Integrated Linguistic Resources for Language Exploitation Technologies

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

Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104

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

Abstract

Linguistic Data Consortium has recently embarked on an effort to create integrated linguistic resources and related infrastructure for language exploitation technologies within the DARPA GALE (Global Autonomous Language Exploitation) Program. GALE targets an end-to-end system consisting of three major engines: Transcription, Translation and Distillation. Multilingual speech or text from a variety of genres is taken as input and English text is given as output, with information of interest presented in an integrated and consolidated fashion to the end user. GALE's goals requires a quantum leap in the performance of human language technology, while also demanding solutions that are more intelligent, more robust, more adaptable, more efficient and more integrated. LDC has responded to this challenge with a comprehensive approach to linguistic resource development designed to support GALE's research and evaluation needs and to provide lasting resources for the larger Human Language Technology community.

1. Introduction

The goal of the DARPA GALE program is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. GALE consists of three major engines: Transcription, Translation and Distillation. The output of each engine is English text. The input to the transcription engine is speech and to the translation engine, text. Engines will pass along pointers to relevant source language data that will be available to humans and downstream processes. The distillation engine integrates information of interest to its user from multiple sources and documents. Linguistic Data Consortium supports the GALE Program by providing linguistic resources – data, annotations, tools, standards and best practices – for system training, development and evaluation.

2. Collection

At its outset, GALE targets Arabic, Chinese and English data in traditional genres like broadcast news and newswire, and adds broadcast conversations – that is, talk shows, roundtable discussions, call-in shows and other interactive-style broadcasts. Weblogs and newsgroups (electronic bulletin boards, Usenet newsgroups, discussion groups and similar forums) are also targeted. Conversational telephone speech will be incorporated in later phases of the program. Required data volume is high: approximately 1000 hours per language and genre for broadcast sources, and over a million words of web text per language/genre are planned for Phase 1.

2.1. Broadcast News and Conversation

LDC's broadcast collection focuses primarily on Chinese and Arabic because of the relative shortage of available recordings of these languages when compared to English. But because collection infrastructure is integrated rather than separated across the three languages and multiple genres, English broadcasts continue to be collected at negligible expense, which means English recordings from the same time epoch will be available if and when they are required by GALE or another program.

LDC's existing broadcast collection system was extended in multiple ways to meet GALE program demands. At the program's outset, our collection platform was modified to include more static storage, an upgraded AV digitizer and better monitoring capabilities. Target programs come in via satellite; audio is captured and stored on local servers as .wav files, and where applicable video is stripped out and saved separately.

Automatic speech recognition output is then generated. ASR assists with downstream data selection and provides an additional sanity check on the recording process. Generally, ASR output is produced automatically the day after programs are recorded. Further, audio files in the broadcast archive that do not already have ASR output are queued manually into the ASR system and processed. Where possible, Closed Caption output (CCAP) is also produced for English audio recordings. CCAP is generated by the "SoftTouch Hubcap" system that decodes North American line-21 closed captions, the only CCAP system used in the United States. CCAP output is generated for English-language programs with CCAP at the time of transmission. If CCAP output cannot be collected at the time of transmission because of technical problems in the transmission or because of problems with LDC's system, CCAP cannot be "re-generated" after transmission.

Once audio, ASR and CCAP files have been created and the recording database has been updated, native speakers of each language manually audit each recording via a specialized GUI with a database backend. The broadcast auditing interface links audio files from the servers where they are stored into a list that can be filtered by date, source, program, language, and audit status. Auditors listen to three randomly selected thirty-second sound bites from the beginning, middle and end of each recording. For each segment auditors answer the following questions: Is there a recording? Is the audio quality OK? What is the language? Is it speech from the right program? What is the data type? What is the topic? Capturing information on data content and format at this

stage aids data selection for downstream annotation tasks. Audit logs are reviewed daily as a sanity check on the collection process, in order to catch things like unannounced broadcast schedule changes.

Several new broadcast programs were added to LDC's recording inventory based on GALE's requirements, particularly Chinese and Arabic talk shows and interviews. Native speakers at LDC suggested new programs, and samples of each were recorded for evaluation while our external relations manager pursued intellectual property rights arrangements with the data providers. However, there is a limit on the number of Chinese and Arabic broadcasts (especially talk shows and other conversational programs) on the airwaves and satellite feeds available in Philadelphia. Meeting the program's demands for volume and variety therefore also required novel approaches in the realm of data collection. To this end, LDC has developed portable broadcast data collection platforms for use by partner sites. The platforms are fully functional versions of the hardware and collection, auditing, annotation and workflow management software used successfully for local data collections. The platform is a compact TiVO-like DVR capable of recording two streams of AV material simultaneously. Each system supports several broadcast standards (NTSC-M and PAL D/K for Y1). The systems are portable, weighing under 30 pounds, with a footprint of less than 30cm x 30cm x 10cm. Available recording formats are MPEG-1 and MPEG-2. Each system runs Ubuntu Linux with Mysql and the ivtv software package. The first portable system will be installed at HKUST in March 2006, with a second platform planned for installation in Tunisia or Morocco later in 2006.

2.2. Newswire

No new large-scale newswire collection is required for GALE. But because newswire data is targeted for downstream annotation, and because the newswire genre is at least initially included in the annual system performance targets, LDC continues to maintain core newswire collections in each language. This represents a significant downscaling of effort in this genre from previous programs (like TIDES); for instance, expensive and lower-yield sources were removed from the inventory. The resulting GALE newswire collection relies on existing infrastructure and requires minimal supervision. Rather than making quarterly newswire deliveries (on par with deliveries for other data types), newswire data is distributed to GALE via LDC's annual Gigaword corpus updates that are part of our regular catalog.

2.3. Web Text

The new interest in web domains has also meant a rapid refocusing of LDC's data collection activities, and the introduction of a new task, data scouting. Two web genres are targeted in Phase 1: newsgroups consist of posts to electronic bulletin boards, Usenet newsgroups, discussion groups and similar forums; and weblogs which are posts to informal web-based journals of varying topical content. Data distributed to GALE includes both the source data format (HTML) and a standardized XML formatted version of each file.

The newly-defined data scouting task provides a formalized mechanism for quickly and effectively locating, evaluating, harvesting and processing a web data in large volumes across multiple languages. Infrastructure like a customized user interface facilitates the process and provides structure for a task that might otherwise devolve into simply surfing the web. Data scouts are assigned particular topics and genres along with a production target in order to focus their web surfing. For instance, a scout might be asked to find 10 Arabic newsgroup threads and 15 Arabic blog posts on the topic of "car bombings in the Middle East" during one work session. Formal annotation guidelines provide necessary structure for the task, and a customized annotation toolkit provides a front-end to help annotators manage the process and track progress. Annotation decisions are logged to a database.

A nightly process queries the annotation database and harvests all designated URLs. While early stages of data scouting are entirely manual, once it has been established that a particular site has significant usable content, it is harvested on an ongoing basis. Whenever possible, the entire site is downloaded, not just the individual thread or post identified by the annotator. A scheduler program then iterates threads/posts in the database to download the thread if it has not been downloaded already, or re-downloads if there are new postings for the thread. Download schedules are variable, based on the level of activity on the thread and feedback from the annotators.

Data on the web occurs in numerous formats, with highly variable (and inconsistently-applied) markup. We have developed a number of scripts to standardize formatting so data can be more easily fed into downstream annotation processes. Original-format versions of each document are also preserved. Typically a new script is required for each new domain name that has been identified. After scripts are run, an optional manual process corrects any remaining formatting problems.

Automated harvesting results in a large data pool that contains "bad" as well as "good" documents (as defined by their information content). Before documents are earmarked for downstream annotation tasks they still must be reviewed for content suitability; failure to do so might result in for instance blogs about knitting patterns or vegan recipes being translated or Treebanked. Data selection is a semi-automated process. Documents and text passages already labeled as "good" provide input to a statistical analysis of token frequency for good/bad documents for each topic or data type. The analysis is used to generate a list of positively- and negatively-weighted keywords to help in the identification of additional "good" documents from the data pool. The list of keywords is then fed through LDC's custom search engine to generate relevancy rankings for each document. Some additional processes exclude easily-identified "junk" documents. Finally, an annotator reviews the list of relevance-ranked documents and selects those that are suitable for a particular annotation task or for annotation in general. These newly-judged documents in turn provide additional input for generation of new ranked lists. Both data types – manually selected and automatically harvested (whether "audited" or not) – are released to GALE sites. To date, over 4.5 million words of weblog and newsgroup text has been distributed. IPR for both genres has been secured under fair use.

3. Data Processing, Selection and Annotation

A variety of annotation tasks follow collection. Because annotation is both costly and time-consuming, careful selection of data to be annotated is critical to avoid wasting time marking up data of little value. We have developed interactive search and indexing tools that allows users to target data for annotation based on content of interest like concentration of entities and events or presence of spoken phenomena like disfluencies.

So that data tagged in one task is readily available for exploitation in the next, we have also implemented an end-to-end annotation strategy that results in selected data being consecutively transcribed, translated, word-aligned and treebanked, then tagged for information content under the distillation task. This pipelined approach not only saves costs but also offers the opportunity to learn from multiple annotations on the same source data.

3.1. Transcription

Nearly all spoken resources collected under GALE will be transcribed, both to supply training data to transcription engines and to provide input to downstream annotation tasks. The default mode of transcription is Quick Transcription (QTR), based upon previous experiments demonstrating that this type of transcript is most cost-effective for system training. To support more agile development of systems that output readable transcripts, transcripts are further labeled according to a Rich Transcript Markup specification developed for GALE. We have turned to partner sites and commercial agencies who posses the infrastructure and know-how to rapidly produce high-volume, low-cost training transcripts of sufficient quality. Much smaller amounts of speech data are carefully transcribed to create test sets and provide material for consistency analysis.

3.1.1. Web transcripts and scripts

Because volume demands are high and resources limited, LDC has expended considerable effort to capture existing transcripts from the web for sources that we record (and particularly for sources for which we have IPR agreements in place). In a few cases, large transcript archives are available and these are downloaded and processed. Scripts have also been developed harvest new transcripts on an ongoing basis.

The problem of identifying matches between harvested transcripts and LDC's audio recordings is not a trivial one. Web indexes do not always provide detailed information about the broadcast time/date of the transcript, and in many cases only snippets of programs are transcribed. LDC relies on time/date information from our recording database and ASR output for locally collected recordings to identify likely matches, but some manual effort is required to confirm a real match. Furthermore, sources vary greatly in the quality and type of web transcript provided. Occasionally the transcripts are verbatim, containing a complete word-for-word transcript for all speakers in the recording plus complete speaker identification. In other cases the transcript is incomplete, with only portions of the show transcribed or speakers missing, or sections summarized rather than transcribed verbatim. Typically, features of spontaneous speech like filled pauses and disfluencies are not present even in (near-)verbatim web transcripts. For the majority of sources, what is available for download are scripts rather than transcripts, which contain talking points or outlines of issues (to be) covered in the program, rather than an actual record of what was spoken. In no cases are web (tran)scripts time-aligned with the audio, nor have sentence units (SUs) been explicitly identified.

Initial data releases to GALE include minimally processed versions of the web transcripts. The transcripts are extracted from the downloaded HTML files, converted to UTF-8 plain text format, and divided into (possible) sentences based on the punctuation characters. The transcripts are then added to LDC's regular annotation pipeline where at minimum transcripts are time-aligned with the audio. As time and funding allow, transcripts are also slated for additional quality control in which content and markup are added (including SUs and complete speaker identification).

3.1.2. Quick (Rich) Transcription

The majority of transcripts distributed to GALE are produced specifically for the program, either at LDC or by a professional transcription agency under contract to LDC. Final transcripts for GALE conform to the quick transcription (QTR) specification whose elements include accurate transcription of content words, segmentation and time-alignment at least to the level of sections and speaker turns, speaker identification, and standardized punctuation and orthography but no additional markup. A portion of the data is further subject to the quick rich transcription (QRTR) specification, which adds time-aligned sentences plus sentence type identification (SUs). A small amount of data will be dually transcribed and discrepancies adjudicated to provide data for analysis of human transcription variation. If necessary, small amounts of data may also be transcribed using Careful Transcription (CTR) specifications in order to provide material for evaluating system performance.

Beyond the sheer volume of transcripts required for GALE, one of the challenges in meeting the program's demands has been identifying commercial transcription agencies that can produce time-aligned, SU and speaker-labeled transcripts of adequate quality with rapid turnaround (and for a price the program can accommodate). The problem is less acute for Chinese where LDC has long-standing relationships with several transcription agencies who were willing to adapt their existing methods to meet GALE's demands. For Arabic the situation is much worse. LDC circulated sample kits (containing transcription guidelines, tools and 5 hours of Arabic data samples) for over 25 Arabic transcription agencies before identifying only two who were able to perform the required task; even then Arabic rates are much more expensive than Chinese. Because of this difficulty, more transcription is being done locally at LDC than originally planned (though at least 75% of the data continues to be outsourced for both languages). LDC has developed a new transcription tool, XTrans (Maeda et. al. 2006), to facilitate in-house transcription. The tool is freely available and we have also distributed it (along with training guidelines) to the commercial agencies providing transcripts for GALE.

All incoming transcripts created by commercial agencies are put through LDC's quality control pipeline where critical problems (bad timestamping,

misinterpretation of guidelines, missing speaker ID) are corrected before transcripts are circulated to GALE sites. Minor errors are overlooked, in keeping with the "quick" transcription target.

3.2. Translation Resources

To support intensive research in Translation, LDC is harvesting large volumes of parallel text from the Internet and, using text and transcripts described above, is producing singly- and multiply-translated text under subcontract to translation agencies. Phase 1 of GALE targets approximately 500Kwords of parallel web text per language plus 125 hours of parallel broadcast transcripts (including both news and conversations).

Much of this data is acquired through professional translation agencies, as in previous programs like TIDES. However, the high data volumes targeted in GALE demand improved processes. One such improvement is the translation guidelines provided to agencies. Existing guidelines were adapted to incorporate specialized instructions for each language and genre, including treatment of genre-specific issues (e.g., handling of disfluencies in spoken genres). The guidelines also incorporate new examples of good and poor translations, and provide more explicit information about what constitutes a translation error. New quality control procedures have also been implemented. Translation samples from every new agency (and from each delivery for existing agencies) are subject to a manual review/scoring process, adapted from the NSA translation quality assessment guidelines, that deducts points for adequacy and fluency problems.[1] To make QC more efficient and consistent across agencies, for GALE we also embed a "test passage" in each batch of data to be translated and will then compare its translation to a pre-existing gold standard. No agency sees the same test passage more than once. This approach will save effort for LDC and will also allow for fair comparison across translation agencies.

In addition to creating new translations manually, LDC continues to harvest parallel text from the web. In some cases, we rely on existing (and expanding) archives of parallel text from data providers like Ummah, HKSAR and the UN. The GALE program manager has also negotiated use of translation resources from FBIS for official use only within GALE, and this source is expected to provide at least a million words of parallel Arabic newswire in Phase 1. New sources of parallel text are regularly identified via BITS (the Bilingual Internet Text Search tool developed at LDC), a tool for finding parallel text over the Internet without human intervention. The input for BITS is a list of web sites that possibly contain parallel text of the pertinent language pair. The output is parallel text aligned at the document level.

LDC is also creating manually word aligned parallel text for training and evaluating word alignment algorithms, since the performance of these algorithms has been shown to have a direct impact on the performance of machine translation systems. Word alignment is a new task for LDC, so we initially planned to use existing infrastructure (tools, guidelines) adapted from peer institutions; but we found that existing tools were not well-suited to the challenges of GALE, so new customized tools were developed instead. The input to the word alignment task is sentence aligned parallel text; the output is standoff word alignment annotation. The user interface presents the results of automatic alignment (by GIZA++ or similar) to bilingual annotators, who then correct the alignment. A subset of the data is dually annotated and adjudicated to build a gold standard corpus as well as to provide input for consistency analysis. Volumes of word aligned data are fairly modest in Phase 1 (200Kwords of word aligned parallel text from text sources and 6 hours from broadcast sources per language).

3.3. Distillation Resources

LDC is performing a set of innovative, integrated annotation tasks to support development of Distillation engines that will gather information at different levels of granularity distributed across various sources, languages and domains. Given contextualized user queries produced by GALE's targeted end users, LDC annotators extract core elements like entities, events and other facts, eliminate repetition and redundancy, flag contradiction, and consolidate the extracted knowledge into coherent, condensed representations.

The annotation task involves responding to a series of user queries developed by government intelligence analysts[2]. Junior annotators use a search-guided relevance assessment technique to find and label documents in the source language (English, Chinese or Arabic) that are relevant to each query[3]. The annotator then extracts snippets from each relevant document. A snippet is defined as a continuous string of text that contains an answer to the query. There are no restrictions on snippet size, granularity or potential overlap between snippets. Because snippets are extracted directly from a document, they may contain pronouns, locatives, temporal expressions or other words whose reference is ambiguous within the selected string. In such cases, annotators must disambiguate the term based on information provided elsewhere in the document.

After snippets have been extracted, annotators then create a nugget for each fact expressed in the snippet. Nuggets describe any piece of information that an annotator considers a valid answer to the query. Annotators follow a series of rules (with examples) to decompose snippets into nuggets, guided by the general requirement that the meaning of each nugget is entailed by the meaning of the snippet. (That is, if the snippet is true, the fact asserted by each nugget must also be true.)

All nuggets created for a given query are then reviewed by senior annotators for that language. Semantically equivalent nuggets (those that mutually

[1] For instance see http://www.ldc.upenn.edu/Projects/GALE/Translation/specs/ar_speech_translation_guidelines_v1.1.pdf

[2] English queries and contexts are provided by the Distillation evaluation coordinator, BAE Systems; LDC then has each query professionally translated into Chinese and Arabic prior to annotation.

[3] This process is similar to the process used for search-guided annotation in LDC's Topic Detection and Tracking efforts; see for instance: http://www.ldc.upenn.edu/Projects/TDT5/Annotation/TDT2004V1.2.pdf

entail, or can be inferred from, one other) are then clustered into "nugs". After all the nugs for a given query, regardless of language or source document, have been created, lead annotators for each language work in committee to build "supernugs"; that is, the cluster of nugs that are semantically equivalent across languages and documents. Each supernug is then translated into English. The resulting list of supernugs for each query should represent a complete list of facts in English, drawn from all of the (multilingual, multi-source) documents that were considered to contain responses to this query. Annotation is aided by a customized user interface with database backend.

For Phase 1 of GALE, LDC plans to provide snippets and nuggets for up to 450 queries (in all three languages), plus nugs and supernugs for a subset of approximately 150 of the queries, for use as Distillation training data.

3.4. XBanks

XBanking refers to syntactic and semantic annotation including part-of-speech tagging, Treebanking and Propbanking. Databases of syntactically and semantically annotated text and speech have proven their use in many speech and language technologies, particularly information extraction. To support this critical component of GALE, LDC is producing parallel English-Arabic Treebanks and is extending Arabic Treebanking to spoken language for the first time.

While previous Arabic XBank efforts focused mainly on newswire data, for GALE Phase 1 the focus shifts to broadcast news. These consist primarily of Modern Standard Arabic (MSA) text, but upwards of 5% of the data includes code-mixing from colloquial Arabic dialects. Using a spoken genre raises new challenges for LDC's traditional Arabic Treebank approach; for instance, ambiguities in Arabic orthography (especially with respect to case endings) will present difficulties for syntactic analysis. In response, LDC's Treebank annotation tool has been modified to allow annotators to access the audio that corresponds to the text file they are tagging. This approach has been shown to have negligible impact on annotation rates while allowing for more accurate Treebanking. In Phase 1, LDC plans to distribute a total of 30 hours of new Arabic Treebank from broadcast news programs recorded and transcribed for GALE.

The focus at LDC in Phase 1 for English is creation of the first parallel Arabic-English newswire Treebank. The data is drawn from the English translation of ATB Part 3 v 2.0 (LDC2005T20), which consists of 500K words of An-Nahar newswire. The data has already been sentence aligned and translated into English. The initial approach was established during production of a 52K-word pilot English/Arabic parallel Treebank prior to GALE's outset. That effort raised several issues requiring new approaches. For instance, the widespread use of clitics in Arabic leads to a word count in the English translation that is roughly 25% higher than Arabic source text. Additionally, there are translations errors as well as tokens that should be disregarded (e.g., extra determiners). Such items are marked with an X node in the tree, allowing annotators to move past them quickly.

In addition to providing English and Arabic Treebanks, LDC is working in collaboration with other GALE sites to produce the first large scale Arabic Propbank.

4. Data Coordination

Given the volume and complexity of the program's data needs and the necessary integration of previously distinct tasks, careful data management is essential. We must ensure that linguistic resources being developed are coordinated across areas, that effort is not duplicated, that the needs of multiple technology areas are met and that the allocation of effort matches the estimated benefit to the program. The data matrix[4] is the primary tool for data management, listing resources to be created for each technology area or annotation task, and describing provenance, characteristics, annotations, distribution schedules and access methods for each piece of data.

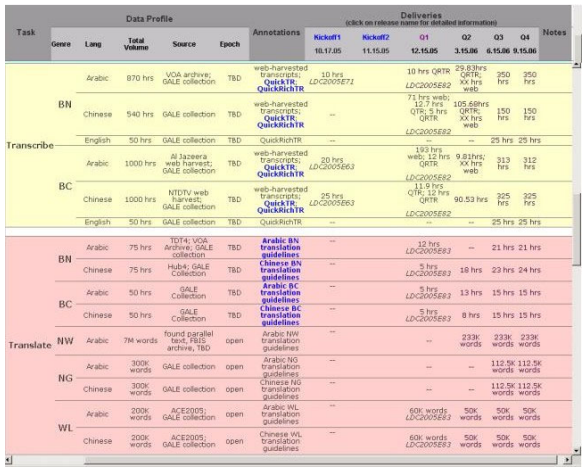

Figure 1: GALE Data Matrix

Links embedded in the matrix for each activity point to full task specifications which state needs and assumptions for each task, describe the process for collecting and/or selecting data for that task, define annotation and quality control procedures associated with the task and describe the distribution formats for the resulting data. Input to the data matrix is facilitated by a web interface and database backend, ensuring that the matrix is carefully maintained and updated regularly.

LDC also maintains mailing lists for each of the three tasks (Transcription, Translation and Distillation) as well as an announce-only list used to circulate information to GALE sites about current and upcoming data releases. We also provide a formal task specification for each GALE activity, which states needs and assumptions for the task, describes the process for collecting and/or selecting data for that task, defines annotation and quality control procedures associated with the task, and describes the distribution formats for the resulting data. Pointers to mailing lists, the data matrix and task specifications for each GALE task are linked to LDC's main GALE website[5], which serves as a clearinghouse for GALE-

[4] http://www.ldc.upenn.edu/Projects/GALE/data/DataMatrix.html

[5] http://www.ldc.upenn.edu/Projects/GALE

related information (even including a "Frequently Asked Questions" page).

5. Data Distribution

The majority of data created for GALE and described above is released to sites on a quarterly basis. Two "kickoff" releases were also scheduled early in Phase 1; and a limited number of ad hoc releases are added as required. There are various methods for data distribution in GALE depending on the size of the dataset. All methods require the participant be on the list of approved GALE sites provided by the program manager. At the program's outset, recipients were required to sign licenses that recognize the copyright of the original data providers and limits recipients' use of the data to linguistic education, research and technology development.

Individual GALE task managers at LDC perform considerable data validation before releasing data to sites. LDC's publications group also provides a series of final sanity checks prior to release, for instance scanning text data for tag and character encoding consistency, and auditing small samples of speech data for audio quality and signal clipping.

There are two types of data distributed to GALE sites: corpora created specifically for the program (i.e., the resources described above), and previously-released corpora that have been designated as GALE-relevant. New and planned GALE releases (kickoff and quarterly) are described on the GALE data matrix. Authorized GALE sites automatically receive copies of these releases once their appointed data contact person has signed the GALE user agreement and provided contact information. Previously-released corpora that have been designated as GALE-relevant are described on the GALE website, and authorized users may request copies of this data from LDC's membership office at any time.

GALE kickoff and quarterly releases are distributed in one of two formats. Text and annotations are distributed via web download. On the release date, each site's data contact person receives an email containing a URL where each corpus can be downloaded. Audio data is generally too large for web download and is therefore released on media. On the release date, LDC prepares a shipment containing CDs, DVDs and/or hard drives, depending on the size of the corpus. Packages are then shipped via DHL or FedEx. All authorized GALE sites receive text and annotation corpora (web downloads). Given the high cost of creating and shipping hard drives, audio data is not distributed to sites who are not participating in GALE evaluations.

In some cases, GALE researchers may need to redistribute LDC data (for instance, inline annotations of some LDC corpus) across sites, an activity which is prohibited by GALE user agreements and by LDC's agreements with our data providers. We have therefore also established a Secure Copy (SCP) Server to allow the data coordinator at each GALE site to upload and download data and work products (e.g., inline annotations) that contain data requiring LDC distribution. Each participant is a member of a group and can modify or share files with all members of their group. Groups are set up so that no group member can see the data from another group. For security reasons, access to the server is by means of public/private keypairs only.

As the linguistic resources described above are distributed to GALE program participants, LDC will wherever possible distribute the data more broadly, for example to its members and licensees, through the usual mechanisms. A small number of resources have been designated "For Official Use Only" and are available only to authorized GALE sites, for use only within the GALE program. In most cases, use of these resources has been negotiated directly by government data providers by the program manager and usage terms are non-negotiable. However, in most cases, the material created for GALE will be made generally available as regular LDC publications over time.

In addition to serving the primary goal of improved performance for GALE Transcription, Translation and Distillation engines, our efforts will lead to substantial corpora with durable value to the worldwide Human Language Technology community and the technology users who benefit from HLT development.

6. References

DARPA Information Processing Technology Office (2005). GALE Program Website. www.darpa.mil/ipto/programs/gale/.

Linguistic Data Consortium (2005). GALE Website. http://www.ldc.upenn.edu/Projects/GALE

Maamouri, M., Bies, A., Buckwalter, T., Jin, H. and Mekki, W. (2005). Arabic Treebank: Part 3 V2.0; LDC2005T20.

Maamouri, M. and Bies, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004.

Maeda, K., Lee, H., Medero, J. and Strassel, S. (2006). A New Phase in Annotation Tool Development at the Linguistic Data Consortium: The Evolution of the Annotation Graph Toolkit. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006).