Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development

Kevin Walker, Christopher Caruso, Denise DiPersio
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
E-mail: walkerk@ldc.upenn.edu, carusocr@ldc.upenn.edu, dipersio@ldc.upenn.edu

Abstract

The development of technologies to address machine translation and distillation of multilingual broadcast data depends heavily on the collection of large volumes of material from modern data providers. To address the needs of GALE researchers, the Linguistic Data Consortium (LDC) developed a system for collecting broadcast news and conversation from a variety of Arabic, Chinese and English broadcasters. The system is highly automated, easily extensible and robust and is capable of collecting, processing and evaluating hundreds of hours of content from several dozen sources per day. In addition to this extensive system, LDC manages three remote collection sites to maximize the variety of available broadcast data and has designed a portable broadcast collection platform to facilitate remote collection. This paper will present a detailed a description of the design and implementation of LDC's collection system, the technical challenges and solutions to large scale broadcast data collection efforts and an overview of the system's operation. This paper will also discuss the challenges of managing remote collections, in particular, the strategies used to normalize data formats, naming conventions and delivery methods to achieve optimal integration of remotely-collected data into LDC's collection database and downstream tasking workflow.

1. LDC's Local Collection

The GALE program[1] established annual collection goals of 1000 hours each of broadcast news (BN) and broadcast conversation (BC) in each of Arabic, Chinese and English. At the commencement of GALE, LDC was already collecting recordings (mostly in the BN genre) in the target languages for various projects, including the EARS program and the 2004-2005 TRECVID programs.[2] The technical challenges related to expanding the collection for GALE included integrating various collection modalities (satellite systems, satellite dishes, receivers); managing multiple audio/video streams as they are collected; routing those streams to their assigned system location; scheduling programs to begin and end as directed; accounting for simultaneous broadcasts from different sources; managing the recordings' processing rate; and making recordings promptly available for downstream tasks (including auditing, automatic speech recognition (ASR) and machine translation output and data selection).

LDC's pre-GALE broadcast collection consisted of the following sources: Aljazeera and Lebanese Broadcasting Corp. (LBC) for Arabic; China Central TV (CCTV), New Tang Dynasty TV (NTDTV) and Phoenix TV for Chinese; NBC/MSNBC and CNN for English; and Voice of America (VOA) for Arabic (via Radio Sawa), Chinese and English. For GALE Phase 1, LDC quickly determined that additional sources and hours could be accommodated in the existing collection infrastructure through cable sources. This was necessary because there were a limited number of receivers in the existing collection, and those receivers were deployed for cable sources. Additional programming from Ajazeera and a new source, Al Arabiya, were added for Arabic, focusing mainly on the BC genre. For Chinese, new BC programs for CCTV, NTDTV and Phoenix were selected.

LDC added additional receivers to increase the range of its Arabic collection in Phase 2, responding to sponsor requests for greater representation of programming across the Arabic-speaking region, particularly the Gulf region and Iraq. A number of Arabic sources were available from free-to-air (FTA) satellites transmitting over the Philadelphia area. LDC designed program surveys of the various sources that ran roughly twice per hour for several days. The collection manager, assisted by native speakers, reviewed the survey recordings and established a recording schedule. Unlike cable sources, the FTA sources do not typically maintain scheduling information. The new Arabic collection boasted thirteen sources with broad coverage including Iraq (Al Iraqiyah), various Emirate states (Abu Dhabi TV, Oman TV, Saudi TV), Iran (Al Alam News Channel) and Syria (Syria TV).

In early Phase 4, LDC added several regional Chinese sources in response to sponsor requests for an increased variety of that material. LDC upgraded its DISH Network

[1] "GALE" refers to the program sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) whose full name is Global Autonomous Language Exploitation. As of the writing of this paper, the program is currently in Phase 4 (Year 4).

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-003. The content of this paper does not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred.

[2] The EARS (Effective, Affordable, Reusable Speech-to-Text) program was sponsored by DARPA and was conducted from 2003 through 2005. TRECVID is the video retrieval evaluation of the TREC (Text REtrieval Conference) program sponsored by the US National Institute of Standards and Technology (NIST).

Chinese subscription to access more regional programming and added the necessary equipment to commence that collection. New sources included Beijing TV, Dragon TV, Fujian TV, Guangdong TV, Hunan TV and Jiangsu TV. As of the writing of this paper in late Phase 4, LDC is enhancing its Arabic broadcast conversation collection at the sponsor's request, targeting programs containing Iraqi and North African Arabic dialects.

Throughout the GALE program, LDC has continued to refine and normalize the local collection schedule. LDC currently collects approximately 205 hours/week of programming from 27 broadcast sources which break down by language as follows: Arabic (15 sources, 115 hours/week) Chinese (9 sources, 63 hours/week) and English (3 sources, 27 hours/week). The combined local and outsourced collection generates approximately 335 hours of programming weekly from 41 broadcast sources.

2. Broadcast Collection System Design and Operation

Part of the design intent driving the development of LDC's broadcast collection system was that it be modular and regularized. That meant that all of the recording nodes should be interchangeable, that filenames and database fields should follow consistent, formal rules and that signal interconnects should be consistent. The receivers feed into an audio/video (A/V) matrix switch so that any source can be routed to any receiver simply by changing an entry in the schedule. The audio/video streams are first digitized as DV25[3] video stream plus stereo audio (the common denominator) which can then be used to derive whatever container and compression method is required for a given project.

The broadcast material is served to the system by a set of FTA satellite receivers, commercial direct satellite systems (DSS) such as DirecTV, direct broadcast satellite (DBS) receivers, and cable television (CATV) feeds. The mapping between receivers and records is dynamic and modular; all signal routing is performed under computer control, using a 256x64 A/V matrix switch. Programs are recorded in a high bandwidth A/V format and are then processed to extract audio, to generate key frames and compressed audio/video, to produce time-synchronized closed captions (in the case of North American English) and to generate ASR output. The recordings and their extracted content are stored in a high performance and highly reliable storage solution which is accessible to LDC teams for auditing, data selection and annotation.

Initial recordings consist of video, stereo audio, and in case of English sources, closed captions. LDC collects both audio and video data for each recording so that this material can be reusable for a variety of research purposes and because having access to the video portion of a given broadcast aids troubleshooting system functions and makes auditing more reliable, more efficient and less error-prone. Recordings are typically transcoded to MPEG-4/AVC at 1 Mbps shortly after capture.

The collection system is illustrated in the block diagram below:

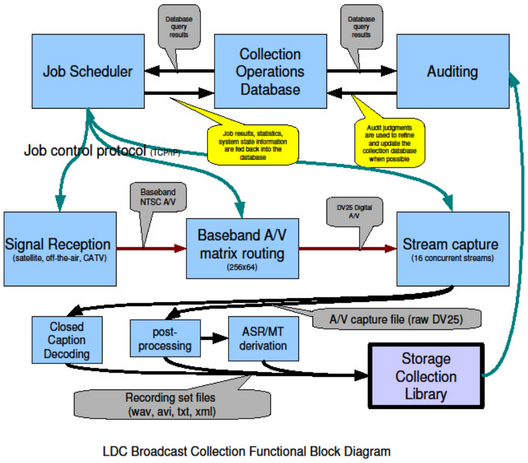
Database query results
Database query results
Job Scheduler
Collection Operations Database
Auditing
Job results, statistics, system state information are fed back into the database
Audit judgments are used to refine and update the collection database when possible
Job control protocol (TCP/IP)
Baseband NTSC A/V
DV25 Digital A/V
Signal Reception (satellite, off-the-air, CATV)
Baseband A/V matrix routing (256x64)
Stream capture (16 concurrent streams)
A/V capture file (raw DV25)
Closed Caption Decoding
post-processing
ASR/MT derivation
Storage Collection Library
Recording set files (wav, avi, txt, xml)
LDC Broadcast Collection Functional Block Diagram

Figure 1: LDC Broadcast Collection System Block Diagram

2.1 Collection Database

The collection schedule is stored in a relational database using a Mysql database server. The database contains a history of all of the recordings that have been made; it has configuration and status information for all recorders; it has information about all receivers and associates specific programs of interest with the appropriate receiver; it contains a schedule of all recording jobs that need to be executed and their status; and it stores all audit judgments associated with a given recording.

The collection database consists of the following tables:

bn_source: A "source" refers to a specific content creator associated with a specific reception mode. For example, "CNN Headline News" received via DirecTV receiver 01 would constitute a specific source. Each source has an associated input identifier, which refers to an input port in the routing context of the system.

bn_program: This belongs to a source, has a language and a typical day/time of airing.

bn_recdev: A hardware resource, identified by IP address and device ID. It has a specific output identifier which

[3] "DV25" refers to IEC Digital Video @ 25Mps, SMPTE-314M, IEC-61834.

refers to an output port in the routing context of the system.

bn_sched: Associates a recorder with a program.

bn_recordings: Each recording has an entry in this table with creation date and time, associated filenames with md5 checksums, auditing summary information and other metadata.

bn_audit: Human judgments about the content and signal quality for multiple subsections of each recording are stored in this table. Auditors review recordings for correct language, genre and program and also note technical problems and degraded signal conditions.

2.2 Computer Hardware Schema

The broadcast collection system is organized as a mini-cluster of linux computers with a master node, a fileserver, eight client recording nodes and three transcoding nodes. The master node runs the database server and dispatches recording jobs according to the defined schedule. The client nodes are functionally equivalent. Each node has two Canopus ACEDVio digitizers which appear as firewire interfaces to the node's operating system. The master scheduler on the control computer is called **bn_run_now**. This script actively monitors the database for any scheduled jobs, reserves recording devices and sends job command strings to the appropriate listen ports on client recording nodes.

xinetd runs on client nodes and waits for connections from the scheduler, then starts an instance of **bncap.pl**. This script runs on every recording node, and upon signal from the scheduler, begins raw video capture with the **dvgrab** utility. Once video capture is complete, **bncap.pl** extracts and downsamples the audio channels into separate A and B channels with **SoX** and uses **mencoder** to convert the video into .avi format. The files are uploaded to the collections server, after which **bncap.pl** connects to the scheduler database and resets the program and recording device values to inactive. In addition, the script inserts recording information into the **bn_recordings** table.

The system's hardware configuration is as follows:

Control Computer x1 (thalia)

File Server/Compute Server x1 (kronus)

Mulitprocessor AMD Opteron Servers (Tyan Transport)

7TB Storage running Ubuntu Linux

Recording Nodes (8)

Dual Xeon Workstation (Dell Precision 450)

Ubuntu LinuxTwo ADS DVio Firewire Cards (analog input, digitized to DV25)

Transcoding Nodes (3)

ASR/Indexing Nodes

2.3 Signal Reception

Three satellite dishes provide access to C-Band and KU-Band broadcasts, including international programming from Galaxy-19 and Galaxy IIC via Globecast, SCOLA[4], Phoenix TV and CCTV North America. The equipment consists of one 3 meter satellite dish, C/KU band; one 3.7 meter satellite dish, C/KU band; and one 1.2 meter satellite dish C/KU band.

Satellite signal reception is handled by a set of digital and analog satellite receivers. Twelve conditional access receivers are used for cable and direct satellite broadcasts (i.e., DISH Network, DirectTV and SCOLA programming). FTA programming is handled by Coship, Pansat and Coolsat MPEG-2 DVB-S receivers that tune to a specified transponder and decode the digital transmission into analog video and audio. In order to access local CATV programming, a set of RS-232 controllable CATV demodulators are also included in the system (Contemporary Research 232-STS Stereo TV Tuner). The system relies on CATV service from the University of Pennsylvania for access to local programming from US broadcasters NBC, CBS, ABC and CNN.

2.4 Signal Routing

One aspect of the modular design of the collection system is the ability to route any input signal stream to any recording device. To accomplish this, LDC uses a computer controllable Knox Chameleon RS256X64 A/V matrix switch. The switch has 256 inputs and 64 outputs, and a single input signal can be distributed to multiple outputs simultaneously. Each connection (input and output) consists of s-video and stereo audio.

[4] SCOLA is a non-profit educational organization that receives and re-transmits in their native languages television broadcasts aired around the world. LDC's collection includes SCOLA retransmissions for Arabic programming from Dubai TV, Saudi TV, Nile TV (Egypt) and Al Ordiniyah (Jordan).

Control commands are transmitted to the A/V matrix switch over RS-232. All control signals originate as TCP/IP at the master computer node and are translated to RS-232/422/485 by the Comtrol Devicemaster RTS.

The routing schema is as follows:

Closed-Caption Decoder x16

Softtouch MagHubcap Closed Captions decoder (RS-232)

Ethernet to RS-232 Bridge x2

Comtrol Devicemaster RTS

The ethernet to RS-232 bridge allows connections of up to 32 serial devices. The input/output of each device is assigned to a specific port, and connections to the port can be made using a utility such as **netcat** or **telnet**. In practice a custom perl script is prepared for each type of device that implements the particular device's control protocol. For example, the Knox Chameleon AV Matrix Switch can receive routing instructions by sending an ASCII string. The string "B1601/r/n" would route both audio and video from input 01 to output 16.

2.5 Automatic Speech Recognition (ASR) Technology

LDC has integrated three ASR client systems into its daily collection process for the duration of the GALE program. They were developed by GALE research sites BBN Technologies (BBN), International Business Machines (IBM) and SRI International (SRI). ASR output is automatically generated on the BBN and IBM systems for all locally-collected Arabic and Chinese audio data. The text output is used for downstream data selection. Audio data from LDC's remote broadcast collection sites is not processed daily, but is run as needed as part of the data selection process. Data is processed on the SRI system by request.

The IBM and SRI applications are installed on local machines at LDC and administered by LDC staff. Data is processed on the BBN system via a connection to remote servers which necessitates an additional file anonymization step.

Locally-collected broadcasts are automatically pooled onto a centralized server as they are recorded and processed. This server then supplies the extracted audio portions to the different ASR systems which run daily cronjobs to generate ASR output for the previous day's broadcasts.

Since each system has different requirements, wrapper scripts were created by LDC to preprocess the input .wav audio file, to run the actual ASR application on each file and to perform additional postprocessing and formatting before copying the output back to the collections server.

2.6 Utilities

LDC's broadcast collection system relies heavily on open source software created by generous software developers around the world. In particular, audio/video signal collection, transcoding, manipulation and metadata management use software from the following projects:

dvgrab (Version 1.8)[5]: This is used to capture the DV25 signal from the Canopus ACEDVio converter direct to disk and to extract keyframes from DV25 recordings. These critical steps serve as the backbone of the recording nodes' collection procedure.

mencoder (Version 1.0rc1-3.4.5): This is a command line video decoding used to convert the raw DV25 data into MPEG-4 compressed .avi files and to extract the audio tracks.

x264 (Version 10rc1-3.4.5): Built into **mencoder**, x264 encodes and decodes (compresses and decompresses) a digital signal into a specific format (-H.264 in this case). This is essentially a variety of MPEG-4 and can be shared with other programs. **mencoder** is the front-end application that uses the codec's programs and rules to determine how the video/audio stream is encoded.

FFmpeg (Version SVN-r9282): This is a complete cross-platform solution to record, convert and stream audio and video. It includes **libavcodec**, an audio/video codec library.

SoX (Version 12.17.9 (recnodes); Version 14.3.0 (thalia, kronus)): This cross-platform utility converts various formats of computer audio files into other formats, applies a range of effects to these files, and plays and records audio files on most platforms.

rsync (Version 3.0.05): This utility provides fast incremental file transfer. In the collection system, it transfers the recordings from the recording nodes to the file server and checks the files as they are transferred.

3. LDC's Outsourced/Remote Collection

GALE's goal to create end-to-end language systems meant that the broadcast collection should incorporate programming from a variety of sources. LDC understood that even with the most efficient collection system, it

[5] All software versions in this section are as of February 2010.

would still be necessary to establish remote collections to assure the availability of data that represented a range of contemporary sources in each language. Accordingly, LDC obtained data collection assistance from several groups located within geographical areas of interest with access to targeted satellite and local programming. Those sites are Hong Kong University of Science and Technology (HKUST), Hong Kong, for Mandarin Chinese broadcasts from Mainland China and Voice of America (VOA) Mandarin broadcasts;[6] Medianet, a video and internet company in Tunis, Tunisia, for regional Arabic broadcasts; and MTC, a language resource creation facility in Rabat, Morocco for Arabic VOA-style programming. [7]

3.1 Technical Challenges

LDC and its collection partners agreed that data collected remotely should be compatible with LDC's recording and formatting requirements and that data delivery should be accomplished quickly and efficiently so that outsourced data could be optimally integrated into the GALE downstream workflow. Remotely-collected files typically follow GALE-specific naming conventions and LDC-preferred file formats with some variations by site. LDC's collection programmers perform further semi-automatic postprocessing on each delivery, such as renaming/translating program names, creating audio .wav files from .avi/.mpeg deliveries and expanding compressed files. The principal challenges of managing the remote collection include enforcing file standardization, automating the delivery and post processing steps and ensuring that remotely-collected recordings enter the data pipeline with contemporaneous locally-collected materials.

HKUST collects audio and video for each recording, but only delivers audio .wav files to LDC via FTP server, retaining the video at its facility. LDC established a file naming convention for HKUST data after the master naming convention developed for the GALE program: date_recordingtime_source_programID. The program ID is typically in pinyin. LDC assigned English names for each pinyin program and as part of its processing of HKUST files, LDC translates the pinyin names into English.

Medianet collects audio and video for each recording and delivers both audio and video to LDC via FTP server. Medianet initially recorded all programs in MPEG-2 format which meant that uploading and downloading deliveries were time-consuming and impacted LDC's storage capacity. LDC and Medianet agreed that Medianet would use video compression software (**Blaze Media Pro**) to convert MPEG-2 files to MPEG-4 files at a standardized 1.25 kb/s bitrate. Medianet uses transliterated Arabic names for programIDs, but the spellings often vary across deliveries. LDC developed scripts to check alternate spellings and to standardize program names and also developed a master name list to be implemented in the collection schedule.

MTC adopted the GALE file name convention and generally uses English names for the programIDs. It delivers audio files in .wav format to LDC via FTP server.

3.2 Content Selection

LDC collaborated with its remote partners on source and program selection. Generally, the sites gathered sample recordings of potential programs for LDC's review. The parties then developed a schedule to meet collection targets.

HKUST has collected, and continues to collect, broadcast programming from CCTV, VOA and regional sources Anhui TV, Hubei TV, Beijing TV and Jiangsu TV.[8] LDC's broadcast collection programmer designed a portable collection platform that was deployed at HKUST's collection facility in 2006 and which is used primarily for CCTV programming.

Medianet's collection reflects the vast range of Arabic language programming broadcast across the Middle East. Sources from this collection include Al Baghdadya, Al Iraqiyah, Al Fayhaa, Al Forat, Al Sharqiya (Iraq); Al Ordiniyah (Jordan); Al Hiwar (UK); Oman TV, Kuwait TV, Abu Dhabi TV, Dubai TV, Bahrain TV, Qatar TV (Emirates); Al Manar (Lebanon); Palestine Satellite Channel; Saudi TV; and Tunis TV.

In addition to Arabic VOA-style programming (i.e., Alhurra and Radio Sawa), MTC collects broadcasts from Al Baghdadya, Yemen TV, and Moroccan sources Al Maghribia and Arabiaa.

4. Broadcast Auditing

All broadcast programming collected for GALE by LDC and by the remote collection sites managed by LDC are manually audited by Arabic, Chinese and English speakers for language, program and quality. The

[6] LDC had collected multilingual data from various Voice of America (VOA) broadcasts under a special arrangement with the U.S. government from 1996 through 2006. LDC lost the ability to collect that data when the U.S. government removed VOA channels transmitting from DOMSAT (a satellite transmitting over the United States).

[7] MTC is affiliated with the European Language Resources Association (ELRA).

[8] As of the writing of this paper, Beijing TV is the only regional source collected by HKUST. HKUST continues to collect CCTV and VOA broadcasts.

broadcast auditing process serves three principal goals: as a check on the operation of LDC's broadcast collection system equipment by identifying failed, incomplete or faulty recordings; as an indicator of broadcast schedule changes by detecting instances where the incorrect program was recorded; and as a guide for data selection by retaining information about a program's genre, data type and topic.

LDC has developed a tool that it uses to audit its local collection. Each remote collection site uses a form of audit procedure based on the LDC model.

4.1 LDC Broadcast Audit Tool

LDC's broadcast audit tool collects files from the servers where they are stored into a list that can be filtered by date, source, program, language and status (null (not audited), audited, auditing, problem). Once presented with the program list, auditors choose the program to be audited. They then listen to randomly selected thirty second sound bites from the beginning (3-8 minutes from the beginning), end (3-8 minutes from the end) and middle of each recording. If auditors cannot make the required judgments from any one of those segments, they have the option to listen to additional portions of a segment and/or to choose different segments from the same random time range.

Auditors answer the following questions for each discrete program segment:

(1) Is there a recording? Auditors choose yes, no or "partial" from a drop-down list. Auditors can enable a text box to add a comment about their decision.

(2) Is the audio quality ok? Auditors choose yes, no or "partial" from a drop-down list. Auditors can enable a text box to add a comment about their decision.

(3) What is the language? Auditors choose from a drop-down list of Mandarin Chinese, Modern Standard Arabic, Other (colloquial) Arabic, English and Other. Auditors can enable a text box to add a comment about their decision.

(4) Is it speech from the right program? Auditors choose yes or no from a drop-down list, and they can enable a text box to add a comment about their decision.

(5) What is the data type? Auditors use check boxes to choose their answers(s) from the following list: talking head; interview; call-in; roundtable; other hard news; other conversational; commercial, music, etc.; field report; in-depth report. The tool allows auditors to choose more than one data type for a particular segment if applicable.

(6) What is the topic? Auditors use check boxes to choose their answers(s) from the following list: current events; human interest; sports report; weather report; celebrity; other. The tool allows auditors to choose more than one data type for a particular segment if applicable. If auditors choose "other", they can enable a text box to add a comment about their decision.

The auditing information is stored in the **bn_audit** table in the collection database. A question-answer pair is stored for each segment of each recording. Each time a decision is made, the username of the current auditor as well as the current timestamp is also recorded.

4.2 Audit Procedures at Remote Collection Sites

Each remote collection site audits the recordings it collects for language, program and quality using procedures based on the LDC auditing model, and each site delivers to LDC only those files that have passed the applicable audit procedure.

HKUST and MTC generate English-language .xml and .html audit reports for the Chinese and Arabic programming they collect, respectively. Those reports contain auditors' judgments from three portions of each program (beginning, middle and end), including whether a recording occurred, the audio quality, language, whether the correct program was recorded, the data type and topic.

Medianet generates English-language .xls reports for the Arabic programming it collects. Those reports contain one set of auditors' judgments for an entire program including audio quality; genre; data format; percentage of Modern Standard Arabic; dialect type and percentage; topic; and comments.

5. Portable Broadcast Collection Platform

LDC's portable broadcast collection platform is a TiVO-style digital video recording (DVR) system that records two streams of A/V material simultaneously. It supports analog CATV (NTSC and PAL) and FTA DVB-S satellite programming and can operate outside of the United States. It has a small footprint and can be transported as carry-on luggage.

The portable platform and the LDC collection system share the same code base and rely on a modular, unified hardware specification. Improvements in the main collection platform therefore translate into benefits for the portable platform.

The portable system runs Ubuntu linux, using a WinTV-PVR-500 for analog cable and a Technotrend Premium S-2300 PCI DVB-S receiver for DVB satellite reception. **dvbstream** is used for satellite recording, and **ivtv** is used for cable recording.

The portable platform deployed at HKUST's Chinese collection facility collects multiple streams of CCTV programming and is maintained by HKUST technical staff. The platform deployed in Tunisia is maintained by the LDC collections programmer via remote login. Recordings are scheduled at LDC and automatically downloaded into LDC's collection server. In each case, LDC is able to collect high-quality broadcast data with minimal equipment and in the case of data collected in Tunisia, to receive that data immediately.

6. Conclusion

LDC's broadcast collection system represents a significant achievement in delivering volumes of high-quality broadcast data from multiple programming sources and geographic locations. The system has performed impressively. To date, LDC has delivered close to 15,000 hours of broadcast audio to GALE sites. Moreover, because it is robust, flexible and extensible, the system can be quickly deployed for virtually any type of broadcast collection.

7. References

Blaze Media Pro (2010). http://www.blazemp.com/.

DVB tools (2010). http://sourceforge.net/projects/dvbtools/

dvgrab (1) (2010). http://ccrma.stanford.edu/planetccrma/man/man1/dvgrab.1.html

FFmpeg (2010). http://ffmpeg.org/

ivtv (2010). http://ivtv.org/

Linguistic Data Consortium (2010). GALE: Task Specifications and Annotation Guidelines. Audit Procedure Specification Version 2.0 http://projects.ldc.upenn.edu/gale/task_specifications/Audit_Procedure_Specificationv2.0.pdf.

Linguistic Data Consortium (2010). GALE: Task Specifications and Annotation Guidelines. Guidelines for Broadcast Audio Collection Version 2.0 http://projects.ldc.upenn.edu/gale/task_specifications/Collection_Task_Specificationsv2.0.pdf.

MPlayer/Mencoder (2010). http://www.videohelp.com/tools/MPlayer

rsync (2010). http://samba.anu.edu.au/rsync/

SCOLA (2010). http://www.scola.org/Scola/Default.aspx

Sox - Sound eXchange | HomePage (2010). http://sox.sourceforge.net/

telnet.org (2010). http://www.telnet.org/

The GNU Netcat project (2010). http://netcat.sourceforge.net/

x264 (2010). http://x264.nl/