Interoperable Annotation in the Australian National Corpus

Steve Cassidy
Department of Computing
Macquarie University
Sydney, Australia
steve.cassidy@mq.edu.au

Abstract

The Australian National Corpus (AusNC) provides a technical infrastructure for collecting and publishing language resources representing Australian language use. As part of the project we have ingested a wide range of resource types into the system, bringing together the different meta-data and annotations into a single interoperable database. This paper describes the initial collections in AusNC and the procedures used to parse a variety of data types into a single unified annotation store.

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 manage 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.

The pilot project that established the AusNC chose a small number of corpora that were felt to characterise the range of corpora in use by Australian researchers. These include a number of important historical collections that have been used to characterise Australian English in the past. The primary focus of the project was to ingest the corpus text and meta-data into a web accessible form and provide a way of browsing this data and publishing meta-data records to the Research Data Australia directory[1]. However, as a part of the ingestion process, we undertook to parse as much annotation data as possible and convert it to an RDF format (Cassidy, 2010) so that it might be used in a future version of the technical infrastructure.

This paper describes some aspects of the process by which meta-data and annotations were extracted from these corpora and the measures we took to ensure the interoperability of the data in the AusNC platform.

2 Overview of Corpora

The corpora included in the initial collection are drawn from a range of disciplines and contain a varied amount of meta-data and annotation. In summary, the corpora are:

[1] http://researchdata.ands.org.au/australian-national-corpus

- **The Australian Corpus of English (ACE)**: Written language, some simple XML like markup for header, bylines etc.
- **The Australian ICE Corpus**: Written and spoken language, XML like markup following the ICE standards.
- **The Corpus of Oz Early English (COOEE)**: Historical texts with minimal markup.
- **The Monash Corpus of Spoken English**: transcribed audio of conversations in Word format, speaker turn annotation
- **The Griffith Corpus of Australian Spoken English**: transcribed audio of conversations in PDF format with embedded Conversation Analysis markup.
- **The AustLit collection**: TEI formatted samples of Australian fiction.
- **The Mitchell and Delbridge Corpus**: audio recordings with time aligned word and phonetic annotations.
- **The Braided Channels Research Collection**: video recordings with transcriptions in Word format, speaker turn annotations, roughly time aligned with video.

All of these corpora are hand-annotated - the annotation was done as part of the data collection and served the research in a particular discipline. There is clearly scope for adding more machine-generated annotation such as sentence segmentation and POS tagging, but doing so was beyond the scope of the project. The work we report here is about understanding the existing annotation and ingesting it into an interoperable framework.

3 Some End User Goals

The goal of the AusNC is to bring together more collections of Australian language so that researchers can benefit from being able to work with many collections in a uniform way. To illustrate this we will look at two example 'use cases' from the point of view of a Linguistics researcher.

The first case involves a study of utterance final constructions and their effect on the following utterances. Researchers want to identify certain lexical items occurring at the end of a speaker turn (eg. 'is it?', 'can he?'), classify the turns according to the gender of the speaker and then study the turns and those that follow them to look for common patterns.

The second case looks at overlapping speech in dialogue. The researcher is interested in the lexical items that are used in backchannel interjections ('hmm', 'yeah', 'really') and so wants to generate a list of words that occur during overlapping speech ordered by frequency and distinguished by the gender of the speaker.

Each of these tasks can be achieved by researchers on the existing data sets; in fact they are things that have been done already. The main issue is that the variability in the way that meta-data and annotation is represented in the corpora mean that any study that wanted to work over multiple corpora would need to process each one separately with difficult and different manual methods. The three corpora that we'll target in these examples are the Griffith, Monash and ICE-AUS corpora, all of which contain transcriptions of dialogue with some overlap information and which have been identified by researchers as good resources that they would like to be able to make use of.

The two cases are similar in that they both involve identifying speaker turns in dialogue. These are represented differently in the source corpora, with Griffith and Monash using formatting within the Word or PDF document (a line starting with a speaker identifier and a colon) and ICE-AUS using XML like markup in the text. In Griffith and Monash, the end of a speaker turn is implicitly marked as the newline before the start of the next turn and so searching for words at the end of turns is problematic.

Speaker meta-data is available in all three corpora but in very different forms. In ICE-AUS it is in a separate spreadsheet; in Griffith and Monash it is at the head of each transcript in a table. Essentially, finding the gender of each speaker is a manual process of tabulating the available data, except for Monash which encodes gender in the speaker identifier.

The third kind of annotation we need to look at is overlap. This is handled very differently in each

case. Monash and ICE-AUS use explicit markup for regions of overlapped speech - in the case of Monash the text is enclosed in square brackets. Griffith's CA style of annotation uses an open square bracket to mark the start of overlap and vertical alignment to mark the relationship between the two speaker's utterances, but the end of overlap is not marked explicitly. ICE-AUS has an explicit mechanism for linking two overlapping segments but Monash relies on the reader to line up multiple segments. So if we have three speakers:

```
BH4M:    [whats that]
BH4MMo:  [what] did he do?
BH4MFa:  .. well we were going to
         the milkbar on Sunday
BH4MMo:  [oh]
BH4M:    [oh] here we go
```

we need to be very careful to keep track of the overlaps from the start of the discourse to be able to identify what overlaps with what.

A final consideration is document selection. Both the Monash and Griffith corpora represent a single kind of language use - conversation. However, the ICE-AUS corpus contains samples of conversation alongside monologues, newspaper text and fiction. Clearly in carrying out any study over multiple corpora, a researcher needs to be able to select appropriate documents based on their descriptive meta-data.

Based on this review, it is clear that if a researcher is to be able to perform queries on more than one data set, the main thing standing in their way is the diversity of representations of the phenomena that are annotated. In this case, the meaning of the annotations is aligned in each case (speaker turns, overlap) but their realisation is quite distinct. In addition, the link to meta-data about the speaker and the kind of language represented in each document needs to be clear.

4 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.

The meta-data and annotation stores are built on an RDF triple store. The use of RDF for meta-data is well understood and our implementation makes use of standard vocabularies as far as possible to describe corpora and their contents. Modelling annotation data as RDF is less well established but our earlier work has shown that the data model and query language are well suited to the task. Among the challenges in this project are managing the scale of data resulting from ingesting annotations from a large number of corpora and dealing with the issues that arise in storing many different corpora in a single annotation store.

4.1 Parsing Annotation

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.

To generate the markup-free based versions of the text we have developed a parsing library that is able to handle the variety of markup that we have found in our target corpora. The library, based on the Python `pyparsing`[2] module, is written such that new parsers can be built by chaining together primitive parser elements. The output of the parsing process is twofold – the plain text without markup and a stream of annotation objects that reference character offsets in the plain text stream. An example of calling a simple parsing procedure is shown in Figure 1.

The output from these parsing procedures is combined to produce the plain text version of the document and a collection of annotations that are then converted to RDF.

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

[2] `http://pyparsing.wikispaces.com/`

```
>>> markupParser('h', 'heading').parseString("<h>some stuff</h>")
([@(some stuff,[heading:  0 -> 10])], {})
```

Figure 1: An example call to one of the parser procedures, in this case parsing an XML style header from the ACE corpus. The result is a representation of the plain text and the annotation with character offsets.

```
RF3:  [Okay]
BH1M: [Im fifteen] years old.
RF3:  Fifteen?
BH1M: Yes.
RF3:  How do I spell your surname?
```

Figure 2: Sample of the original text from the Monash corpus

```
Okay
Im fifteen years old.
Fifteen?
Yes.
How do I spell your surname?
```

Figure 3: Sample of plain text from the Monash corpus corresponding to the raw text in Figure 2

for ingestion.

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 (ISO 24612, 2012) and the intention is that this system is a realisation of that standard as an annotation database, rather than a data exchange format.

4.2 Parsing Speaker Turns and Overlaps

An example of the text version of a document from the Monash corpus is shown in Figure 2; this contains examples of both of the phenomena mentioned in Section 3: speaker turns and overlap. The parsing process removes all markup (in this case, the speaker identifiers and the square bracket overlap notation) and generates the text shown in Figure 3 and a collection of RDF annotations which will be discussed below.

A second part of the ingestion process is to read and normalise the meta-data that is associated with the primary data. This is found in different forms:

```
monash:speaker/BH1M a foaf:Person;
    monashp:role "primary";
    monashp:school "BH";
    foaf:age "15";
    foaf:gender "male" .
```

Figure 4: Part of the meta-data for the sample of Figure 2 describing the speaker BH1M.

spreadsheets, text files and in the case of the Monash and Griffith corpora, in tables at the start of each transcription file. This data is parsed as part of processing the document and normalised to standard vocabularies where possible. Items like speaker identifiers are treated specially to ensure we maintain the link between speaker data and annotations on speaker turns, and that speaker identifiers are unique across the different corpora. Figure 4 shows the description of one speaker which uses the standard `foaf` namespace[3] commonly used to describe individuals. Since the same property names are always used, we can filter speakers by gender or age (where available) irrespective of the corpus they contributed to.

A sample speaker turn annotation is shown in Figure 5 in the RDF format used by the DADA system. This is basically a set of descriptions of objects via attribute-value pairs. In this case, the object `monash:5514A` is an instance of the class `dada:Annotation` and has properties `dada:type` etc. The colon notation denotes namespaced identifiers which can be described by a formal vocabulary (ontology). The RDF descriptions of annotations can reference parts of the meta-data as seen in the `ausnc:speakerid` property in the example which references the speaker described in Figure 4.

The text in Figure 2 also contains an example of overlapping speech marked as square bracketed text. This is also recognised as part of the parsing process

[3] `http://www.foaf-project.org/`

```
monash:5514A a dada:Annotation;
    dada:type ausnc:speaker;
    dada:partof monash:10cdaedc;
    dada:targets monash:5514L;
    ausnc:speakerid monash:speaker/BH1M .

monash:5514L a dada:UTF8Region;
    dada:start 91;
    dada:end 113 .
```

Figure 5: Part of the RDF annotation generated from the raw text in Figure 2. The first part describes the annotation object itself which has a number of properties, this *targets* a locator object described in the second part as a region bounded by UTF8 character offsets. This represents the second line in Figure 2.

and annotations marking this region as overlap are generated. In this case it would be useful to also record the relationship between these two instances of overlap - that 'Okay' is spoken at the same time as 'Im Fifteen'; however, our parser is not yet capable of doing this for the Monash data. We have done this for another corpus, ICE-AUS as part of the work reported in (Wong et al., 2011) but in this case, instances of overlap were numbered to allow the correspondence to be made explicit. However, we found that since the annotators were unable to validate the markup they were writing (it was XML like but didn't conform to any formal system), there were many deviations from the stated rules that needed to be corrected before a useable parse could be completed. We suspect that this will be the case with the Monash data as well.

There are also examples of overlap in the Griffith corpus, marked up with the CA convention of an open square bracket, vertically aligned with the corresponding text from the second speaker. Here's an example:

```
11  H:  [family gen[der book two
12  S:  [can-      [can I borrow
13      that?
```

Given the involvement of vertical alignment and the lack of explicit end markers for the overlap, we've not yet been able to successfully parse this markup, however we are confident that we should be able to recover most of the information here with further work.

```
monash:5513A a dada:Annotation;
    dada:type ausnc:overlap;
    dada:partof monash:10cdaedc;
    dada:targets monash:5513L .
monash:5513L a dada:UTF8Region;
    dada:start 91;
    dada:end 102 .
```

Figure 6: Part of the RDF annotation generated from the raw text in Figure 2 showing an overlap annotation corresponding to the text 'Im Fifteen'

5 Discussion

5.1 Achieving User Goals

In Section 3 we presented two example tasks that users had identified as targets for the work we were doing in building the AusNC. These relied on having a more uniform annotation model that would allow queries over speaker turns and overlapping speech when the source corpora have quite different ways of expressing this markup.

We have described the ingest process for the AusNC which aims to build this uniform representation of annotation. An important part of this is the use of common labels for annotation types such that the same phenomena in different corpora can be identified in the same way. While the examples we chose were quite simple (and not particularly 'semantic'), they illustrate the concept of using standard types to describe kinds of annotation.

The solution that we have describe only goes part of the way towards solving the problems presented in Section 3 however. We've built a model but we need to build the query tools and analysis engines that can make use of the data to answer questions from researchers. We are currently involved in a follow-on project that aims to do just this, adding infrastructure for running tools that will support query and analysis of corpus data from the AusNC as well as generating new annotations by running automatic processes such as parser and POS taggers.

5.2 Annotation Types

Though the annotation data model is standardised across the different corpora, the types and contents of the annotations is different. The `dada:type`

property of each annotation denotes an *annotation type* while the `ausnc:val` property is used to carry a value or label for the annotation. Other feature values can be expressed as additional RDF properties on the annotation node.

The concept of annotation type is not directly expressed in the ISO-LAF standard but is realised in most examples as a non-distinguished property of each annotation or via the AnnotationSpace property. The main point being that there is no *requirement* in ISO-LAF for any kind of type system but that there are a couple of mechanisms by which one could be implemented which would be equivalent to the model used here.

The use of the type system allows us to assert that certain kinds of annotation are semantically equivalent - in this case the speaker turns and overlaps in different corpora. This is a key to the interoperability of annotations because without this we cannot reliably treat the annotations as having the same meaning. The use of RDF makes it natural to use a schema to describe the annotation types, meaning that we can generate schemas to describe different styles of annotation - from transcribed dialogue to Penn Treebank style parse trees.

In order to make any type system useful, the way that it is used needs to be standardised. The DADA vocabulary makes one suggestion that is compatible with the ISO-LAF framework; while there may be other options to consider, it would be an important next step to discuss how this should be realised within the standard.

5.3 Other Annotation Types in AusNC

As the ingest scripts were developed for the different corpora in AusNC, common type names were used for annotations where possible. However, since the focus of the project was on the ingestion of primary data and meta-data, there were only a small number of types that were identified as common over more than one corpus.

In all other cases, annotation type names, values and other properties were derived from the names used in the individual corpora or where appropriate in the documentation for the corpora. A good example is the Griffith corpus which uses Conversational Analysis markup embedded in the text. The documentation for this annotation style was taken from (Lerner, 2004) which contains a glossary of transcription symbols with an informal description of their use and meaning. Table 1 lists the types that we have parsed with some examples of their use (there are a few other types that are used in the corpus that we are still working on parsing correctly).

Type Name	**Example**
micropause	(.)
pause	(1.2)
elongation	fo:r commu:nicating
intonation	if ↑I couldnt bo↓rrow,
latched-utterance	7 H: sexuality=
	8 S: =ah
speaker	5 S: I'm glad I saw you
volume	business °cause° I missed
uncertain	S: (,) this morning,

Table 1: Annotation types and examples from the Griffith corpus

6 Summary

This paper has tried to summarise some of our experiences in taking source data in many different formats and generating a single, interoperable annotation store that can hold annotations on many resources from different collections. The current system is able to present these resources via the web[4] and we are now starting to develop tools to work with the annotated data to help answer research questions for the diverse communities who make use of this data.

References

Steve Cassidy. 2010. An RDF Realisation of LAF in the DADA Annotation Server. In *Proceedings of ISA-5*, Hong Kong, January.

ISO 24612. 2012. Language Resource Management – Linguistic Annotation Framework.

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

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

[4] `http://ausnc.org.au/`