The Value of Relational Databases for Time-Aligned Annotation

Tyler Kendall
Northwestern University & North Carolina State University
t-kendall@northwestern.edu


AACL 2009, University of Alberta, Oct. 10th 2009

Time-Aligned Transcription

• Recent work in a number of linguistic disciplines has stressed the importance and utility of time-aligned annotation for linguistic corpora (cf. Bird and Liberman 2001)

• (It’s not very controversial to claim that time-aligned annotation methods are important.)

XML v. (Relational) Databases

• At the same time, XML has emerged as a popular technology for structuring various sorts of corpus data (used by Text Encoding Initiative (TEI), Corpus Encoding Standard (CES), etc.), more so than any other data management technology (cf. Gries 2009, McEnery et al. 2006; e.g., Simons et al. 2004).

• With few exceptions (namely, Davies 2005), relational database engines, such as MySQL or PostgreSQL, have not been discussed in the linguistic literature as useful for the storage and manipulation of linguistic corpora.

Davies 2005

• “The advantage of using relational databases for large corpora” (IJCL 10.3)
  – Davies’ paper focuses on the size-speed benefits of relational databases in particular with large (e.g., > 100 million word) corpora.
  – He outlines an approach which stores pre-compiled queries across the corpus data, such as n-grams and frequency tables
  – 4 goals:
    1. Size
    2. Levels of annotation
    3. Speed
    4. Modularity
Davies 2005, 2

- I agree, but don’t mean for this talk to simply reiterate that paper.
- I will focus on some other areas of advantage from a different perspective
  - Using smaller, spoken language datasets
  - Considering transcript data (and their linkage to their source audio) as the data of interest
    - Still though, the approach I am discussing centers on the storage of “the fundamental units of the corpus ... in sequential rows of a database” (Davies 2005: 309).

The Sociolinguistic Archive and Analysis Project (SLAAP)

- An initiative at North Carolina State University to digitize a large sociolinguistic interview collection (and increasingly other collections) for preservation and accessibility
  - We are making the collection web accessible, so (with adequate permissions) scholars can access their data from anywhere in the world

- But SLAAP is more than an archive:
  - It is web-based software that seeks to enhance linguistic data through the development of analytic tools and data-models
  - Through this, we are exploring new, computer-enhanced techniques for interacting with the collection and for conducting sociolinguistic and linguistic analyses
  - In Poplack’s (2007) terms, SLAAP is a tool with no projected end-product

Goals

- My goal today is two-part:
  1. To demonstrate and describe an approach to linguistic data management and time-aligned annotation based heavily on a database engine design
     - The Sociolinguistic Archive and Analysis Project (SLAAP; http://ncsiaap.lib.ncsu.edu/)
  2. And to wonder aloud about a possible extension to broader areas of corpus design and sharing
     - I am interested in a broader conversation about the strengths and weaknesses of relational DBs v. XML
     - I hope to remember to disclaim later that I’m not particularly married to this approach — I mostly believe it raises some interesting ideas and I’d love to hear feedback...

The (Current) Archive

- Currently (October 2009), SLAAP houses:
  - Over 1,450 interviews
  - Over 2,450 media files (> 1,125 hours of audio)
  - Not just from North Carolina...

- Transcription is ongoing (and slow):
  - Over 34 hours of time-aligned transcripts (~340,000 words)
Adding a Transcript to the Archive

“Raw” Time-Stamped Transcript
Typically Praat TextGrid-based

\[ \text{SLAAP upload webpage}^* \]

Processing...
(e.g., linked to media file, speakers associated with databased entities)

\[ \text{Resides in SLAAP’s database} \]

* For non-SLAAP users, there’s a public tool to convert TextGrids to plain text on the website

Kendall, AACL 2009

---

SLAAP itself works by...

- All web pages are generated by PHP scripts
- Acoustic analysis features work by:
  - Extracting relevant information from the database (timestamps, media file location, etc.)
  - Sending this information to Praat via customized Praat scripts
  - Reading Praat output (and optionally post-processing, via software like ImageMagick & LAME)
  - Compiling and formatting this output and sending it back to the user.

E.g., extracting pitch info:

<table>
<thead>
<tr>
<th>Start Time</th>
<th>Pitch Ref</th>
<th>Simple Ortho Rep</th>
<th>End Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>858.936</td>
<td>PEO</td>
<td>I became a commissioner</td>
<td>860.049</td>
</tr>
</tbody>
</table>

---

Storage in DB precedes XML

**MySQL database**

XML Version 1

XML Version 2

Kendall, AACL 2009

---

The value of relational databases

- This is not to say that SLAAP is possible because of its use of relational databases
- Or that it would be impossible with an XML backend

- However, the benefits of the database model seem clear in SLAAP
  - E.g., the ability to dynamically generate multiple versions of a “transcript”, both text-based and graphical is straightforwardly implemented with the DB backend. It’s harder (for me) to envision a simple way to do this in XML...

Kendall, AACL 2009
A broad question

- Can we envision public corpora stored entirely in relational databases?
  - Accessed remotely through SQL, or a web-based (e.g., PHP) interface, or scripts in R, etc...
    - Variable interfaces, a separation of form and content ("XML")

- Corpus developers could grant free and open access to their materials, without “giving” away the actual corpus documents.
  - ~ Davies’ excellent resources (http://corpus.byu.edu/)
  - ~ http://corpus.byu.edu/copyright.asp

E.g.,

Example from Santa Barbara Corpus (Du Bois et al. 2005) stored in MySQL DB

But could also access directly from remote MySQL clients

Or from R using the RMySQL library

Example from Santa Barbara Corpus (Du Bois et al. 2005) stored in MySQL DB

E.g., Can create SLAAP-like custom interfaces

Thank you

- Reactions / Thoughts?

Thanks to support from the North Carolina State University Libraries, the North Carolina Language and Life Project, and the William C. Friday Endowment at NC State University.
References
