Annotated Corpora in Linguistic Research

Steven Bird & Stephanie Strassel

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104
www.ldc.upenn.edu
The Linguistic Data Consortium

- a not-for-profit organization
- serving:
  - researchers, educators, technology developers
- in language-related fields by:
  - creating/collecting, archiving, distributing language resources
Continuing Need

Research, development and teaching

- vast amounts of data, many languages
- ‘ecologically valid’ data
- range of specialized skills
  - processing of newswire, closed caption video, VOA, ...
  - intellectual property rights
  - data transformation and publication: CD, Web
- collaborative research
- replication of results

Annotated Corpora in Linguistic Research
LDC Principles

Publish the data that researchers need
- sponsored programs (TDT, Hub-4, OLEADA,...)
- community initiatives (ACL/DCI, Unipen...)
- non-LDC projects (CSAE, CELEX, Trains...)
- LDC-funded (Treebank, COMLEX, WordNet...)

Make data available to everyone
- consortium membership is open to all
- most databases available to non-members

Promote the idea of shared resources
- IPR intermediary
- advice on collection, publication and IPR issues
- develop standards and tools
140 Publications + 12 in press

- 93 speech corpora + 7 in press
- 33 text corpora + 5 in press
- 14 lexicons
- 250 Gb

Distribution

- 267 members
- 520 nonmembers
<table>
<thead>
<tr>
<th>Language</th>
<th>Speech / Transcripts</th>
<th>Parallel Text</th>
<th>News/Corpus</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (Egyptian)</td>
<td>Broadcast</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>Telephone</td>
<td>WarmBand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandarin</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serbo-Croatian</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwanese</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thai</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>Telephone</td>
<td>Parallel Text</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

▶ Afrikaans, Bamileke, Basque, Estonian, Hungarian, Italian, Kazakh, Kurdish, Latvian, Manding, Polish, Slovene, Ukrainian, Uzbek, Xhosa, Yoruba
Annotated Corpora in Linguistic Research

Cameroon
Switchboard Corpus Indexing

Define the corpus you want to search

- Sex: Female
- Education level: 5
- Year of birth: 1940-50

- State: Western
- NIST: 1000
- Call time: 10:11

Search results sorted by

- Time margin (seconds): 10
- Concordance width: 30
You can check the home user to the works review several works at the same time. By selecting the file, you can get the whole contents. If you find a wrong speech segment, labeling, Report, etc, will automatically report the problem. In case of query, just go through the corresponding part.
Annotated Corpora in Linguistic Research
Switchboard Tagged Corpus Search: Concordance

Text concordances for N P Y1 like:

Search result for corpus subw

article+ Speaker 1: I do not remember, but it was like it was, but it was
article+ Speaker 2: open to turn that it was like: every year my email:
article+ Speaker 1: it was like, and I think... I think, in 2001, they sent
article+ Speaker 2: kind of like, and they sent like, it was like kind
different to open and do. They were this kind of their like.
article+ Speaker 1: they were doing it and doing it...
Switchboard Tagged Corpus
Search: Results

Annotated Corpora in Linguistic Research
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>Race</th>
<th>Education</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Jane</td>
<td>25</td>
<td>Female</td>
<td>White</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>002</td>
<td>John</td>
<td>30</td>
<td>Male</td>
<td>Black</td>
<td>High School</td>
<td>Worker</td>
</tr>
<tr>
<td>003</td>
<td>Emily</td>
<td>22</td>
<td>Female</td>
<td>Asian</td>
<td>PhD</td>
<td>Researcher</td>
</tr>
<tr>
<td>004</td>
<td>Adam</td>
<td>28</td>
<td>Male</td>
<td>African</td>
<td>High School</td>
<td>Salesman</td>
</tr>
<tr>
<td>005</td>
<td>Sarah</td>
<td>35</td>
<td>Female</td>
<td>Caucasian</td>
<td>College</td>
<td>Teacher</td>
</tr>
</tbody>
</table>
Regional Variation in Flapping

- Southern
- Other
- Southern
- Other
- Midland & Western
- Other

(From Smood 1996)