

## **Annotation graphs: theory and applications**

**Steven Bird**

University of Melbourne

Annotated corpora have been a critical component of research in the speech and language sciences for some years. Today, these corpora are being created and deployed for a rapidly expanding set of languages, disciplines and technologies. A wealth of formats and tools have sprung up around this enterprise, many of which are documented on the Linguistic Annotation page:

[<http://www ldc.upenn.edu/annotation/>].

"Linguistic annotation" is a term which covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings - or it may be textual. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, and so on.

This tutorial will focus on a model of linguistic annotation which provides a simple framework for representing and manipulating complex, heterogeneous, multi-layered annotations. The model uses directed acyclic graphs having labels on the edges and time-offsets on the nodes, so-called annotation graphs. The tutorial will cover the formalism, the software infrastructure, and practical applications. Participants will learn about the steps involved in building their own special-purpose annotation tools.

As we create new language resources, such as annotated corpora and the associated annotation software, there needs to be a standard way to describe them so that they can be found and re-used by others. The tutorial will cover a new framework for sharing language resources, the Open Language Archives Community [<http://www.language-archives.org/>].

### **Topics:**

- + introduction to linguistic annotation
- + a survey of annotation tools and formats, requirements analysis
- + annotation graphs: a formal model for annotation
- + the ATLAS architecture
- + software infrastructure for building annotation tools
- + the annotation graph API; the XML interchange format
- + wrapping existing software components
- + database representation, annotation servers, query
- + design principles for annotation tools
- + case studies
- + language resource metadata
- + OLAC, the Open Language Archives Community