The AusNC Project: Plans, Progress and Implications for Language Technology

Simon Musgrave
Linguistics Program
Monash University
VIC 3800 Australia
Simon.Musgrave@arts.monash.edu.au

Michael Haugh
School of Languages and Linguistics
Griffith University
QLD 4111 Australia
m.haugh@griffith.edu.au

Abstract

In the last eighteen months, 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. The progress of this initiative is presented in this paper, along with discussion of some important design issues and a consideration of how the initiative relates to the field of language technology in Australia.

1 Introduction

Large-scale corpora are becoming an increasingly important resource in language research, including many sub-disciplines within language technology. An initiative has developed over the last year or more which aims to construct such a corpus as a key element of research infrastructure for Australia. A national repository of language data would have significant value as research infrastructure for a number of research communities in Australia and overseas, thereby increasing access to Australian language data and widening the global integration of research on language in Australia. It would facilitate collaborative ventures in collecting new language data to support multimodal research in human communication and it would consolidate presently scattered and relatively inaccessible collections of historical language data where possible within the Australian National Corpus (AusNC). Such data is of interest not only to researchers in linguistics and applied linguistics, but also to members of the wider Humanities and Social Sciences and informatics research communities who have an interest in Australian society. Such a large annotated language dataset would also provide invaluable training data for work in natural language processing, speech recognition, and the further development of semi-automated annotation.

In this paper, we will give a short overview of the progress of the initiative to date. This will be followed by an introduction to some design questions which have been the subject of discussion in the preliminary phase of the project, and a consideration of some implications of the project for the field of language technology.

2 History and Recent Progress

At least two projects can be considered to have made substantial contributions to corpus-building in Australia before the present initiative was launched. From 1986, the Australian Corpus of English was compiled at Macquarie University.[1] This corpus consists of 500 text samples of (minimally) 2,000 words each, giving a total size of approximately 1,000,000 words. This corpus has been integrated into the International Corpus of English project. The Australian National Database of Spoken Language was collected in the years between 1991 and 1995.[2] This corpus consists of recordings of various types of spoken language plus associated transcripts. Although some of the material in this collection is taken from pairs of speakers collaborating on a map task, the majority of the corpus consists of recordings of speakers reading carefully chosen prompts. The number of native speakers of Australian English who were recorded is 108. Additionally, 96 speakers from two migrant groups were recorded for a subset of the material, and a speaker from each of nine other migrant groups was also recorded. Given the nature of the material, it is not useful to attempt an estimate of the size of this corpus in terms of number of words.

Since more than a decade has now passed since these two corpora were constructed, members of various language-based research communities in Australia have come to see the importance of establishing a national corpus as an essential aspect of research infrastructure. Although primary interest in

[1] Australian Corpus of English: < http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM>
International Corpus of English: < http://ice-corpora.net/ice/index.htm>
[2] ANDOSL: < http://andosl.anu.edu.au/andosl/>

such a resource would come from linguists and applied linguists, it was also clear that a number of other groups of researchers would derive value from such a resource. In July 2008, the Australian Linguistic Society (ALS) and the Applied Linguistics Association of Australia both held their annual conferences in Sydney, and the opportunity was taken to hold a meeting of scholars interested in the development of a national corpus. The outcome of this meeting was a "Statement of Common Purpose" which includes the following wording:[3]

> the aim of developing a freely available national corpus is that it can become an ongoing resource not only for linguists, but also historians, sociologists, social psychologists, and those working in cultural studies with an interest in Australian society or culture. We therefore see such a corpus as an important part of the development of research infrastructure for humanities researchers in Australia.

The initial list of signatories to this statement has expanded since that meeting and now has 45 names on it.

The existence of the Statement of Common Purpose and of the demonstrated support for it allowed the leaders of the initiative to approach the Australian Academy of the Humanities and the ARC Network in Human Communication Science (HCSNet) to seek funding. This approach was successful, and has made possible an initial phase of planning activity. Firstly, a workshop entitled 'Designing the Australian National Corpus' was held in December 2008 as part of the HCSNet Summerfest 2008. Selected papers from this workshop will appear shortly (Haugh et al, in press [2009]). A second workshop supported by HCSNet will be held (at the same time as the ALTA 2009 Workshop) and this meeting will concentrate on questions about data sources and tools.

Another workshop, supported by the Australian Academy of the Humanities ,was held in Brisbane in May 2009 concentrating on legal and ethical issues. As discussed in the following section, the current plan for AusNC is that the corpus will include significant amounts of material from the World Wide Web and other types of computer-mediated communication. But issues of copyright and, in some cases, of confidentiality arise in relation to such data (Lampert, in press [2009]), and these issues must be resolved before data collection can begin. There may also be confidentiality and copyright problems in making available existing data which has not previously been openly accessible.

The Statement of Common Purpose discussed previously was the outcome of a meeting of interested parties and not a formal activity of the learned societies from which those parties were drawn. However, the 2009 meeting of the ALS committed the society's formal support to the initiative, with the following statement appearing in the minutes of the annual general meeting:

> The meeting expressed its strong support for this initiative to develop an Australian National Corpus, which will stand out as a significant national resource and which will contribute to the research strength of this country.[4]

In addition, that meeting voted to contribute $2500 to the initiative to support the conduct of an audit of existing language data in Australia.

This audit will commence in late 2009 and will have several aspects to it. Firstly, a survey will be sent to individuals and organizations which might be expected to have relevant holdings of data, such as linguistics departments of universities and other research bodies. Secondly, contact will be made with bodies which are known to have significant holdings such as the Australian Broadcasting Corporation and the National Film and Sound Archive. And finally, information about privately held data will be sought by making a request in the mass media.

These various activities are being directed by a steering committee which was formed following the workshop in December 2008. A list of the members of this committee can be found in the Appendix to the current paper.

3 Planning the AusNC

Initial discussions concerning a possible AusNC have emphasized the diversity of research agendas which it might support and the corresponding diversity of content which might be desirable. In this section, we will present some of the issues which have been raised in these discussions, concentrating 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 technological change should influence this design. Secondly, there is 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.

[3] The full text is available at http://blogs.usyd.edu.au/elac/2008/08/australian_national_corpus_ini.html

[4] < http://www.als.asn.au/newsletters/alsnews200908.html>

3.1 Core corpus design

A corpus is planned 'to represent a language or some portion of a language' (Biber, Conrad and Reppen, 1998: 246). In the case of an AusNC, one intention is to represent the English language as used in Australia. However, it would not be sensible to attempt to achieve this goal without taking into account comparable existing corpora. One possible strategy is that adopted by the International Corpus of English project, which has one basic design which is followed as closely as possible by all the contributing sub-corpora (Nelson, 1996). An Australian component of ICE already exists, as discussed in section 2, but the ambition of the AusNC project is to achieve a corpus which is bigger than that (1 million words) by at least an order of magnitude. The benchmark for comparability then becomes either the British National Corpus (BNC, Leech, 1992) or the American National Corpus (ANC, Ide and Macleod, 2001). These two corpora are not identical in design; although ANC was initially based on the design of the BNC, it has diverged in the course of its development. Therefore if direct comparability is sought, it is necessary to make a choice between these two. BNC is recognised as a crucial project in the history of corpus linguistics, but it is also now almost twenty years old and therefore has limitations which will be discussed shortly. ANC is also not an ideal model, as its design has evolved over time in response to various pressures (Ide, in press [2009]).

The design of the AusNC has not yet been finalized, 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 BNC or the ANC. 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 21st 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 be questions about 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 time when 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 (see, for example, Thieberger and Musgrave, 2007). Concurrent with these developments, and interdependent with them, has been an increasing focus on multimodal data as the basis for comprehensive language research and this change is in turn interdependent with the emergence of language documentation as a sub-field of linguistics (Haugh, in press [2009], Musgrave and Cutfield, in press [2009]). A major corpus being designed now must take these developments into account, and this means that the AusNC will very likely include a substantial component of recordings of actual language use of various types. For such material, the actual multimodal material will be the basic data, in contrast to 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 directly comparable with other major corpora (see section 3.2), but this, on the other hand, raises extremely interesting research possibilities (see section 4).

3.2 Other material

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 (Dixon 2002). A small part of that diversity remains and the indigenous people of Australia also speak distinctive varieties of English (scarcely represented in written texts) and various contact varieties (McConvell and Meakins, 2005, Sandefur, 1986, Shnukal, 1996). In addition to the language use of indigenous people, there has also been a huge change to the language picture of Australia as a result of migration in the last half century (Clyne 2005). Ideally, all of this diversity will be represented in the AusNC.

Initially, at least, this is unlikely to result in any new data collection. The intention is instead that the AusNC should have at least two major divisions.

One of these will be the carefully planned core component discussed in the previous section, while the second will have more of the nature of a text archive (See Peters, in press [2009] for discussion of this term). This component of the AusNC will be relatively unplanned and opportunistic in its accession of data, but the guiding aim will be to enable access to data about language in Australia in the widest sense. This will include, in addition to more standard varieties of English, indigenous languages, languages of migrant communities as used in Australia, indigenous varieties of English and contact varieties, varieties of English specific to different ethnic groups, and varieties of spoken English.

The 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) which is in a condition suitable for inclusion, or where the data can be brought to meet 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.

3.3 Technical issues

The discussion of the preceding sections already implies that one 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 for discovery, the discovery and access systems to be used, and a storage architecture (Cassidy, 2008).

One part of the Statement of Common Purpose from 2008 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, Davies, 2009).

Two crucial pieces in 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 (e.g. Bird and Simons, 2003). 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.[5]

In order to ensure that data from a diverse range of sources can be stored in a way which that makes that data maximally useable for as many people as possible the use of a design based on stand-off annotation (Ide & Suderman, 2007) is a crucial design principle for the AusNC. 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 stand-off 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, 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 stand-off annotation makes the two possibilities conceptually equivalent.

4 Implications for Language Technology

One of the research communities which will be serviced by an AusNC is the language technology community. The purpose of this section is to sketch some of the areas in which the project may be expected to impact on research in language technology. An important component of this resource is that it be sufficiently similar to the BNC and the ANC so that meaningful comparative work can be undertaken. The AusNC will also aim to include good samples of recently emerging genres, including computer-mediated communication, an increasingly important dimension of any type of language research. Such data will be freely accessible with copyright and ethical issues settled in advance. In some cases, this may mean that some data will have access or usage restrictions imposed on it, but these will be clearly indicated in metadata records and provided as part of the discovery tools.

Firstly, and most obviously, an AusNC will provide an easily accessible source of language samples taken from Australian usage which can be used for testing hypotheses and tools. In developing more accurate speech recognition systems, for instance, an AusNC will hold spoken language data that has been annotated not only instrumentally, as traditionally undertaken in speech recognition science, but also for what is "hearable" in the sense of

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

being interactionally meaningful according to language use researchers, in particular conversation analysts. A detailed comparison of these different approaches to the annotation of the same set of language data is likely to be mutually beneficial for both fields.

In human-computer interaction studies as well a large collection of annotated human-human interactions, and subsequent comparisons with newly developed human-computer interactional systems, will allow for the kinds of statistical analysis that are so important to the field (Dale, 2005), as well as enabling closer analysis of differences between human-human and human-computer communication (Viethen and Dale, 2009).

Current plans emphasize a dynamic structure for AusNC, with data being added to the collection over time. Ideally this will lead to a collection which can be used to answer questions about changes in language use across time. The static nature of the BNC is becoming a significant issue, as research based on that resource does not necessarily generalise to contemporary usage. Ongoing maintenance and expansion will be included as part of the corpus design for AusNC but any solution depends on the level of funding which is available for ongoing work, and this is not a variable whose value can be foreseen.

One particular use of the AusNC flowing from this component will be in localization research (see for example Shreve 2006). The availability of a large corpus of specifically Australian English will be of great value in, for example, establishing local usage in respect of terminology and in the detailed investigation of other conventions specific to Australian English. Although the available resources will be less extensive, the AusNC will also be of use where localization of other languages for an Australian audience is at issue.

The preceding paragraphs have discussed some of the ways in which an AusNC would provide access to relevant data for language technologists. But language technologists would also have an important role in developing the tools which would provide that access. Various aspects of the design discussed in section 3 pose interesting problems in this regard, especially the inclusion of large quantities of multimodal data. Access to such data via rich metadata is straightforward, but ultimately direct access to the media would be enormously desirable. Some steps in this direction are being taken (e.g. Gaudi: Google Audio Indexing, Alberti et al. 2009), but there is great potential for research in this area (Baker et al., 2009). In addition to the problems of discovery, there are also problems in delivering specified segments of audio or video to a web browser on demand. Again, this is an area in which some research has taken place, including in Australia (Annodex, Pfeiffer et al., 2003), but it is also an area with great scope for further work.

These last two examples illustrate a more general point. The development of the technical infrastructure of any project such as an AusNC will offer a wide variety of possibilities for language technology research. The design of metadata standards and of discovery and access software will all require a great deal of new research and much of this will crucially depend on work in language technology.

5 Conclusion

In this paper, we have described the current state of the AusNC initiative, the plans which have been made to date and the first steps which have been taken towards implementing those plans. The community of language technology researchers in Australia is a community which must have a considerable stake in any such project, and we have also tried to set out some of the areas in which the field of language technology could contribute to and benefit from a resource such as AusNC.

Perhaps the most important point to take from this paper is that, although some general principles are emerging, the design of an AusNC is still very much negotiable. Language technologists can and should make their needs and preferences known. Such input can influence the shape of any project which does finally eventuate and it is in the interests of everybody that any project should be designed to be as useful as possible to as many different research communities as possible.

Appendix

Members of the AusNC Steering Commmittee:

Associate Professor Linda Barwick (Sydney)
Professor Kate Burridge (Monash)
Associate Professor Steve Cassidy (Macquarie University)
Professor Michael Clyne (Monash/Melbourne)
Associate Professor Peter Collins (UNSW)
Professor Alan Dench (UWA)
Professor Cliff Goddard (UNE)
Dr Michael Haugh (Griffith)
Professor Bruce Moore (ANU)
Dr Simon Musgrave (Monash)
Professor Pam Peters (Macquarie)
Professor Roly Sussex (Queensland)
Dr Nick Thieberger (Melbourne/Hawai'i at Manoa)

Wiki at: https://sakai-vre.its.monash.edu.au/access/wiki/site/89e714f1-79dd-4f1c-b031-2591b9d0a9fb/home.html

References

Alberti, Christopher, Michiel Bacchiani, Ari Bezman, Ciprian Chelba, Anastassia Drofa, Hank Liao, Pedro Moreno, Ted Power, Arnaud Sahuguet, Maria Shugrina, Olivier Siohan. (2009). 'An audio indexing system for election video material.' In *Proceedings of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp.4873-4876).

Baker, Janet, Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan and Douglas O'Shaugnessy. (2009). Research developments and directions in speech recognition and understanding. Part 1. *IEEE Signal Processing Magazine* 26(3):75-80.

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

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

Cassidy, S. (2008) Building infrastructure to support collaborative corpus research, paper presented at the HSCNet Workshop on Designing the Australian National Corpus, UNSW, 4-5 December 2008.

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

Dale, Robert. (2005). Human communication from the perspective of natural language processing. Paper presented at ConCom05, Conceptualising Communication. University of New England, 8-9 December 2005.

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

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

Haugh, M., K. Burridge, J. Mulder, and P. Peters (eds.). (In press [2009]). *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Language*. Somerville, MA: Cascadilla Proceedings Project.

Haugh, Michael. (In press [2009]). 'Designing a multimodal spoken component of the Australian National Corpus.' In Haugh et al. (eds).

Ide, Nancy. (In press [2009]). 'The American National Corpus: Then, now and tomorrow". In Haugh et al. (eds).

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

Ide, Nancy, and Keith Suderman. (2007). 'GrAF: A graph-based format for linguistic annotations'. In Boguraev, B., 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.

Lampert, Andrew. (In press [2009]). 'Email in the Australian National Corpus'. In Haugh et al. (eds).

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

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

Musgrave, Simon and Sarah Cutfield. (In press [2009]) 'Language Documentation and an Australian National Corpus' In Haugh et al. (eds).

Nelson, Gerald (1996) "The design of the corpus". In S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English*, 27-35. Oxford: Clarendon Press.

Peters, Pam (In press [2009]) 'The Architecture of a Multipurpose Australian National Corpus'. In Haugh et al. (eds).

Pfeiffer, Silvia, Conrad Parker and Claudia Schremmer. 2003. Annodex: a simple architecture to enable hyperlinking, search & retrieval of time--continuous data on the Web. In *Proceedings of the 5th ACM SIGMM international Workshop on Multimedia information Retrieval* (pp. 87-93). Berkeley, California.

Sandefur, John R. (1986). *Kriol of North Australia: A language coming of Age*. Work Papers 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.

Shreve, Gregory M. (2006). 'Corpus enhancement and computer-assisted localization and translation'. In Keiran J. Dunne (ed.) *Perspectives on Localization* (pp.309-331). Amsterdam/Philapdelphia: John Benjamins.

Thieberger, Nick and 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.

Viethen, Jette and Robert Dale. (2009). Referring expression generation: what can we learn from human data? In *Proceedings of the Pre-Cogsci Workshop on Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference*, 29 July 2009, Amsterdam, The Netherlands.