# Annotation Tool Development for Large-Scale Corpus Creation Projects at the Linguistic Data Consortium

**Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, Stephanie Strassel**

Linguistic Data Consortium
University of Pennsylvania
3600 Market St., Suite 810
Philadelphia PA, 19104 USA
{maeda, haejoong, smedero, jmedero, parkerrl, strassel}@ldc.upenn.edu

**Abstract**

The Linguistic Data Consortium (LDC) creates a variety of linguistic resources – data, annotations, tools, standards and best practices – for many sponsored projects. The programming staff at LDC has created the tools and technical infrastructures to support the data creation efforts for these projects, creating tools and technical infrastructures for all aspects of data creation projects: data scouting, data collection, data selection, annotation, search, data tracking and workflow management. This paper introduces a number of samples of LDC programming staff's work, with particular focus on the recent additions and updates to the suite of software tools developed by LDC. Tools introduced include the GScout Web Data Scouting Tool, LDC Data Selection Toolkit, ACK - Annotation Collection Kit, XTrans Transcription and Speech Annotation Tool, GALE Distillation Toolkit, and the GALE MT Post Editing Workflow Management System.

## 1. Introduction

The Linguistic Data Consortium (LDC) creates a variety of linguistic resources – data, annotations, tools, standards and best practices – for many sponsored projects. One of our current projects is the DARPA GALE[1] program (Strassel et al., 2006). 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. LDC supports the GALE program by providing linguistic resources for system training, development and evaluation. Another of our recent project is the REFLEX-LCTL (Research on English and Foreign Language Exploitation) program. REFLEX-LCTL is a medium-scale effort in simultaneous creation of basic language resources for several less commonly taught languages (LCTLs) (Simpson et al., 2008).

The programming staff at LDC has created the tools and technical infrastructures to support the data creation efforts for both these programs as well as all other LDC projects. The majority of the annotated data was created with highly customized annotation tools (Maeda et al., 2006; Maeda and Strassel, 2004; Bird et al., 2002). In addition to annotation tools, LDC's programming staff creates tools and technical infrastructures for all aspects of data creation projects: data scouting, data collection, data selection, annotation, search, data tracking and workflow management.

This paper introduces a number of samples of LDC programming staff's work, with particular focus on the recent additions and updates to the suite of software tools developed by LDC. In the following sections, we will present samples of tools/systems created for various aspects of data creation projects, from data scouting, to annotation, to workflow management.

- Data Scouting: *GScout Web Data Scouting Tool*
- Data Selection: *LDC Data Selection Toolkit*
- Flexible Annotation/Judgment tool: *ACK - Annotation Collection Kit*
- Speech Transcription and Annotation: *XTrans Transcription and Speech Annotation Tool*
- Document Search and Annotation: *GALE Distillation Toolkit*
- Workflow Management System: *GALE MT Post Editing Workflow Management System*

## 2. GScout Web Data Scouting Tool

The GALE program requires multiple types of data, including weblogs and newsgroup postings, newswire articles, broadcast conversation audio, and broadcast news audio. In order to harvest weblog and newsgroup postings appropriate for the GALE evaluation tasks, annotators are tasked to scout for weblogs and newsgroups containing contents appropriate for for the project.

The *GScout* Web Data Scouting tool was developed for this purpose, and consists of three components: the user interface written in the XML User Interface Language (XUL) and Java Script; the server-side scripts written in PHP; and the mySQL database which stores the scouting results.

This tool is launched from a web browser equipped with Gecko, a layout engine supporting XUL. The annotator logs onto the system using their login name and password. The annotator then receives a topic assignment, and starts to search for weblogs and newsgroups that are suitable for this topic, using any methods they would normally use when looking for web sites. The *GScout* interface remains in the browser window as a sidebar. Once the annotator finds a web page that may be related to the given topic, the annotator records various judgments about it, and saves some quoted text from the page.

[1] Global Autonomous Language Exploitation (http://projects.ldc.upenn.edu/gale/)

The collected lists of weblogs and newsgroups are then added to our regular web data harvesting processes. Postings from these weblogs and newsgroups are regularly harvested and formatted into a predefined SGML format.
Figure 1 shows a screenshot of the GScout tool shown within the FireFox web browser.

## 3. LDC Data Selection Toolkit

In creating manual transcriptions and translations, it is very important to select the right source materials. Requirements for content vary among projects, but normally certain documents, such as sports scores, weather forecasts, advertisements and horoscopes are not in the target document types and should be excluded from the translation set. Other factors that should be checked are sound quality and background noises (for speech), file formatting or encoding problems (for text), dialects and writing systems (e.g., romanization vs. native script, traditional vs. simplified, etc.).
During the initial phase of GALE, LDC used a simple GUI-based data selection tool for making judgments about documents for text data. For GALE Phase 2 Evaluation, LDC needed to develop a new tool that allows annotators to select snippets of text or audio that are suitable as evaluation materials. LDC designed, developed and implemented a new tool that allows annotators to select a snippet of text or audio from a GUI interface. If the source data is audio, corresponding ASR output is also displayed in order to facilitate faster scanning of the contents.
This tool uses the QWave module that was developed for displaying and playing back audio files in XTrans. The tool also includes a text display and a panel for storing additional information about the selected snippets. The selection results are stored in an XML format.

## 4. ACK - Annotation Collection Kit

ACK, the Annotation Collection Kit, is a new addition to LDC's suite of annotation tools. This is a web-based system which is implemented using PHP, CodeIgniter, and MySql. This tool allows software developers, project managers and project members to quickly develop annotation kits for variety of annotation tasks. Questions to be answered – or items to be annotated – are stored in CSV, comma separated value, files. Widgets for annotations, such as text boxes and radio buttons, are stored with the entries. The following is an example of question csv files.

```
qid,qtype,qlabel,qleadin,qvalues,qcomment
1,radio,pataan,"choose pos",noun::pron::adv:...
2,radio,pagsasalita,"choose pos",noun::pron:...
3,radio,paldas,"choose pos",noun::pron::adv:...
...
```

This tool was used extensively for the Less-Commonly Taught Languages (LCTL) project[2] (Simpson et al., 2008). For LCTL, LDC has created resources such as lexicons, annotated text, named entity taggers, part-of-speech taggers and morphological analyzers for multiple diverse languages such as Bengali, Berber, Panjabi, Pashto, Tagalog, Tamil, Thai, Tigrinya, Urdu, and Yoruba. ACK was extremely useful in an environment where we had to ask for native speakers judgments for a variety of tasks in short timelines. It also allowed remote native speaker annotators to complete annotation task via the Internet.
As of the writing, more than 700 annotation kits have been created and more than 800,000 annotations have been collected.
Figure 2 shows a screenshot of ACK.

## 5. XTrans Transcription and Speech Annotation Tool

XTrans is LDC's ne*X*t generation *Trans*cription tool developed by LDC's programming staff. XTrans is a transcription tool that allows transcription of overlapping speech, multiple speakers and multiple channels. During the past two years, LDC has used this tool extensively for creating transcripts and annotated speech data for projects such as DARPA GALE, Mixer[3], NIST RT (Rich Transcripts) Meeting Recognition Evaluation[4], and Phanotics (Phonetic Annotation of Typicality in Conversational Speech - a forensic speech research sponsored by the US Department of Homeland Security and the US Secret Service), and made improvements to the tools based on feedback from the transcribers/annotators. Thousands of hours of speech has been transcribed for these projects using XTrans.
In addition to LDC's in-house transcribers, the XTrans tool was distributed to the transcription agencies who were tasked to work on LDC's outsourced transcription effort for GALE. XTrans directly saves files in the TDF (tab-delimited fields) format used in the GALE Transcription task, and there was no need to convert the file format. The XTrans tool was particularly useful for creating the Quick Rich Transcripts (LDC, 2006), which includes the semantic-unit (SU) type annotations.
Variants of this tools were also created by adding components to the XTrans tool. The first variant is called *SU-Trans*, and was developed for performing manual sentence segmentation of text, or correction of the output from an automatic segmentation tool for translation purposes. Weblog and newsgroup postings are often written without proper punctuation, and incomplete sentences are often observed. Automatic segmentation tools are often not sufficient for preparing the data for sentence-aligned translations.
The second variant of XTrans is called *QCTrans*. This tool was developed for checking and correcting manual translations created by translation agencies. *QCTrans* allows the user to view and compare multiple editions of the translations sentence by sentence. The specifications for this tool were developed by LDC's translation task manager and translation team members in order to include all required functionalities and maximize its usability for the task.
Figure 3 shows a screenshot of XTrans.

## 6. GALE Distillation Toolkit

GALE Distillation Annotation tool was developed for creating training data for the DARPA GALE Distillation task (LDC, 2007).

[2] http://projects.ldc.upenn.edu/lctl

[3] http://projects.ldc.upenn.edu/mixer/

[4] http://www.nist.gov/speech/test_beds/mr_proj/

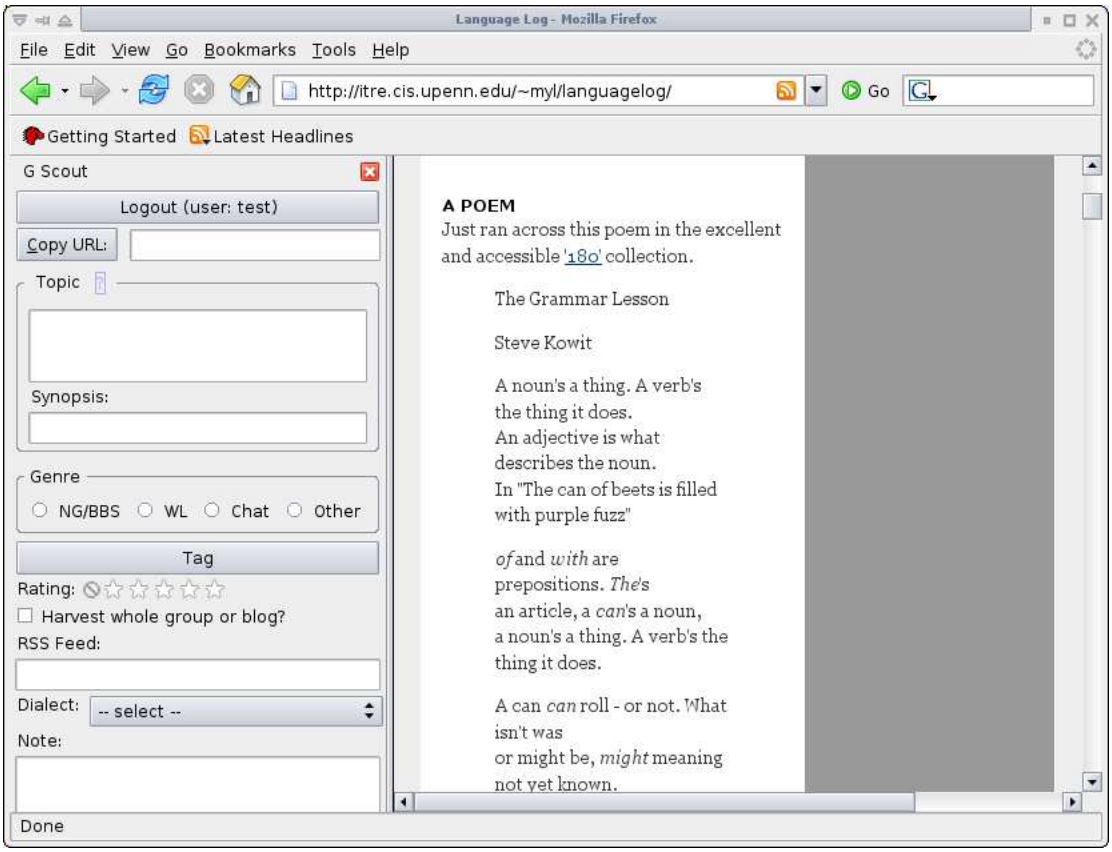


Figure 1: GScout Web Data Scouting Tool

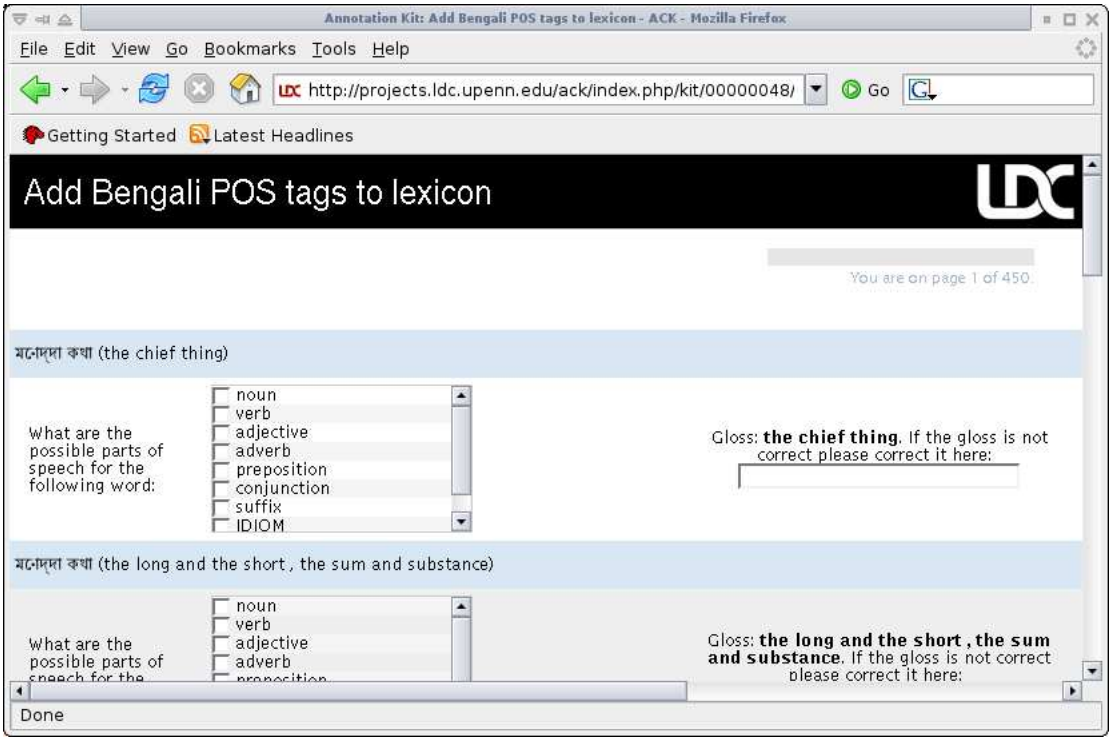


Figure 2: ACK - Annotation Collection Kit

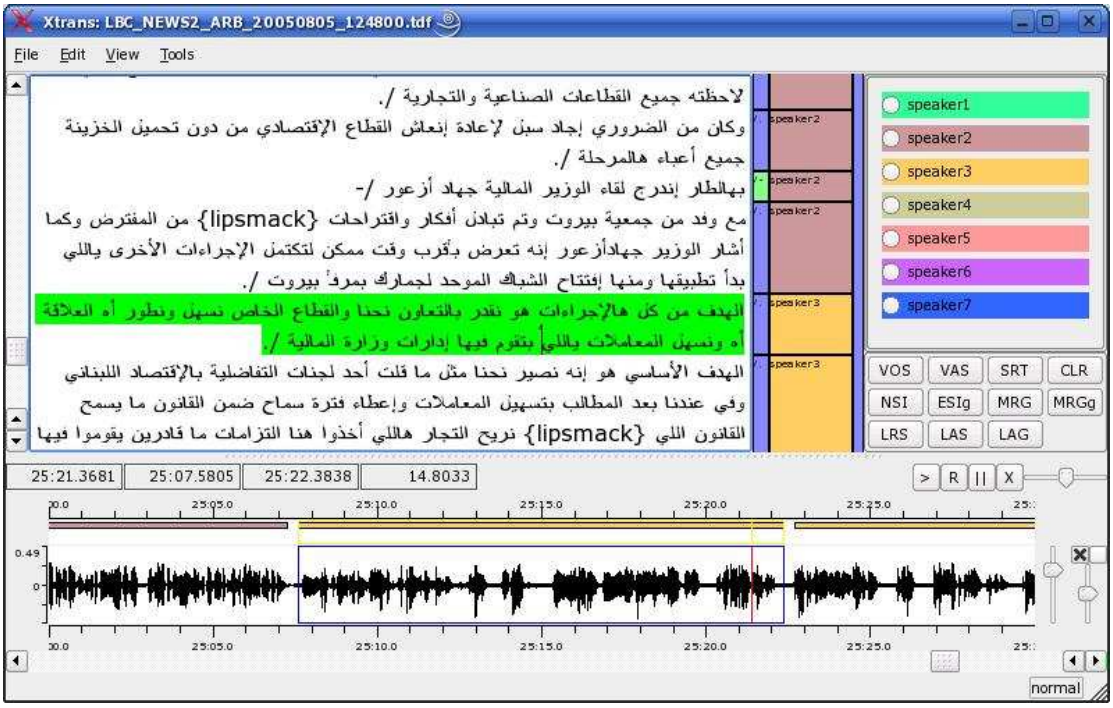


Figure 3: XTrans

The distillation annotation task involves responding to a series of user queries. For each query, annotators first identify snippets of text which contains relevant information in the Arabic, Chinese or English source document. For each snippet, they create a *nugget*, which is a fact or a statement extracted from the snippet.

The annotation task is done using LDC's GALE Distillation Annotation Tool. The query module of the tool uses a search engine that was created by LDC. This search engine is already used by the LDC Online[5], and other projects. The tool is implemented in Python, using Qt and PyQt; the annotations are saved in a MySQL database. The snippets and nuggets are dumped into an XML file format for data deliveries.

## 7. GALE MT Post Editing Workflow Management System

One of LDC's important tasks for the GALE MT Evaluation is the MT post editing – this is a process where skilled editors edit the MT output so that it has the *same meaning* as the gold-standard human translation and is *understandable*. The system output is then automatically scored in terms of edit distance from the gold-standard translation.

LDC has developed a web-based workflow management system for managing task assignments for the post editors who are working outside of the LDC facility. The system managed the assignment of *kits*, sets of gold-standard translations and MT output, to post editors.

Each post editor has an account, and logs onto their account to get their assigned kit. The editor works on the kit, and submits the kit through the workflow management interface. The editor has an option of marking the kit as "broken", indicating that there were some technical problems with the kit. Quality control (QC) measures are applied upon check-in and kits that did not pass the QC are returned to the editors. The scorer automatically runs on the checked-in kits, and the scores are saved.

The editors are divided into the first-pass editors and the second-pass editors, who have more experience than the first-pass editors. The roles of the first-pass and second-pass editors are defined in the guidelines. The system provides methods to communicate between the first-pass editor and the second-pass editor who work on the same kit. There is a comment function and there is a *nudge* function which sends a reminder to the first-pass editor that the second-pass editor is waiting for the kit to be first-passed.

This system was initially created for the GALE Phase 1 Evaluation, and then was completely redesigned and reimplemented for GALE Phase 2 Evaluation. The workflow and detailed functionalities for the second version were designed by the developer and the project manager, with input from other project members.

## 8. Summary

LDC creates linguistic data resources for large-scale projects, such as DARPA GALE, LCTL, NIST RT, NIST Open MT, Mixer and Phanotics. LDC's research programming staff has created an extensive array of customized, and generalized annotation tools for supporting the corpus creation effort for these projects. Most of the tools introduced in this paper are, or will be in the near future, available as open-source software to the research community via our source distribution web site[6].

[5] http://online.ldc.upenn.edu

[6] http://tools.ldc.upenn.edu

## 9. Acknowledgment

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## 10. References

Steven Bird, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, Beth Randall, and Salim Zayat. 2002. TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse tools built on the annotation graph toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

LDC. 2006. English broadcast QRTR guidelines. `http://projects.ldc.upenn.edu/gale/Transcription/`.

LDC. 2007. Distillation training data annotation guidelines v2.3. `http://projects.ldc.upenn.edu/gale/Distillation/`.

Kazuaki Maeda and Stephanie Strassel. 2004. Annotation tools for large-scale corpus development: Using AGTK at the Linguistic Data Consortium. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Kazuaki Maeda, Haejoong Lee, Julie Medero, and Stephanie Strassel. 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*.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. In *Proceedings of LREC 2008 Workshop - Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages*.

Stephanie Strassel, Christopher Cieri, Andy Cole, Denise DiPersio, Mark Liberman, Xiaoyi Ma, Mohamed Maamouri, and Kazuaki Maeda. 2006. Integrated linguistic resources for language exploitation technologies. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.