Ingesting the Auslan Corpus into the DADA Annotation Store

Steve Cassidy
Centre for Language Technology,
Macquarie University, Sydney, Australia
Steve.Cassidy@mq.edu.au

Trevor Johnston
Department of Linguistics
Macquarie University, Sydney, Australia
Trevor.Johnston@mq.edu.au

Abstract

The DADA system is being developed to support collaborative access to and annotation of language resources over the web. DADA implements an abstract model of annotation suitable for storing many kinds of data from a wide range of language resources. This paper describes the process of ingesting data from a corpus of Australian Sign Language (Auslan) into the DADA system. We describe the format of the RDF data used by DADA and the issues raised in converting the ELAN annotations from the corpus.

1 Background

The DADA system is being developed to support collaborative access to and annotation of language resources over the web. DADA provides a web accessible annotation store that delivers both a human browsable version of a corpus and a machine accessible API for reading and writing annotations. DADA is able to ingest data from a number of different annotation formats and the data model it supports is intended to be a general model of annotation data. This paper reports on our efforts to ingest data from the Australian Sign Language (Auslan) corpus which has been annotated with the ELAN tool[1]. The primary goal of this project is to provide a read-only web-accessible version of the corpus but a longer term goal is to allow research groups to collaborate in extending the annotation.

DADA follows the principle of *linked data* (Bizer et al., 2008), every object (e.g. annotation) within the store is associated with a URL and accessing that URL generates a description of the object which includes links to the URLs of related objects. So, accessing the URL of an annotation might return a web page describing that annotation linked to its parent or the annotation set or corpus that it is part of. Linked data is an emerging design pattern in semantic web research which is being used to enable data discovery and general purpose browsing tools. For our purposes, the idea that every component of a corpus has a web-accessible URL is very powerful; it means that individual annotations could be cited in publications and opens up a whole new range of possibilities in communicating results and analysis.

There have been a number of earlier projects that aimed to develop web accessible versions of data annotated in ELAN. EOPAS (Thieberger and Schroeter, 2006) aimed to provide a means of exploring ethnographic data on the web. Various kinds of annotation data, including ELAN, could be ingested into the EOPAS XML format using stylesheets. The flexibility of an XML database was used to allow the web views of data to be generated via calls to XSLT and XQuery scripts. Because of the nature of the data being displayed, EOPAS developed views particularly suited to interlinear text although the same infrastructure could be used to generate other kind of display.

Like EOPAS, DADA makes use of an industry standard data store, however we choose RDF instead of XML because of the very close fit between the RDF data model (a directed graph) and the data model that has been shown to be needed to represent annotation data (Bird and Liberman, 2001).

The choice of RDF also allows us to leverage existing work on annotation on the web. The Annotea project at the W3C and the later Vannotea project (R.Schroeter et al., 2003) define an RDF format for storing annotations on media on the web. The models developed for DADA owe a lot to these earlier systems but build on them to provide an appropriate data model for linguistic annotation.

1.1 The Auslan Corpus

The Auslan corpus is a digital video archive of Australian Sign Language (Auslan) (Johnston and Schembri, 2006). The archive is the product of an Endangered Languages Documentation Project funded through the Hans Rausing Endangered Languages Documentation Program (ELDP) at the

[1] http://www.lat-mpi.eu/tools/elan/

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 154–157,
Suntec, Singapore, 6-7 August 2009.

School of Oriental and African Studies (SOAS), University of London (grant #MDP0088 awarded to Trevor Johnston). The corpus brings together into one digital archive a representative sample of Auslan in video recordings to which are added metadata files and annotation files created in ELAN. It is consists of two sub-corpora: data collected through the ELDP and data collected as part of the Sociolinguistic Variation in Auslan Project (SVIAP) conducted by Adam Schembri and Trevor Johnston (ARC #LP0346973). Both datasets are based on language recording sessions conducted with deaf native or early learner/near-native users of Auslan.

Many tiers are needed in an ELAN file to annotate a text in a signed language because sign languages can have several simultaneous levels of linguistically significant behavior. For example, each hand may utter a separate manual sign at the same time, or grammatically important body movements and facial expressions (which are not unlike prosody) may co-occur with the production of individual manual signs. All this needs to be identified and time aligned.

2 Mapping ELAN to RDF

RDF, the Resource Description Framework, is the core language of the semantic web used to make assertions about resources, describing them in terms of properties and relations to other resources. DADA stores annotations as RDF in a dedicated database called a *triple store* and uses semantic web technologies to manipulate and present data. To represent annotations, DADA defines a core ontology that maps to the data structures inherent in annotation data. The ontology is designed to be able to represent many kinds of annotation data and as such owes much to similar *lingua franca* efforts such as Annotation Graphs (Bird and Liberman, 2001) and the Linguistic Annotation Format (Ide and Suderman, 2007).

To ingest the annotations from the Auslan Corpus into DADA requires transcoding of ELAN XML annotation files into the RDF format. This section provides an overview of the DADA RDF ontology and then discusses the issues raised by mapping ELAN data.

The core object types within the DADA ontology are: the **corpus**, a collection of annotation sets; the **annotation set**, a collection of annotations on one or more media files denoting a single event or stimulus; the **annotation**, the basic unit of annotation associated with a region within the source media and the **anchor**, an abstraction of the idea of a location within a source media file. Each of these written in this paper as, for example, `dada:Annotation` but this is shorthand for a URL (`http://purl.org/dada/schema/0.1#Annotation`) which provides a unique name for the property. Each of these object types can have arbitrary properties and relations defined from the DADA or other ontologies. DADA properties define the basic structure of annotations; an example is given in Figure 1. In the figure the lines between nodes define the named relations; for example, the offset times of the anchors are defined by relations denoting the units of measurement (`time:seconds`). The data associated with the annotation is encoded by one or more relations (e.g. `auslan:RH_ID_gloss`); in this way, each annotation is associated with a *feature structure* that can encode multiple properties of the annotation.

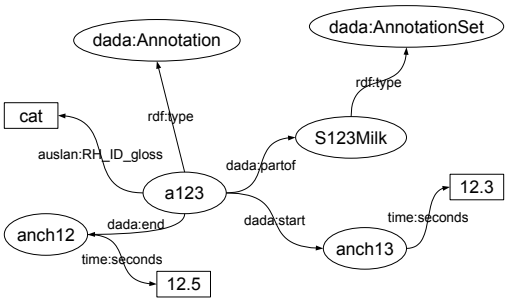

Figure 1: An example annotation structure in RDF.

The mapping between the ELAN EAF format used by the Auslan Corpus and the DADA RDF format is relatively straightforward. EAF stores annotations for a single media file (or group of related media) in an XML format which can be transformed into the RDF/XML format via an XSLT stylesheet. ELAN annotations exist on a set of *tiers* which have associated type information; for example, Auslan uses the *RH ID Gloss* tier to represent the sign being made by the right hand.

The type of annotation on a tier is defined by the associated *linguistic type* which gives a name for the type and defines it as one of five *stereo-*

types that describe how the annotation divides the timeline and relates to other annotations. There are a number of stereotypes defined by ELAN but the Auslan corpus only makes use of two: a simple time aligned type and a *symbolic association* type. The simple time aligned tiers form the base of the annotation and denote the start and end of signs and other events in the video stream. Symbolic association tiers provide additional information about these base level annotations; each annotation on one of these tiers is associated with a base level annotation which defines its start and end points. This is modeled in RDF by additional properties denoting the associated annotations. For example, Auslan defines the *RH ID Gloss* tier as a base segmentation of the video timeline and has associated tiers *RH gram cls* and *RH loc* among others. Instead of building separate annotations for each of these, they are modeled in RDF as three properties of a single annotation as illustrated in Figure 2.

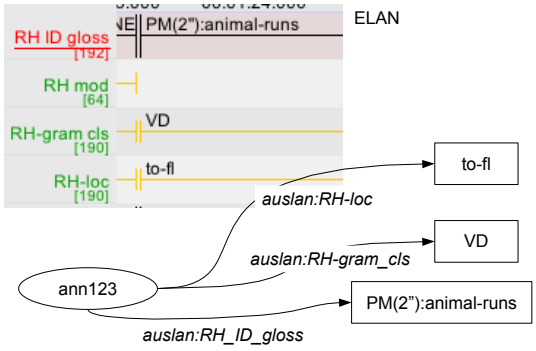

Figure 2: Conversion of associated tiers in ELAN to RDF properties

ELAN does support other types of inter-tier relationships, in particular one tier can be a *symbolic subdivision* of its parent. In this case, additional annotations must be made for each tier and the RDF model becomes a little more complex. This is not discussed further here as it is not required for modeling annotations in the Auslan corpus.

Since the RDF representation of annotations requires us to use formal relation names for properties corresponding to tiers, we are required to define these names in an ontology specific to the style of annotation being used in the corpus. While ELAN does not provide a mechanism to define a *schema* – definitions of a set of tiers – for a corpus, most corpora will use the same tiers in every file. As a side effect of importing a set of ELAN files into the DADA RDF format we generate an RDF schema that defines the linguistic types being used. These types form a point of reference for the corpus and can form a useful part of the documentation of the annotation system being used. In the longer term, the availability of such public documented linguistic types might lead to more effective sharing of annotation types between corpora. While these are stored as RDF documents, it is easy to generate template ELAN annotation files or indeed templates for other annotation tools from the same data.

While the current definitions of linguistic types are generated entirely from the ELAN source file, there is scope to link these to external ontologies being developed for linguistic analysis. Relatedly, ELAN supports linking to external controlled vocabularies (Sloetjes and Wittenburg, 2008) such as the ISO Data Category Registry (ISO 12620) which allows sharing of terms (such as Verb, Noun) that might be used as annotation labels.

3 Publishing on the Web

Once ingested into the RDF store, the ELAN annotations can be manipulated by the DADA system using standard interfaces such as the SPARQL query language. The primary interface to DADA is via the web, either through a standard web browser or via the well defined HTTP based interface for application programs. This interface presents machine-readable versions of the annotation data in well known formats; for example, returning an ELAN or Transcriber XML formatted version of a set of annotations or even a lower level XML representation. The application would not generally see the annotations in raw RDF form although such an interface can be supported if needed.

The primary view of the annotation data on the web is via HTML pages generated by the server that can contain any data derived from the annotation store. We have developed a set of generic displays for each kind of object (corpus, annotation set, etc) that are generated in response to a request for the appropriate URI from a web browser. These display all of the relevant data for the object and could be customised to provide an appropriate view of different kinds of data.

The web browser is not the only kind of client

that can retrieve data from the DADA server over the web. DADA makes use of HTTP *content negotiation* between the client and the server to enable a client to request one of a number of alternate forms of data. For example, the server can generate an ELAN format XML file which closely mirrors the original data ingested into the system. Since the output is generated via templates, other formats could also be generated to feed into alternate tools. In addition to generating existing XML formats it is also useful to generate data in a form that is easily consumed by custom applications. JSON (Javascript Object Notation[2]) is a data format that is frequently used to transport data in modern web applications and is easily parsed by libraries in many target languages. The DADA JSON interface will deliver descriptions of any kind of object in the data store in a way that makes it easy to implement clients that present interactive displays of the data.

4 A Javascript Client

As a demonstration of the web based API allowing remote clients to read annotation data from the server, we have implemented a Javascript based browser for annotation data that is able to show data aligned with the source video data. The Javascript client can be hosted on a website unrelated to the DADA server since it gains access to data purely via HTTP requests for JSON formatted data.

The client provides a web interface that is entirely dynamic, allowing the user to choose from a list of corpora hosted on the server, then choose an annotation set and finally select a type of annotation to display. The client also queries the server for details of the media files associated with the annotation set and embeds the web accessible FLV formatted video in the page. The user is able to interact with the page and navigate through the video via links from the annotation.

5 Summary

The DADA system aims to provide general purpose infrastructure for collaborative annotation of language resources. Built on core semantic web technologies it provides a server based solution that is able to ingest annotation data from a number of sources and deliver them both to human browsers and to client applications. In the first phase of development the emphasis is on delivering views of existing corpora on the web.

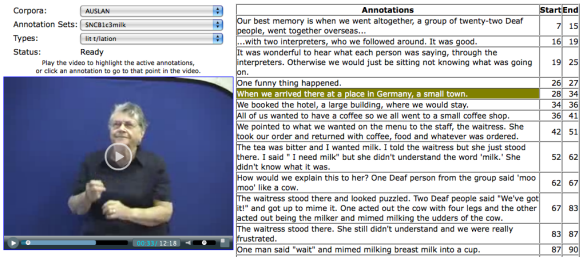

Figure 3: A screenshot from the Javascript client

A demonstration version of the DADA server is hosted at `http://dada.ics.mq.edu.au/` and contains a link to the Auslan data described here. More information on the Auslan corpus can be found at `http://www.auslan.org.au/`.

References

S. Bird and M. Liberman. 2001. A Formal Framework for Linguistics Annotation. *Speech Communication*.

Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. 2008. Linked data on the web (ldow2008). In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1265–1266, New York, NY, USA. ACM.

N. Ide and K. Suderman. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007*, Prague. `http://www.cs.vassar.edu/~ide/papers/LAW.pdf`.

T. Johnston and A. Schembri. 2006. Issues in the creation of a digital archive of a signed language. In L. Barwick and N. Thieberger, editors, *Sustainable data from digital fieldwork: Proceedings of the conference held at the University of Sydney*, pages 7–16, Sydney, December. Sydney University Press.

R.Schroeter, J.Hunter, and D.Kosovic. 2003. Vannotea A Collaborative Video Indexing, Annotation and Discussion System for Broadband Networks. In *Proceedings of the Knowledge Markup and Semantic Annotation Workshop, K-CAP*, Sanibel, Florida, Oct.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: Elan and iso dcr. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Nicholas Thieberger and Ronald Schroeter. 2006. EOPAS, the EthnoER online representation of interlinear text. In Linda Barwick and Nicholas Thieberger, editors, *Sustainable Data from Digital Fieldwork*, pages 99–124, University of Sydney, December.

[2] `http://www.json.org/`