Evaluation Initiatives:
US Activities in HLT Evaluation

Christopher Cieri
ccieri@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and Department of Linguistics
3600 Market Street, Philadelphia, PA 19104 U.S.A.

www.ldc.upenn.edu

♦ ELRA HLT Evaluation Workshop, December 1 & 2, 2005
Recent Activities

- **TIDES**: Translingual Information Detection, Extraction and Summarization
  - Goal: build technologies to support a news understanding system that accepts queries in user’s language, finds relevant spoken or written documents in many other languages, extracts important information, translates and summarizes.
  - English, Mandarin Chinese, Modern Standard Arabic with surprises in Cebuano, Hindi

- **EARS**
  - Goal: improve speed, accuracy and readability/usability of STT technologies
  - MDE: tag SUs, disfluency for presentation, further processing
  - English, Mandarin Chinese, Egyptian and Levantine Colloquial Arabic
Global Autonomous Language Exploitation
- Technologies to absorb, analyze, interpret huge volumes of speech, text in multiple languages. Engines convert, distill data, delivering pertinent, consolidated information in easy-to-understand forms.

Engines
- Transcription: Multilingual Speech > English Text (includes translation)
- Translation: Multilingual Text > English Text
- Distillation: integrates information of interest to user from multiple sources and documents

Data Types
- newswire, newsgroups, blogs,
- broadcast news, talk shows, conversations

Annotations
- Transcripts with some Arabic vowelized
- Translations with some aligned at word level
- Treebanks, Propbanks
- “Distillation” Annotation

New approach to evaluation post-facto, scenarios based
New MT evaluation metric based on post-editing and edit distance
NIST MT, RT evaluations continue independently, allow non-US sites.

ELRA HLT Evaluation Workshop, December 1 & 2, 2005
Research on English & Foreign Language Processing
- Machine Translation, Content Extraction (ACE)
- Less Commonly Taught Languages

LCTL => 1M native speakers, scarce resources
- Bengali, Punjabi, Hungarian, Tagalog, Tamil, Thai, Yoruba, Urdu, Uzbek

Resources
- Monolingual News Text
- LCTL -> Eng Parallel Text
- Eng -> LCTL News, Elicitation, Conversation
- Lexicon
- Encoding Converter
- Word, Sentence Segmenters
- POS Tagset, Tagger, Tagged Text
- Morphological Analyzer, Morphologically Analyzed Text
- Named Entity Tagged Text, Tagger
- Name Transliterator
- Grammatical Sketch

ELRA HLT Evaluation Workshop, December 1 & 2, 2005
LVD ID

- Supports Language ID Research, Technology Development & Evaluation
- Language Variety and Dialect ID
  - English: British, Indian, Australian/New Zealand and American English
  - Chinese: Standard Mandarin, Shanghai Wu, Southern Min/Taiwanese Cantonese (Guangdong Hua/Yue Yu)
  - Aceh, Amharic, Bengali, Burmese, Chechen, Farsi, Georgian, Guarani, Hindi, Italian, Japanese, Khmer, Korean, Lao, Punjabi, Russian, Spanish, Tagalog, Tamil, Thai, Tigrigna, Urdu, Uzbek, Vietnamese
- Resource Sharing with Speaker Recognition
- Speaker Recognition
  - Increasing impact of forensic applications
  - Shift toward multilingual, and multichannel tasks
  - Mixer Corpora

ELRA HLT Evaluation Workshop, December 1 & 2, 2005
Mixer Corpora Phases I & II

Arabic: 100x4
Mandarin: 100x4
Russian: 100x4
Spanish: 100x4

Unique Handsets: 100x4
Extended: 550x20

XC: 200x4
TR: 100X4

600 Speakers x 10 Calls

♦ ELRA HLT Evaluation Workshop, December 1 & 2, 2005
## Mixer Corpora Yields

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Targeted</th>
<th>Recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (x10 Calls)</td>
<td>650</td>
<td>1150</td>
</tr>
<tr>
<td>Arabic (x4 Calls)</td>
<td>100</td>
<td>129</td>
</tr>
<tr>
<td>Mandarin (x4 Calls)</td>
<td>100</td>
<td>115</td>
</tr>
<tr>
<td>Russian (x4 Calls)</td>
<td>100</td>
<td>107</td>
</tr>
<tr>
<td>Spanish (x4 Calls)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Extended (x20 Calls)</td>
<td>550</td>
<td>611</td>
</tr>
<tr>
<td>Super-Extended (x30 Calls)</td>
<td>0</td>
<td>382</td>
</tr>
<tr>
<td>Cross Channel (x4 Calls)</td>
<td>200</td>
<td>201</td>
</tr>
<tr>
<td>Transcript Reading</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

◆ ELRA HLT Evaluation Workshop, December 1 & 2, 2005