Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 00 (2017) 1–27

www.elsevier.com/locate/csl

Toward a format-neutral annotation store

Robert Fromont

New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, New Zealand

Received 30 August 2016; received in revised form 21 December 2016; accepted 20 January 2017

Abstract

Sharing speech corpora and their annotations is desirable, in order to maximise the value gained from the expense and hard work involved in transcribing and annotating them. However, differences in conventions and format are barriers to sharing of data; text conventions conflict, file formats differ, and annotation ontologies do not match up. Using a 'pivot' form to store annotations in a tool and format neutral manner can alleviate many of these difficulties. There are several possibilities for the pivot form, including the Annotation Graph model, which meets most of the requirements to be a pivot. The LaBB-CAT software's implementation of Annotation Graphs incorporates some extensions to the model, which handle the remaining unmet requirements, and create the possibility of defining an annotation API that makes automation of conversion, querying, and manipulation of annotations easier.

Keywords: Speech corpora; Languane annotation; Annotation graph; Interoperability

1. Introduction

Linguists and others who study language have long collected examples of actual language usage, both written, and increasingly spoken, and in the course of their investigations have found it useful to annotate their examples with features that are relevant to their particular research question. Often these annotations are devised and conducted by an individual researcher and are only used for a single research project. But there is increasing desire to share and re-use not only examples of language, but also the annotations that accompany them.

Speech corpora of increasing diversity and size are collected in many different domains of research, and are explored and processed by an overwhelming variety of software tools that aid manual and automatic analysis of speech, for example in the domain of child speech research the Child Language Data Exchange System (CHILDES) project (MacWhinney, 1984) annotated with CLAN (Spektor and Chen, 2012), in the corpus linguistics domain the British National Corpus (BNC) (BNC Consortium, 2007) transcribed using a version of the Text Encoding Initiative (TEI) guidelines (Burnard and Bauman, 2012), in the domain of phonetics the Buckeye corpus (Pitt et al., 2007) released in the XWaves (Hawkings, 2008) format, and in the discourse analysis domain the AMI Corpus (Carletta et al., 2006) released in the NXT format (Kilgour and Carletta, 2006), to name but a few. Recording, transcribing, and annotating speech is an expensive and time-consuming process, and so sharing these resources is desirable.

Brian MacWhinney, one of the driving forces behind the CHILDES project, has explained that the sharing of such language data is important to facilitate further study from the same raw data, and also to promote

E-mail address: robert.fromont@canterbury.ac.nz

http://dx.doi.org/10.1016/j.csl.2017.01.004
0885-2308/

academic rigour, as findings can be checked against the data from which they are derived (MacWhinney, 2012, Section 3.1 pp. 14–15).

In addition to promoting academic rigour, a great deal of research can be facilitated by the re-use of existing recordings, transcriptions, and annotations. This applies not only to 'open' corpora that are available to researchers in different institutions and for different purposes, but also to 'closed' corpora that, for participant consent or other reasons, cannot be shared outside their originating institution; new research projects can build on the work of previous projects, allowing corpora to accumulate annotations that are increasingly diverse or refined.

There are many challenges to re-using and sharing linguistic annotation data, and different approaches have been used in order to address these challenges. This paper discusses some of these challenges and solutions.

The structure of this article is as follows: Section 2 discusses in some detail, barriers to the sharing and re-use of linguistic annotations, and some solutions to the problems raised, including the use of a 'pivot' annotation model. Section 3 describes various pivot models that have been proposed in the literature, including "Annotation Graphs", an extension of which we have found useful in the development of a corpus annotation system called "LaBB-CAT". This software is then described in Section 4, which explains our pivot model extensions, how they solve some outstanding problems, and provide further benefits for annotation processing. Finally, the discussion is summarised and future work is proposed in Section 5.

2. Sharing, converting, and re-using

When linguistic annotations are created the focus is often on the particular needs of a specific research project. However, facilitating the sharing of annotation data is an increasingly important goal of annotation. New research projects may involve comparisons with past projects' data, building on the annotation work previously carried out, or merging different corpora into larger collections. Annotation sharing also encompasses the translation of data between two inter-operating systems.

Differences in conventions and format are barriers to sharing of data; text *conventions* (the labels and textual constructions edited by human annotators) conflict, file *formats* (the formal, computer-processed file structure) differ, and annotation *ontologies* (the collection of entities of interest to the researcher) do not match up. This problem is neither new nor completely solved yet. Both the CHILDES' project's CHAT format, and the TEI guidelines (Burnard and Bauman, 2012)[1] were conceived as solutions, and both have their roots in the 1980s, but thirty years later (Draxler et al., 2011; van Gompel and Reynaert, 2013) authors are still lamenting the existence of ad-hoc, undocumented, and unstandardised formats.

These problems, and various approaches to data sharing, are discussed below.

2.1. Using a single format

One way to facilitate language data sharing is to ensure that everybody who wants to use it is using the same tools, the same formats, and the same conventions; i.e. total standardisation of data. This approach has achieved quite some success in child language studies, many of which use the SALT software (Miller, 2012), and the CHILDES project (MacWhinney, 1984) uses the CLAN (Spektor and Chen, 2012) software with the CHAT (MacWhinney, 2012) format. With the TalkBank project (MacWhinney, 1999), the approach has also broadened into other research domains.

Both SALT and CHAT have carefully devised prosodic annotation conventions, enumerations of standard spellings for contractions, rules for bound morpheme annotation and other coding, and predefined dependent tiers, which are all designed to achieve this kind of total standardisation. This can clearly work well within a specific community of research where tools and standards are very well established and the range of annotations can be very clearly defined in advance. They are necessarily highly prescriptive about what can be annotated, how it should be annotated, and under what circumstances, so annotations not envisaged in the format definitions cannot easily be made.

[1] TEI is described in Section 3.1.

Please cite this article as: R. Fromont, Toward a format-neutral annotation store, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.004