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

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015

Culicover (2013)

- investigate relations between bound pronouns and their antecedents

(1)  a. \[Who\] offended \[his\] 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. *Who did \[his\] mother offend \[it\]?
    b. Tom’s mother offended Tom. And Jerry’s mother offended Jerry.

- mainstream opinion: weak crossover not possible for structural reasons

- Culicover: there exist some cases

- operationalization: find counter-examples in attested language

Pattern

<table>
<thead>
<tr>
<th>Comment</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>relative clause</td>
<td>noun pronoun_{wh} pronoun_{poss}</td>
</tr>
<tr>
<td>(non-restrictive) relative clause</td>
<td>noun , pronoun_{wh} pronoun_{poss}</td>
</tr>
<tr>
<td>nominal wh-question</td>
<td>pronoun_{wh} verb_{aux} pronoun_{poss}</td>
</tr>
<tr>
<td>nominal wh-question</td>
<td>which noun pronoun_{poss}</td>
</tr>
<tr>
<td>embedded question</td>
<td>which noun pronoun_{poss}</td>
</tr>
</tbody>
</table>

Culicover

noun pronoun_{wh} pronoun_{poss}:

(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.

Culicover

noun , pronoun_{wh} pronoun_{poss}:

(4)  a. John, Jr., whom his family called John John, seldom talked with reporters about himself or his memories?
    b. The farmer, whom his parents had located to hide them, was a Christian.
    c. Gary has remarried, and Greg and his wife have had two children, whom their grandfather barely knows.
    d. 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?
Bresnan et al. (2007)

- question: which linguistic factors influence the choice of expression in cases of dative alternation structures

(5) a. Susan [gave] [NP the children [Recipient]] [NP toys [Theme]]
b. Susan [gave] [NP toys [Theme]] [PP to the children [Recipient]]

(6) a. That movie gave me the creeps. (= causing a change of state (possession))
b. *That movie gave the creeps to me. (= 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

Louis & Nenkova (2010)

- 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

Louis & Nenkova (2010)


(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.

Louis & Nenkova

- 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.
Louis & Nenkova

<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 1</td>
<td>s1, s2</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>wsj_0126 1</td>
<td>s2, s3</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>wsj_0126 1</td>
<td>s3, s4</td>
<td>core</td>
<td>implicit</td>
<td>shared entities</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>wsj_0126 2</td>
<td>s5, s6</td>
<td>weak</td>
<td>explicit</td>
<td>shared entities</td>
<td></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