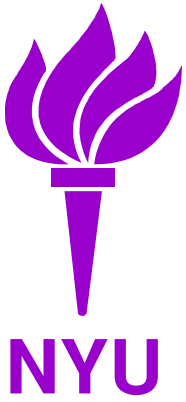

SSWL Wiki: Some comments from a computational linguist

Ralph Grishman

November 2007

Our Perspective

- Production of annotated corpora and dictionaries
  - Adam Meyers will discuss further tomorrow
- Participation in design of corpora annotated by LDC [Linguistic Data Consortium]

- Objectives:
  - Improved language analysis performance on major languages
  - Rapid ramp-up of basic capabilities for other languages
    - Use of 'language packs'

- As much data as we can afford to annotate
  - Typically 10^5 to 10^6 words

Need for Quality Control

- Very hard to learn from noisy data
  - Especially for automatic learning, but true also for manual data analysis
- Detailed criteria for annotation & property assignment
  - Tested and revised several times
  - Lots of examples
  - Nothing is obvious
- Annotators are tested on training corpora

Even with all this, still hard to get consistent annotation

Wikipedia Model

- Distinguish tools (wiki) from social model (Wikipedia)

- Wikipedia entries are meant to be read, not processed as a data base (DB), so we forgive a lot of inconsistencies
  - Using Wikipedia as a DB requires a lot of ‘clean up’
  - Less uniformity than commercial encyclopedia
  - If we want to build a DB on a Wikipedia model, need
    - Strong templates (format constraints)
    - Substantial initial model
    - Editors

How to maintain DB quality?

- Focus on SSWL as an *example base*
  - Each property assignment must be supported by multiple examples
  - Language snapshots to aid understanding of examples and uniform mark-up and glossing
    - Gloss dictionary assembled from examples
    - Morphology rules
    - Rendering of standard set of short sentences

How to maintain DB quality?

- Encourage detailed criteria for properties
  - Include several examples of hard cases
  - Questionnaire (tax return preparation) model is attractive, but ambitious … can be hard to develop and maintain

Need for Incremental Prototypes

- Start small, build in several stages
  - Initially use a small group of volunteers who can discuss design every few weeks
    - Regular discussion essential
    - Multiple coders for same features to test adequacy of specifications
  - Too hard to restructure once a lot of data is assembled
  - Focus on content first (before interface)
  - Reveal unexpected problems early
  - Success provides a good grant proposal base

Even the Smallest Prototype is not So Small …

- C.C.: 30 languages * 100 properties = 3000 values
- Need 5 examples to attest typical property: 15,000 glossed and translated examples
  - Maybe 20,000 including some dual-entry
  - Maybe 200,000 words of toy sentences (or 500,000 words of real sentences)