Best practice in empirically-based dialogue research

Laurent Romary, Michael Strube, David Traum

Laurent.Romary@loria.fr

Michael.Strube@eml.villa-bosch.de

traum@ict.usc.edu

Outline (main presenter)

1. **Introduction (Traum)**
2. **The Museum of Annotation (Strube)**
3. **Corpus development and use life-cycle (Traum)**
4. **Developing an Annotation Scheme (Traum)**
5. **Representation, Data Format, Standards (Romary)**
6. **Annotation Tools (Strube)**
7. **Using annotated Data (Strube)**
8. **Conclusions (All)**

Empirical Topics in Semantics/Pragmatics of Dialogue

- Dialogue acts
- Semantic Representations
- Discourse structure
- Dependent reference
- Grounding
- Initiative
- Turn-taking
- Coherence relations
- Entrainment
- …

Data collection vs manufactured examples

- **Made-up examples can be more succinctly illustrative of particular phenomena, but,..**
  - Doesn't show whether phenomena really appear in natural dialogue
  - Doesn't show frequency of phenomena
    - Is it so infrequent/unimportant that it can be safely ignored
    - How common is it
    - Does it conform to pre-conceived normative rules
- **Corpus data can help discover**
  - What features might be correlated or explain presence/absence
  - Explanatory models of data
  - Learned taggers

Question-oriented data analysis

- **Hypothesis of features of discourse**
  - E.g, are pronoun antecedents usually in initiative part of initiative-response pairs?
- **Is there data that can support/disprove the hypothesis?**
  - Corpus of the appropriate kind of dialogue
  - Annotated for pronouns, antecedents, and initiative-response units
  - If so, count co-occurrence, measure significance
- **Annotate existing data**
  - Annotation algorithm or annotation manual
  - Verification of validity (Carletta 1996)
  - Separate annotation of distinct phenomena (theory neutral)
- **Collect new data**
  - Design scenario, features
  - recording

Corpus Desiderata

- **corpus should be extensible**
- **annotations should follow guidelines which others can understand (and you the designer understand two years later)**
- **you should be able to put new annotations on top of the old ones**
- **you should focus on a particular research question but your annotations should not be restricted to that**

Does corpus have "all" the relevant data?

- **Is communication condition carefully described?**
  - Face to face, telephone/radio, computer mediated,…
    - Could they see each other? (gaze, deictics)
    - Could they see things in common?
  - what was in visual environment?
  - What instructions were given (if any)?
  - What went on before the recording?
  - Did participants know each other?
  - Demographics of participants
- **If spoken corpus, is speech signal available (including prosody, timings, etc) or just transcripts?**
  - How accurate are transcripts (non-standard pronunciations, timings, etc)
- **Is visual communication recorded?**
  - Video recording
  - Transcripts of action

How to interpret results

- **Are they significant?**
- **Do they generalize**
  - Beyond this dialogue
  - Beyond speakers
  - Beyond this task
  - Beyond this genre
- **Need large enough, diverse enough corpus to justify claims**