

# The Preposition Project

<http://www.clres.com/prepositions.html>

Ken Litkowski  
CL Research  
9208 Gue Road  
Damascus, Maryland USA  
[ken@clres.com](mailto:ken@clres.com)

Orin Hargraves  
Lexicographer  
5130 Band Hall Hill Road  
Westminster, Maryland USA  
[orink@carr.org](mailto:orink@carr.org)

# Project Objective

- Create publicly available databases that characterize English prepositions based on
  - Disambiguated prepositions in FrameNet sentences
  - Using a lexicographically-based sense inventory (Oxford Dictionary of English)
  - Within a semantic framework of traditional English grammar (Quirk et al., 1985)

# The Sense Inventory

- *Oxford Dictionary of English*
  - 373 prepositions (including many phrasal prepositions)
  - About 847 senses (grows as evidence indicates)
- *Each sense described by*
  - Semantic role
  - Syntactic and semantic properties of its complement and attachment point
  - Syntactic function and meaning (from Quirk)
  - Similar prepositions
  - FrameNet frames and frame elements

# FrameNet Instances

- Obtained from 7500 FrameNet lexical unit files
  - Uses publicly available CL Research FrameNet Explorer (FNE)
  - Selects all sentences in FN subcorpora with **ppprep** in name (this is not exhaustive of all preposition instances)
- Creates text file with one line for each instance
  - Frame, frame element, lexical unit, subcorpus name, and sentence identifier
  - Imported into Excel spreadsheet with a column added to record the sense number

# Lexicographer's Tasks

- Professional lexicographer being paid to perform lexicographic task of “sorting” corpus instances into sense groupings
- Using ODE sense divisions, FN instances, and Quirk analysis as a start, begins characterizing senses
  - Assigns arbitrary semantic role name (i.e., a tag)
  - Identifies Quirk syntactic function and “similar” prepositions for each sense from Quirk paragraphs and intuition
  - Uses FNE to view each sentence and sense-tags each instance
  - Begins and continues refining syntactic and semantic properties of complement and attachment point

# Tagging Progress To Date

- Instances completed for *by* (1314), *through* (842), *with* (2219), *for* (1766), and *of* (1/4 of 4759)
- Tagging rate of 1500 instances per 10 hours
- At completion of tagging for a preposition, lexicographer has produced
  - Excel spreadsheet of tagged instances
  - Excel spreadsheet of sense characterization, including characterization of subsense relation of subsenses to core senses (more specific or more general, with difference in meaning component)
  - Word document summarizing analysis, including identification of idiomatic phrases using preposition

# Initial Use of Tagging Results

- After sorting by sense assignments, a list of (Frame,FrameElement) pairs is generated for each sense
- Each instance is a seed into the FrameNet database (9309 lines for *by* and 5440 lines for *through*)
  - Used to generate a list of other prepositions with the same (Frame,FrameElement) pair
  - Used to generate a list of syntactic realizations with the same (Frame,FrameElement) pair
    - FrameNet's grammatical forms and phrase types

# Analysis of Semantic Role Names

- Senses are labeled intuitively based on lexicographer's assessment of the definition
  - Overtly not based on linguistic principles, but serve as a placeholder for later analysis
  - Usually revised as instances are tagged, generalized to cover instances, frequently a compromise between ODE and FN
- List of (Frame,FrameElement) pairs can be compared to lexicographer's names
  - Broad consistency in names, but:
  - Raises questions of granularity in both ODE sense divisions and FrameNet level of detail
  - Is Gildea & Jurafsky (2002) mapping into (18) higher level semantic roles the way to go for NLP?

# Analysis of Other Prepositions

- Lexicographer identified a couple for each sense based on intuition and Quirk
- Using tagged-instances as seeds, tens of other prepositions were often identified for a sense
  - What does this mean? Inherent polysemy of prepositions or something else going on?
- Examination of senses for other prepositions sometimes revealed oversights in lexicographer's intuition, but even using OED leaves most unexplained
- Hypothesis to be examined: very small components of meaning are at work and are closely tied with meanings of words in context (e.g., meanings of verbs and their nominalizations)

# Analysis of Alternation Patterns

- How are frame elements realized?
- The seed method yields a large database of semantic role (frame element) alternations
  - E.g., **Cure:Treatment** as seeded by *through* can be realized as the subject of a verb, in gerundials governed by a preposition, adverbially (*pharmacologically*), or anaphorically
- Semantic role alternations seem to be at a different level than, e.g., diathesis alternations
- Hypothesis to be examined: Alternations are strongly associated with lexical groups sharing meaning components

# Complement and Attachment Properties

- Lexicographer's characterizations are only starting points and need to be refined for NLP use
- When instances are generated, an accompanying XML file of sentences is generated
  - In Senseval format, suitable for use as “lexical sample” file (a task for Senseval-4?)
  - Now serving as the basis for development of preposition disambiguation routines in CL Research question-answering and summarization software
- Sentence files are suitable for application of other preposition analysis approaches (e.g., O'Hara & Wiebe, 2003)

# Available at Project Web Site

- Evolving web page with detailed descriptions of methods and data
  - Will move in directions by our own judgments and by comments and suggestions of others
- Summary tables of preposition analysis for each preposition
- All files generated in the project
  - Sense analyses, tagged instances, sentences, and alternation data

# Usefulness of Available Data

- Only 5 (of 373) prepositions and about 90 (of 847) senses, but
- Focus on core prepositions means that results are transferable to
  - Other prepositions with similar meanings
  - Other prepositions that might inherit from these core prepositions (Litkowski, 2002)
- Large amount of data generated for each preposition can serve as the basis for further study by any and all

# Future Directions

- Seeking funding to speed up the tagging and basic generation of files
  - Extend the tagging to American National Corpus instances
- Seeking collaborators with additional ideas for analysis
- Seeking comments and suggestions
  - Will include attributions and links on the web page
- Linking to the preposition literature and to work of other projects and investigators

# Summary and Conclusions

- The Preposition Project is creating databases of useful characterizations of preposition behavior
  - Based on professional lexicographic principles
  - Using an extensive database of instances developed for unrelated purposes (FrameNet)
- The Preposition Project is designed to be comprehensive in coverage and to be freely available

# Further Work

- Y'all come and play
- Come in your work clothes
- Be prepared to muck in data up to your eyeballs
- Thanks!!!!