Annotated Corpora in Linguistic Research

Steven Bird & Stephanie Strassel

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

Annotated Corpora in Linguistic Research
The Linguistic Data Consortium

- a not-for-profit organization
- serving:
  - researchers, educators, technology developers
- in language-related fields by:
  - creating/collection, 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
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
Distribution

# Sales

1 (12)
2 (9)
3 (3)
4 (1)
5 (2)
6 (1)
7 (1)
8 (2)
9 (1)
13 (1)
14 (1)
15 (1)
18 (1)
21 (1)
24 (1)
25 (1)
48 (1)
225 (1)
<table>
<thead>
<tr>
<th>Language</th>
<th>Broadcast</th>
<th>Telephone</th>
<th>Web/Email</th>
<th>Parallel Text</th>
<th>News/Dir/Other Text</th>
<th>Lexicon</th>
<th>Traditional Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (Egyptian)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandarin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serbo-Croatian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telugu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> Afrikaans, Bamileke, Basque, Estonian, Hungarian, Italian, Kazakh, Kurdish, Latvian, Manding, Polish, Slovene, Ukrainian, Uzbek, Xhosa, Yoruba
Switchboard Corpus Indexing

Search criteria you want to search:

Sex: Male Female
Education level: Low Medium High

Location:
- All
- MNE
- NEW ENGLAND
- SOUTHERN
- SWELLERS
- SMITH MND
- DAK

Terminology within query:

Time margin (seconds): 0 4 Concordance width: 20

Annotated Corpora in Linguistic Research
Annotated Corpora in Linguistic Research
Speaker A: [somewhere on a computer screen]

Speaker A: [typing on the keyboard]

Speaker B: I think now they're a lot more expensive than that. How much money have you spent so far?

Speaker A: It's a [insert blinking cursor]

Speaker A: I have [insert typing on keyboard]... [typing on keyboard]

Speaker B: And... does the computer let you print your own labels?

Speaker A: Yes.

Speaker B: And... does the computer let you print your own labels?
<table>
<thead>
<tr>
<th>Type</th>
<th>Gender</th>
<th>Age</th>
<th>Ethnicity</th>
<th>Region</th>
<th>Education</th>
<th>Income</th>
<th>Occupation</th>
<th>Marital Status</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>25</td>
<td>African</td>
<td>West</td>
<td>High</td>
<td>Medium</td>
<td>Sales</td>
<td>Single</td>
<td>1234567890</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>30</td>
<td>Caucasian</td>
<td>East</td>
<td>Low</td>
<td>Low</td>
<td>Teacher</td>
<td>Married</td>
<td>0987654321</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>40</td>
<td>American</td>
<td>Midwest</td>
<td>None</td>
<td>None</td>
<td>Engineer</td>
<td>Divorced</td>
<td>1122334455</td>
</tr>
</tbody>
</table>

Annotated Corpora in Linguistic Research
Regional Variation in Flapping

(From Sum and 1960)

Annotated Corpora in Linguistic Research