Introduction: Mustering Languages in Australia

Michael Haugh[a], Kate Burridge[b], Jean Mulder[c], and Pam Peters[d]
[a]Griffith University, [b]Monash University, [c]The University of Melbourne, [d]Macquarie University

This volume presents a set of benchmark papers for the proposed Australian National Corpus (AusNC). The Australian National Corpus Initiative was launched amid the Australian Linguistic Society conference held 2-4 July, 2008 as part of LINGFEST at the University of Sydney. It was followed by the HCSNet Workshop on Designing the Australian National Corpus, held as part of the annual HCSNet SummerFest at UNSW, 4-5 December 2008. The papers in this volume include a selection of those presented at that workshop, together with invited papers from selected participants. These are being published now to foster wider discussion of the AusNC Initiative.

The first three papers included serve as "platform" papers in terms of the conception and design of the AusNC. The overall shape or "architecture" of the corpus should support as many linguistic, humanistic and computational research agendas as possible. Peters discusses the data that might be built into its core structure to serve that range of purposes, especially English in all its varieties and forms in Australia, but also languages other than English. This is complemented by Musgrave and Cutfield's paper which explores the inter-relationship between the fields of corpus linguistics and language documentation, and how the experiences in the latter might guide the design of the AusNC, particularly in regards to the place of Aboriginal language data in the AusNC. The issue of corpus size is addressed by Bednarek who proposes that the AusNC should be designed so that not only large-scale computerized corpus analysis, but also automated small-scale corpus analysis and manual analysis of individual texts would be enabled.

The next five papers focus on language in Australia, and the implications of the diversity to be found for the design of the AusNC. Phuong focuses in her paper on the place of mainstream Australian English in the AusNC. She argues that it should be designed with reference to other large English corpora to enable comparative research to be undertaken. Ferguson, Craig and Spencer suggest in their paper that for speech pathologists, who collect disordered speech from clinical settings, it is important that the core AusNC contain sufficient data on normal adult Australian speech, to act as a foil for their specialised corpora of pathological speech. Lising's paper meanwhile puts the case for collecting ethnic varieties of English in Australia, at least for the main immigrant groups, to reflect the multicultural grounding of Australian society. Clyne argues in his paper for the inclusion of data in the AusNC from Australian studies of bi- and trilingualism, so as to better understand the place of migrant or community languages in the Australian linguistic landscape. Lampert's particular interest in emails is a reminder that the AusNC should embrace electronic data in an age where communication is increasingly carried out online.

The next four papers discuss the range of material, other than traditional texts and transcripts, which should be collected and included in the AusNC. While they involve mediums and modalities beyond the orthographic text historically found in corpora, and bring technical and other formal challenges with including them, they are vital if the AusNC is to support the emerging linguistic and computational frontiers of the twenty-first century. Mulder, Penry Williams and Loakes explore the question of how to design a quality spoken component of the AusNC, focusing in particular on the collection and transcription of spoken data. This theme is further explored in the paper by Haugh, who proposes that the spoken component of the AusNC should include audio(visual) data alongside the spoken data, before outlining some of the key challenges of designing a multimodal spoken component of the AusNC. The theme of multimodality is further expanded upon by Johnston who discusses issues surrounding the inclusion of visual data in the AusNC, which is needed to document Australian Sign Language (Auslan). Finally, the overlap between the Australian Speech Science Infrastructure (ASSI) and the AusNC is explored in the multi-authored paper led by Burnham, which outlines the plan for a Big Australian English Corpus, a collection of audio recordings and transcriptions of spoken language data collected in highly controlled settings for the purposes of instrumental analysis.

The collection is rounded off by Ide's paper on the advances in annotation associated with the American National Corpus. It is a timely reminder that the building of any corpus should be made with reference to the broader international context of language data management.

While the papers in the volume represent a range of positions as to how the AusNC Initiative might best proceed, a tentative architecture for the AusNC can nevertheless be seen to emerge from this collection. It appears that the Australian National Corpus is jointly conceptualized as consisting of a set of multimodal and multilingual resources distributed across a network of hosting institutions which meet leading-edge technical standards. By multimodal it is envisaged that not only plain text, but also visual, audio and audiovisual language data will feature in the corpus, while by multilingual it is intended that the corpus incorporate significant collections of English in Australia (including Australian English and migrant Englishes), indigenous languages, community languages, and Australian Sign Language. These collections of language data will include both historical collections (that is, language data sets that have already been gathered) as well as being a point of departure for current and future collections of relevant language data.

In closing, we would like to thank all who participated in the workshop, as well as the generous support of HCSNet in helping to organize and fund the workshop. We would also like to thank the Griffith Institute of Educational Research for the support they provided in copyediting this volume, in particular the work of Andrea Kittila.

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:

Haugh, Michael, Kate Burridge, Jean Mulder, and Pam Peters. 2009. Introduction: Mustering Languages in Australia. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., v-vi. Somerville, MA: Cascadilla Proceedings Project.

or:

Haugh, Michael, Kate Burridge, Jean Mulder, and Pam Peters. 2009. Introduction: Mustering Languages in Australia. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., v-vi. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2280.