LSA.343
Precision Grammar Implementation for Linguistic Hypothesis Testing

Emily M. Bender
University of Washington
ebender@u.washington.edu

Dan Flickinger
CSLI Stanford
danf@csli.stanford.edu

Stephan Oepen
Universitet i Oslo & CSLI Stanford
oe@csli.stanford.edu
Linguistic Motivations for Large, Deep Grammars

• Want to know the frequency of linguistic phenomena in a corpus
  Scientific curiosity
  Planning for research and applications
Linguistic Motivations for Large, Deep Grammars

- Want to know the frequency of linguistic phenomena in a corpus
  - Scientific curiosity
  - Planning for research and applications
- Want to find examples of some phenomena in real language use
  - In isolation
  - In combination with other phenomena
Linguistic Motivations for Large, Deep Grammars

- Want to know the frequency of linguistic phenomena in a corpus
  - Scientific curiosity
  - Planning for research and applications
- Want to find examples of some phenomena in real language use
  - In isolation
  - In combination with other phenomena
- Want to evaluate an NLP system’s behavior on some phenomena
  - Coverage, accuracy, performance
Some existing annotated corpora

- Manually constructed treebanks of natural corpora
  - e.g. Penn Treebank (English): Labelled bracketing, POS tags
  - e.g. PropBank (English): Semantic dependencies
  - e.g. TIGER Treebank (German)
  - e.g. Alpino dependency bank (Dutch)

- Manually classified phenomena in linguistic test suites
  - e.g. HP/CSLI for English (1200 items)
  - e.g. TSNLP for German, French, English (5000 items each)

- Automatically annotated corpora
  - POS tagging, chunking, shallow parse trees, dependencies

- Minimal linking of structural annotations to phenomena
  - e.g. TIGERSearch, tgrep/tgrep2
Grammar-driven Analysis of a Corpus

- Parse the corpus using an existing deep grammar
- Construct a treebank by manual disambiguation
  - Record derivation tree and associated semantics for each item
    - Includes fine-grained lexical and syntactic types
- Define correlations between grammar and phenomena
- Quantify frequency of occurrence by phenomena
**LOGON portion of Redwoods Treebank**

- LOGON: Norwegian-English machine translation project
  - Collaboration among universities in Bergen, Oslo, Trondheim
  - Domain of back-country tourism in Norway
  - Grammars of Norwegian (LFG/XLE) and English (HPSG/DELPH-IN)
  - Translation via semantic transfer (MRS)

- Redwoods: Treebank of manually disambiguated analyses
  - 20,000 sentences in multiple domains (scheduling, ecommerce)
  - Candidate analyses produced by ERG (English Resource Grammar)
  - Discriminant-based tool used in full manual disambiguation

- LOGON portion of Redwoods
  - Published tourism booklets for Jotunheimen, in Norwegian
  - Three professional translations into English
  - 90,000 words of text, with average 14 words per sentence
  - 5000 of 6400 sentences correctly parsed and treebanked
The start of the hike follows the trail to Gjendesheim through the saddle between eastern and western Hestlægerhø. This is a varied and pleasant trip that runs both along the shoreline of Bygdin and up to 1700 meters between the great peaks north of the lake. The launch originally was bought for use on Tyin, and the intent was to transport it first to Bygdin and from there over the isthmus from Eidsbugarden to Tyin, but the task took too much time and cost too much. Aside from Vestfjorddalen at Rjukan, Gjende, Norway’s most beautiful mountain lake, was DNT’s principal development area in the early years. The remains of the Englishmen’s stay can be seen on the incline below the tourist lodge; the stone oven they used for baking is still there. In 1867 a log cabin was put up next to an old stone hut at Nybua, about halfway down the 17.5-mile-long Bygdin, the biggest lake in Jotunheimen. Owners: Charlotte and Eiliv Sulheim. Marked trails from Morkabu and Ingjerdbu to Vetti, Skogadalsbøen, Tyinhølmen, and Slettningsbu.
## Coverage profile for Jotenheimen Booklet 4 (English)

<table>
<thead>
<tr>
<th>Aggregate</th>
<th>total items</th>
<th>word string</th>
<th>lexical items</th>
<th>distinct analyses</th>
<th>total results</th>
<th>overall coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>φ</td>
<td>#</td>
<td>φ</td>
<td>#</td>
<td>φ</td>
</tr>
<tr>
<td>55 – 60</td>
<td>3</td>
<td>56.67</td>
<td>251.67</td>
<td>23848.00</td>
<td>1</td>
<td>33.3</td>
</tr>
<tr>
<td>50 – 54</td>
<td>2</td>
<td>50.00</td>
<td>238.00</td>
<td>0.00</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>45 – 49</td>
<td>6</td>
<td>46.50</td>
<td>242.17</td>
<td>150758.40</td>
<td>5</td>
<td>83.3</td>
</tr>
<tr>
<td>40 – 44</td>
<td>17</td>
<td>42.00</td>
<td>221.73</td>
<td>66813.30</td>
<td>10</td>
<td>58.8</td>
</tr>
<tr>
<td>35 – 39</td>
<td>26</td>
<td>36.69</td>
<td>178.56</td>
<td>61130.45</td>
<td>20</td>
<td>76.9</td>
</tr>
<tr>
<td>30 – 34</td>
<td>61</td>
<td>31.80</td>
<td>163.43</td>
<td>29185.54</td>
<td>44</td>
<td>72.1</td>
</tr>
<tr>
<td>25 – 29</td>
<td>116</td>
<td>26.86</td>
<td>128.11</td>
<td>17409.33</td>
<td>98</td>
<td>84.5</td>
</tr>
<tr>
<td>20 – 24</td>
<td>195</td>
<td>21.72</td>
<td>110.25</td>
<td>6131.30</td>
<td>164</td>
<td>84.1</td>
</tr>
<tr>
<td>15 – 19</td>
<td>271</td>
<td>17.06</td>
<td>91.20</td>
<td>1005.57</td>
<td>245</td>
<td>90.4</td>
</tr>
<tr>
<td>10 – 14</td>
<td>241</td>
<td>11.77</td>
<td>56.03</td>
<td>101.74</td>
<td>234</td>
<td>97.1</td>
</tr>
<tr>
<td>5 – 9</td>
<td>295</td>
<td>6.71</td>
<td>31.63</td>
<td>19.06</td>
<td>285</td>
<td>96.6</td>
</tr>
<tr>
<td>0 – 4</td>
<td>366</td>
<td>2.16</td>
<td>6.02</td>
<td>2.60</td>
<td>356</td>
<td>97.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1599</strong></td>
<td><strong>13.60</strong></td>
<td><strong>67.05</strong></td>
<td><strong>4747.42</strong></td>
<td><strong>1462</strong></td>
<td><strong>91.4</strong></td>
</tr>
</tbody>
</table>
The start of the hike follows the trail to Gjendesheim through the saddle between eastern and western Hestlægerhø. This is a varied and pleasant trip that runs both along the shoreline of Bygdin and up to 1700 meters between the great peaks north of the lake. The launch originally was bought for use on Tyin, and the intent was to transport it first to Bygdin and from there over the isthmus from Eidsbugarden to Tyin, but the task took too much time and cost too much.

Aside from Vestfjorddalen at Rjukan, Gjende, Norway’s most beautiful mountain lake, was DNT’s principal development area in the early years. The remains of the Englishmen’s stay can be seen on the incline below the tourist lodge; the stone oven they used for baking is still there.

In 1867 a log cabin was put up next to an old stone hut at Nybua, about halfway down the 17.5-mile-long Bygdin, the biggest lake in Jotunheimen. Owners: Charlotte and Eiliv Sulheim.

Marked trails from Morkabu and Ingjerdbu to Vetti, Skogadalsbøen, Tyinholmen, and Slettningsbu.
the stone oven they used for baking is still there.
Derivation tree: the stone oven they used for baking is still there.

(SUBJH
(HSPEC
 (THE_1 ("the" 17 18))
(NADJ_RC
 (NOUN_N_CMPND
 (STONE_N2 ("stone" 18 19))
 (SING_NOUN_IRULE (OVEN_N1 ("oven" 19 20))))))

=> (FIN_NON_WH_REL
 (SUBJH_NMC
 (NOPTCOMP (THEY ("they" 20 21)))
 (HADJ_I_UNS
 (EXTRACOMP
 (PAST_VERB_ORULE (USE_V1 ("used" 21 22))))
 (HCOMP (FOR ("for" 22 23))
 (BARE_VGER (VGERING (PRP_VERB_ORULE (BAKE_V1 ("baking" 23 24)))))))))

(HCOMP
 (HCOMP
 (ADVADD (BE_C ("is" 24 25)))
 (STILL_ADV1 ("still" 25 26))))
 (NPADV (PUNCT_PERIOD_ORULE (THERE_NOM ("there." 26 27))))
Semantics: the stone oven they used for baking is still there.
Mapping from corpus annotations to phenomena

- Syntactic rule ==> phenomenon
  e.g. measure-NPs (17.5-mile-long)
  e.g. appositives (the Englishman Charles Bamford)
- Lexical type ==> phenomenon
  e.g. subordination (... because the route is easily seen)
- Set of related rules ==> phenomenon
  e.g. relative clauses
  e.g. unbounded dependencies
- Interactions among rules (and lexical types)
  e.g. ’across-the-board’ extraction in coordination

This tour, we will begin and end in Bergen.
# Phenomenon Frequency in LOGON Corpus (English)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th># Items</th>
<th>% Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure NPs</td>
<td>279</td>
<td>4.4</td>
</tr>
<tr>
<td>Appositives</td>
<td>275</td>
<td>4.4</td>
</tr>
<tr>
<td>NP Fragments</td>
<td>1588</td>
<td>24.8</td>
</tr>
<tr>
<td>NP Coordination</td>
<td>987</td>
<td>15.4</td>
</tr>
<tr>
<td>Multi-NP Coord</td>
<td>265</td>
<td>4.1</td>
</tr>
<tr>
<td>VP Coordination</td>
<td>411</td>
<td>6.4</td>
</tr>
<tr>
<td>S Coordination</td>
<td>588</td>
<td>9.1</td>
</tr>
<tr>
<td>Relative Clauses</td>
<td>486</td>
<td>7.6</td>
</tr>
<tr>
<td>Unbounded Deps</td>
<td>1168</td>
<td>18.2</td>
</tr>
<tr>
<td>Yes-No Questions</td>
<td>7</td>
<td>0.1</td>
</tr>
<tr>
<td>WH Questions</td>
<td>42</td>
<td>0.6</td>
</tr>
<tr>
<td>Imperatives</td>
<td>219</td>
<td>3.4</td>
</tr>
<tr>
<td>Free relatives</td>
<td>101</td>
<td>1.6</td>
</tr>
<tr>
<td>Passives</td>
<td>1072</td>
<td>16.7</td>
</tr>
</tbody>
</table>
Illuminated linguistic phenomena

- Fragments
- Coordination
- Punctuation
- 'Peripheral' constructions
  - Measure-NP phrases
  - NP-adverbs
  - Appositives
  - Degree phrases
  - N-ed phrases
  - Depictives
  - Free relatives