Towards the Design of the Australian National Corpus

Phuong Dzung Pho
Monash University

1. Introduction

Corpora are becoming more and more important as a research tool for linguists as they are large collections of authentic text. However, not every researcher has the time and resources to compile their own corpus. Large corpora in the world such as the BNC, the ANC or the International Corpus of English (ICE) have been widely used for research on the English language in general or an English dialect in particular. Although there are several corpora of Australian English at the moment, researchers who would like to study an aspect of Australian English or to compare Australian English with other dialects generally have difficulty finding a corpus that meets their needs. One of the reasons for this is that the corpora that are currently available on Australian English are small in size and scope, and most of them are outdated. The Australian Corpus of English (ACE), for example, was created in 1986, more than two decades ago. It consists of one million words and only covers written texts. A more recent corpus, the Australian subcorpus of the ICE, was developed between 1991 and 1995. Although it contains both spoken and written language data, it has only one million words. Other corpora are even smaller in size and cover only certain genres, as suggested by their names. For instance, the Electronic Documents corpus (EDOC) consists of only 200,000 words. The same number of words can be found in the Talkback Radio corpus. The Corpus of High School Writing has only 160,000 words.

In contrast, large corpora in the world have or are aiming to have at least 100 million words. For example, the well-known BNC, originally developed between 1991 and 1994, consists of over 100 million words (http://www.natcorp.ox.ac.uk). The ANC is aiming at 100 million words and so far has 22 million words (http://americannationalcorpus.org/SecondRelease). Another widely available corpus of American English, the Corpus of Contemporary American English (COCA), has an even larger size with more than 385 million words collected since 1990 (http://www.americancorpus.org). Not only are these corpora large in size, they contain both spoken and written language data and have similar designs. All these characteristics make them attractive to corpus studies of one particular dialect of English or comparative studies of the two main English dialects in the world. A corpus of Australian English as another major English dialect in the world would thus be beneficial to researchers of this specific dialect or researchers who need to compare Australian English with other Englishes, especially British and American English.

In order to make the Australian National Corpus (AusNC) useful for comparative studies, it is important that the corpus be comparable to the BNC, ANC, and COCA in terms of both size and design. This study therefore aims at reviewing the design of these three main corpora currently widely used by researchers before proposing a design for the AusNC.

2. Designs of the BNC, ANC, and COCA

Although both the ANC and COCA are claimed to follow the design of the BNC, there are differences in various aspects between these corpora. Whereas the BNC comprises texts from 1960 to 1994 (although most texts were written in 1990s), the more recent corpora of ANC and COCA consist of texts from 1990 onwards. The COCA even specifies an amount of 20 million words per year. Despite the difference in size as mentioned above, the three corpora have more or less similar proportions of the written component and the spoken component. Ninety percent of the texts in the

 Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus, ed. Michael Haugh et al., 25-29. Somerville, MA: Cascadilla Proceedings Project.

BNC are written; only 10 percent are spoken. Similarly, the proportions of written and spoken are 83 versus 17 in the ANC and 80 versus 20 in the COCA. The fact that written texts are preferred by corpus compilers is understandable, since, as Baker (2006) points out, spoken data are much more difficult to collect. It is also time-consuming to transcribe spoken data, whereas many of the written texts are already in electronic formats.

The spoken component of the BNC is more structured and varied than that of the ANC or COCA. The BNC contains 700 hours of recordings by 124 British English speakers in the United Kingdom spread across various genders, age groups, social classes, and regions. The recordings also cover a wide range of text types and domains, from educational and informative (e.g., lectures, news), business (e.g., meetings), public/institutional (e.g., speeches), to leisure (e.g., chat shows). Such an ideal design for a spoken data collection, however, is not easy to obtain without considerable funding. COCA, for example, has to settle for a limited range of texts – television and radio programs. Similarly, the spoken part of the ANC is made up of three small corpora: telephone conversations, face to face / spontaneous conversations, and academic discourse. Although the range of text types is not as wide as that of the BNC, the ANC consists of samples of both formal and informal speech, which can be useful for sociolinguistic studies.

Unlike the spoken component, the written components of the three large corpora are more compatible in terms of text types. All of them contain texts in the two main mediums: books and periodicals. However, only the BNC and ANC include what the BNC calls 'miscellaneous published' (e.g., brochures, leaflets, advertisements), 'miscellaneous unpublished' (e.g., letters, memorandums, minutes, essays) and 'written-to-be-spoken' (e.g., play scripts) (Burnard, 2000). Although not as extensive as the BNC, the ANC contains some of these genres – travel guides and letters. Interestingly, the ANC written component also includes web logs, which cannot be found in the BNC. One obvious reason is that web logs did not exist in the early 1990s, when the BNC was created. Web logs and emails, however, should be included in modern corpora as these are now more common than the traditional letter, memorandum, or essay. Although both the ANC and COCA include both journals and newspapers, which the BNC groups as 'periodicals', the COCA seems to have only fiction books, whereas the ANC and BNC include both fiction and nonfiction books. Attempts should be made to include a range of text types as wide as possible to make the corpus more representative (Biber, 2004). The domains of the texts are also taken into consideration in the compilation of the three corpora. For example, the written texts in the BNC cover the following fields: applied science, arts, belief and thought, commerce and finance, imaginative, leisure, natural and pure science, social science, and world affairs.

One of the features that distinguish the corpora, or at least distinguish between the BNC and the ANC – COCA, is that each corpus is representative of a main English dialect, either British English or American English. This is also a reason why it is important to have a corpus of Australian English since large corpora such as the BNC or ANC are not suitable for studies of Australian English or for studies that compare language use in different regions. The designers of all of the three corpora claim that the written texts were selected on the basis that they were written by authors who are native to the particular dialect and published by local publishers. Likewise, the spoken texts were produced by native speakers. However, as admitted by Ide (2003), it is hard to ensure that an author is a native speaker of the language involved. Similar issues will have to be considered for the AusNC as, like the United States, Australia is a multicultural country.

The levels of annotation also vary across the three corpora. The data in all of the three corpora were tagged for part of speech (POS-tagged) using CLAWS (the Constituent Likelihood Automatic Word-tagging System), a tagging program developed at Lancaster University. In addition to CLAWS tags, the ANC used two other sets of POS tags, namely the Biber tags and the Penn tags (Ide & Suderman, 2006). The annotation of the ANC is also different from the BNC and COCA in that the ANC also provides annotations of noun chunks and verb chunks, sentence tagging, and syntactic bracketing. The approach adopted by the ANC, as stated by Ide (2003), is a 'stand-off' one, that is, annotations are in separate documents from the raw texts.

3. Proposed Design for the Australian National Corpus (AusNC)

As mentioned earlier, in order to make the AusNC useful for comparative studies and widely acceptable, the design of the AusNC should be as close as possible to that of the BNC or ANC. In terms of the size of the corpus, I propose that the AusNC should have at least 100 million words. The timeline, however, should not follow the BNC, which includes data collected before 1992. Even the BNC World version released in 2001 and the BNC XML version released in 2007 do not contain updated data (Hoffmann, Evert, Smith, Lee, & Prytz, 2008). As Baker (2006) points out, this limitation may cause problems for studies of contemporary English written and spoken in the UK. On the other hand, since the BNC is still the most widely used corpus for research on British English, the AusNC could have a core corpus consisting of texts from the 1990s and an up-to-date part with data from the new century. In fact, the AusNC can have a 'static' component and a 'dynamic' component as proposed by Ide and Macleod (2001) for the ANC. The static component would consist of texts from the 1990s and 2000s (ideally five million words per year), and the dynamic component would consist of texts added to the corpus every five years after the corpus is released. This will keep the corpus up-to-date while researchers can still use it for comparative studies with British English or American English. Furthermore, researchers who would like to study how the use of Australian English has changed over time will also find such a corpus as AusNC useful. In terms of the medium of texts, the AusNC should follow the proportion of spoken and written data in the BNC and the ANC, which is 10% spoken and 90% written.

3.1. The Spoken Component

The spoken data should consist of recordings of Australian English speakers of different genders, age groups, social class, and region. The age groups, as for the BNC, should include 0-14, 15-24, 25-34, 35-44, 45-59, and over 60 years. As for social class, we could follow the classification used by the BNC, namely AB (managerial and professional), C1 (supervisory and clerical), C2 (skilled manual), and DE (unskilled manual and unemployed) (Burnard, 2000). This system seems to be equivalent to the social class classification of upper class, upper-middle class, lower-middle class, and lower class in Australia. In terms of region, only small parts of the country have been surveyed to date (largely on the eastern coast). We are in need of data from different states and from both urban and rural areas of each state. Such information would be useful for studies of regional variation in terms of vocabulary or accent.

As far as text type is considered, the design of the BNC in this regard should be followed as closely as possible to allow for comparison.

Table 1.
Text Types of Spoken Data in the BNC (adapted from Burnard, 2000)

Domain	Text types
Educational and informative	lectures, talks, educational demonstrations, news commentaries, classroom interactions, etc.
Business	company talks and interviews, sales demonstrations, business meetings, consultations, etc.
Public / institutional	political speeches, sermons, public/government talks, council meetings, religious meetings, Parliament proceedings, legal proceedings, etc.
Leisure	speeches, sports commentaries, talks to clubs, broadcast chat shows and phone-ins, club meetings, etc.

3.2. The Written Component

Like the spoken component, the written data in the AusNC should satisfy at least the following two selection criteria: published by an Australian publisher and written by Australian authors (authors who are native speakers of Australian English). For each written text, the following information should be collected when available for classification purposes: author type (multiple / sole), gender, age group (child / teenager / adult), place of publication, and sampling method (whole text, beginning, middle, or ending section). The information regarding the age and gender would be useful for sociolinguistic studies, while information about the sampling method might be useful for discourse studies. However, unless a text is too long (over 40,000 words), the whole text should be collected.

In regards to text types, the written component of the AusNC should strive for the following proportions: books (55%), periodicals (30%), published miscellanea (5%), unpublished miscellanea (5%), and written-to-be-spoken (5%). Texts should be collected from various domains: natural and pure science, applied science, social science, arts, belief and thought, commerce and finance, imaginative, leisure, and world affairs.

3.3. Annotation of the Corpus

For a corpus to be useful for researchers, a large collection of words is not enough; it has to be annotated. Two kinds of annotation might be useful for research purposes – linguistic tagging and demographic information annotation. A tagged corpus will be useful for morphosyntactic studies. One basic type of tags is part of speech (POS) tagging, for which CLAWS7 should be used as it is used for all of the three large corpora, BNC, ANC, and COCA. Moreover, CLAWS7 is found to be the most accurate POS tagger (van Rooy & Schäfer, 2003). Lemmatization, sentence, paragraph, and section marking (for written texts), utterance and turn marking (for transcriptions of spoken data), or, at a higher level, syntactic parsing of the corpus, will make the corpus even more valuable for linguistic research. On the other hand, recording information about the author or the speaker (e.g., age, gender, region) will benefit sociolinguistic researchers.

However, the annotated version of the corpus should be separated from the original text. If researchers do not need the annotation, the presence of such information can make the original text difficult to read.

3.4. Other Parallel Subcorpora

In addition to the main corpus of Australian English, I propose that the AusNC, being a national corpus, include samples from other major languages used in Australia as well. One group of these is Aboriginal languages. Other main community languages, according to the most recent census in 2006 (Australian Bureau of Statistics, 2006), are Greek, Cantonese, Arabic, Mandarin, Vietnamese, and Spanish. Since these languages are much less commonly used than English, we cannot apply the same design that we use for the construction of the main corpus of Australian English. For instance, we cannot expect to be able to collect written and spoken data of all the text types and domains as for the main corpus. In fact, spoken data might be easier to collect than written data for these subcorpora. One possible source of spoken data is the community language broadcasts on television or the radio. More informal spoken data can come from recordings of face-to-face or telephone conversations. Written data can be collected from newspapers, magazines, websites, letters, emails, or web logs written in the target language. Some of such data may have been collected by researchers in various fields in Australia; therefore, a request for donation of these individual corpora will save a lot of effort and will benefit the whole research community.

4. Conclusion

A large corpus of contemporary Australian English will be valuable to researchers, not only those in formal linguistics but also those in sociolinguistics, psycholinguistics, computational linguistics, education, and sociology. A corpus that is comparable to such large corpora in the world as the BNC

and the ANC will also increase the importance of Australian English as a major dialect of English in the world. In addition, a corpus of Australian English that is parallel to a corpus of British English or American English will make it particularly useful to linguists and people from other fields to compare across corpora.

This paper has discussed some of the basic aspects that constructors of the Australian National Corpus should consider. This design, once agreed on, should be made more detailed, for example, the estimated number of words for each text type should be specified. It will then provide a guideline for potential contributors to the corpus to collect the right kind of data or to refine their corpora before donating them to the Australian National Corpus. A careful design will make the task of corpus compiling more manageable and avoid waste of resources (i.e., the gathering of data that cannot be used or too much of the same kind of data).

References

Australian Bureau of Statistics. (2006). *2006 Census*. Available at http://www.abs.gov.au/.

Baker, Paul. (2006). *Using corpora in discourse analysis.* London: Continuum.

Biber, Douglas. (2004). Representativeness in corpus design. In Geoffrey Sampson & Diana McCarthy (Eds.), *Corpus linguistics: Readings in a widening discipline* (pp. 174-197). London, New York: Continuum.

Burnard, Lou. (Ed.). (2000). *Reference guide for the British National Corpus (world edition)*. Oxford: Oxford University Computing Services.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee, & Ylva Berglund Prytz. (2008). *Corpus linguistics with Bncweb: A practical guide.* Frankfurt am Main: Peter Lang.

Ide, Nancy. (2003) The American National Corpus: Everything you always wanted to know ... and weren't afraid to ask. *Invited Keynote, Corpus Linguistics 2003 Conference, Lancaster University (UK), 28-31 March 2003.* Available at http://www.americannationalcorpus.org/bib.html/.

Ide, Nancy & Catherine Macleod. (2001). The American National Corpus: A standardized resource of American English. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, & Shereen Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference, Lancaster University (UK)* (pp. 274-280).

Ide, Nancy & Keith Suderman. (2006). Integrating linguistic resources: The American National Corpus model. *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy*. Available at http://www.cs.vassar.edu/~ide/pubs.html/.

van Rooy, Bertus, & Lande Schäfer. (2003). An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference, Lancaster University (UK),* (pp. 835-844).

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:

Pho, Phuong Dzung. 2009. Towards the Design of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 25-29. Somerville, MA: Cascadilla Proceedings Project.

or:

Pho, Phuong Dzung. 2009. Towards the Design of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 25-29. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2284.