English Translation Treebanks at LDC: Update on Current Efforts

Ann Bies
Linguistic Data Consortium
bies@ldc.upenn.edu

EATB Phase 3 Accomplishments

- Site-requested hyphenation revisions
  - Revised hyphenation, tokenization guidelines
  - 1.2Mw EATB, ECTB targeted for revision
    - 997Kw completed to date
- New annotation also underway
  - Broadcast news (EATB)
    - 120Kw released; 30Kw in progress
  - Web text (EATB)
    - New guidelines and software in place
    - 52Kw complete; 15Kw in progress

Technical Improvements

- Developed technical infrastructure to address hyphenation tokenization changes in AG annotation files
- On-going development of technical infrastructure to convert legacy corpora to AG annotation format
- Continued refinement of error detection and search processes

EATB Hyphenation Revision

- Guidelines for new hyphen and tokenization policy developed per BAC with Ann Taylor in fall 2008
- Hyphen and tokenization corrections underway: 997 Kw complete
- Complete hyphenation revisions by end February 2009, pending issues with legacy corpora

ETTB Treebank-PropBank Merge Changes

- Complete revision of Treebank-PropBank merge changes, as determined by Banks Advisory Committee
  - Update all existing corpora to current guidelines
- EATB
  - BN done (parallel to ATB5)
  - AFP, Annahar, webtext to follow
- ECTB prioritized by request for OntoNotes
  - Sinorama in progress now
  - Xinhua to follow
    - Resolve technical issues with legacy data
- 1.2 Mw to be completed by May 2009

EATB Phase 4 New Data

- Release new pilot BN and WB corpora in early 2009
- Continue BN, BC, WB annotation
  - 200Kw targeted
  - Final makeup under negotiation with BAC