The Australian National Corpus: National Infrastructure for Language Resources

Steve Cassidy, Michael Haugh, Pam Peters, Mark Fallu

Department of Computing, School of Languages and Linguistics, Department of Linguistics, eResearch Services
Macquarie University, Griffith University, Macquarie University, Griffith University
Steve.Cassidy@mq.edu.au

Abstract

The Australian National Corpus has been established in an effort to make currently scattered and relatively inaccessible data available to researchers through an online portal. In contrast to other national corpora, it is conceptualised as a linked collection of many existing and future language resources representing language use in Australia, unified through common technical standards. This approach allows us to bootstrap a significant collection and add value to existing resources by providing a unified, online tool-set to support research in a number of disciplines. This paper provides an outline of the technical platform being developed to support the corpus and a brief overview of some of the collections that form part of the initial version of the Australian National Corpus.
Keywords: national corpus, annotation, meta-data

1. Introduction

The Australian National Corpus (AusNC) is a new project to create a wide ranging resource for research on language in Australia. In contrast to other National Corpora, it is not a new, targeted collection of language data. Instead, the AusNC will provide a common technical infrastructure for a range of collections of language use in Australia that will be unified by common meta-data, data and annotation standards and formats. This approach allows us to curate existing important collections and incorporate new collections into a larger whole that may prove more useful than the sum of its parts.

In the long term, 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 Aboriginal and Torres-Strait Islander languages, AUSLAN, and the community languages of immigrants. The Corpus also aims to serve a wide range of research disciplines from grammatical and lexical studies to sociolinguistic research and language technology. By including audio and video sources the Corpus hopes to be able to serve researchers interested in acoustics and gesture as well as language technology applications that require this kind of data to train and test computational models.

These are broad and far-reaching goals and we are a long way from achieving them in the initial instantiation of the project. Our initial round of funding has covered the establishment of the technical infrastructure and ingestion of 6-10 existing collections representing a range of data types, formats and disciplines. All of the initial collections represent Australian English and were chosen with the goal of illustrating the value of this resource to the research community.

2. Initial Collections

This section provides an overview of the initial collections that are included in the AusNC as of the launch in March 2012. As was mentioned earlier, these were selected to represent a range of different data types, formats and disciplines to allow us to learn as much as possible about the problems we will face in developing this resource. All of the initial collections are existing corpora that have been used in research on Australian language or that provide a useful resource for such research.

Data types in the collections cover text, audio and video with examples of audio collected for close phonetic analysis and for higher level sociolinguistic analysis. The **data formats** were determined by the time in which they were collected and the disciplinary research context that has made use of them. They vary from simple text files to audio and video files with associated transcripts and aligned annotation. In many cases the source data was only ever intended for manual analysis and so the format is informally defined. In other cases, a precise XML notation such as TEI is used. We think these formats are quite representative of the kinds of legacy data that is in use.

The collections represent very different **disciplines** and even include data not originally intended for linguistic analysis. There are two goals in this case, to make data available to disciplines that is familiar and hence encourage the use of the facility, but also to illustrate the value in providing access to cross-disciplinary data sets to encourage sharing of data between disciplines.

The remainder of this section summarises the different collections in the initial version of the AusNC.

2.1. Written Corpora

The initial set includes a number of collections that represent written Australian English. The Australian Corpus of English (ACE – (Green and Peters, 1991)) has a structure similar to the Brown and LOB corpora of US and British English. The Corpus of Oz Early English (COOEE – (Fritz, 2007)) is a collection of historical letters representing language use from the early days of the Australian colony until the end of the 19th century. These collections include minimal annotation but have useful meta-data.

The Austlist collection is drawn from the holdings of AustLit, a research repository in the Humanities which col-

lects Australian literature who's goal *"is to support research in and the teaching of Australian literary, narrative, and print cultures"*. The Austlit archive contains 744,075 works, however for the initial ingest into AusNC, a small sample of out of copyright texts has been drawn from the collection. The data is stored in TEI XML format and so acts as a useful sample for this data format. The samples have not been selected based on any specific criteria but they are all described by rich metadata that could be used to select texts of interest for analysis.

2.2. Corpora of Spoken Language

The Australian component of the International Corpus of English (ICE) includes a collection of written material but has a significant component of transcribed spoken language. This corpus, collected in the mid-1980s, is heavily annotated using the ICE tag-set (Wong et al., 2011) for things like speaker turns, overlaps and other speech events. The audio files for this corpus are available in digital form but the terms of the original data collection prevents these being made available online via the AusNC.

The Monash Corpus of Spoken English and Griffith Corpus of Australian Spoken English are small collections of transcripts of conversational speech including some annotation of conversational features, in the case of the Griffith corpus this uses the common Conversational Analysis style of markup (Lerner, 2004) embedded in the text and stored as Microsoft Word files. Both of these corpora include the original audio recordings although the transcripts are not time-aligned in any way. These corpora are representative of many corpora collected for small scale projects that are marked up informally in Microsoft Word using various embedded annotation schemes. We hope that by developing an ingest process for this kind of data we will be able to encourage Linguists who have similar collections to contribute them to AusNC.

2.3. The Mitchell and Delbridge Corpus

The Mitchell and Delbridge Corpus (Mitchell and Delbridge, 1965) is a collection of audio recordings of Australian speech collected on reel-to-reel tape around 1960. This is a significant National collection that has been digitised and for which various levels of annotation exist, for example word and phonetic segmentations for some recordings and transcripts for others. This represents a very different kind of data to the other corpora in the initial collection since the primary research goal was to characterise the Australian English accent, rather than looking at lexical or higher level features of language. However, the presence of interviews with school children in the 1960s means that possible contrasts with modern day data might be possible if the data is drawn into the same analytical framework.

2.4. Braided Channels

The Briaded 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. AusNC has ingested the transcripts of these interviews along with the video recordings. Transcripts are time aligned and segmented into speaker turns but have no additional annotation.

This is an interesting case of a data set collected outside of the linguistic domain being re-purposed by including it in the AusNC. The availability of high quality audio along with detailed transcripts opens up the data to a number of possible research disciplines.

```
10      J:    [(is that your      )
11            (.)
12      N:    this a- (.) media studie...
13            education subject?
14      J:    ah[:
15      N:        [but we still they we...
16            she promised, th[at the fo...
```

Figure 1: Sample of the original text from the Griffith corpus, lines have been shortened to fit in the figure

3. Technical Architecture

The goal of the project is to establish a unified technical platform that can store the source media (text, audio, video), meta-data and annotations from these different corpora and provide not only online access to the resources but value-added services that make them more useful to the research community. The technical architecture builds on the DADA system (Cassidy, 2010) and integrates separate data stores for the source media, meta-data and annotation behind a web based presentation and analysis layer based on the Plone content management system.

In order to ingest the original data from the variety of formats it is contributed in we have developed a toolkit for parsing a variety of document types and generating the required data formats for the system.

3.1. Source Media

All annotation in the corpus is stored as stand-off annotation, so the source media, be it text, audio or video, is stored separately in a web accessible location that will be referenced by the meta-data and annotation stores. For audio and video resources this is standard practice; for the text based corpora this has meant generating markup-free versions of the text to act as the source media.

The plain text version of the document is generated by removing any annotation from the original documents leaving only words and punctuation; this includes removing any speaker turn markers.

The process differs depending on the source form of the documents in the corpus. For documents that originate in Word format, we use the antiword document reader `http://www.winfield.demon.nl/` to generate a plain text version. Similarly for PDF documents we use the pdftotext utility from the poppler package `http://poppler.freedesktop.org/`. This text version of the document with markup is then processed to identify the meta-data and annotation and generate the plain text version.

An example of the text version of a document from the Griffith corpus is shown in Figure 1. This is then reduced to the

```
is that your
this a-   media studies it's a firs...
         education subject
ah
but we still they were supposed...
         she promised that the fo...
```

Figure 2: Sample of plain text from the Griffith corpus, lines have been shortened to fit in the figure

text shown in Figure 2. In this example, the alignment of text used to indicate overlaps is maintained but line numbers, speaker labels and all other annotation is removed. It is this text that forms the source document for this item in the data store. Annotations are recorded relative to character offsets in this document.

In other cases. the separation of plain text from meta-data and annotation is a little easier, for example when the original data uses an XML or simple text based format (ACE, COOEE, Austlit). In the case of the ICE corpus, we drew on earlier work on a validating parser for ICE markup (Wong et al., 2011) which was able to convert the validated ICE markup to a standoff annotation format suitable for ingestion.

The project now has experience in processing a wide range of text types to generate standoff markup on a plain text source. The methods used are encapsulated in a parsing library that can be applied to new corpora as they are contributed to the AusNC.

3.2. Meta-data Framework

The meta-data in the original collections is stored in a variety of formats ranging from spreadsheets to text embedded in tables at the start of Microsoft Word files. As part of the ingestion process, we parse this and generate a standard format that can be processed by XML tools. Meta-data is converted to RDF format to be stored in a Sesame triple store on the server.

A significant challenge has been defining a common vocabulary that meta-data fields can be mapped to across the entire corpus. The design of this vocabulary has been mediated by the desire to remain compatible with existing standards such as Dublin Core and OLAC but also by a need to provide collection level meta-data to the Australian National Data Service (the funding body) in the RIFCS format. The resulting vocabulary is necessarily a hybrid of these different vocabularies and will expand as new collections are added that have their own unique meta-data fields.

One of the strong aims of the AusNC is to support selection of data from multiple corpora using a consistent set of criteria. Therefore, we want to support as broad a common set of meta-data fields as possible to allow for this. Unfortunately the original collectors of the data have only recorded the fields that they saw as being useful and in most cases the semantic distinctions that were made in defining categories are different between corpora. This is a common problem when merging meta-data descriptions. We were able to define a small set of fields that could be given values for the majority of the items in the corpus. These include title, date of creation, location and where there are identified speakers or authors, the gender, age and some kind of place information for each person. In addition to this, we developed a multi-faceted classification to describe what some corpora called 'genre'. This expands the simple genre tag (which might have said "Popular Fiction", "Newspaper" or "Broadcast News" to a more orthogonal description in terms of distinctions such as written/spoken and published/unpublished. This classification goes some way towards being able to select data of a similar type from the different corpora that make up the AusNC and allow cross-corpus analysis.

3.3. Annotation Standards

Most of the component collections include some kind of annotation ranging from simple speaker turn boundaries to time aligned phonetic annotation and embedded Conversational Analysis markup. Our goal is to store this annotation in a unified format in a way that will support rich queries against both the text and the annotation structure.

As described in earlier papers on the DADA system (Cassidy, 2010), annotations are modelled as RDF and stored on the server in a Sesame triple store. The annotation model used is now closely aligned with the proposed ISO Linguistic Annotation Framework (rev00, 2008) and the intention is that this system is a realisation of that standard as an annotation database, rather than a data exchange format.

All annotations are converted to stand-off form by the ingestion process which, as described above, separates out the markup from the plain text version of the original document. As part of this process, embedded annotations such as speaker turns or pause markers are converted to offset annotations represented as RDF. The example in Figure 3 shows a single speaker turn annotation that has been converted to RDF. In the ISO LAF model, an annotation is attached to a Node in the annotation graph and can contain one or more properties; the node is anchored into the source document via a region defined by start and end markers. In this case the region is defined by UTF8 character offsets, but other examples might use millisecond times (for audio source documents).

Note that the speaker identifier used in the example of Figure 3 is a unique identifier for this speaker, rather than just a single letter identifier as used in the original file. As part of the ingest process, we link the annotation values representing speakers with the meta-data descriptions of these speakers. In this way, it will be possible to identify utterances via the meta-data properties of the speaker recorded in the annotation in a uniform way across the entire AusNC.

3.4. End User Capabilities

A primary goal of the project is to bring together these varied collections under a common technical framework so that a rich set of end-user tools can be built to make them a more useful resource to the research community. In the first instance we are concentrating on providing a rich search and browse interface which makes use of the data and meta-data stored in the system.

The end user facing system is implemented using the Plone content management system. Plone provides a rich set of

```
gcause:1008A a graf:Annotation;
    ausnc:speakerid gcause:speaker/GCAuse18#2;
    graf:annotates gcsause:1008;
    graf:type "speaker" .

gcsause:1008 a graf:Node;
    graf:partof gcsause:79111690-07b9;
    graf:targets gcsause:1008L .

gcsause:1008L a graf:UTF8Region;
    graf:end 3479;
    graf:start 3286 .
```

Figure 3: An example speaker turn annotation from the Griffith corpus

facilities for building web applications backed by large data stores. In this case we have used adaptors to interface the Sesame RDF store to Plone so that it can directly reflect the data and meta-data in the interface. The individual items in the corpus (an item is all of the data associated with a single source media file) appear as the equivalent of pages in the Plone CMS. As such they can be grouped into collections based on search criteria using any of the meta-data fields stored with the item. This allows us to present the original corpora (ACE, ICE, etc) to users but it also allows them to define their own collections for analysis. For example, a user could select transcripts of conversations that involve female participants under the age of 20 from across the corpus. This might draw data from the Monash, Griffith and ICE corpora. The user can then browse or perform an analysis on this collection.

The Plone front end along with the SOLR full text search engine has been used to build a full text search capability on the textual material in the corpus. The full text search facility can be use to find instances of words or phrases in a collection, displaying the results in a number of ways. The architecture of the system allows new analysis methods and new display methods to be developed and added to the system incrementally.

Search based on annotation data is not currently available due to the limitations of the initial funding for the project. We are now planning extensions to the current search facility that combines structural search of annotation data with the full text and meta-data search to provide a richer search facility.

These end user tools are designed to support the most common kinds of analysis that our user community required to make use of the AusNC data. However, we are aware that this provides only a small part of the support that would be useful for the broader language research community. The architecture of the system is such that new search and analysis facilities can be built on top of the existing data stores and deployed as web services. We expect to be able to develop new tools as funding becomes available to support research in particular disciplines.

One relatively easy to implement facility has been somewhat contentious in some cases. This is the ability to download the source documents for some or all of a collection to allow offline processing of the data. For some of the collections, the contributors of the data are uncomfortable with this as it effectively 'liberates' the data, removing it from from the protection of the web based system. For corpora that were collected some time ago when online access was not even considered, this is a significant change from the original terms of use of the data. Consequently, this facility will not been made available to end users at this time, although we are planning to allow it for those collections which can be distributed in this way.

4. AusNC as a Corpus

As it is described, AusNC is not a corpus in the normal sense of the word in the field of Corpus Linguistics where that term is used to describe a designed collection of samples, balanced over a given set of categories or contrasts. Parts of the AusNC are certainly corpora by this definition, but putting them together gives us something different. One might ask then ask how useful research can be done on such a heterogeneous collection of data.

To some extent, we don't know the answer to this question. Our hope in bringing these data sets together is that they might combine to be more than the bare sum of their parts. While it may not be possible to query over the entire AusNC collection and get balanced results, it should be possible to use the meta-data available to select your own balanced categories on which to carry out an analysis. Some pairs of corpora within AusNC will obviously be more compatible with each other. For example, the ACE and ICE corpora both represent snapshots of written Australian English at different times that might be used together to study changes in the language or just combined to provide a larger, longer timespan sample.

One request that we have had from researchers is to be able to upload their own data in order to make use of the technical infrastructure and compare their analysis with one of the AusNC corpora. So, for example, a researcher with a collection of transcripts of speech from Perth in Western Australia might upload their data and contrast it with the Monash (Melbourne) and Griffith (Brisbane) collections. This would be an excellent way of growing the holdings of the AusNC as well as making use of the existing holdings.

In summary then, while the AusNC might not be a corpus in the usual sense of the word, there may be opportunities and new ways of looking at data that arise from having so much

diverse data available under a common technical platform. At the very least, the component corpora don't lose any of their cohesion by being part of AusNC, so a researcher can still carry out a study on ACE, ICE or any of the other collections.

5. Summary

The Australian National Corpus is a major new initiative that explores a new way to provide national level infrastructure for language resources, building a National corpus from the many existing collections and serving a broad range of disciplines. The initial funding has allowed us to incorporate a small number of diverse collections into a single unified platform and established a set of tools that will enable the corpus to grow as new collections are contributed.

The AusNC is available for use at `http://www.ausnc.org.au/`.

6. Acknowledgements

This work was funded by a grant from the Australian National Data Service. A pilot study was supported by the Australian Research Council Network in Human Communication Sciences - HCSNet.

7. References

Steve Cassidy. 2010. An rdf realisation of laf in the dada annotation server. In *Proceedings of ISA-5*, Hong Kong, January.

C.W.A. Fritz. 2007. *From English in Australia to Australian English: 1788-1900*. English corpus linguistics. Lang.

E. Green and P. Peters. 1991. The australian corpus project and australian english. *ICAME JOURNAL*, 15:37–53.

G.H. Lerner. 2004. *Conversation analysis: studies from the first generation*. Pragmatics & beyond. John Benjamins Pub.

A. G. Mitchell and A. Delbridge. 1965. *The pronunciation of English in Australia*. Angus and Robertson.

ISO/TC 37/SC 4 N463 rev00. 2008. Language resource management - linguistic annotation framework. Technical report, International Organization for Standardization, Geneva, Switzerland.

Deanna Wong, Steve Cassidy, and Pam Peters. 2011. Updating the ice annotation system: tagging, parsing and validation. *Corpora*, 6(2):115–144.