The Architecture of a Multipurpose Australian National Corpus

Pam Peters
Macquarie University

1. Introduction

Collaborative planning by Australian researchers for a large national corpus presents us with quite diverse views on:

(a) what a corpus is,
(b) what research agendas it should support,
(c) which varieties of discourse it should contain,
(d) how many languages it could include,
(e) how the material might be collected, and
(f) what kinds of annotation are needed to add value to the texts.

The challenge for us all is to develop a comprehensive plan and architecture for a corpus which will encompass as many research agendas as possible. A modular design which allows independent compilation of segments of the notional whole recommends itself, so long as common systems of metadata and annotation can be set down at the start.

2. What is a Corpus?

Divergent understandings of what a corpus is reflect some aspects of the word's history, and the different academic disciplines which have made use of them.

- In literary studies since the earlier C18, the Latin word *corpus* has been used to refer to the body of work by a single author, for example, Shakespeare, or set of them, such as the romantic poets. It implies the total output of those authors, not a sampling.
- In linguistic science since the 1960s, *corpus* has referred to a structured collection of texts sampled from various types of discourse, including written and – since the 1980s – spoken as well (e.g., Collins Cobuild). It is thus deliberately heterogenous, and designed to be in some sense 'representative' of a standard language or variety of it. The International Corpus of English (ICE) consists of a set of regional corpora, from Australia, New Zealand, Singapore, India etc. A linguistic corpus is usually compiled out of texts from a specified time period, in order to control or reduce temporal variation in the language of the samples.
- In discourse-based humanistic studies, the term *corpus* is applied by some to a large ad hoc assemblage of texts which may be mined for data on a particular topic. Collections of out-of-copyright works, such as Project Gutenburg, may be used as an example. Those who reserve the term *corpus* for a structured selection of texts (like the ICE corpora) would refer to collections like Project Gutenburg as a 'text archive'.

The size of the 'corpus' varies considerably. If it comprises the complete oeuvre of the writer concerned (= definition 1 above), it varies with their productivity, but is essentially finite. Researchers who use the word *corpus* to refer to their finite collections of homogenous texts are typically speaking of a small corpus containing anything from 5000 to 500 000 words.

The second type of corpus (= definition 2) is structured around samples of many different registers and text-types, and will have a target size, as did the 1 million-word corpora used to study regional varieties of English from the 1970s (e.g., Brown, LOB etc., and the ICE corpora, all with 2000 word samples); and the 100 million-word British National Corpus, created in the 1990s with samples of

 Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus, ed. Michael Haugh et al., 1-9. Somerville, MA: Cascadilla Proceedings Project.

varying length. The use of standard-sized samples is clearly more important in a small corpus, to prevent skewing of the data in favor of particular authorial or registerial features.

The text-archive *(= corpus* definition 3) is open-ended on all parameters, with no particular targets in terms of text-types, or the number and size of texts to be included in it. "Bigger is better" is its essential principle, and diversity may or may not be achieved through ad hoc acquisitions of digitized texts. The Collins Cobuild Bank of English is one such text archive, oriented towards nonfiction material with large quantities of current newspaper text; whereas the Project Gutenberg and the Oxford Text Archive consist largely of literary fiction.

The term *corpus* is now almost always used in the sense of a digitized collection of texts which can be interrogated by means of computer tools, at least by the string-search type, although more sophisticated searches can be achieved with the aid of corpus markup (see Section 5).

3. Corpora and Research Agendas

3.1. Text-Based Corpora

The data provided by the different types of text corpus varies in quantity and quality, depending on the nature of the inquiry. The very large volumes of data extractable from a text archive (e.g., the IBM corpus) support computational research into NLP, where the quantity compensates for the lower quality of data extracted by automatic methods, and the fact that it may over- or under-represent certain kinds of linguistic variation. There is a tacit assumption that such surface variation is immaterial, especially if the purpose is to induce the underlying systems of the language.

Very large text archives are now routinely used by lexicographers in making dictionaries of the general language (e.g., Collins Cobuild, Macquarie, Oxford), where the data extracted can always be reviewed by the human researcher in terms of its discoursal context of time, place, and genre, in order to evaluate its quality as citational evidence.

Large general-language text archives also serve as a foil to corpora of a single type of specialized discourse (= definition 1), especially for terminologists and terminographers seeking to establish the boundaries between words and terms (Chung & Nation, 2004). Substantial divergences in the frequencies of words and terms derived from the archive and from the specialized microcorpus of, for example, cell biology, help to identify the key terms embedded in the latter (Peters et al., 2008).

By contrast, the corpus or rather sets of corpora (definition 2) have typically been designed to support research into linguistic variation in English, on regional, temporal, generic, and social parameters. The earliest corpora (Brown corpus of standard American English and LOB of standard British English) consisted of 15 different written genres or text categories (nonfiction and fiction), which allowed researchers to compare usage in more and less formal styles. Temporal comparisons in the constituents of these text-types have been made possible through the second generation of these corpora (Frown and FLOB, where the "F" says that they were made by researchers at Freiburg University in Germany, in the 1990s; see Mair, 1997).

The ICE corpora updated the Brown/LOB model with the addition of several types of speech, dialogic and monologic, so that for the first time, written and spoken norms in different regional varieties of English could be compared (Greenbaum & Nelson, 1996). A new anthology of Australian and New Zealand English capitalizes on regional data from the ICE corpora (Peters, Collins, & Smith, 2009). The intercomparability of the ICE corpora, in terms of their structure and time frame (1990s), gives them considerable value beyond the 'sum of their parts'.

Sociolinguistic information is included with the spoken data contained in the ICE corpora, although neither sociolinguistic nor socioeconomic parameters were used, and they could not be adequately represented within such small corpora, along with all the other parameters. In the much larger British National Corpus, socioeconomic sampling of speakers was much more extensive. The researchers used a market survey company to collect data from speakers all over Britain, in the standard consumer categories A to E, with some remarkable findings in terms of the high-frequency words used by speakers of different SES groups, genders, and age brackets.

Table 1.
Words Most Characteristic of Male Speech (A) and Female Speech (B), Adapted from Rayson, Leech, and Hodges (1997)

Words (A)	Males	M%	Females	F%	Chi-sq
Fucking	1401	0.08	325	0.01	1233.1
Er	9589	0.56	9307	0.36	945.4
The	44617	2.60	57128	2.20	698
Yeah	22050	1.29	28485	1.10	310.3
Aye	1214	0.07	876	0.03	291.8
Words (B)					
She	7134	0.42	22623	0.87	3109.7
Her	2333	0.14	7275	0.28	965.4
Said	4965	0.29	12280	0.47	872.0
N't	24653	1.44	44087	1.70	443.9
I	55516	1.44	44087	1/70	357.9

Linguistic data like these from the 'demographic' samples of private conversation in the BNC will continue to be mined for insights into social variation in British English; while the 'institutional' speech samples support research on more public patterns of speaking. The BNC also supplies benchmarks for the linguistic and stylistic properties of the many written text-types included.

Historical variation in English has so far been most specifically designed into the ARCHER corpus (A Representative Corpus of Historical English Registers) by Biber and other international corpus researchers (Biber, Finegan, & Atkinson, 1994). ARCHER includes extracts from various types of English texts written from 1650 on, including journals, diaries and letters, fiction and drama, and medical/scientific writing, all sampled in 50-year periods.

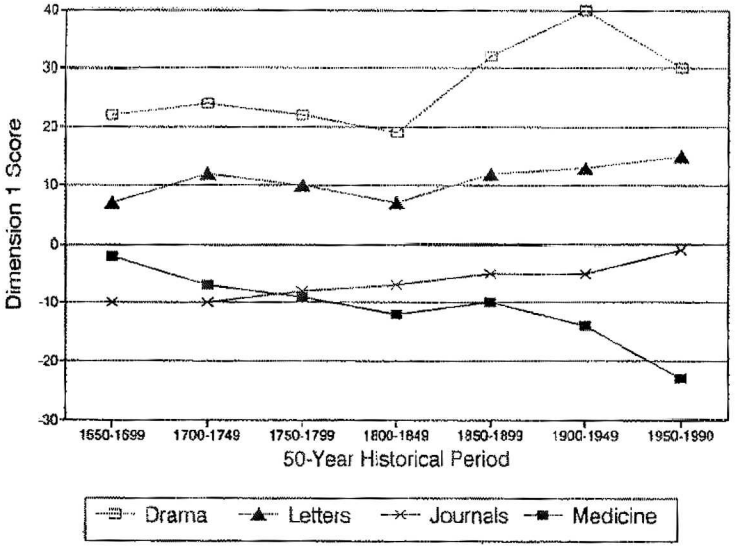

Figure 1. Changing values for several historical genres on Dimension I ('involved v. informational production'), following Biber et al. (1994).

ARCHER data analyzed along several linguistic dimensions by factorial analysis provide very remarkable insights into the gradual changes in style within distinct text-types. While the samples contained within ARCHER are relatively small, the fact that they are anchored in time is invaluable in mapping historical developments in English lexicogrammar. Research based on the ARCHER corpus reminds researchers of the fact that the reference genres themselves are products of culture, evolving over the course of time.

Sample corpora, whether large or small, all provide benchmarks of 'normal' discourse in a given time and social setting. When calibrated for age, the spoken samples in the BNC or ICE, generated under known contexts of speech production, can provide a foil for researchers dealing with pathological speech patterns (see Section 4.3). Fresh samples of spoken discourse by both children and adults need to be collected in the AusNC, whereas only adult speech was included in ICE-AUS.

3.2. Corpora of Other (Non-Text) Material

Although the spoken material contained in the text corpora discussed in 3.1 is based on sound recordings, few of them are of sufficient quality to support phonetic research. By their nature they were made in natural, noisy settings, and their purpose is to illustrate spoken discourse produced by one or more speakers, rather than to support acoustic analysis of phonetic segments.

Australian phonetic research calls for a kind of subcorpus, one which includes:

(a) high-quality (studio-recorded) audio material, which submits to acoustic analysis by customized tools such as EMU (Cassidy & Harrington, 2001), and

(b) very short samples (words, phrases) of phonetically varied contexts for each phoneme and intoneme, to benchmark the inventory of sounds produced by speakers, as in the ANDOSL database (Millar, Dermody, Harrington, & Vonwiller, 1990; Cox & Palethorpe, 2008)

Speech from different sociolinguistic varieties is increasingly needed for such individual inventories, again so as to provide the benchmarks for normal speech sought by speech pathologists and audiologists.

Radically different types of subcorpus are needed to embrace the needs of gestural researchers and include video material within the AusNC plan. In documenting the use of AUSLAN and other kinds of sign language, the video material is central, as are the complex systems of annotation within ELAN, which allow multilevel searching. Gestural data are also needed for multidimensional analysis of dyadic communication, so as to analyze the paralinguistic elements in parallel with the linguistic. Gestural plus locational data are sometimes important in researching the modes of communicating in Aboriginal discourse, which in its natural setting is sometimes accompanied by drawing in the sand. Again multilayered annotation systems would need to be built into complex data of this kind.

4. What Languages to Include?

4.1. Languages Other than English

Apart from the corpora of Aboriginal languages and AUSLAN in the previous paragraph, all those mentioned so far have been concerned with English. Clearly an Australian National corpus needs to embrace the needs of researchers of indigenous languages: they are per se a very significant dimension of linguistics in Australia and for the Australian Linguistic Society. The AusNC plan needs to include such indigenous languages, or at least be able to interface with dedicated corpora/databases which already exist, such as Paradisec and the AUSLAN database. Because they are designed according to parameters appropriate to their particular functions and types of discourse, they may or may not line up with those for the analysis of English and its variation in Australia. If not, they could still be associated as with the AusNC as adjunct corpora through a computer interface, provided that their metadata and annotation systems were compatible (see below, Section 5; and Cassidy's 2009 paper (in this volume). The compatibility of annotations in any external corpora with those of AusNC would be paramount.

Corpus data collected for community languages are also candidates for inclusion within AusNC, in a separately structured subcorpus or microcorpus. Again such data may have been collected according to independent parameters, which suggests that using compatible annotation for external interface

would be more appropriate. Any legacy material which requires transcription or additional annotation to facilitate the interface with AusNC would entail costs to be factored in to the AusNC budget.

4.2. Translation Corpora

Also worth considering, although probably outside the design of the main AusNC, is a purposeful translation corpus, for use in the training of professional translators and interpreters (Slatyer, 2006). Overseas researchers concerned with the quality of translation and avoiding 'translationese' have developed parallel corpora, so that the crosslingual effects of translating between particular pairs of languages can be studied; and comparable original texts in the same target language can be tapped for optimizing translation equivalence. The first calls for 'translation' texts, the second for 'comparable' texts in the same genre (McEnery, Xiao, & Tono, 2006). One corpus structured along these lines is the Oslo translation corpus, designed to track translation between Norwegian, English, and German texts, as shown in Figure 2.

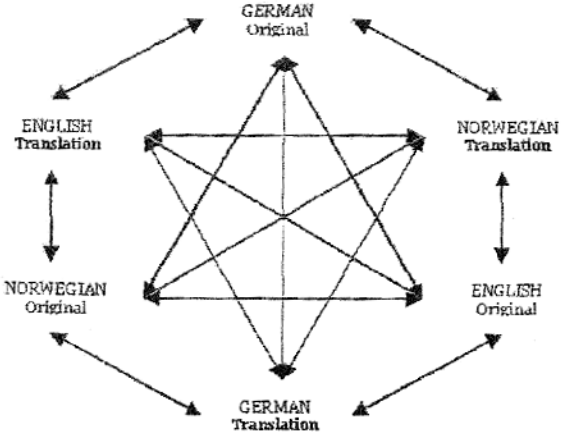

Figure 2. Schematic diagram of Oslo multilingual corpus (Johansson, 2003).

Translation corpora designed in this way, so as to consist of both translation texts and comparable texts, are invaluable for research in contrastive linguistics and the best resource for producing high-quality translations (Johansson, 2007). They could again be considered as a type of subcorpus, while the English originals and comparable texts are contained in the main AusNC corpus, with an interface to the translated texts and originals in languages other than English. The model lends itself to a subcorpus of translations of Aboriginal texts.

Translation corpora of languages with non-alphabetic scripts, such as Arabic, Chinese, Hindi, Korean, can be processed using Unicode 8. They present large challenges of alignment, but also enormous scope for research.

4.3. Pathological Speech

Collections of discourse produced by clinical patients (in any language or mix of languages) which have been made for the purposes of diagnosis and research would be a further useful dimension for the AusNC. It goes without saying that such data would have to come with the necessary ethics clearances. As a microcorpus within the AusNC, such data could be readily compared with 'normal' discourse in the same genre or register, and enlarge the applications of the corpus. It would be particularly useful for the training of interpreters in multilingual contexts, where the challenges of understanding the patient's talk in terms of one or more languages are great, as are the linguistic demands made on the interpreter to diagnose the pathological elements (Roger, 2004).

5. Annotation and Metadata

5.1. Corpus Structure and Metadata

A large heterogeneous corpus depends for its usability on interoperable systems. The AusNC would need a comprehensive platform to allow researchers to move easily between the various sections of the corpus, including its subcorpora (structured subsets) and microcorpora (specialized corpora of homogenous material). Controlled access to associated adjunct corpora (externally managed) also needs to be included in the software planning.

Common sets of headers and corpus text identifications are needed to facilitate access to the different parts of the corpus and to texts within them. This is important not only within a structured corpus with heterogenous material, but also to be able to 'talk' to external corpora on other software platforms with other software systems. While facilitating the interface, the software would also need to be able to block access to material in particular subcorpora, microcorpora, or adjunct corpora, when general access is not permitted under the conditions of text collection.

Also needed is a metadata template which will indicate to researchers the linguistic parameters of any corpus or text they might wish to access, for example, language(s), time/place of collection, sociolinguistic profiles of speakers, text-type/genre of discourse, and level of annotation. Computational means of cross tabulating these variables for the texts available would be highly desirable.

5.2. Types of Annotation/Markup

Although words and phrases can be extracted from a corpus by means of simple string searches, some levels of annotation are needed to provide contextualization for them, as is important in linguistic research. Corpus annotation begins with identifying the text-type and text unit (by coded numbers), within which sentences can be automatically numbered, at least in written material. But the marking up of most important structural aspects of the text (written or spoken) must be added by hand (e.g., paragraphs, headings, speaker turns). Contextual information like this enables corpus users to see that the verb *wed* is mostly found in newspaper headlines, or that Speaker A in a particular text is particularly fond of using *absolutely.*

Apart from identifying the structural components of the discourse, corpus annotation may be used to identify discoural elements and functions embedded within the text, for example, types of speech act or speech functions (Stenstrom 1994; Eggins & Slade, 1997) which can then be extracted from large stretches of discourse within the corpus. These days such annotation should be XML-compatible (see Section 5.3).

Grammatical tagging is the most intensive kind of corpus annotation, used to establish the word classes (parts of speech) of every word used in the corpus texts. This procedure is usually done by means of automatic taggers, with human readers correcting the grammatical tags assigned by the software where necessary. A high level of successful tagging (above 96%) is claimed by automatic taggers such as the Lancaster CLAWS (Garside, 1987) and Eric Brill's 'trainable' tagger (Brill, 1995), whose default software is designed for grammar of written discourse but can be adapted for speech, and for languages other than English. This capacity lends itself to the translation corpus, where parallel texts from languages with divergent syntax can be better aligned.

Grammatical tags serve as input to syntactic analysis, though the interface between them is far from straightforward and very labor-intensive, witness the large investment of the ICE-GB team in these processes over two decades. It is of course possible for corpus builders to grammatically tag and syntactically parse selected texts, rather than aim for syntactic analysis of the whole corpus.

5.3. Annotation and Validation

Any linguistic annotation inserted by human hands into a text needs to be checked and validated at regular intervals. This process ensures the accuracy and consistency of the corpus markup, and is essential to making it amenable to common computational search tools, most of which are XML-based.

XML is capable of handling individually worded annotations of almost any kind within its conventional markup units (Wong, Cassidy, & Peters, in preparation). Thus the markup already contained in legacy corpus materials (see Section 6.2) could/should be checked by means of validation tools, and amended where necessary. XML validation tools are available as freeware from the internet.

6. Collection of Data for the AusNC

6.1. Corpus Design and Collection Priorities

In creating an AusNC and collecting the data for it, some priorities need to be set, especially in relation to the structured corpus. A text archive whose composition is entirely open can of course be compiled as opportunities arise, and by serendipity. But a structured corpus with maximal value as a reference corpus requires a systematic program of collection, and modular development which contributes to the design of the whole. The most demanding and costly kind of data to collect is spoken discourse, but the need for it was articulated at the inaugural AusNC meeting (at ALS, July 2008). It can be accorded top priority because of its intrinsic importance in sociolinguistic analysis, as well discoursal and phonetic research (provided that the recordings are of high-quality). The text of conversational speech is expensive to transcribe, but transcriptions of public monologues and dialogues may come ready-made from the institutions that stage them, and can be used for –emic analyses of grammar and discourse. A set of protocols for levels of transcription for different types of linguistic research needs to be compiled for the AusNC's collection of speech.

The collection of written material for AusNC could take second priority, although a comprehensive plan of the types and quantities of written texts should be drawn up at the start, to ensure a balance of written discourse types – from mass media as well as more specialized publications. Written texts collected for the AusNC must be digitized, to avoid transcription costs wherever possible. There will still be some annotation costs in adding metadata and structural annotation.

The question of including electronic texts of various types – from personal email to discussion lists and institutional information – needs to be discussed within the architecture of AusNC. The fundamental question is whether these text-types count as a third medium (alongside written and spoken discourse), or as types of content (comparable to personal letters, conversations, and government documents etc.) which are already there in the corpus design. Support for researching the electronic medium should certainly be considered in the design of the AusNC. Sourcing texts from the internet is relatively easy, though there are issues (see 6.3).

6.2. Donations of Previously Collected Texts

Contributions to the AusNC in the form of previously collected texts could be included, at least as microcorpora. Apart from their intrinsic value as samples of English or other languages, such legacy material represents substantial amounts of labor by previous linguistic researchers, which adds to the resources of their successors. However it would be vital to check (a) what kinds of permissions have been obtained for the use of such texts; and (b) what format the texts are in. Older audio recordings would require digitization, and verbal texts already digitized may or may not be in compatible computer formats. Additional annotation will probably be required to make them interoperable with the main AusNC, with attendant costs.

6.3. Uses of the Internet and Internet Material

The internet is an enormous resource and, for some, the ultimate corpus. String searches are easily performed by Google and other search engines, and the statistics provided give broad-brush frequencies to use for comparative purposes. But there is no way of knowing what range of texts supply the summary statistics, and whether they are in some way conflated with Google normalizations of your search (as sometimes signaled by its question: Did you mean?).

Selected texts can of course be extracted from the internet, for use in either the main corpus or an associated text archive. There is plenty of variety out there, already digitized for the more complex computer search tools if annotated appropriately. It may not however be free of copyright: *caveat*

corporis creator! The internet can also be used to download audio files in digital form from streaming media (as was done for the ART corpus). The need to transcribe such texts and the copyright caveat still apply.

7. Conclusions

Strategic planning is needed to ensure the core AusNC collection can be completed. The highest priority would be material which supports the widest range of linguistic research. The core corpus provides the benchmark for the range of written and spoken (and internet) material. Its known provenance and sociolinguistic properties are crucial to its validity as evidence. As a whole, the core corpus provides a useful foil for specialized microcorpora, and its particular categories of text complement the design of subcorpora, adding value to them. Though demanding in terms of planning, funding, and management, an AusNC of this type is eminently worthwhile.

An AusNC of this type needs to be adequately funded, against the cautionary tale of the American National Corpus (ANC), which was never fully funded like the British National Corpus (BNC), from a mix of public and private funding. The ANC has suspended its collection at around 20—30 million words, well short of the target 100 million to match the BNC (Ide, 2009, see paper in this volume), and is now harnessed to computational rather than linguistic research, as its source of funding and means of survival. For computational purposes, a large text archive of opportunistically gathered material is quite satisfactory and much less demanding to compile. It does not, however, provide a research resource which is sensitive to the linguistic variability of individual texts,informed about their provenance, and able to supply rich insights into language use in context.

References

Biber, Douglas, Edward Finegan, & Dwight Atkinson. (1994). ARCHER and its challenges. In Udo Fries, Gunnel Tottie, & Peter Schneider (Eds.), *Creating and using English language corpora. Papers from the 14th ICAME conference* (pp. 1-13). Amsterdam: Rodopi.

Brill, Eric. (1995). Transformation-based, error-driven learning in natural language processing. *Computational Linguistics 21*: 4, 543-565.

Cassidy, Steve, & Jonathan Harrington. (2001). Multilevel annotation in EMU speech database management system. *Speech Communication, 33*, 61-77.

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

Chung, Theresa, & Paul Nation. (2004). Identifying technical vocabulary. *System, 32*, 251-263.

Cox, Felicity, & Sallyanne Palethorpe. (2008). Reversal of short front vowel raising in Australian English. *Proceedings of Interspeech*, 22-26 September (pp. 342-345). Brisbane, 342-345.

Eggins, Suzanne, & Dianna Slade. (1997). *Analysing casual conversation.* London: Cassell.

Garside, Roger. (1987). The CLAWS word-tagging system. In Roger Garside, Geoffrey Leech, & Geoffrey Sampson (Eds.), *Computational analysis of English.* London: Longman.

Greenbaum, Sidney, & Gerald Nelson. (1996). *Comparing English worldwide: The International Corpus of English.* Oxford: Oxford University Press.

Ide, Nancy. (2008). *The American National Corpus: Then, now and tomorrow*. Keynote paper presented at the HCSNet Workshop on Designing the Australian National Corpus, 4-5 December, UNSW, Sydney, Australia.

Johansson, Stig. (2003). *Multilingual corpora: Models, methods, uses* (pp. 1-14). Online at http://cst.dk.

Johansson, Stig. (2007). *Seeing through corpora: On the use of corpora in contrastive studies.* Amsterdam: John Benjamins.

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

Mair, Christian. (1997). Parallel corpora: A real-time approach to the study of language change in progress. In M. Ljung (Ed.), *Corpus-based Studies in English. Papers from the Seventeenth ICAME Conference* (pp. 195-209). Amsterdam: Rodopi.

Millar, Bruce, Phillip Dermody, Jonathan Harrington, & Julie Vonwiller. (1990). A national cluster of spoken language databases. *Proceedings of the 3rd International Conference on Speech Science and Technology.* Melbourne Australia.

Peters, Pam, Peter Collins, & Adam Smith. (2009). *Comparative studies of Australian and New Zealand English: Grammar and beyond.* Amsterdam: John Benjamins.

Peters, Pam, Alan Jones, Adam Smith, Theresa Winchester-Seeto, Jennifer Middledorp, & Peter Petocz. (2008). TermFinder: Creating online termbanks of technical terms for early university study. *Journal of Applied Linguistics, 3* (2), 219-248.

Rayson, Paul, Geoffrey Leech, & Mary Hodges. (1997). Social differentiation in the use of English vocabulary. *International Journal of Corpus Linguistics, 2* (1), 1333-152.

Roger, Peter. (2004). *Language assessment by remote control.* Paper presented at the Style Council Conference, Sydney, NSW.

Slatyer, Helen. (2006). *Standardisation in community interpreter training.* Paper presented at the IATSIS Conference, Capetown, South Africa.

Stenstrom, Anna-Brita. (1994). *An introduction to spoken interaction.* London: Longman.

Wong, Deanna, Steve Cassidy, & Pam Peters. (In preparation). Updating the ICE annotation system: Tagging, parsing and validation.

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:

Peters, Pam. 2009. The Architecture of a Multipurpose Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 1-9. Somerville, MA: Cascadilla Proceedings Project.

or:

Peters, Pam. 2009. The Architecture of a Multipurpose Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 1-9. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2281.