Treebanking and the Theoretical Linguist
The 2011 Linguistic Summer Institute
University of Colorado at Boulder
July 7 – Aug 2, 2011

Course No: LING7800-076
Class time: Tues and Fri: 10:30am-12:15pm
Location: CHEM 133

Instructors:

- Rajesh Bhatt, UMass Amherst, bhatt@linguist.umass.edu
- Fei Xia, Univ. of Washington, fxia@uw.edu

Office hours: by appointment

Description:
Recent work on creating treebanks (i.e., annotated corpora with syntactic structure) is one of the main contributors to rapid progress in the field of computational linguistics in the past two decades. The course has two goals: the first is to introduce our audience to the particular challenges involved in treebanking, with the hope that some will get involved in the creation of annotated corpora for their languages. The second is to expose the audience to the annotated corpora that already exists and provide them with the tools to use these resources.

This course has three main components: (1) an overview of treebanks and their usages in computational linguistics and theoretical linguistics, (2) main issues in creating Multi-representational and multi-layered treebanks, and (3) syntactic and semantic phenomena and the factors that should be taken into consideration in treebank design.

Course objectives:

- Introduce students to the design of treebanks
- Demonstrate how treebanks can inform research in theoretical linguistics
- Show that treebanks play an important role in CL

"Verbs" machines:

- ssh your_id@verbs.colorado.edu
- course directory: /home/verbs/shared/LSA7800_076/
- Basic unix/linux commands (e.g., ls, cd, ln, sort, head): tutorials on unix

Grading:

- Attendance and class participation (70%)
- Assignments (30%): due at noon on Thursdays, submitted via email.

Course plan:

Session	Date	Topic	Reading	Homework	Due date
1	7/08(F)	Course overview Introduction to Treebanks Introduction to CL	PTB guidelines (ch 1-2, etc.)	Hw1	7/14 (R)
2	7/12(T)	Using treebanks in linguistic studies	Tgrep2 user manual	Optional hw: play with tgrep2	
3	7/15(F)	Creating a Treebank	(Xue et al., 2005) (Xia, 2000a) (Xia, 2000b) (Xue and Xia, 2000)	Hw2	7/21 (R)
4	7/19(T)	PropBanks (by Ashwini Vaidya)	(Palmer et al., 2005) (Bonial et al., 2011)		
5	7/22(F)	Finish Lecture 3 Historical treebanks	 (Santorini, 2010) (Santorini and Wallenberg)	Hw3	7/28 (R)
6	7/26(T)	Finish historical treebanks Treebanking a Blackfoot corpus (by Joel Dunham)			
7	7/29(F)	Hindi-Urdu Treebank DS-to-PS conversion	(Palmer et al., 2009)	Reminder	
8	8/02(T)	Discourse Treebanks (by Nianwen Xue)	(Prasad et al., 2007)		

Reading:

- Bies et al., 1995. Bracketing Guidelines for Treebank II style Penn Treebank Project. http://faculty.washington.edu/fxia/lsa2011/readings/ptb2-guidelines.pdf

- Douglas Rohde, 2005. Tgrep2 user manual. http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf

- Nianwen Xue, Fei Xia, Fu-dong Chiou, and Martha Palmer, 2005. "The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus", Journal of Natural Language Engineering, 11(2): 207-238, 2005. Cambridge University Press. [pdf]

- Fei Xia, 2000. "The Segmentation Guidelines for the Penn Chinese Treebank (3.0)", IRCS Report 00-06, University of Pennsylvania, Oct 2000. [pdf]

- Fei Xia, 2000. "The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0)", IRCS Report 00-07, University of Pennsylvania, Oct 2000. [pdf]

- Nianwen Xue and Fei Xia, 2000. "The Bracketing Guidelines for the Penn Chinese Treebank (3.0)", IRCS Report 00-08, University of Pennsylvania, Oct 2000. [pdf]

- Martha Palmer, Daniel Gildea, and Paul Kingsbury, 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1). http://acl.ldc.upenn.edu/J/J05/J05-1004.pdf

- Claire Bonial, Olga Babko-Malaya, Jinho Choi, Jena Hwang, and Martha Palmer, 2011. Propbank Annotation Guidelines (version 3.0). http://faculty.washington.edu/fxia/lsa2011/readings/pb_guideline_v3.0.pdf

- Penn Parsed Corpora of Historical English. http://www.ling.upenn.edu/hist-corpora/

- Beatrice Santorini, 2010. Annotated manual for the Penn Historical Corpora and the PCEEC (Release 2, Jan 2010). http://www.ling.upenn.edu/hist-corpora/annotation/index.html

- Beatrice Santorini and Joel C. Wallenberg. “WorkBook: Using Corpora for Linguistic Research”. DIGS 13 Workshop. http://faculty.washington.edu/fxia/lsa2011/readings/digsworkbook.pdf

- Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia, 2009. "Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure", Proceedings of the 7th International Conference on Natural Language

Processing (ICON-2009), pages 259-268, Hyderabad, India, Dec 14-17, 2009.
http://ltrc.iiit.ac.in/icon_archives/ICON2009/Papers/pdf/28.pdf

- Prasad et al., 2007. The Penn Discourse Treebank 2.0 Annotation Manual.
http://faculty.washington.edu/fxia/lsa2011/readings/pdtb-annotation-manual.pdf