A New Phase in Annotation Tool Development at the Linguistic Data Consortium: The Evolution of the Annotation Graph Toolkit

Kazuaki Maeda, Haejoong Lee, Julie Medero, Stephanie Strassel

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

Abstract

The Linguistic Data Consortium (LDC) has created various annotated linguistic data for a variety of common task evaluation programs and projects to create shared linguistic resources. The majority of these annotated linguistic data were created with highly customized annotation tools developed at LDC. The Annotation Graph Toolkit (AGTK) has been used as a primary infrastructure for annotation tool development at LDC in recent years. Thanks to the direct feedback from annotation task designers and annotators in-house, annotation tool development at LDC has entered a new, more mature and productive phase. This paper describes recent additions to LDC's annotation tools that are newly developed or significantly improved since our last report at the Fourth International Conference on Language Resource and Evaluation Conference in 2004. These tools are either directly based on AGTK or share a common philosophy with other AGTK tools.

1. Introduction

The Linguistic Data Consortium (LDC) has created annotated linguistic data for various common task evaluation programs as well as for projects to created shared language resources, including DARPA TIDES (Translingual Information Detection, Extraction and Summarization)[1], DARPA EARS (Effective, Affordable, Reusable Speech-to-Text)[2], DARPA GALE (Global Autonomous Language Exploitation)[3], ACE (Automatic Content Extraction)[4] and LCTL (Less Commonly Taught Languages)[5]. These projects required annotation tools that were highly customized for the specific and evolving annotation specifications, and for the maximum efficiency of the manual annotation process. The software developers at LDC have created annotation tools using a common infrastructure, and a common philosophy.

2. History of Annotation Tool Development at LDC

2.1. Annotation tool development in mid to late 1990's at LDC

In the mid to late 1990's, LDC added the creation and distribution of annotated linguistic data to its primary goals. Early annotation tools created at LDC utilized existing software tools, such as the GNU Emacs/Mule multilingual editor and the Entropic ESPS/Waves tools. The Emacs/Mule editor was highly configurable, and allowed customized key-bindings, macros and other functionalities via its Emacs Lisp programming interface. The ESPS/Waves toolkit had an interprocess communication capability, which was useful in annotation tools. Also used during this time period was the Tcl/Tk scripting language, which allowed the rapid prototyping of annotation tools that required graphical user interfaces.

2.2. The Annotation Graph Toolkit (AGTK)

In the early 2000's, the Annotation Graph Toolkit (AGTK) was created. The main component of AGTK is the AG Library, which implements the Annotation Graph model proposed by Bird and Liberman Bird and Liberman (2001). The AG File I/O modules allowed the import and export of various file format used for storing linguistic annotations. AGTK has become a primary infrastructure for annotation tool development at LDC (Maeda and Strassel, 2004; Bird et al., 2002; Maeda et al., 2002).

2.3. A new phase in annotation tool development at LDC

The past few years has been a very busy, yet productive time period for the tool developers at LDC. During this time period, we have created various annotation tools on very tight time lines in order to support annotation projects which have newly defined and/or highly evolving annotation specifications. At the same time, this was a very valuable time period, with direct feedback from both the annotation task designers and the annotators. We have learned what factors are important in the annotation tool design and development processes. We have learned what issues still remain to be resolved or improved.

In the rest of the paper, we will introduce the annotation tools that have been developed at LDC in the recent years, and will discuss the philosophy behind the design and implementation of these tools and the lessons we have learned.

3. XTrans Transcription and Speech Annotation Toolkit

3.1. Overview

To support the demand for rapid, efficient and consistent transcription, LDC has created a next-generation speech annotation tool, XTrans, to directly support a full range of

[1] http://projects.ldc.upenn.edu/TIDES

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

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

[4] http://projects.ldc.upenn.edu/ACE

[5] http://projects.ldc.upenn.edu/LCTL

speech annotation tasks including quick-, careful- and rich-transcription of broadcasts, telephone calls and meetings.

3.2. Virtual Speaker Channel

LDC has typically used Transcriber (Barras et al., 2001) for broadcast news transcription tasks, and MultiTrans for telephone speech conversation transcription tasks. Another kind of recordings LDC has started to transcribe in recent years is meeting recordings, which typically involve multi-channel sound files or multiple single-channel sound files covering the same meeting. In the past, we transcribed each channel individually using MultiTrans, and combined the completed transcripts afterwards; while this approach works in many cases, it has some shortcomings.

Meeting recordings are typically performed with multiple microphones. Sometimes, lapel microphones and headset-mounted microphones are used to record each speaker independently; sometimes, one microphone is used to record multiple speakers. Unlike conversational telephone speech recordings, in which two speakers are recorded in separate channels, speakers are not always separated in meeting recordings. Most existing transcription tools are not well-suited for the transcription of recordings in which multiple speakers are recorded in one channel; speech from multiple speakers may overlap within one channel, as the following example illustrates.

```
A: I think it is a good idea
B:                   Right. I mean..
```

In order to address this issue, the XTrans transcription tool uses the concept of *virtual speaker channels (VSCs)*. Each VSC corresponds to one speaker, rather than any particular channel in a sound file. One VSC may be used for background noises. Multiple VSCs may be assigned to one channel in a sound file when there are multiple speakers recorded in the channel. One VSC may be assigned to multiple channels in a single sound file or multiple sound files.

3.3. Bidirectional Text

Another common shortcoming of the existing transcribing tools is the support for bidirectional (bidi) text, such as Arabic, Farsi, Urdu and Hebrew. Both Transcriber and MultiTrans use the Tk GUI (Graphical User Interface) toolkit, which does not provide complete support for bidi rendering. Even though there are ways to work around this issue, for our purposes, we preferred to use a GUI toolkit that provided complete support for bidi rendering. We have tested both the Gtk+ GUI toolkit and the Qt GUI toolkit, and have chosen Qt as the GUI toolkit for XTrans.

3.4. Speech Annotation

Another goal of this new tool was to incorporate speech annotation functionalities into the transcribing tool. The MDE (Metadata Extraction) annotation tool allowed annotation of disfluencies, fillers, discourse markers and semantic units on existing transcripts (Maeda and Strassel, 2004). Adding this functionality to the transcription tool allows these annotations to be created while the transcriber is transcribing the speech, facilitating an efficient transcription and annotation workflow.

3.5. QWave Waveform Display Module

For the XTrans tool, we have developed a sound waveform display module named QWave, which is based on the Qt GUI toolkit. In most of our past speech annotation tools, we have utilized the Snack sound library (Sjölander, 2000) and the WaveSurfer module (Sjölander and Beskow, 2000), which are based on the Tk GUI toolkit. Snack and WaveSurfer provide excellent sound handling and displaying capabilities, with additional attractive features, such as the spectrogram display. However, since the text display module of XTrans is based on the Qt GUI toolkit, these Tk-based components do not fit well within XTrans. The QWave module is optimized for fast display and playback of any portions of single-channel and multi-channel sound files. If more sophisticated sound handling capabilities, such as a pitch track display, are required for annotation purposes, we plan to use Snack and WaveSurfer as external modules, and have the XTrans tool communicate with them via an interprocess communication method.

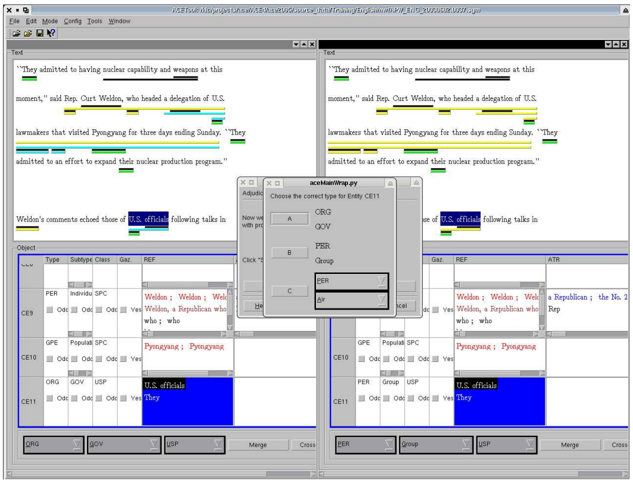
3.6. XTrans

Figure 1 shows a screen shot of the XTrans tool. XTrans is currently used for all in-house transcription tasks, including the GALE rich transcription task, the Mixer[6] transcription task and a meeting speech transcription task at LDC.

4. ACE 2005 Annotation Tool

4.1. ACE 2005 Annotation Tool

LDC has used an AGTK-based annotation tool for creating the training and evaluation data for the ACE program for the past three years. For the ACE 2005 evaluation program, major improvements were added to the ACE annotation tool. These include an event annotation interface, an xml-based configuration system for the annotation type and subtype inventory, and a decision point based adjudication wizard to facilitate efficient adjudication of dual annotation. Figure 2 shows a screen shot of the ACE 2005 annotation tool, in the adjudication mode.



ACETool: /ldc/projects/ace/ACE4/ace2005/source_data/Training/English/nw/APW_ENG_20030602.0037.sgm

``They admitted to having nuclear capability and weapons at this moment,'' said Rep. Curt Weldon, who headed a delegation of U.S. lawmakers that visited Pyongyang for three days ending Sunday. ``They admitted to an effort to expand their nuclear production program.''

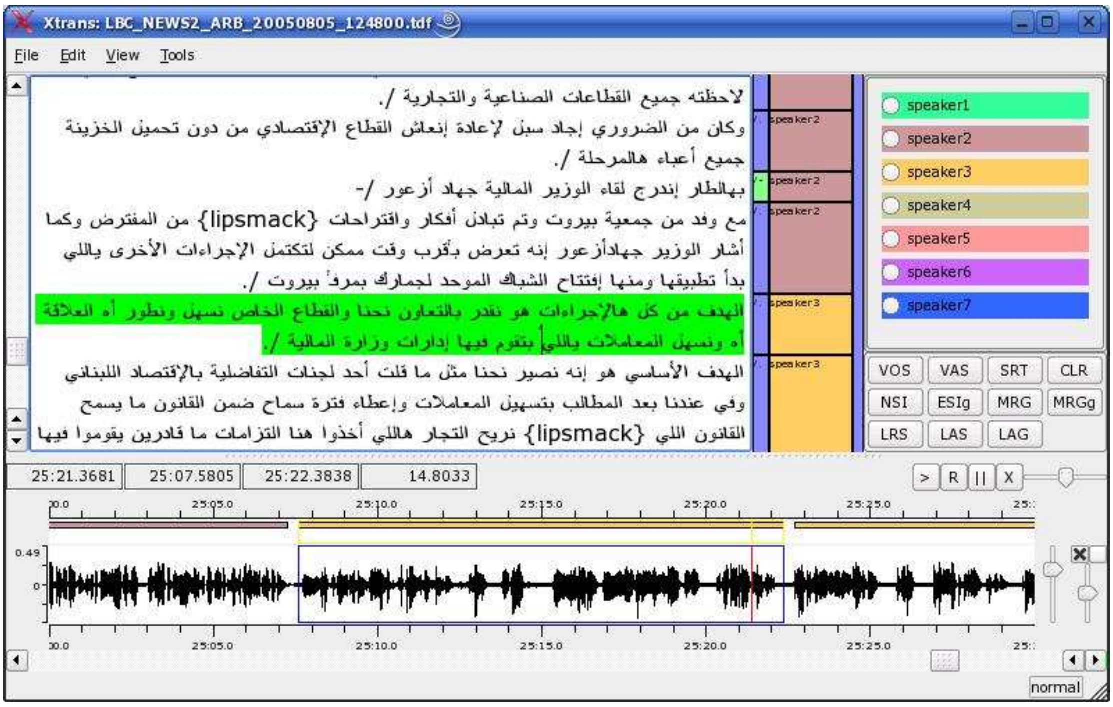
Weldon's comments echoed those of U.S. officials following talks in

aceMainWrap.py

Choose the correct type for Entity CE11

Figure 2: ACE 2005 Annotation Tool

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



لاحظته جميع القطاعات الصناعية والتجارية /.
وكان من الضروري إجاد سبل لإعادة إنعاش القطاع الإقتصادي من دون تحميل الخزينة
جميع أعباء هالمرحلة /.
بهالطار إندرج لقاء الوزير المالية جهاد أزعور /-
مع وفد من جمعية بيروت وتم تبادل أفكار واقتراحات {lipsmack} من المفترض وكما
أشار الوزير جهادأزعور إنه تعرض بأقرب وقت ممكن لتكتمل الإجراءات الأخرى ياللي
بدأ تطبيقها ومنها إفتتاح الشباك الموحد لجمارك بمرفأ بيروت /.
الهدف من كل هالإجراءات هو نقدر بالتعاون نحنا والقطاع الخاص نسهل ونطور أه العلاقة
أه ونسهل المعاملات باللي بتقوم فيها إدارات وزارة المالية /.
الهدف الأساسي هو إنه نصير نحنا مثل ما قلت أحد لجنات التفاضلية بالإقتصاد اللبناني
وفي عندنا بعد المطالب بتسهيل المعاملات وإعطاء فترة سماح ضمن القانون ما يسمح
القانون اللي {lipsmack} نريح التجار هاللي أخذوا هنا التزامات ما قادرين يقوموا فيها

Figure 1: XTrans Transcription and Speech Annotation Tool

4.2. Adjudication Wizard

One of the most important requirements for the ACE 2005 annotation effort was to create a large amount of adjudicated annotation. This means, for each file, two annotators perform independent first-pass annotation, and a senior annotator adjudicates the differences in the two first-pass annotation files in order to create high-quality annotated data. A smaller amount of adjudication was also performed in the ACE 2004 annotation effort; however, adjudication was performed on all of the data in ACE 2005. The ACE 2005 annotation tool included an adjudication wizard for "decision point" based adjudication allowing a very efficient and intuitive adjudication process (Medero et al., 2006).

5. TDT5, HARD2004 and GALE Distillation Toolkits

The tool developers at LDC created annotation tools for the fifth phase of the Topic Detection and Tracking (TDT) project, and the High Accuracy Retrieval from Documents (HARD) project in 2004. These tools used the same development infrastructure as the other annotation tools; however, unlike the other tools, the TDT5 and HARD2004 annotation tools used the MySQL open source database, and a search engine developed at LDC. In 2006, a similarly designed tool for the GALE Distillation annotation task is being developed. These tools do not use the AG library; however, they are prime examples of integrating database technologies into annotation tools.

6. Simple NET Annotation Tool for LoDL and LCTL:

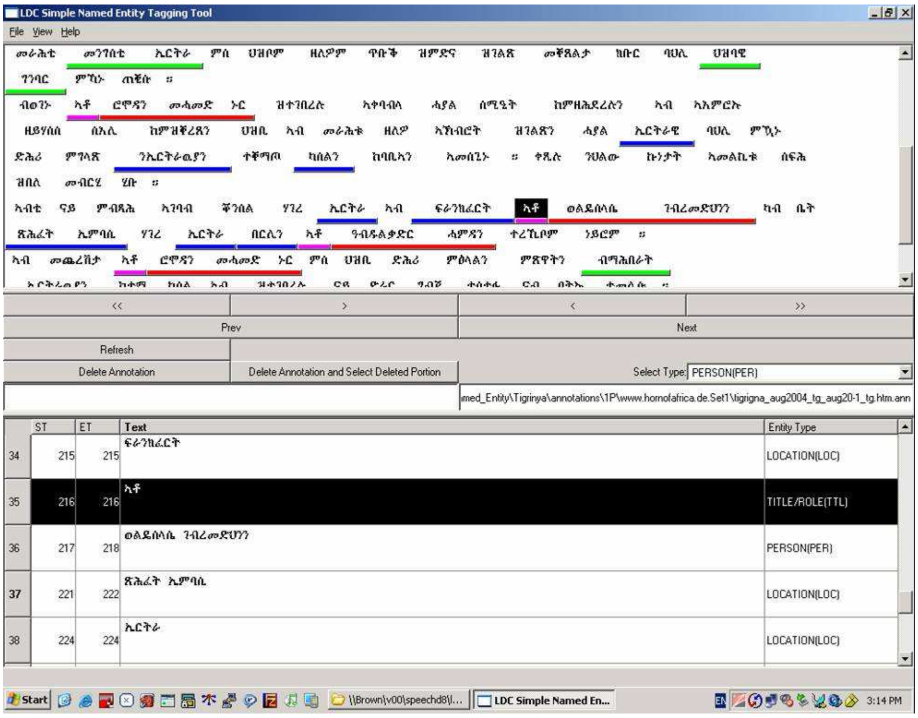
The SimpleNET named entity annotation tool is an application that was initially developed for the DARPA TIDES Surprise Language exercise. The primary focus of this tool is to maximize the usability for annotators without requiring extensive training. Most of the annotators for the Surprise Language exercise were native speakers of the languages who are not necessarily computer experts or language experts. The operation of the tool was very simple. Tokenization is performed prior to annotation, and annotators simply need to mark an extent with the mouse or the keyboard, and select a type of the named entity, such as organization, person, location, time/date and title/role.

Despite the short amount of time spent of the development of this tool, SimpleNET continued to be used for the annotation tasks for subsequent large scale annotation projects for the Low Density Language (LoDL) program and the Less Commonly Taught Languages (LCTL) program. The languages annotated with this tool include Cebuano, Hindi, Tigrigna, Urdu, Uzbek, Thai, Hungarian, Bengali, Punjabi, Tamil and Yoruba. Figure 3 shows a screen shot of the simpleNET annotation tool with Tigrigna text.

The ACE annotation tool and the SimpleNET annotation tool are at the opposite ends of the spectrum. The ACE annotation tool is able to handle very complex annotation specifications, and was designed to be very flexible. The SimpleNET annotation tool was designed for one specific purpose. However, both tools turned out to be useful — this suggests that having the right tool is more important than having the most sophisticated tool.

7. Future Plans

One of our most important future plans is to maximize the usability of these tools for the users outside of LDC. This not only includes the enhancement of tool usabilities, but also includes the ease of installation on various platforms.



መራሕቲ መንግስቲ ኤርትራ ምስ ህዝቦም ዘለዎም ጥቡቕ ዝምድና ዝገልጽ መቐጸልታ ክቡር ባህሊ ህዝባዊ
ግንባር ምኳኑ ጠቒሱ ።
ብወገኑ ኣቶ ሮሞዳን መሓመድ ኑር ዝተገበረሉ ኣቀባብላ ሓያል ሰሚዒት ከምዘሕደረሎን ኣብ ኣእምሮኡ
ዘይሃሰስ ስእሊ ከምዝቐረጸን ህዝቢ ኣብ መራሕቱ ዘለዎ ኣኽብሮት ዝገልጽን ሓያል ኤርትራዊ ባህሊ ምኻኑ
ድሕሪ ምግላጽ ንኤርትራውያን ተቐማጦ ካሰልን ከባቢኣን ኣመስጊኑ ። ቀጺሉ ንህልው ኩነታት ኣመልኪቱ ሰፊሕ
ዝበለ መብርሂ ሂቡ ።
ኣብቲ ናይ ምብጻሕ ኣገባብ ቐንበል ሃገረ ኤርትራ ኣብ ፍራንክፈርት ኣቶ ወልደሰላሴ ገብረመድህን ካብ ቤት
ጽሕፈት ኢምባሲ ሃገረ ኤርትራ በርሊን ኣቶ ዓብዱልቃድር ሓምዳን ተረኺቦም ነይሮም ።
ኣብ መጨረሽታ ኣቶ ሮሞዳን መሓመድ ኑር ምስ ህዝቢ ድሕሪ ምዕላልን ምጽዋትን ብማሕበራት

Select Type: PERSON(PER)

med_Entity\Tigrinya\annotations\1P\www.homofafrica.de.Set1\tigrigna_aug2004_tg_aug20-1_tg.htm.ann

	ST	ET	Text	Entity Type
34	215	215	ፍራንክፈርት	LOCATION(LOC)
35	216	216	ኣቶ	TITLE/ROLE(TTL)
36	217	218	ወልደሰላሴ ገብረመድህን	PERSON(PER)
37	221	222	ጽሕፈት ኢምባሲ	LOCATION(LOC)
38	224	224	ኤርትራ	LOCATION(LOC)

Figure 3: SimpleNET

Well-written user manuals and installation manuals are crucial for this plan. In short, our goal is to provide better "out-of-the-box" experience for the users outside of LDC.

Another plan is to incorporate the results of recent research on linguistic annotation, such as the Querying Linguistic Databases (QLDB) project[7]. The QLDB project investigates efficient data models for trees, interlinear texts, lexicon and linguistic paradigms. This research could have an immediate impact on ways to store these kinds of linguistic data.

Also, we plan to incorporate annotation and transcription assisting technologies, such as an audio segmenter, a spell checker and a POS tagger, into the annotation tools.

8. Conclusion

The annotation tool developers at LDC have benefited from feedback from the end users in house. We hope to continue this tradition. At the same time, we would like to enhance the support for users outside the LDC. All of the software components and tools described in this paper will be distributed as open-source software.[8]. We hope that the various annotation tools and components introduced in this paper are useful to both the creators and users of linguistics resources.

9. References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33:5–22.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

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*.

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, Steven Bird, Xiaoyi Ma, and Haejoong Lee. 2002. Creating annotation tools with the Annotation Graph Toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Julie Medero, Kazuaki Maeda, Stephanie Strassel, and Christopher Walker. 2006. An efficient approach to gold-standard annotation: Decision points for complex tasks. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Kåre Sjölander and Jonas Beskow. 2000. WaveSurfer – an open source speech tool. In *Proceedings of the 6th International Conference on Spoken Language Processing*. `http://www.speech.kth.se/wavesurfer/`.

Kåre Sjölander. 2000. The Snack sound toolkit. `http://www.speech.kth.se/snack/`.

[7] http://projects.ldc.upenn.edu/QLDB

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