LING 519: Technologies for Language Documentation
Syllabus and Policy Statement

Damir Cavar

dcavar1@emich.edu

September 2011

PRAY-H 313, 6:30-9:10 PM

Office hours: Monday 4-5 PM, 7:30-9 PM, Wednesday 4-6:30; and by appointment. Don't hesitate to stop by if there's anything at all that you'd like to discuss. And don't hesitate to call or email me. It is perfectly all right to call me on my mobile number if you can't reach me at school.

Course web page: http://www.cavar.me/damir/ling519-11/

Important: When you email me, please put LING 519 in the subject line, so that I recognize it easier.

Equipment needed: You will need access to a computer and a website that you can publish to. Relevant content for this class will be shared over a web page, and either Google Docs or Dropbox. The relevant accounts can be created freely. I recommend that you register such an account. Alternatively, all material will be made available and distributed on demand via email. For work during class time, you can use the computers in Cooper; but these are not available during the day. So you may want to bring a laptop, so that you can start and finish projects on the same machine. We will set up a scratch area on the LL or another server that you can use as a website if you do not have one. On your personal machine you might need an installation of:

- Java Runtime

You may need any of those interpreters and environments, if you want to actively write scripts for data manipulation:

- Python 3.x
- Racket (DrRacket, Scheme)
- Perl

I will make use of additional software packages that will be made available during the class (free of charge, free license). I will provide help and individual support during office hours for installations on your laptops or mobile computers.

Readings: Most of the texts are on the web or will be scanned and placed on the class website. A bibliography giving the complete references has been uploaded to the class website. As we go along, we will amplify this bibliography with other readings, especially readings suitable for your reports.

Requirements:

Annotation projects (Assignment 1a-b)	20%
Lexicon project (Assignment 2)	25%
A/V presentation project (Assignment 3)	20%
Class report & handout (Assignment 4)	15%
Exercises & class participation	20%

Exercises and class work: Since the best way to become acquainted with an application or standard is to try it out, I will ask you to do short exercises at irregular intervals. You might, for example, be asked to write metadata in Dublin Core format, write an RDF fragment, etc. For each exercise, 2 people will be asked to xerox enough copies of their homework to give out to everyone, so that we can discuss it in class.

Conferences: I hope you will feel free to confer with me at any time. But if you are having difficulty with the reading or coursework, you should see me as soon as possible.

Assignments

Assignment 1 (Annotation projects): You will get a raw text for conversion and annotation in a TEI compatible XML format. Additionally, I will provide a short video or audio file containing language data. You may use your own file or one that Veronica Grondona has provided for us. Transcribe the file if necessary. (Veronica's are already transcribed, but if you want to transcribe a new file, you can read something on transcription, do the transcribing, and tell us about this experience as Assignment 4.) Put the text into interlinear glossed text format, and align it with the audio or video using Elan, TasX, or Transcriber. You may work in pairs to do these assignments. (Warning: both members of the pair will receive the same grade on it, no matter who does most of the work.)

Assignment 2 (Lexicon project): Using unstructured word lists and dictionaries, you will format a well formed dictionary in XML in any of the given and discussed standards.

Assignment 3 (A/V presentation format): Create a webpage displaying language resources in multimedia format, using SMIL or another open standard. (This assignment is tentative and may change, depending on time constraints and the interests of the class.)

Assignment 4 (software demo or report): Each class period we will have either a report or a software demo. You may choose from these two options:
Teach the class how to use a piece of software useful in data creation, management, or display, e.g., Toolbox, Lexus, Transcriber, Elan, TasX, FIELD

Make a 10-minute report on an article relevant to language endangerment and preservation, or the creation and mobilization of digital language documentation.
In either case, you should prepare a handout for the class. For the report, the handout should give the thesis of the article and examples of the data or supporting details. For the software demo, the handout should describe the functionalities of the software and provide information on access. See examples of software demo handouts on the E-MELD site: http://www.emeld.org/workshop/2003/proceeding03.html#demo
If you choose to demo software, you may work with a partner. If you work with a partner, your handout should explain how to get started with the software and, if possible, you should put together a brief exercise that the class can do using the software, e.g., provide a list of 5 words to enter into LEXUS and some instructions on how to edit them. We will then allow class time for everyone to try out the software.

Outcomes

- Knowledge of data encoding technologies and annotation standards
- Practical experience with XML and related technologies
- Practical knowledge of annotation tools for different types of language data and annotations at different linguistic levels
- Basic knowledge of scripting and language and linguistic data preparation and manipulation using Python, Perl or Scheme

Course plan

The following topics will be covered in the sequence as given over 14 sessions.

Creating and managing language data

1. General introduction and course planing
2. Digital formats and meta information
3. Corpus annotation examples: part-of-speech annotation, treebanks
4. XML
5. XML and TEI
6. Lexicon encoding standards
7. Semantic concepts and relations
8. Conversions of data using XSLT and scripts
9. Using data collections in databases and corpus tools
10. Annotation tools for multimodal, audio and video language data

Extra option:
Every session of class will cover approx. 30 minutes of programming scripts in Python (or alternatively another language, even on an individual level). The scripts will be related to the content of the class and topic.
A special crash course can be organized for interested students outside of class sessions.