How could a Typology Database Fit in with Linguistic Resources Created for Computational Linguistics?

Adam Meyers, New York University

November 10, 2007

Outline

- **Where to get Resources and Resource Information?**
- **What are Computational Lexicons?**
- **What are Linguistically Annotated Corpora?**
- **How might a typology database interact with existing computational linguistic resources?**
- **How might a typology database be constructed using similar methodology?**
- **Concluding Remarks**

Resource Information

- Sources for Lexicons and Corpora
  - Linguistic Data Consortium (LDC) at UPenn
  - European Language Resources Association (ELRA)
- Workshops, Conferences, Special Interest Groups
  - Language Resource and Evaluation Conference (LREC)
  - The Linguistic Annotation Workshop (the LAW)
  - Association for Computational Linguistics SIGS
    - SIGLEX and SIGANN

Computational Lexicons

- Part of Speech and Morphological Lexicons
  - 10-50 languages (LDC, ELRA, etc.)
- Syntactic & Semantic Lexicons
  - Toy Lexicons for many languages
  - Best resources available for English, Japanese, Italian, German, Czech and a few others.
- Our Group at NYU:
  - Comlex Syntax (40,000 lemmas)
    - POS for nouns, adjectives, adverbs, verbs & various closed class words
    - Subcategorization for nouns, adjectives, verbs
    - Modification structures for adverbs (Identifying Modifiees)
    - Other features (including word order) for nouns, adjectives, verbs, adverbs
  - NOMLEX (1000 hand-coded lemmas + 6800 semi-automatic)
    - Argument Structure (including mappings to verb and adjective arguments)

Linguistically Annotated Corpora

- Tagging instances in "real" text with features representing some linguistic phenomena
- Text = newspaper, magazine, blogs, fiction, spoken...
- Types of annotation: semantic classes of proper nouns, numerical expressions, time expressions, part of speech (morphology), syntax (hand-coded syntactic trees), argument structure, senses disambiguation, etc.
- Languages
  - Basic – 10-50 languages
  - High Level – English, German, Chinese, Japanese, Italian, Czech
- NYU – NomBank – Noun Arguments on WSJ Corpus

Quality of Linguistic Resources

- Quality of Specifications & Feasibility of Task
  - Measured by consistency
- Kappa measures consistency between annotators

$$\text{Kappa} = \frac{\text{Agreement - Expected Agreement}}{1 - \text{Expected Agreement}}$$

- F-Score measures accuracy based on a gold standard
  - F-Score = harmonic mean of precision and recall

$$\text{Recall} = \frac{Correct\_Answers}{Possible\_Correct} \qquad \text{Precision} = \frac{Correct\_Answers}{Total\_Answers}$$

Assumptions of The Computational Linguistics Resource Community

- Many Different Theoretical Frameworks
  - Each Set of Specifications Defines a new Framework
  - Influenced by:
    - Local Preferences and Fads
    - Requirements of phenomena being analyzed
  - Dependency trees
    - Internal and leaf nodes labeled with words
    - Head concept is forced, even for conjunctions, names, etc.
  - Phrase Structure Grammars
    - No Head Labeling
    - Layering of Constituents, e.g., N-bar, VP, etc.
- Descriptive Adequacy is the Priority
- Approaches that apply to multiple languages
  - Primarily statistical

Levels of Generalization

- Corpus Annotation
  - Applies to specific instances
  - Generalizations about words/phrases built up statistically from annotated corpora
- Lexicons
  - Applies to words in general
  - In practice, less accurate than corpus-based statistics
  - Used for backing off to generalizations across words
  - Used for Out of Vocabulary items (words not found in training corpus)

Could A Typology Database Provide the Next Level of Generalization?

- More to Less General:
  - Annotation →Lexicon →Typology
- Word Order Examples
  - Adjective Word Order
    - English is mostly adjective first
      - *president elect, time immemorial, eggplant parmigiana*
    - Spanish is more free, but mostly adjective last
      - *buen cuidado prenatal*
        good care prenatal
        "good prenatal care"
  - The language settings adjective-first (English) and adjective-last (Spanish) could be used as defaults
- Quick Ramping Up Data for New Language?

How Consistent are the Specifications?

- Would a sampling of linguists set the features the same way for a few sample well-known languages?
- What would the F-score and/or Kappa be?
- Refinement of Specifications should improve scores.
- Corpus Annotation (after rounds of refining specs)
  - For easy tasks, 93-97% F-scores are possible
  - For difficult tasks, 85% F-score is probably OK
  - NB: F-score is affected by both recall and precision

Constantly changing specifications cannot be consistent

- Evolution of the classification has positive and negative consequences.
- Will additions be vetted?
- Will language entries be updated for new features?
- Will a subset of features be fixed?
- Will there be a core of reliable features?

Which Features Characterize a Language?

- Typological features often seem to apply to specific lexical items, but most typological analyses assume that features classify languages.
- Do word order restrictions apply to languages or words?
  - Lexical word order exceptions (elect, parmigiana, buen).
- Do anaphoric restrictions apply to languages or words?
- How do lexical restrictions determine language features?
  - Statistically? E.g., most adjectives are pre-nominal in English

More Questions

- Is it possible to do typology without an extensive grammar of a language?
- Is it possible to do typology without a substantial lexicon of a language?
- Is it possible to separate syntactic, morphological and semantic topological features?

Concluding Remarks

- Computational Linguistics Resources
  - Community with related goals to this project
  - Some resources may be useful for this project
- It may be worth thinking about how a typological database fits into the world of language resources.
- A large typological database could be a lasting resource that would be useful to many research communities.