XTrans: a speech annotation and transcription tool

Meghan Lammie Glenn, Stephanie M. Strassel, Haejoong Lee
Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania
{mlglenn,strassel,haejoong}@ldc.upenn.edu

Abstract
We present XTrans, a multi-platform, multilingual, multi-channel transcription application designed and developed by Linguistic Data Consortium. XTrans provides new and efficient solutions to many common challenges encountered during the manual transcription process of a wide variety of audio genres, such as supporting multiple audio channels in a meeting recording or right-to-left text directionality for languages like Arabic. To facilitate accurate transcription, XTrans incorporates a number of quality control functions, and provides a user-friendly mechanism for transcribing overlapping speech. This paper will describe the motivation to develop a new transcription tool, and will give an overview of XTrans functionality.

Index Terms: linguistic resources, speech transcription, transcription tool, speech annotation, linguistic annotation

1. Introduction
The past 20 years of human language technology research has taught us that effective linguistic technology is based on statistical modeling of large corpora of linguistic data, and that the most reliable way to improve system performance is to improve the resources upon which it is based. For speech recognition, the availability of large quantities of transcribed, time-aligned speech has been a cornerstone of progress, enabling significant improvements in read and spontaneous speech in broadcast, conversational and meeting domains, for multiple languages. Linguistic Data Consortium was established in 1992 at the University of Pennsylvania to support language-related education, research and technology development by creating and sharing linguistic resources, including data, tools and standards; this mission often includes creation of annotated corpora for technology evaluations including speech recognition. Several evaluation programs, including DARPA TIDES (Translingual Information Detection, Extraction and Summarization), EARS (Effective Affordable Reusable Speech-To-Text), and the current GALE (Global Autonomous Language Exploitation) program, as well as NIST Speech-to-Text and Rich Transcription evaluations, have invested in the creation of manual transcripts at LDC and elsewhere for system training and evaluation. As program requirements for data volume, annotation complexity, range of genres and languages expand over time it becomes increasingly important for corpus developers to strike a balance between annotation efficiency and quality; one key piece of this equation is the choice of tools for human annotation. This paper describes recent technical advances by LDC in the creation of a next-generation speech annotation tool, XTrans, designed to support a wide variety of tasks, genres and languages while improving both efficiency and quality of annotation.

2. Motivation
Since its first transcription project in 1995 LDC has utilized a number of different annotation tools and approaches. Early on LDC developed customized user interfaces built around existing commercial software like Entropic’s ESPS/waves+ and emacs. Over time user-friendly multi-featured tools like Transcriber [1] became more widely available and were incorporated into LDC’s annotation pipeline. To meet data and annotation needs for specific evaluation programs, LDC also developed customized transcription tools within the Annotation Graph Toolkit (AGTK) framework [2], [3] including TableTrans, a spreadsheet-style audio annotation tool, and MultiTrans, a multi-channel transcribing tool [4]. The EARS and GALE programs introduced a new constellation of challenges: manual transcription of thousands of hours of broadcast and telephone speech, in multiple languages, produced on a very tight schedule. While previous transcription projects could be completed locally using trained LDC staff, to meet the demands of these new programs it was necessary for LDC to outsource much of the transcription task and inter-annotator and cross-site consistency became a major concern.

Table 1. Comparison of transcription tools.

<table>
<thead>
<tr>
<th></th>
<th>Entropic ESPS/ waves+ and emacs</th>
<th>Transcriber</th>
<th>TableTrans, MultiTrans</th>
<th>XTrans</th>
</tr>
</thead>
<tbody>
<tr>
<td>max audio channels</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>unlimited</td>
</tr>
<tr>
<td>max speakers/ channel</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>unlimited</td>
</tr>
<tr>
<td>bidi text support</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>text display</td>
<td>emacs</td>
<td>Tk</td>
<td>Tk</td>
<td>Qt</td>
</tr>
<tr>
<td>audio display</td>
<td>xwaves</td>
<td>WaveSurfer</td>
<td>WaveSurfer</td>
<td>Qtwave</td>
</tr>
<tr>
<td>language/ libraries</td>
<td>C. Xlib</td>
<td>Tel/Tk, snatch</td>
<td>Python, AGTK</td>
<td>Python, C++, Qt</td>
</tr>
<tr>
<td>platform</td>
<td>Unix/Linux</td>
<td>Unix/Linux, Windows, Mac OS X</td>
<td>Unix/Linux, Windows</td>
<td>Unix/Linux, Windows</td>
</tr>
</tbody>
</table>

Existing transcription tools addressed some aspects of this challenge, but none provided a complete solution. For instance, existing tools were not optimized for transcribing recordings of multiple speakers on a single audio channel where overlapping speech is prevalent, and lacked rich support for speaker identification markup. No single tool could readily handle both single- and multi-channel transcription across languages including Arabic, Chinese and Russian with full support for bidirectional and non-ascii text input. Other tools were less than intuitive to use, leading to slower than desired throughput. Finally, most tools lacked support for common
quality control procedures like spell checking, identifying gaps in the transcript, and verifying speaker labels. Table 1 summarizes design features and functionality for a number of transcription tools commonly used for manual transcription at LDC and other organizations.

To overcome these issues and better address the demand for rapid, efficient, consistent, high-volume transcription, LDC developed the XTrans tool to support a full range of transcription tasks, including quick, careful and rich transcription of broadcasts, telephone speech and meetings.

3. Data model

3.1. Virtual channel

Transcripts may be created from single channel audio input, as with broadcast news recordings, or multiple channels, as with telephone speech or meeting recordings. We term each separate audio recording a physical channel. A single-channel broadcast recording will typically contain multiple sound sources, such as a primary news anchor and multiple correspondents. In XTrans we assign each of these sound sources to a separate virtual speaker channel (VSC). Each VSC corresponds to one speaker or other sound source, rather than to any particular physical channel in a sound file. By introducing this concept to the transcription model, it becomes possible to transcribe a potentially unlimited number of simultaneous speakers without having to resort to special data representations for handling cases of overlapping speech.

Data models capable of supporting the virtual channel concept have been developed previously. For example, AGTK and ATLAS (Architecture and Tools for Linguistic Analysis Systems) allow multiple layers of different types of annotations on a single time axis. [2] However, few prior transcription tools implement VSCs in their design. The Transcriber tool presents users with one physical channel and employs a special convention for transcribing overlapping speech from two speakers. LDC’s MultiTrans tool supports up to four channels of transcribable audio, but links the transcript of each speaker to a different physical audio channel. In contrast, XTrans does not require multiple information streams to be linked to separate physical audio channels. The data model consists of a set of segments, each of which defines a region for one VSC containing a single speaker’s speech. Two segments for the same VSC cannot temporally overlap, but there are no restrictions on temporal overlap among multiple distinct VSCs.

3.2. File format

The output of XTrans is a Tab Delimited Format (TDF) in which data is represented as a set of “records,” containing fields that identify each transcript segment with a physical speaker channel, speaker, start and end times, transcript, and section and sentence annotations. Given this simple structure, XTrans TDF files are easily converted to other common formats and are readily usable by downstream manual and automatic annotation tasks. XTrans can directly import files in Transcriber (TRS) format. It also exports files in a plain text format designed for maximum human readability, as an added quality control feature.

4. Technical description

The current version of XTrans runs on FreeBSD, Linux and Windows platforms. Most of its components are written in Python with some C++ components, such as the QWave waveform display module [5] based on the Qt GUI (Graphical User Interface) toolkit. Previous annotation tools developed at LDC have relied on the Snack sound library [6] and WaveSurfer [7]. These are founded on the Tk GUI toolkit and while they provide excellent sound handling and display capabilities, Qt adds important functionality. The QWave module is optimized for fast display and playback of any portion of single-channel and multi-channel sound files, and handles both NIST Sphere (.sph) and Windows wave (.wav) formats, among others. QWave does not natively support more advanced audio capabilities such as pitch tracking and spectrogram display, but XTrans can be configured to communicate with Snack, WaveSurfer or other external modules via an interprocess communication method [5]. Qt also offers international language support and portability among various operating systems.

5. User interface design

The XTrans design roughly follows the model-view-controller (MVC) design pattern [8]. The controller component is not explicitly realized in the code, but rather is a part of the view components. The communication between the data model and the view components is facilitated by the signal/slot mechanism of Qt. The flexible virtual channel data model is implemented as a table internally, where rows are the segments and columns are the attributes.

The tool’s physical layout includes separate panels for each major function.

![Figure 1. XTrans layout, with all five panels labeled.](image)

1. **Audio panel**. This panel provides waveform display and playback functionality for the selected audio file(s). In addition to the usual playback features (play, pause, etc.), XTrans allows users to change playback speed, select mono or stereo playback, and mute or activate selected audio channels. The audio panel display is tightly integrated with other XTrans display elements. As users label a segment (utterance, turn, etc.) in the audio file, a rectangle appears above the waveform, marking the duration of that segment. The rectangle is color-coded to indicate which VSC/speaker it is assigned to.

2. **Transcript panel**. The orthographic transcript is created within this panel. Untranscribed but segmented regions are displayed as empty bullet points. Selecting a region of the transcript highlights it in the transcript panel and simultaneously highlights the corresponding region in the audio panel.

3. **Speaker panel**. This panel lists all speakers assigned to segments in the file. Speakers are identified by name or unique numerical identifier, and names are coded with one color per VSC/speaker.
(4) Widget panel. This panel incorporates speaker identification and quality control functions, described in Section 6.3.

(5) Segmentation panel. This region displays segments of different granularity, arranged hierarchically. In the broadcast transcription task, transcribers might first identify sentence units (statements, questions); then turns (multiple consecutive sentences by a single speaker); then stories (multiple consecutive turns that have some topical cohesion). Each segment type is represented by a vertical column, with color-coded boxes of different sizes within each column depicting the labeled, nested segments.

6. Features

XTrans was designed with input from LDC transcribers and transcription managers, to respond to a number of specific challenges.

Figure 2. Arabic broadcast news transcript session in XTrans with overlap of two and three speakers.

6.1. Text directionality

One shortcoming of prior transcription tools is the lack of adequate support for bidirectional (bidi) text for languages like Arabic, Farsi, Hindi, Urdu and Hebrew. Unlike Transcriber and MultiTrans, which both use the Tk GUI toolkit and cannot fully support bidi display, XTrans relies on Qt which has built-in bidi support [5]. In theory, any language with an input method can be transcribed with XTrans. Most commonly, Windows users rely on Microsoft IME or a third-party IME like Google while UNIX users rely on X input method. Figure 2 depicts an Arabic transcription session.

6.2. Overlapping speech annotation

When faced with the task of transcribing thousands of hours of single-channel, multi-speaker news and talk show recordings, an efficient approach to overlapping speech is critical. XTrans’ utilization of the Virtual Speaker Channel concept permits regions of overlapping speech to be treated identically to non-overlapping speech, enabling overlaps of two, three or more speakers.

XTrans visually represents these regions in the audio panel by showing overlapping color-coded horizontal bars for each segment, one per speaker, as shown in Figure 3.

Figure 3. Overlapping segments displayed in audio panel.

6.3. Quality control

Over the years LDC has developed a number of post-hoc quality control scripts and processes to ensure transcript accuracy and completeness. XTrans integrates these functions directly into the user interface, giving transcribers the power to improve their own output. Features supported by XTrans as “hot buttons” include:

- **LRS** – Listen Random Segment – listen to a random segment from a selected speaker, used to disambiguate one from dozens of unique speakers.
- **LAS** – Listen All Segments – listen to all segments of a selected speaker to ensure consistent speaker identification throughout a file.
- **LAG** – Listen All Gaps – listen to all unsegmented audio to identify segmentation errors.

6.4. Multiple audio files

In recent years LDC has experienced increased demand for high quality meeting room transcripts. This domain is particularly challenging given the rapid, conversational nature of the data and the large number of speakers, who are sometimes recorded individually and sometimes recorded with multiple speakers per channel. In the past transcribers were required to run a combination of separate tools in succession to produce one final verbatim, time-aligned transcript that includes all meeting participants regardless of the number of audio input channels. XTrans streamlines the process by enabling transcribers to load a potentially unlimited number of audio files into the tool and transcript them all at once.

Figure 4. Meeting recording with four speakers on individual audio channels.

Transcribers can mute or activate channels as needed in order to focus on a single speaker or listen to all speakers simultaneously, and can execute QC steps as part of the normal transcription process. Figure 4 shows a meeting recording with separated transcripts for the four speakers.

6.5. User-defined functionality

The original design concept for XTrans was informed by senior LDC transcribers who had considerable experience using existing tools. A common request from these users was a single tool that was language and genre independent and could be easily configured to meet individual project requirements. XTrans addresses this need by allowing users or managers to define and/or modify keybindings for common transcription tasks. The tool was also designed with a number of features to
make transcription more efficient. All segmentation and transcription functions can be performed using keyboard shortcuts as well as the mouse. A complete, verbatim, time-aligned, manually segmented transcript can be produced without use of the mouse at all, and experienced transcribers report a dramatic increase in efficiency and comfort when using this approach.

6.6. Adaptability
XTrans was designed to be easily adaptable and extensible for other projects and applications. Recently XTrans has been used for tasks including spoken dialect classification and supralexical annotation which includes markup for disfluencies, filled pauses, named entities and other meta features of the transcript. A recent extension of XTrans is QCTool, which was developed to facilitate production and quality control of sentence-aligned parallel text in multiple domains including broadcast recordings [9]. QCTool allows bilingual translators to review and modify successive versions of a translation and provides integrated access to the source language audio recording and time-aligned transcript where applicable [10].

6.7. Efficiency
XTrans’ accessible design and the fact that the same tool supports multiple domains, languages and tasks leads to more efficient annotator training and supports cross-team, cross-task consistency. Annotators can shift easily from one transcription project to another without having to learn a new approach. XTrans has been the default transcription tool for LDC and our partner transcription sites and vendors for several years, and in that time we’ve witnessed an increase in transcription efficiency at all stages of the pipeline.

As a point of comparison, using the previous multi-tool transcription approach, manual transcription for the NIST RT05 meeting recording evaluation took on average 65 times real time per channel (i.e., 65 hours to transcribe one hour of speech). Using XTrans, real time rates for similar data in the RT06 and RT07 evaluations dropped to under 50 times real time [11]. XTrans has also enabled highly efficient quick transcription (accurate, time aligned transcription with minimal markup) of broadcast news. While previous quick transcription rates for non-English typically exceeded 30 times real time, with XTrans all languages currently in production now achieve rates under 15 times real time.

7. Conclusions and future directions
To date, XTrans has been used to produce over 2000 hours of transcribed broadcast, telephone and meeting speech in at least a half dozen languages: English, Arabic, Mandarin, Spanish, Farsi and Russian. It has been used to support at least six separate human language technology programs in speech-to-text, rich transcription, speech transcription, and speaker recognition [12]. It is the primary transcription tool for LDC as well as for our network of collaborators and transcription vendors, including Evaluations and Language resources Distribution Agency (ELDA), Velda Inc. and Hong Kong University of Science and Technology (HKUST), and a number of sponsor agencies.

Extensive use XTrans tool by LDC and our partners over the past several years has led to a number of suggested improvements and extensions, such as enhanced multimedia support for common audio formats like .mp3 and video playback to assist with speaker ID in multi-speaker broadcast recordings. We also plan to update the text widget to allow layers of annotation over the text transcript, which would resolve lingering problems with bidi rendering.

XTrans is freely distributed by LDC upon request. XTrans for Linux is released under GPL v.3, while XTrans for Windows is available as a developer’s version under GPL v.3. A full Windows version is in development.

8. Acknowledgements
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.

9. References
[8] Gamma E., R. Helm, R. Johnson, J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley, 1995.