Digraph Analysis of Dictionary Preposition Definitions

Ken Litkowski
CL Research
ken@clres.com
http://www.clres.com
http://www.clres.com/siglex02-prep/index.html
Prepositions and Their Definitions

▫ “a word governing, and usually preceding, a noun or pronoun and expressing a relation to another word or element in the clause” (New Oxford Dictionary of English, 1999 - NODE)

▫ **Objective**: Use definitions in NODE to characterize the set of relations, their meanings, and disambiguation criteria for polysemous prepositions

▫ **Types of preposition definitions**
  ▫ Substituting: definition can be substituted where preposition is used (usually a PP and another preposition (around - “on every side of”))
  ▫ Usage note: specifies the type of relation and/or syntactic or semantic information on its use or meaning

▫ Will use digraph analysis to organize the analysis of preposition definitions
Labeled Directed Graphs (Digraphs)

- A labeled directed graph consists of a set of labeled nodes (vertices, points) and directed arcs (edges, paths) between them.
- Modeling something with digraphs entails assigning an interpretation to the nodes and arcs.
- An arc should represent at least a transitive relation.
Modeling Preposition Definitions

- Dictionary digraph analysis
  - **Nodes** represent entire entries (labeled by headwords) or “concepts” (labeled by synonyms and phraseology - close to WordNet synsets)
  - **Arcs** represent defining relations (usually ISA or a simple “is used in the definition of”)

- Preposition definitions
  - Doesn’t quite fit the ISA model
  - Formally, node A “**contributes a typed meaning component with an open slot to**” to node B
    - **around** (”on every side of”) => **(part-of of around)**
What Happens in a Definition Digraph Analysis

- **Objective** is to find primitive words and meanings, develop meaning components for primitives, and propagate meaning components to non-primitives.

- **Digraph analysis** arranges nodes into an inheritance hierarchy, particularly identifying definitional cycles (equivalence classes known as “strong components”).

- **Primitives** are members of the “basis set” of the digraph:
  - Algorithm is an extension of depth-first traversal of the digraph.
  - Starts at an arbitrary node and follows arcs to leaf nodes.
  - Prunes leaf nodes with no outgoing arcs.
  - Identifies strong components, pruning them if the set has no outgoing arcs.
  - Eventually identifies basis set.

- Analysis results can then be transformed into a suitable display of a (possibly tangled) inheritance hierarchy (MIT’s graphviz).
Dictionary Preparation

- Extracted all entries from NODE MRD labeled as prepositions
- Searched all phrasal runons in NODE for definitions adhering to a preposition signature
  - a preposition
  - a prepositional phrase + a preposition
  - (leading phrase)* + present participle of a transitive verb
  - leading phrase + infinitive of a transitive verb
- 155 labeled prepositions, 218 phrasal prepositions, 847 senses
- Put all substitutable senses into sentence frames, parsed the sentences and identified the “final preposition” (used as the “hyponymic” link)
- Usage note definitions treated immediately as primitive
- Cleaned resultant dictionary by hand to ensure good links
Preposition Basis Digraph
Preposition Primitives

- Digraph analysis eliminated 309 of 373 entries, with remaining 64 grouped into 25 equivalence categories.
- Basic set consists of 8 strong components: in, of, than, as, from, as far as, including, (by reason of, because of, on account of).
- Most frequent hypernyms: of (175), to (74), than (44), by (39), from (30), for (22), as (20), in (12).
- Identified 62 non-prepositional verb primitives, suggesting some preposition relations stem from verb relations.
- Identified 155 usage note primitives, with 74 senses in 26 entries as most basic.
A Definition Cycle

- **because of**: on account of; by reason of
- **on account of**: because of
- **by reason of**: because of

- What do they mean? We just know (?)hardwired)
Interpretation of Results

- Mostly follow our intuitions and expectations
  - Consistent with Quirk et al. and UMLS semantic relations hierarchy (which is strongly verb-based)
- Indicates need for digraph analysis at the “meaning” level to tease apart the one strong component with 33 entries
- Surprising prominence of than (spatial and temporal prepositions)
What We Do with Digraph Analysis Results

- Provides an inheritance hierarchy to serve as backbone to more detailed analysis
- Identifies entries and senses worthy of initial focus
  - Build typology of preposition relations
  - Identify meaning components of primitives for inheritance
  - Develop criteria (syntactic and semantic tests) for disambiguation
- Provides a testbed: how well do our typology, meaning representation, and disambiguation criteria play out in the inheritance hierarchy
# Definitions of “of” (19)

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition (Subsense(s))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Partitive</td>
<td>relationship between a part and a whole (part functioning as head; after a number, quantifier, or partitive noun, with the word denoting the whole functioning as the head of the phrase)</td>
</tr>
<tr>
<td>2. Scale-Value</td>
<td>relationship between a scale or measure and a value (an age)</td>
</tr>
<tr>
<td>3. Genitive</td>
<td>association between two entities, typically one of belonging (relationship between an author, artist, or composer and their works collectively)</td>
</tr>
<tr>
<td>4. Direction</td>
<td>relationship between a direction and a point of reference</td>
</tr>
<tr>
<td>5. Hyponym</td>
<td>relationship between a general category and the thing being specified which belongs to such a category (governed by a noun expressing the fact that a category is vague)</td>
</tr>
<tr>
<td>6. Deverbal</td>
<td>relationship between an abstract concept having a verb-like meaning and (a noun denoting the subject of the underlying verb; the second noun denotes the object of the underlying verb; head of the phrase is a predicative adjective)</td>
</tr>
<tr>
<td>7. IndirectObject</td>
<td>relationship between a verb and an indirect object (a verb expressing a mental state; expressing a cause)</td>
</tr>
<tr>
<td>8. Substance</td>
<td>the material or substance constituting something</td>
</tr>
<tr>
<td>9. Time</td>
<td>time in relation to the following hour</td>
</tr>
</tbody>
</table>
Use of Definitions in Discourse Analysis

- **Objective:** Characterize prepositions in text processing
  - Type the use
  - Identify and characterize the arguments

- **Vehicle:** Discourse analysis of texts for use in question answering (encyclopedia, TREC) with XML output
  - Parsing text
  - Identifying discourse entities (e.g., NPs)
  - Analyzing verbs and prepositions to identify and characterize arguments

- **Preposition analysis**
  - Identify and characterize arguments (part of speech, semantic category, root form)
  - Disambiguate preposition based on these arguments: tests for literals, parts of speech, WordNet role (synonym, part-of), WordNet type, thesaurus category
Initial Discourse Analysis Results

- Propagating semrel types based on preposition inheritance hierarchy

- XSLT on XML output to view all instances of disambiguations
  - Small but growing coverage (50% on “of”, others just beginning)
  - Acts like a lexicographer’s concordance list

- **Approach reveals interesting tough semantic questions: new desiderata for lexicons?**
  - Is x a partitive noun?
  - How do you determine that the object of a preposition is the subject of the verb underlying the noun to which the prep phrase is attached?
  - How do you determine that the object of a preposition is the material or substance constituting the noun to which the prep phrase is attached?
  - What are the criteria for recognizing a belonging relationship?
Conclusions

- Digraph analysis of preposition definitions provides a viable mechanism for research into preposition meaning
  - Identifies “working” set of primitives
  - Lays out an inheritance hierarchy
  - Provides data for developing a typology, building meaning representations, and identifying disambiguation criteria

- Attempting to use resulting computational preposition lexicon identifies difficult and important semantic questions
  - Definitions of primitives provide the focus
  - Resulting lexicon provides structure for incorporating studies from other fields (e.g., AI studies of time expressions)

- OED1 has 135 senses for “of”
Other Digraph Analyses

(Continuing Investigations)

- Verbs (20,000 entries, 100,000 definitions) in Webster’s 3rd New International Dictionary (hand)
- Verbs from Amsler’s Merriam-Webster’s Pocket Dictionary
- Subordinating conjunctions across four dictionaries
- Subparagraphs from Macquarie’s Roget-style thesaurus, which is linked to definitions in the Macquarie Dictionary
- Analysis of words assigned to content analysis categories (using WordNet entries) to provide “definition” of category
- LDOCE defining vocabulary