Resources for Arabic Natural Language Processing

Mohamed Maamouri, Christopher Cieri
{maamouri,ccieri}@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and
Department of Linguistics
www.ldc.upenn.edu
Background

- Language resources necessary component to language development
- Language resources expensive to create
  - require special skills/staff, specialized equipment
- Organizations that create language resources may not distribute
  - no interest, no infrastructure, reduce competitive advantage
- Problem: Lack of adequate supply of resources stands as an impediment to language development.
- Solution: Build non-profit language resource center to promote language development through the sharing of resources
  - Acquire specialized equipment, develop specialized staff
  - Build relationships with corpus authors, other data providers, and research communities
  - Maintain permanent data archives with bug reports, re-releases, on-going rights to data
  - Provide standard reference data to evaluate competing algorithms/analyses
LDC Roles

- Founded April 15, 1992 as a non-profit activity of the University of Pennsylvania
- Specialized publisher (>15,000 copies of 209 publications)
  - resources for linguistic education research and technology development
  - activities supported primarily through membership fees
- Open consortium: any organization interested in language resources may join (almost 1400 users)
- Intellectual property intermediary: negotiate agreements between data providers and data users
- Corpus Creator: create and annotate language resources to specification that can be widely shared
  - community initiatives, corporate and government sponsored projects, joint projects
- Research Group:
  - research approach to language resources
  - conduct research on standards and best practices
## Resources by Language

<table>
<thead>
<tr>
<th>Language (Broadcast</th>
<th>Telephone</th>
<th>WideBand</th>
<th>Parallel Text</th>
<th>Newswire/Other Text</th>
<th>Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (Egyptian)</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>
</tr>
<tr>
<td>Dutch</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>
</tr>
<tr>
<td>French</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>
</tr>
<tr>
<td>Hindi</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>
</tr>
<tr>
<td>Korean</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>
</tr>
<tr>
<td>Persian</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>
</tr>
<tr>
<td>Russian</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>
</tr>
<tr>
<td>Spanish</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>
</tr>
<tr>
<td>Thai</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>
</tr>
<tr>
<td>Vietnamese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Albanian, Arabic, Armenian, Azerbaijani, Bangla, Belorussian, Bosnian, Bulgarian, Burmese, Cantonese, Croatian, Czech, Dari, English, Estonian, Farsi, French, Georgian, German, Greek, Hausa, Hindi, Indonesian, Kazakh, Khmer, Kinyarwanda/Kirundi, Korean, Kosovian, Kurdish, Kyrgyz, Laotian, Latvian, Lithuanian, Macedonian, Mandarin, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Tajik, Tatar-Bashkir, Thai, Tibetan, Turkish, Turkmen, Ukrainian, Urdu, Uyghur, Uzbek, Vietnamese
Coordinated Resources

- Focus on major languages: English, Chinese, Arabic, Spanish
- Battery of Resources to meet major research and development needs
- Supporting: language modeling, speech recognition, translation, translingual information retrieval, natural language processing
- Resources also useful for any empirical language study including linguistic analysis, language teaching
- Gigaword News Text Corpora – 1B words, variety of news sources
- Parallel Text – pairs of documents and aligned translations
- Broadcast News – with time-aligned transcripts, important domain for its inherent interest and for its broad vocabulary
- Conversational Speech – telephone conversations and meetings, with time-aligned transcripts
- Pronunciation/Multilingual Lexicons – relate source word forms to:
  - set of target glosses, syntactic and frequency information, pronunciation, morphological analysis, optionally mediated through morphological analysis/synthesis engine
- Treebanks – text annotated to show the morpho-syntactic properties of sentences and their constituents
- Technology-Specific Evaluation Resources – MT & IR
Very Large Text Corpora

- Collecting news text since 1994
- Published Arabic Newswire, 76Mwords, in 2001
- To support robust modeling of rare phenomena need Gigaword News Text Corpora
- English, Chinese and Arabic
- Arabic: 480,000,000 words from Al Hayat, An Nahar, AFP, Xinhua, IRNA - looking for more
- Consistent encoding
- Light XML markup inline
- Other annotations should be stand-off
Broadcast News

• Goal: Database of broadcast news from around the Arabic speaking world, accurately transcribed

• Current 120 hours of Voice of America radio; 60 hours of Nile TV via SCOLA

• Topic Detection and Tracking Corpus – Phase 4 will contain Arabic broadcast news.

• 40 hours will be carefully transcribed and released jointly with ELRA under the NSF-EU funded Networking Data Centers project

• Building capacity to collect additional source locally; also interested in partnerships.
Conversational Arabic

- 1995 began collecting conversations in 18 linguistic varieties to support research in language identification and automatic transcription
- Included >450 telephone calls in Egyptian Colloquial Arabic
- 10 minutes from each of 200 calls transcribed, 120 of those released
- Publications include plain audio, time-aligned transcripts and a pronouncing lexicon
- Lexicon: surface form, romanization, pronunciation, morphological analysis and frequency in 3 data sets
TDT Corpora

- Topic Detection and Tracking Corpora – support development of news understanding system
  - convert speech to text and segment into stories
  - identify new topics in the news and find all stories discussing a selected topic

- TDT-2 and TDT-3 together contain >1000 hours audio, >100K stories, annotated for relevance to 220 topics in English and Chinese

- TDT-4 will add 200 hours of Arabic news audio and transcripts plus newswire totaling 12,000 stories to similar amounts of English and Chinese annotated for 60 new topics

**30002**

**Hurricane Mitch**  
Chinese

**Seminal Event**

**WHAT**: Hurricane Mitch forms over warm ocean waters, killing thousands and causing millions of dollars in damage.  **WHERE**: The Caribbean and surrounding areas, particularly Honduras, Nicaragua and Central America.  **WHEN**: Mitch forms in late September 1998, and lasts through the month of October.

**Topic Explication**

Hurricane Mitch was the most destructive Atlantic hurricane since 1780, killing over 10,000 people in Central America and leaving millions homeless. **On topic**: coverage of the disaster itself, estimates of damage and reports of loss of life; relief efforts by the Red Cross and other aid organizations; impact of the hurricane on the economies of the effected countries.

**Rule of Interpretation**  
**Rule 4. Natural Disasters**
• Text REtrieval Conference
  – organized by NIST, multiple tracks including SDR, CLIR, Q&A
  – broader topics than TDT, assessment replaces annotation

• CLIR 2001 Corpus is LDC Arabic New Corpus, 384,000 stories from Agence France Presse 1994-2000 and 25 topics; CLIR 2002 will add 50 topics

<table>
<thead>
<tr>
<th>Title</th>
<th>YES</th>
<th>NO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performing arts and Islamic institutions</td>
<td>383</td>
<td>471</td>
<td>854</td>
</tr>
<tr>
<td>Arab and western cinema</td>
<td>315</td>
<td>548</td>
<td>863</td>
</tr>
<tr>
<td>Traditional crafts and technology</td>
<td>133</td>
<td>898</td>
<td>1031</td>
</tr>
<tr>
<td>Arab cities and advertising pollution</td>
<td>88</td>
<td>1266</td>
<td>1354</td>
</tr>
<tr>
<td>Polio eradication in the Middle East</td>
<td>57</td>
<td>825</td>
<td>882</td>
</tr>
<tr>
<td>Measles immunization campaigns in the Middle East</td>
<td>17</td>
<td>645</td>
<td>662</td>
</tr>
<tr>
<td>Bilharzia/Schistosomaisis prevention in Egypt</td>
<td>24</td>
<td>949</td>
<td>973</td>
</tr>
<tr>
<td>Environmental protection laws in Egypt</td>
<td>57</td>
<td>668</td>
<td>725</td>
</tr>
<tr>
<td>Egyptian-Libyan relations during the 1990s</td>
<td>321</td>
<td>703</td>
<td>1024</td>
</tr>
<tr>
<td>Tourism in Cairo</td>
<td>242</td>
<td>683</td>
<td>925</td>
</tr>
<tr>
<td>Dead Sea archaeological finds</td>
<td>13</td>
<td>866</td>
<td>879</td>
</tr>
<tr>
<td>Information technology &amp; the Arab world</td>
<td>132</td>
<td>958</td>
<td>1090</td>
</tr>
<tr>
<td>Water resources in the Nile Valley</td>
<td>100</td>
<td>664</td>
<td>764</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>4122</strong></td>
<td><strong>18622</strong></td>
<td><strong>22744</strong></td>
</tr>
</tbody>
</table>
MT Corpora

- MT research lacks a stable metric to evaluate systems
- To support development of a metric, LDC created Multiple Translation Corpora
- >20,000 words, >100 stories in Chinese (Xinhua, Zaobao, VoA), Arabic (AFP, Xinhua)
- Selected from newswire and news broadcast to represent the mode story lengths
- Each story translated by at least 10 human translators, at least 3 systems to represent the range of translation practices and quality
- Translations are sentence aligned to original.
- Translations subsequently assessed by human judges
- Fluency – is the translation grammatical in target language?
- Adequacy – does story convey all information conveyed by idea translation?
- Chinese translations published in 2002; assessments to be added
- Arabic will be published with assessments in 2002.
Conclusions

• We’re keeping busy!
• This was just a survey of some resources; Maamouri will focus on part-of-speech tagged text and Treebanks
• So why is he here?
• Networks of coordinated resources are the way of the future
• Learning more about Arabic Processing
• Looking for additional resources
• Looking for users
• Looking for annotators
• Looking for collaboration that produces concrete results.