Corpus Linguistics (L415/L615)
Case Studies using Annotated Data

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015
investigate relations between bound pronouns and their antecedents

(1)  
   a.  [i Who] offended [i his] mother?  
      b.  Tom offended his – Tom’s – mother. And Jerry also offended his – Jerry’s – mother.

   weak crossover: bound pronoun occurs structurally between its antecedent and the derivational base position of the latter; bound pronoun itself is embedded in another phrase

(2)  
   a.  *[i Who] did [j [i his] mother] offend [i t]?  
      b.  Tom’s mother offended Tom. And Jerry’s mother offended Jerry.
Culicover

- mainstream opinion: weak crossover not possible for structural reasons
- Culicover: there exist some cases
- operationalization: find counter-examples in attested language
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun pronoun_{wh} pronoun_{poss}</td>
<td>relative clause</td>
</tr>
<tr>
<td>noun , pronoun_{wh} pronoun_{poss}</td>
<td>(non-restrictive) relative clause</td>
</tr>
<tr>
<td>pronoun_{wh} verb_{aux} pronoun_{poss}</td>
<td>pronominal wh-question</td>
</tr>
<tr>
<td>which noun verb_{aux} pronoun_{poss}</td>
<td>nominal wh-question</td>
</tr>
<tr>
<td>which noun pronoun_{poss}</td>
<td>embedded question</td>
</tr>
</tbody>
</table>
noun pronoun<sub>wh</sub> pronoun<sub>poss</sub>:

(3)  a. . . will play on HBO later this year) presents a complex portrait of a man who his biographer, the Pulitzer Prize-winning political historian Samantha Power, describes in . . .

b. Have you ever thought as to why people are so vituperative about this president who his critics say doesn’t have a philosophy – you can’t call him liberal?

c. Wojciechowski has become a symbol for Duke’s revival, the gritty little guy whom his fellow students bark at in appreciation for his dogged defensive style.
noun, pronoun\textit{wh} pronoun\textit{poss}:

(4)  
\begin{enumerate}
  \item John, Jr., whom his family called John John, seldom talked with reporters about himself or his memories?
  \item The farmer, whom his parents had located to hide them, was a Christian.
  \item Gary has remarried, and Greg and his wife have had two children, whom their grandfather barely knows.
  \item It might be that a sluggish bond-servant, or an undutiful child, whom his parents had given over to the civil authority, was to be corrected at?
question: which linguistic factors influence the choice of expression in cases of dative alternation structures

(5)  a. Susan \([V \text{ gave}] \left[ \text{NP the children}^{\text{Recipient}} \right] \left[ \text{NP toys}^{\text{Theme}} \right]\)
    b. Susan \([V \text{ gave}] \left[ \text{NP toys}^{\text{Theme}} \right] \left[ PP \text{ to the children}^{\text{Recipient}} \right]\)

(6)  a. That movie gave me the creeps.
     \([= \text{causing a change of state (possession)}]\)
    b. *That movie gave the creeps to me.
     \([= \text{causing a change of place (movement to goal)}]\)

(7)  This story is designed to give the creeps to people who hate spiders, but is not true.
Bresnan

- ideal annotation?
- doable with syntactic annotation? POS tags? Text?
- they used syntactically annotated / text corpora:
  - Switchboard in Penn Treebank
  - full Switchboard (not annotated)
  - Penn Treebank
- google
- annotate data: length of the object constituents
  (operationalization: number of words), their definiteness
  (definite vs. indefinite), their animacy (animate vs. inanimate), and accessibility (accessible vs. given vs.
  new)
- use statistical tests to find out what factors influence choice
Bresnan
investigate local coherence: smooth transitions from one sentence to the next (cognitive perspective: transition is smooth if recipient can easily integrate information in second sentence into current discourse model)

two types of local coherence:
- entity-based coherence: subsequent sentences refer to the same referents
- discourse relation-based coherence: logical relations (e.g. temporal or causal relations) hold between adjacent sentences

(9) There is [an incredible pressure on [school systems] and teachers]] to raise [test scores]. [Causal So] [efforts to beat [the tests]] are also on the rise.
Use two different corpora, sharing some textual material

- Hypothesis 1: Adjacent sentences that do not share entities are related by non-elaboration (i.e., “strong”) discourse relations.

- Hypothesis 2: Adjacent sentences joined by non-elaboration (i.e., “strong”) discourse relations have lower entity coherence: such pairs are less likely to mention the same entities.

- Hypothesis 3: Almost all pairs of sentences in a coherent text either share entities or participate in non-elaboration (i.e. “strong”) discourse relations.
<table>
<thead>
<tr>
<th>ID</th>
<th>Text</th>
<th>Para</th>
<th>SentPair</th>
<th>DiscRel</th>
<th>Explicit</th>
<th>SharedEnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wsj_0126</td>
<td>1</td>
<td>s1_s2</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
</tr>
<tr>
<td>2</td>
<td>wsj_0126</td>
<td>1</td>
<td>s2_s3</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
</tr>
<tr>
<td>3</td>
<td>wsj_0126</td>
<td>1</td>
<td>s3_s4</td>
<td>core</td>
<td>implicit</td>
<td>shared entities</td>
</tr>
<tr>
<td>4</td>
<td>wsj_0126</td>
<td>2</td>
<td>s5_s6</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Louis & Nenkova

<table>
<thead>
<tr>
<th></th>
<th>Shared entities</th>
<th>No sharing</th>
</tr>
</thead>
<tbody>
<tr>
<td>core</td>
<td>1 832 (21.80 %)</td>
<td>1 580 (18.80 %)</td>
</tr>
<tr>
<td>weak</td>
<td>2 485 (29.56 %)</td>
<td>2 508 (29.84 %)</td>
</tr>
</tbody>
</table>
Web as Corpus: Using Google

- the web is the largest collection of texts available as of now
- search engines (e.g. google) can be (ab-)used as linguistic query tools
- google offers "word counts"
- google offers advanced search: exclude terms, include terms, restrict language, region
Web as Corpus: Problems

- data is not clean: wikipedia entries, commercials, blog data
- webpages include metadata: "colorless green ideas sleep furiously": 58,000 hits
- google counts are not reliable: "the": 1,160,000 hits
  "the the": 1,270,000,000 hits
- web pages may contain multilingual content
- web content may be copied ("retweets")
- no annotation available