Corpus Development & Use cycle

- **Data Acquisition**
  - Collecting existing material
  - Recording experiences (that are happening independently)
  - Designing experiences (especially for data-collection)
- **Data Representation**
  - Transcripts, event logs, visualizations, indexed recordings
- **Data Annotation**
  - Dialogue specific: Turn-taking, initiative, dialogue acts, participant ID, disfluency & repair, …
  - Shared with text-processing: POS, parse, semantics, reference, discourse coherence,…
- **Data Analysis**
  - Statistics, rules, principles, language/dialogue models
- **Data Use**
  - Stats, rules, phrases, etc used in systems

Data Acquisition: Existing Material

- **Issues**
  - How to find it? (access, permissions, privacy etc)
  - How good is it?
  - Can we understand it?
    - what are the contextual features underlying interaction
    - Is all (relevant) interaction captured
- **Tools**
  - Forms, procedures for description
  - Web-based search/processing?
- **Product: Raw Dialogue Data**

Data Acquisition: Recording

- **Issues**
  - Does recording influence interaction experience?
  - Nature of participants: humans, computers, mixed, how many?
  - Nature of modality: face to face, speech, computer graphical interaction
- **Tools**
  - Capture: Cameras, microphones, screen capture, event capture, logging tools, …
  - Synchronization
- **Product: Raw Dialogue Data**

Data Representation

- **Issues**
  - Transcription: non-standard orthography
  - Representation of non-verbal aspects
  - Audio/video browsing
- **Tools**
  - ASR transcribers
  - Human transcription aids (e.g., Praat, Transcriber)
  - Gesture coders
  - browsers
- **Product:**
  - More accessible/analyzable data

Data Annotation

- **Issues**
  - What to annotate? Who? How many?
  - Agreement on annotation schemes?
  - Inter-coder reliability, genuinely ambiguous/vague phenomena
- **Tools**
  - Automatic taggers
  - Annotation schemes
  - Human annotation aids (e.g., MMAX, DAT, ANVIL,…)
  - Hybrid taggers
  - Evaluation of annotation schemes and tagging efforts (Kappa,…)
- **Product**
  - Useful data, ready for analysis of dialogue
  - Gold standards

Data Analysis

- **Issues**
  - How big a corpus?
  - How representative a corpus?
- **Tools**
  - Learning packages (TBL taggers, Ripper, EM, Bayes,…
- **Product**
  - Theories
  - Taggers
  - Language/dialogue models

Data Use

- **Issues**
  - How are various models of aspects of dialogue combined
  - What models are relevant to performance
  - How to combine multiple models
  - How to build multi-corpus (genre) models?
- **Tools**
  - Dialogue managers
  - Theory testers
- **Product**
  - Better dialogue systems
  - Better understanding of (aspects of) dialogue