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

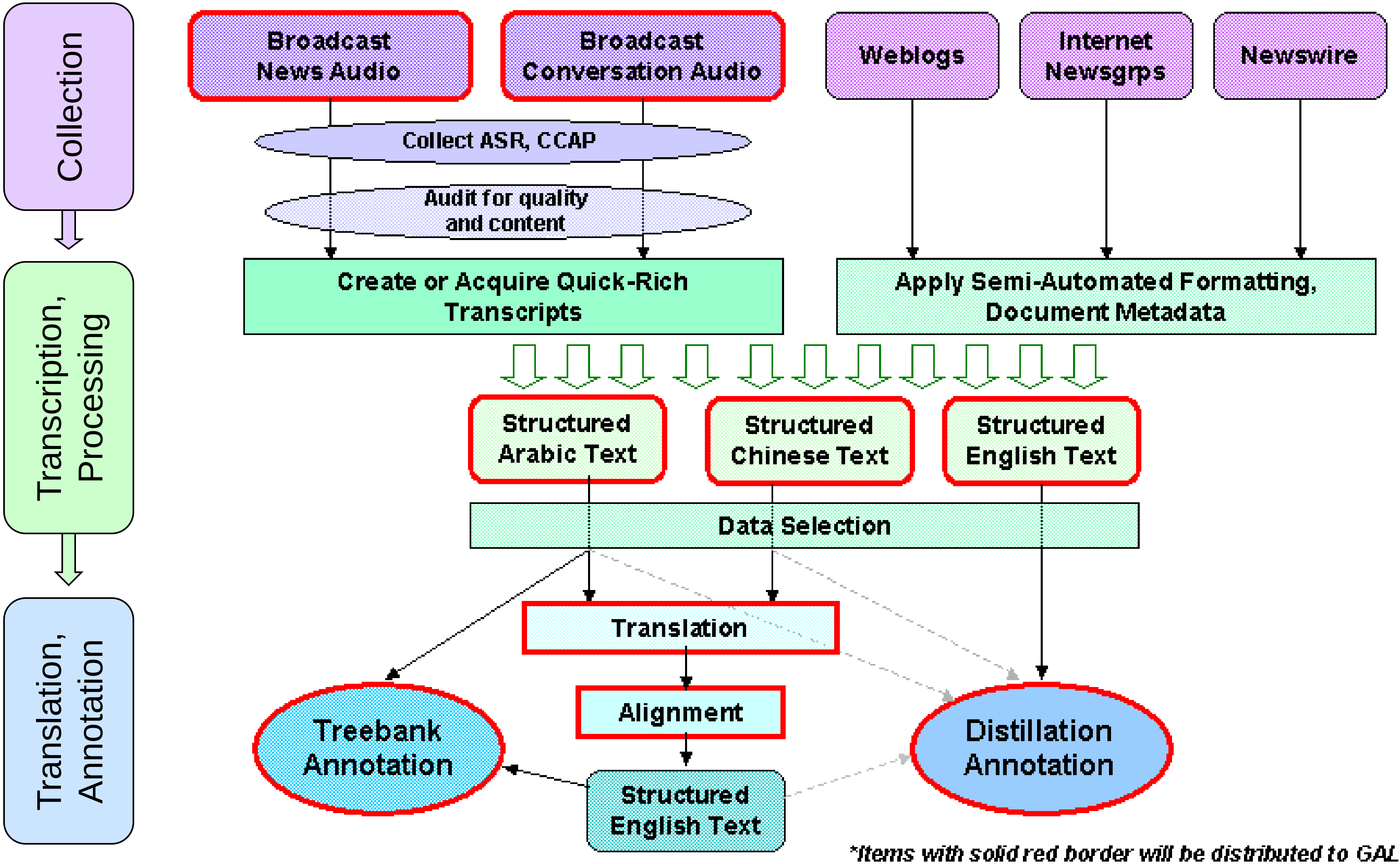
♦ **LREC-06 Genoa, Italy – May 24, 2006**

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

Overview of LDC Activities



Collection

Transcription, Processing

Translation, Annotation

Broadcast News Audio

Broadcast Conversation Audio

Weblogs

Internet Newsgrps

Newswire

Collect ASR, CCAP

Audit for quality and content

Create or Acquire Quick-Rich Transcripts

Apply Semi-Automated Formatting, Document Metadata

Structured Arabic Text

Structured Chinese Text

Structured English Text

Data Selection

Translation

Alignment

Structured English Text

Treebank Annotation

Distillation Annotation

*Items with solid red border will be distributed to GALE

Data Needs, LDC Targets

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

- **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**

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 installtation 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 newsiwre
    - 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

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

Distillation Annotation

All queries

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

Subset of queries

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

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)

English-Arabic Parallel TB

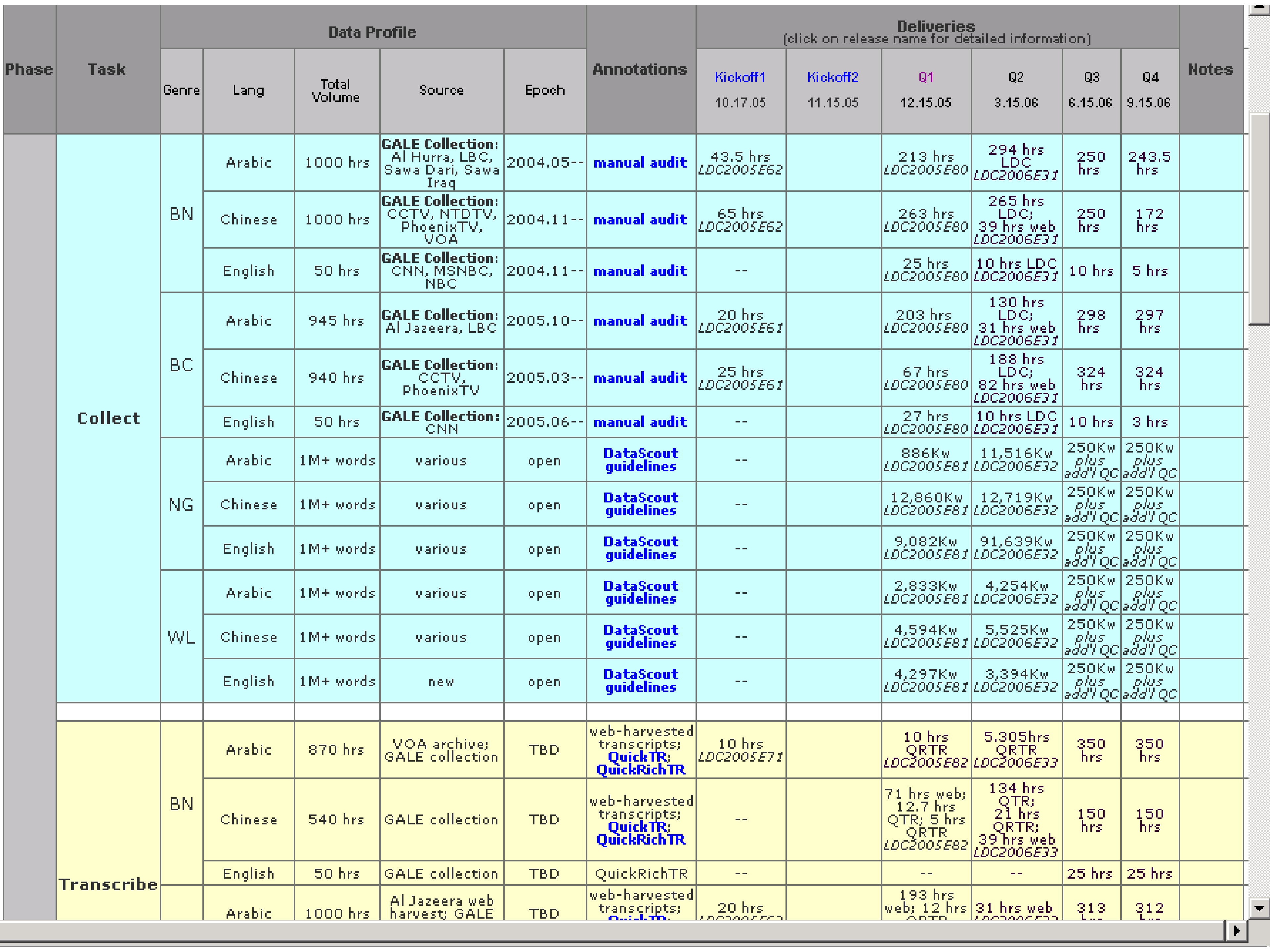
- 500 Kw sentence-aligned English newswire translations from Arabic TB Part 3
- Challenges
  - Large volume (500Kw) requires new staff
    - Steep learning curve for TB annotation
  - Translation errors and tokens to be disregarded slow down annotation
    - E.g., extra determiners
  - Improved parsing 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



Phase	Task	Data Profile					Annotations	Deliveries (click on release name for detailed information)						Notes
		Genre	Lang	Total Volume	Source	Epoch		Kickoff1 10.17.05	Kickoff2 11.15.05	Q1 12.15.05	Q2 3.15.06	Q3 6.15.06	Q4 9.15.06	
	Collect	BN	Arabic	1000 hrs	**GALE Collection:** Al Hurra, LBC, Sawa Dari, Sawa Iraq	2004.05--	**manual audit**	43.5 hrs *LDC2005E62*		213 hrs *LDC2005E80*	294 hrs LDC *LDC2006E31*	250 hrs	243.5 hrs	
			Chinese	1000 hrs	**GALE Collection:** CCTV, NTDTV, PhoenixTV, VOA	2004.11--	**manual audit**	65 hrs *LDC2005E62*		263 hrs *LDC2005E80*	265 hrs LDC; 39 hrs web *LDC2006E31*	250 hrs	172 hrs	
			English	50 hrs	**GALE Collection:** CNN, MSNBC, NBC	2004.11--	**manual audit**	--		25 hrs *LDC2005E80*	10 hrs LDC *LDC2006E31*	10 hrs	5 hrs	
		BC	Arabic	945 hrs	**GALE Collection:** Al Jazeera, LBC	2005.10--	**manual audit**	20 hrs *LDC2005E61*		203 hrs *LDC2005E80*	130 hrs LDC; 31 hrs web *LDC2006E31*	298 hrs	297 hrs	
			Chinese	940 hrs	**GALE Collection:** CCTV, PhoenixTV	2005.03--	**manual audit**	25 hrs *LDC2005E61*		67 hrs *LDC2005E80*	188 hrs LDC; 82 hrs web *LDC2006E31*	324 hrs	324 hrs	
			English	50 hrs	**GALE Collection:** CNN	2005.06--	**manual audit**	--		27 hrs *LDC2005E80*	10 hrs LDC *LDC2006E31*	10 hrs	3 hrs	
		NG	Arabic	1M+ words	various	open	**DataScout guidelines**	--		886Kw *LDC2005E81*	11,516Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
			Chinese	1M+ words	various	open	**DataScout guidelines**	--		12,860Kw *LDC2005E81*	12,719Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
			English	1M+ words	various	open	**DataScout guidelines**	--		9,082Kw *LDC2005E81*	91,639Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
		WL	Arabic	1M+ words	various	open	**DataScout guidelines**	--		2,833Kw *LDC2005E81*	4,254Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
			Chinese	1M+ words	various	open	**DataScout guidelines**	--		4,594Kw *LDC2005E81*	5,525Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
			English	1M+ words	new	open	**DataScout guidelines**	--		4,297Kw *LDC2005E81*	3,394Kw *LDC2006E32*	250Kw *plus add'l QC*	250Kw *plus add'l QC*	
	Transcribe	BN	Arabic	870 hrs	VOA archive; GALE collection	TBD	web-harvested transcripts; **QuickTR; QuickRichTR**	10 hrs *LDC2005E71*		10 hrs QRTR *LDC2005E82*	5.305hrs QRTR *LDC2006E33*	350 hrs	350 hrs	
			Chinese	540 hrs	GALE collection	TBD	web-harvested transcripts; **QuickTR; QuickRichTR**	--		71 hrs web; 12.7 hrs QTR; 5 hrs QRTR *LDC2005E82*	134 hrs QTR; 21 hrs QRTR; 39 hrs web *LDC2006E33*	150 hrs	150 hrs	
			English	50 hrs	GALE collection	TBD	QuickRichTR	--		--	--	25 hrs	25 hrs	
			Arabic	1000 hrs	Al Jazeera web harvest; GALE	TBD	web-harvested transcripts;	20 hrs		193 hrs web; 12 hrs	31 hrs web	313	312	

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
  - 500Kwords of translations & harvested parallel text
  - 300Kwords 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