaction (video)
- electronic texts including email, blogs and social media

LEFT

AUSNC webpage screengrab.



ABOVE

A KeyWordInContext search using COOEE data. From research by S. Musgrave and K. Burridge on Irish features in early Australian English.

statement reads: 'We further propose that such a corpus should be freely accessible and useful to the maximum number of interested parties', and this commitment leads naturally to a conception of the AusNC as a distributed group of resources meeting common standards which allow them to be linked by a set of network services. In most cases, users will interact with the corpus via a network connection (cf. the Corpus of Contemporary American English which is only available online).[13]

Two things crucial to ensuring that such an architecture is possible will be well-understood metadata standards and a coherent approach to annotation. Metadata for linguistics resources has received a good deal of attention over the last decade.[14] There are currently two well-developed standards which can be used at least as a basis for new projects: the Open Language Archives Community metadata scheme, and the IMDI metadata scheme.[15] The corpus is organised conceptually as consisting of *collections*, that is, groups of data created by a person or group as a coherent resource. Collections consist of individual *items* of data which may or may not be accompanied by *annotations*, that is, linguistic and social information attached to the data. The current stage of the project has developed a metadata scheme which accurately describes the corpus at the collection level and at least partially at the item level.[16]

In order to ensure that data from a diverse range of sources can be stored in a way that makes that data maximally useable for as many people as possible, the use of standoff annotation is a crucial design principle for the AusNC.[17] Treating annotation as distinct from primary data will ensure that data is multi-purpose and maximally accessible for diverse types of research. This approach will also have the advantage of making multimodal data tractable. The data to which standoff 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 might be a section of an audio recording specified by time codes, and the annotation can be a transcript of the specified section of a recording, just as tagging for parts of speech might be the annotation for a specified segment of text. The use of standoff annotation makes the two possibilities conceptually equivalent.

An audit of existing data (which has begun) will seek to identify holdings of any type of language data (English or other languages, text or multimodal) in a condition suitable for inclusion, as well as data that can be brought to the technical standards of AusNC with a relatively small investment. In the future, researchers across all aspects of language in Australia will be encouraged to create data and metadata which meet the standards of AusNC so that such data can be added to the collection relatively easily.

In the initial stages of the project, ten existing collections of data will be made available via a common web portal. The collections are listed and briefly described below; they should be available via AusNC early in 2012.

Australian Corpus of English (ACE)

The ACE corpus was compiled to match Australian data from 1986 with the standard American and British corpora (Brown and LOB) from the 1960s. It includes one million words of published text in 500 samples from 15 categories of nonfiction and fiction.

Australian Radio Talkback (ART)

ART is a set of samples of Australian talkback radio (2004-2006), totalling just over 200 000 words, from national, regional and commercial radio. It was collected in connection with an ARC-funded project: Australian English Grammar.

AustLit

AustLit provides full-text access to hundreds of examples of out of copyright poetry, fiction and criticism ranging from 1795 to the 1930s. The collection includes literature intended for popular audiences as well as literature intended for audiences concerned with literary quality or the establishment of a national canon. The bibliographical information associated with these records enables researchers to investigate the relationships between texts and particular publishers or to track the first publication of each text in newspapers, magazines or journals. This provides indirect evidence of the original audience for each text and the evolution of reception over time if the texts were subsequently republished in other contexts.

Braided Channels

The Braided Channels research collection includes materials collected on Australia women, land and history in the Channel country. The collection is constructed from some 70 hours of oral history interviews with women from Australia's Channel Country, together with archival film, transcripts, photos and music. It includes examples of Aboriginal English.

Corpus of Oz Early English (COOEE)

This material, collected by Clemens Fritz, had to meet a regional and a temporal criterion. Texts had to be produced between 1788 and 1900 and written in Australia, New Zealand or Norfolk Island. But in a few cases, other localities were allowed.

> THE CORPUS OF OZ EARLY ENGLISH (COOEE) OFFERS FASCINATING POSSIBILITIES FOR RESEARCH ON THE LANGUAGE OF NINETEENTH-CENTURY AUSTRALIAN LITERATURE

Email Australia

10,000 emails submitted for a project sponsored by nineMSN and the Powerhouse Museum called Email Australia in which people submitted their favourite emails to be included in Australia's first email archive.

Griffith Corpus of Spoken English (GCSAusE)

GCSAusE comprises a collection of transcribed and annotated recordings of spoken interaction amongst Australian speakers of English, as well as users of English in Australia more generally, collected by staff and students at Griffith University.

International Corpus of English (Australia's contribution is ICE-AUS)

The ICE-AUS is a one million-word corpus of transcribed spoken and written Australian English from 1992 to1995. Its internal structure with 500 samples (60% speech, 40% writing) matches that of other ICE corpora (associated with the International Corpus of English).

Mitchell & Delbridge

The Mitchell and Delbridge database contains recordings of Australian English as spoken by 7736 students at 330 schools across Australia, mostly collected in 1960. The tapes were digitised and reissued as an online database by the University of Sydney in 1997-98.

Monash Corpus of Spoken English (MCE)

MCE consists of a collection of recordings and transcriptions of interviews made in Melbourne c. 1997. The subjects of the interviews were adolescents from a variety of schools. The data were collected and transcribed by staff of the Linguistics Programme at Monash University.

These collections provide a sample which illustrates the aims of the project almost completely. There is a substantial amount of text-based material, but some of that represents a computer-based text type. There are also collections of multimodal material; the only area which AusNC aims to cover, not present in the initial collections, is data on non-English languages used in Australia. All of the collections will be accessible and searchable from the AusNC website.

The Australian National Corpus is intended to meet a crucial need by providing access to large amounts of data on language and language use in Australia. It aims to become an essential component of the infrastructure available for e-research in Australia (and more widely). Although the initial impetus for the project has come from linguists and applied linguists, the project is designed to provide a resource of use to researchers across many disciplines in the humanities. For example, the Corpus of Oz Early English (COOEE) offers fascinating possibilities for research on the language of nineteenth-century Australian literature. COOEE includes some literary sources, but also contains data from many other types of writing which can provide valuable comparisons to the language being used in published literature at the time.

The Management Committee of the Australian National Corpus Incorporated welcome input from individual researchers or research communities with suggestions as to the types of data which should be included and the ways in which data can be made maximally useful for their purposes. We also welcome information about existing bodies of data which should be recognised in our ongoing audit and which might be candidates for inclusion in the AusNC, but this should not be taken as an indication that the primary orientation of the project is archival. AusNC intends to take a role in encouraging the collection of new language data in Australia by making it easy for researchers to identify gaps in existing coverage and by providing a service for making data easily accessible. One area given high priority in initial discussions about AusNC is that of spoken discourse. The initial collections include two historical sources of such data (Mitchell and Delbridge from the 1960s and ICE-AUS from the 1990s), but the value of these resources would be increased by the addition of current data. Such data is being collected by the AusTalk project,[18] and it is to be hoped that this data will be made available via AusNC in the future. The comparison of spoken Australian English across time which would then be possible would stimulate new research possibilities, and this is the role AusNC will have in the long-term: not only gathering in data created by previous research but also fostering new kinds of research.

SIMON MUSGRAVE is a lecturer in the School of Languages, Cultures and Linguistics at Monash University. His research interests include Austronesian languages, language documentation and language endangerment, African languages in Australia, communication in medical interactions, and the use of technology in linguistic research.

[1] I am grateful to Kate Burridge, Michael Haugh, Pam Peters and Robyn Rebollo, all of whom read a draft of this article and made suggestions which improved it.

[2] The project website is at http://www.ausnc.org.au; *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, ed. by Michael Haugh, Kate Burridge, Jean Mulder and Pam Peters (Somerville, MA: Cascadilla Proceedings Project, 2009). <http://www.lingref.com/cpp/ausnc/2008/index.html>

[3] Geoffrey Leech, '100 Million Words of English: The British National Corpus (BNC)', *Language Research*, 28 (1992), 1-13.

[4] Nancy Ide and Catherine Macleod, 'The American National Corpus: A Standardized Resource of American English', in *Proceedings of Corpus Linguistics 2001* (Lancaster, 2001).

[5] See for example Nick Thieberger and Simon Musgrave, 'Documentary Linguistics and Ethical Issues', *Documentary and Descriptive Linguistics*, 4 (2007), 26-37.

[6] Michael Haugh, 'Designing a Multi-modal Spoken Component of the Australian National Corpus', in *Selected Proceedings of the 2008 HCSNet Workshop*, ed. by Haugh et al., pp. 74-86; Simon Musgrave and Sarah Cutfield, 'Language Documentation and an Australian National Corpus', in *Selected Proceedings of the 2008 HCSNet Workshop*, ed. by Haugh et al., pp. 10-18.

[7] R. M. W. Dixon, *Australian Languages* (Cambridge: Cambridge University Press, 2002).

[8] Patrick McConvell and Felicity Meakins, 'Gurindji Kriol: A Mixed Language Emerges from Code-Switching', in *Australian Journal of Linguistics*, 25 (2005), 9-30; see also John R. Sandefur, 'Kriol of North Australia: A Language Coming of Age', *Work Papers of SIL-AAB: Series A*, 10 (1986); Anna Shnukal, 'Torres Strait Creole', in *Macquarie Aboriginal Words*, ed. by Nick Thieberger and William McGregor (Sydney: The Macquarie Library Pty Ltd, 1994), pp. 374-398.

[9] Michael Clyne, *Australia's Language Potential* (Sydney: University of New South Wales Press, 2005).

[10] T. Johnston and A. Schembri, *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics* (Cambridge: Cambridge University Press, 2007).

[11] Steve Cassidy, 'Building Infrastructure to Support Collaborative Corpus Research', paper presented at the HSCNet Workshop on Designing the Australian National Corpus, University of New South Wales, 4-5 December 2008.

[12] See <http://www.ausnc.org.au/about-1/AusNCFramework_NovConf2.docx/view>

[13] Mark Davies, 'The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights', *International Journal of Corpus Linguistics*, 14 (2009), 159-90.

[14] See for example Steven Bird and Gary Simons, 'Seven Dimensions of Portability for Language Documentation and Description', *Language*, 79 (2003), 557-582.

[15] Open Language Archives Community: <http://www.language-archives.org/>; IMDI: <http://www.mpi.nl/IMDI/>

[16] See <http://www.ausnc.org.au/about-1/ausnc-data-model>

[17] Nancy Ide and Keith Suderman, 'GrAF: A Graph-based Format for Linguistic Annotations', in *The LAW: Proceedings of the Linguistic Annotation Workshop*, ed. by B. Boguraev, N. Ide, A. Meyers, S. Nariyama, M. Stede, J. Wiebe et al. (Stroudburg, PA: Association for Computational Linguistics, 2008), pp. 1-8.

[18] See <http://austalk.edu.au/>