Integrated Linguistic Resources for Language Exploitation Technologies

Stephanie Strassel, Christopher Cieri, Andy Cole, Denise DiPersio, Mark Liberman, Xiaoyi Ma, Mohamed Maamouri, Kazuaki Maeda

{strassel, ccieri, acole2, dipersio, myl, xma, maamouri, maeda}@ldc.upenn.edu

Linguistic Data Consortium
http://www.ldc.upenn.edu/Projects/GALE

♦ LREC-06 Genoa, Italy – May 24, 2006
Overview
GALE Program

- DARPA funded
  - Three multi-site international research teams
  - Multiple infrastructure providers (data, evals)
- Targets an end-to-end system with 3 components
  - **Transcription** converts speech in any language to text
  - **Translation** converts foreign text to English
  - **Distillation** consolidates information
- Translation and distillation go/no-go evaluations assess technology improvements
- Additional utility evaluation judges how well GALE engines help end user perform tasks
- Phase 1: Arabic, Chinese, English
  - Newswire, weblogs, newsgroups
  - Broadcast news, broadcast conversation

LREC-06 Genoa, Italy – May 24, 2006
Overview of LDC Activities

Collection

Broadcast News Audio
Broadcast Conversation Audio

Collect ASR, CCAP
Audit for quality and content

Create or Acquire Quick-Rich Transcripts

Weblogs
Internet Newsgroups
Newswire

Apply Semi-Automated Formatting, Document Metadata

Structured Arabic Text
Structured Chinese Text
Structured English Text

Data Selection

Translation

Treebank Annotation

Alignment

Structured English Text

Distillation Annotation

*Items with solid red border will be distributed to GALE

LREC-06 Genoa, Italy – May 24, 2006
Data Needs, LDC Targets

<table>
<thead>
<tr>
<th>Medium</th>
<th>Language</th>
<th>Collect exist</th>
<th>Collect need/develop</th>
<th>Transcribe green exist</th>
<th>Transcribe yellow need/develop</th>
<th>Translate exist</th>
<th>Translate need/develop</th>
<th>Align exist</th>
<th>Align need/develop</th>
<th>Treebank source exist</th>
<th>Treebank source need/develop</th>
<th>Treebank target exist</th>
<th>Treebank target need/develop</th>
<th>Propbank source exist</th>
<th>Propbank source need/develop</th>
<th>Distillation source exist</th>
<th>Distillation source need/develop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours Broadcast News</td>
<td>Arabic</td>
<td>5100</td>
<td>1000</td>
<td>130</td>
<td>870</td>
<td>50</td>
<td>75</td>
<td>3</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>4800</td>
<td>1000</td>
<td>460</td>
<td>540</td>
<td>75</td>
<td>3</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>11500</td>
<td>50</td>
<td>9350</td>
<td>50</td>
<td>75</td>
<td>30</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hours Talk Shows</td>
<td>Arabic</td>
<td>55</td>
<td>945</td>
<td>1000</td>
<td>50</td>
<td>50</td>
<td>3</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>60</td>
<td>940</td>
<td>1000</td>
<td>50</td>
<td>3</td>
<td>5</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>260</td>
<td>50</td>
<td>20</td>
<td>50</td>
<td></td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words Newswire</td>
<td>Arabic</td>
<td>527M</td>
<td></td>
<td></td>
<td></td>
<td>3M</td>
<td>7M</td>
<td>100K</td>
<td>820K</td>
<td>500K</td>
<td>500K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>1045M</td>
<td></td>
<td></td>
<td></td>
<td>15M</td>
<td>100K</td>
<td>500K</td>
<td></td>
<td>250K</td>
<td>250K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>2817M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.78M</td>
<td>1.1M</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words News Groups</td>
<td>Arabic</td>
<td>80K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td>300K</td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>60K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td>300K</td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>120K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words Weblogs</td>
<td>Arabic</td>
<td>50K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td>200K</td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>50K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td>200K</td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>13K</td>
<td>1M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50K</td>
<td></td>
<td>50K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Exist** – part of LDC catalog or TIDES/EARS resource
- **Need/Develop** – needs as stated at GALE Planning meeting in July 2005
  - **blue**: LDC will provide data to meet stated need
  - **red**: LDC will attempt to match stated need
  - **green**: LDC will provide beyond stated need with negligible cost to GALE
  - **brown**: proposal for what LDC will provide based on estimated program needs
  - **grey**: other site will develop resource

♦ LREC-06 Genoa, Italy – May 24, 2006
Collection Activities
Text Collection

- **Newswire**
  - Not a focus, maintenance-level collection only
    - Some data required for annotation, evaluation
    - All NW data licensed through data providers
  - Annual Gigaword corpus updates

- **Web text: newsgroups and blogs**
  - Initial manual scouting to identify sites with suitable content
  - Followed by automatic harvesting of all pages within a domain
  - Followed by manual audit to weed out the junk
  - IPR secured under fair use
Broadcast Collection

- **Local collection platform**
  - Existing infrastructure improved for GALE
    - More static storage, upgraded AV digitizer, better monitoring
  - Also collect or produce CCAP, ASR wherever possible
  - Each recording subject to manual auditing

- **Portable collection platform**
  - TiVO-like digital video recorder
  - Records two AV streams simultaneously
  - Supports multiple broadcast standards
  - Recording formats: MPEG-1, MPEG-2
  - Ubuntu Linux with MySQL and ivtv software package
  - One portable platform installed at HKUST (Chinese) in March
  - Second planned for installation in North Africa (Arabic) for GALE Phase 2
Data Formatting & Selection

- Collected data from a variety of sources, in a wide range of formats

- Some standardization required at start of pipeline
  - File naming conventions standardized across tasks, collections
  - Audio standard: 16KHz, single-channel, 16-bit PCM .sph audio files (web audio varies)
  - Text standard: .tdf for transcripts, .sgm for web and newswire
    - Also distribute .html for web-harvested data
  - Downstream annotation tasks add standoff markup
    - Source files remain stable for all downstream tasks

- Careful data selection for end-to-end annotation
  - Selected files are annotated for most or all tasks
  - Best data is rich in targeted content (entities, topics, etc.) and is representative of targeted genre
Transcription and Translation Activities
Transcription

❖ Create or collect transcripts for 1000 hours/language/genre (less for English)

❖ Goal: Quick Rich Transcription (QRTR)
  ❖ Time-aligned content-accurate transcript, some markup, limited QC
  ❖ Sentence unit (SU) labels, speakerID, story boundaries

❖ Reality: high volume requires flexibility
  ❖ Web harvested
    • Verbatim transcripts or scripts; some good, some poor quality
    • No time alignment or speaker turns
  ❖ QTR (from some transcription agencies)
    • Quick transcription; speaker turns & speaker ID
    • No SU annotation or SU time alignment
  ❖ QRTR (from agencies and LDC)
    • Quick transcription; speaker turns & speaker ID
    • SU annotation and SU time alignment

❖ Good news: new XTrans transcription toolkit
  ❖ Improved manual transcription rates and quality

♦ LREC-06 Genoa, Italy – May 24, 2006
Translation

- Create manually translated text for
  - 300K words/language newsgroups
  - 200K words/language weblogs
  - 75 hours/language broadcast news
  - 50 hours/language broadcast conversation
  - All sentence-aligned with manual QC
  - Subset manually word-aligned

- Harvest millions of words of parallel text
  - Through found or donated resources
    - FBIS website, Ummah, UN, HKSAR
  - Through regular BITS web searches
Translation Strategies

Translation agencies
- Using best teams from earlier programs, plus some newcomers
- Improved quality and better prices than in past

Translation guidelines
- Increased emphasis on correct translation of proper names
- Required to use conventional English punctuation (not source text punctuation)
- Additional guidelines for spoken genres
  - Filled pauses, restarts, repetitions are translated
  - Partial words are marked as “%pw”
  - Correct mispronounced words and typos
  - Do not correct factual errors

Quality control
- Samples extracted from each agency delivery
- Translation quality assessed by bilingual LDC staff
  - Adopted NSA grading system for translator assessment
Translation Evaluation

- Evaluation uses edit distance metric rather than BLEU or similar
- Requires assessment by human post-editors at LDC
  - Compare MT system output from each team with human-created gold standard reference
    - Gold standard based on multiple human translations created by NVTC
    - Adjudicated, but alternatives indicated
  - Editors alter MT output for content discrepancies only
  - Editors trained to minimize number of edits while preserving meaning
Distillation Activities
Distillation

- **10 query templates**
  - FIND STATEMENTS MADE BY OR ATTRIBUTED TO [person] ON [topic(s)]
  - IDENTIFY PERSONS ARRESTED FROM [organization] AND GIVE THEIR NAME AND ROLE IN ORGANIZATION AND TIME AND LOCATION OF ARREST

- **LDC providing training data**
  - Up to 40 queries and responses per type
  - All responses in English
  - Subset of responses also in Chinese, Arabic

- **Eval answer keys will be created by eval coordinator (BAE) using somewhat different process**

---

LREC-06 Genoa, Italy – May 24, 2006
Distillation Annotation

1. Identify relevant documents
   - Search-guided document retrieval (like TDT, HARD)
   - Distribute both on- and off-topic docs, plus “dup” docs

2. Extract snippets
   - Strings from source text that answer the query
     - Use BAE guidelines for determining relevance of snippets
     - Pronoun, temporal, locative resolution

3. Build nuggets
   - “Facts” extracted from each snippet

4. Create supernugs
   - “Co-reference” on semantically equivalent nuggets
     - Defined by mutual entailment
     - Across documents, across languages
   - Create English gloss

5. BAE provides relevance judgments for supernugs

*LREC-06 Genoa, Italy – May 24, 2006*
Core Linguistic Resources
Arabic Treebank

- 30 hours of GALE broadcast news
  - Primarily MSA
    - About 5% dialectal (Egyptian Arabic, Levantine, etc.)
  - Using established UPenn ATB approach

- Challenges
  - Code-mixing/switching MSA-Dialectal Arabic
    - Difficulty of defining specifications for dialectal
  - Transcription errors from QRTR impact TB annotation
  - No case endings, wrong case endings in spoken genres
  - Disfluencies (mostly in the dialectal segments)
500 Kw sentence-aligned English newswire translations from Arabic TB Part 3

Challenges

- Large volume (500Kw) requires new staff
  - Steep learning curve for TB annotation
- Translation errors and tokens to be disregarded slow down annotation
  - E.g., extra determiners
- Improved parsing technology affects both annotation speed and annotation quality
Distribution and Data Management
Data Management

http://www.ldc.upenn.edu/Projects/GALE/Data

- **Catalog query** lists all GALE-relevant corpora
- **Data matrix** describes new resources under development for current phase of program
  - Which task/need each resource addresses
  - Language, volume, source, epoch information
  - Links to specifications/guidelines
  - Access instructions for each resource

- **Task Definitions**
  - State needs, assumptions for each task
  - Describes data characteristics
    - Language, genre, type, source, how selected, …
  - Defines annotation task and quality control
  - Describes distribution formats

♦ LREC-06 Genoa, Italy – May 24, 2006
<table>
<thead>
<tr>
<th>Phase</th>
<th>Task</th>
<th>Genre</th>
<th>Lang</th>
<th>Total Volume</th>
<th>Source</th>
<th>Epoch</th>
<th>Annotations</th>
<th>Deliveries (click on release name for detailed information)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Collect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>1000 hrs</td>
<td>GALE Collection: Al Hurra, LBC, Sawa Dari, Sawa Iraq</td>
<td>2004.05</td>
<td>manual audit</td>
<td>43.5 hrs</td>
<td>213 hrs LDC2005E80, 294 hrs LDC2006E91</td>
<td>250 hrs, 243.5 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>945 hrs</td>
<td>GALE Collection: Al Jazeera, LBC</td>
<td>2005.10</td>
<td>manual audit</td>
<td>20 hrs</td>
<td>203 hrs LDC2005E80, 130 hrs LDC2006E91</td>
<td>298 hrs, 297 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>50 hrs</td>
<td>GALE Collection: CNN</td>
<td>2005.06</td>
<td>manual audit</td>
<td>--</td>
<td>27 hrs LDC2005E80, 10 hrs LDC2006E91</td>
<td>10 hrs, 3 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>1M+ words</td>
<td>various</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>886Kw LDC2005E81, 11,516Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chinese</td>
<td>1M+ words</td>
<td>various</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>12,860Kw LDC2005E81, 12,719Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>1M+ words</td>
<td>various</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>9,082Kw LDC2005E81, 91,639Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>1M+ words</td>
<td>various</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>2,833Kw LDC2005E81, 4,254Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chinese</td>
<td>1M+ words</td>
<td>various</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>4,594Kw LDC2005E81, 5,525Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>1M+ words</td>
<td>new</td>
<td>open</td>
<td>DataScout guidelines</td>
<td>--</td>
<td>4,297Kw LDC2005E81, 3,394Kw LDC2006E32</td>
<td>250Kw plus add! QC, 250Kw plus add! QC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>870 hrs</td>
<td>VOA archive; GALE collection</td>
<td>TBD</td>
<td>web-harvested transcripts; QuickTR, QuickRichTR</td>
<td>10 hrs LDC2005E71</td>
<td>10 hrs QTR LDC2005E82, 5,305hrs QTR LDC2006E33</td>
<td>350 hrs, 350 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chinese</td>
<td>540 hrs</td>
<td>GALE collection</td>
<td>TBD</td>
<td>web-harvested transcripts; QuickTR, QuickRichTR</td>
<td>--</td>
<td>71 hrs web; 12.7 hrs QTR; 5 hrs QTR OR TR LDC2005E82, 134 hrs QTR; 21 hrs QTR LDC2006E33</td>
<td>150 hrs, 150 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>50 hrs</td>
<td>GALE collection</td>
<td>TBD</td>
<td>QuickRichTR</td>
<td>--</td>
<td>--</td>
<td>25 hrs, 25 hrs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabic</td>
<td>1000 hrs</td>
<td>Al Jazeera web harvest; GALE</td>
<td>TBD</td>
<td>web-harvested transcripts; QuickTR</td>
<td>20 hrs</td>
<td>193 hrs web; 12 hrs QTR LDC2005E82, 31 hrs web LDC2006E33</td>
<td>313 hrs, 312 hrs</td>
</tr>
</tbody>
</table>
Data Distribution

● GALE sites receive
  ✷ GALE-relevant LDC catalog items (*by request*)
  ✷ Large quarterly releases of new data and occasional ad hoc releases (*automatic delivery to all sites*)
    • Web download (replaces earlier FTP delivery system)
    • CD, DVD for larger corpora (typically over 100 MB)
    • Hard drives for audio

● GALE evaluation license provides free access to resources for duration of program
  ✷ Ongoing access through LDC Membership
  ✷ One license per site
  ✷ No re-distribution of data permitted
    • LDC-hosted SCP server allows secure exchange data (e.g. inline annotations of copyrighted data) within multi-site teams

● For small number of resources, FOUO license
  ✷ No use outside of GALE (even by GALE sites)
Conclusion

- Data plan, approaches continue to be refined as program needs evolve
- To date, GALE resources encompass over:
  - 1500 hours of collected audio, 5Mw collected web text
  - 800 hours of new transcripts
  - 500K words of translations & harvested parallel text
  - 300K words of treebanks
  - 200 distillation query responses
  - Additional “ad hoc” resources
  - New web scouting, transcription, word alignment, distillation, treebank annotation tools and processes
- Vast majority of GALE data will be published as part of LDC’s regular catalog, starting in late 2006

http://www.ldc.upenn.edu/Projects/GALE

LREC-06 Genoa, Italy – May 24, 2006