Linguistic Resources for Meeting Recognition

Meghan Glenn, Stephanie Strassel
Linguistic Data Consortium

{mlglenn, strassel@ldc.upenn.edu}

http://projects.ldc.upenn.edu/Transcription/NISTMeet

Overview

- Training data distribution
- New corpus creation
  - Transcription team
  - NIST Phase 2 Corpus
    - Quick transcription
  - Conference room test data
    - Careful transcription
    - Quality control
  - Unique challenges of meeting transcription
- Infrastructure
  - XTrans Toolkit
    - Existing features for meetings
    - Future features for meetings
- Inter-annotator consistency study (first pass transcription)
  - Approach
  - Results

RT-07 Training Data
provided by LDC

Title	Speech	Transcripts	Volume	Domain
Fisher English Part 1	LDC2004S13	LDC2004T19	750+ hours	CTS
Fisher English Part 2	LDC2005S13	LDC2005T19	750+ hours	CTS
ICSI Meeting Corpus	LDC2004S02	LDC2004T04	72 hours	Meeting
ISL Meeting Corpus	LDC2004S05	LDC2004T10	10 hours	Meeting
NIST Meeting Pilot Corpus	LDC2004S09	LDC2004T13	13 hours	Meeting
RT-04S Dev-Eval Meeting Room Data	LDC2005S09	LDC2005S09	14.5 hours	Meeting
RT-06 Spring Meeting Speech Evaluation Data		LDC2006E16	3 hours	Meeting
TDT4 Multilingual Broadcast News Corpus	LDC2005S11	LDC2005T16	300+ hours	BN

RT-07 Transcriber team

- Team makeup
  - Largely the same team from RT-06
    - Experienced with meeting recording transcription
    - Very well-versed in XTrans
  - Transcribers from GALE QRTR team
    - Quick transcription approach
    - SU annotation
    - Single channel
- NIST Phase 2 corpus
  - Topic preference
    - Reviewed files briefly for topic content, number of speakers
    - Transcribers selected topics that fit personal interests, based on topic content descriptions like the following:
      - *Instructional presentation on flying planes; 5 speakers, 1 via telephone*
      - *Mary Kay Makeup Presentation; 4 speakers: 1 presenter, 3 testers*
- Eval corpus
  - Transcription training still fresh from Phase 2 corpus creation
  - Modified CTR approach to suit XTrans better
  - Random file assignment
  - File reassignment based on topic expertise (financial, literary, etc) or personal preference if necessary
    - Careful transcription requires multiple, independent quality control passes, so each transcriber sees every file eventually
  - File reassignment based on task expertise and preference, if necessary

NIST Phase 2 Corpus

- Data profile
  - 18 hours
  - 3 - 6 speakers per session
  - Sessions approximately 1 hour each (some 40 minutes, one close to 2 hours)
  - Native and non-native speakers
  - “Ambient” speakers (2 via telephone)
    - Extend annotation rate
  - Varied topic content
    - Business meetings
    - Role playing
    - Product presentations

NIST Phase 2 Corpus – QTR

- Quick transcription approach
  - IHM recordings
  - Automatic segmentation
    - LDC AutoSegmenter
      - Segmentation of utterances using Entropi’s ESPS library
      - Applied on individual channels
  - Manual review of segmentation
  - First pass transcription
    - Targets content words
    - Transcribers listen to segments once or twice
    - Markup of acronyms and spoken letters
    - No markup of filled pauses or proper nouns
    - Optional (additional) modification of segmentation
  - Quality control to resolve and standardize proper nouns, “uncertain transcription”

RT-07 Test data

- Conference room data
  - Eight meeting sessions, nine excerpts
  - 4 - 6 speakers per session
  - Contributed by four organizations
  - Multiple recording conditions for each session
  - 22 minutes each
  - Topic content: primarily business, very similar to RT-06
    - Military briefings
    - Product design
    - Memorial preservation
    - Business planning

RT-07 Test data - CTR

- Careful transcription
  - IHM recordings, one speaker per channel
  - *Stage 1: Manual Segmentation*
    - Segments are breath groups
      - Average 3-8 seconds, primarily for ease of transcription
      - ~10 ms padding at edges of segment boundaries
    - 1 X RT
  - *Stage 2: Verbatim Transcription*
    - Slow, very careful orthographic transcription
    - No time limit
  - *Stage 3: Transcription Verification and Markup*
    - Add markup for filled pauses, proper names etc.
    - Verify segmentation & transcription accuracy
    - Revisit difficult sections
    - 3 X RT

Quality Control (CTR)

- After transcription of all speaker channels in a meeting
  - Use mixed IHM recordings
- Add speaker and background noise
  - Stage 4: transcribers revisited files and inserted disruptive noise
    - Speaker noise
      - Vocalized noise – limited to 5 sounds
        - » Ignored consistent heavy breathing
        - » Concentrated on coughs, laughs, sighs, sharp in- or exhalations, sneezes
      - Speaker-generated noise
        - » Paper rustling, microphone tapping, etc.
    - Background noise is generic
      - Experimented with adding “virtual speaker” for background noise instead of assigning it to a speaker

Quality Control (CTR) (2)

- Additional QC pass by senior transcribers checks for
  - Transcription & segmentation accuracy, completeness
  - Speaker ID consistency
  - Consistency, accuracy of names, acronyms, terminology
  - Examine silence (untranscribed) regions for missed speech using customized tool functions
  - Markup consistency
  - Final spell check
  - Export to CTS (.txt) format
  - Expand contractions

Unique Challenges

- Multiple speakers
  - Overlapping speech
  - Asides
- Meeting content
  - Acronyms (military briefings)
  - Project discussion groups
  - Role playing meetings
- Meeting spaces
  - Ambient noise
- Varying levels of speaker participation
  - Often no speech but other speaker/ background noise
- No video access
  - In the works for future versions of XTrans
  - Improve speaker ID, especially for ambient speakers

Infrastructure

- Previously, LDC lacked good tools for meeting transcription
  - Cobbled together multiple tools and scripts for different stages
- In late 2005, we created XTrans to address this and other issues
  - Generalized speech annotation tool
  - Multi-platform, multi-lingual, multi-domain
  - Based on QT, implemented in Python and C++
  - Component-based, reconfigurable for new tasks
  - Built-in support for common LDC tasks
    - Quick and careful transcription, structural spoken metadata annotation
    - Meetings, conversational telephone and broadcast speech
    - Quality control: translation QC, compare trans
- Windows and Linux versions available on LDC website
  - projects.ldc.upenn.edu/gale/Transcription/download_xtrans-windows-latest.php
    or projects.ldc.upenn.edu/gale/Transcription/download_xtrans-linux-latest.php
  - (user: xtrans / password: download)

XTrans Features

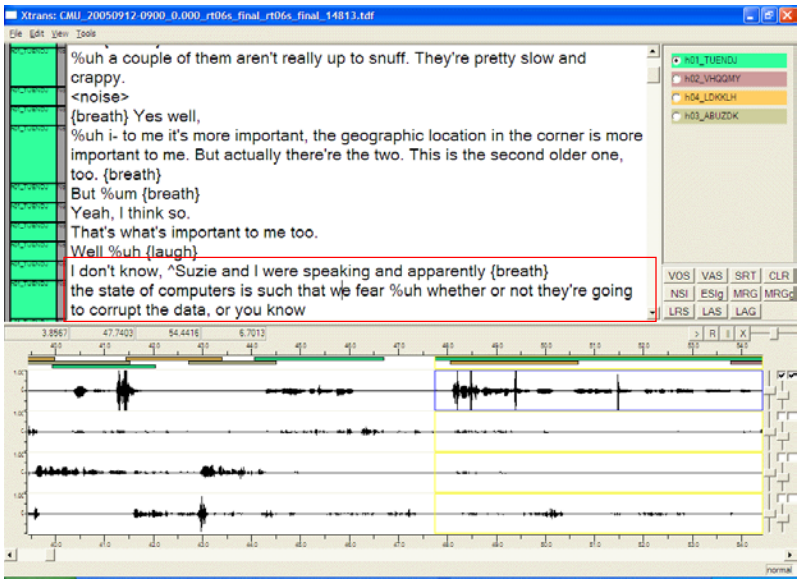
- User-friendly GUI
  - All commands can be issued from keyboard or from mouse
  - Keyboard-only is much faster
  - User-configurable keybindings for common tasks
- Bi-directional text input
  - Critical for languages like Arabic
- SpeakerID verification functions include
  - LRS: Listen to a random segment from this speaker to verify voice
  - LAS: Listen to all segments from this speaker in the file
  - NSI: Assign new speaker ID for this segment
- Waveform display/playback components
  - QWave, based on QT
    - Variable speed playback (no pitch control)
    - Relative volume control for individual channel
    - Amplitude control
- Inter-gap playback
  - LAG: Listen to the unsegmented audio "gaps" (helpful for doing quality control, to catch unsegmented speech)

XTrans Features (2)

- Virtual speaker channels
  - One VSC per *speaker*, not per audio channel
    - Enables easy handling of overlapping speech in single-channel audio
- Fluid single vs. multiple speaker focus
  - Arbitrary number of audio channels can be loaded at once
  - Toggle between multiple playback functions
    - Merged IHM
    - Multiple individual IHMs
    - Single IHM for one speaker
    - Any channel can be muted
  - Toggle between merged, multi-speaker transcript view and single-speaker view
    - Use complete transcript for context
- Waveform markup display makes speaker interaction obvious
- Easy creation/modification of configuration files makes transcription more efficient
  - Frequently-used terms

Multi-Speaker Focus

Single Speaker Focus



Impact of XTrans

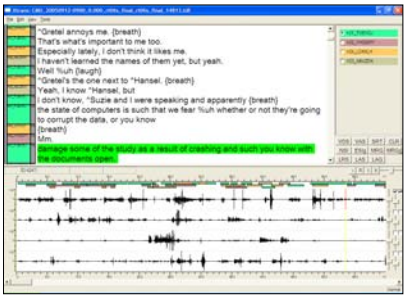
- Annotator feedback
  - Easy to use, well-designed
  - Look and feel is familiar from other projects, tasks
  - Ability to toggle between single and multi-speaker focus is great
- Quality control
  - Better integrated into tool itself rather than stand-alone post-process
  - Customized features support QC at all stages
  - Adjudication module for comparing two versions transcripts
- Real time transcription rates
  - RT-05: over 65 x realtime for QTR
  - RT-07: 50 x realtime for QTR (all channels)

Future plans for XTrans

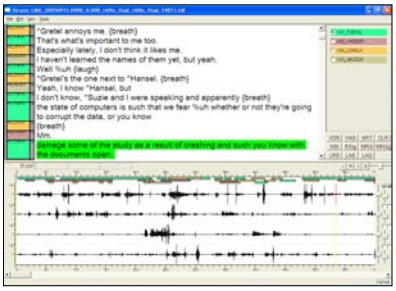
- Legacy data format input and/or output
  - Currently supports .trs (Transcriber) and .tdf (XTrans) format
  - Add RTTM, older LDC formats
- MP3 audio capability
- Video playback capability
  - Easier speaker ID
  - Easier meeting “contextualization”
- Integrate additional annotation functions
  - Currently stand-alone modules
    - Contraction expansion
  - New text display component for MDE annotation
    - Will enable transcript correction during annotation tasks
- Better non-English (non-Roman) input methods
  - Currently rely on SCIM and other external protocols

Research Task Question

- 1) Consistency study
  - How consistent are the first pass transcripts from the careful transcription process?
    - Qualitative analysis



=

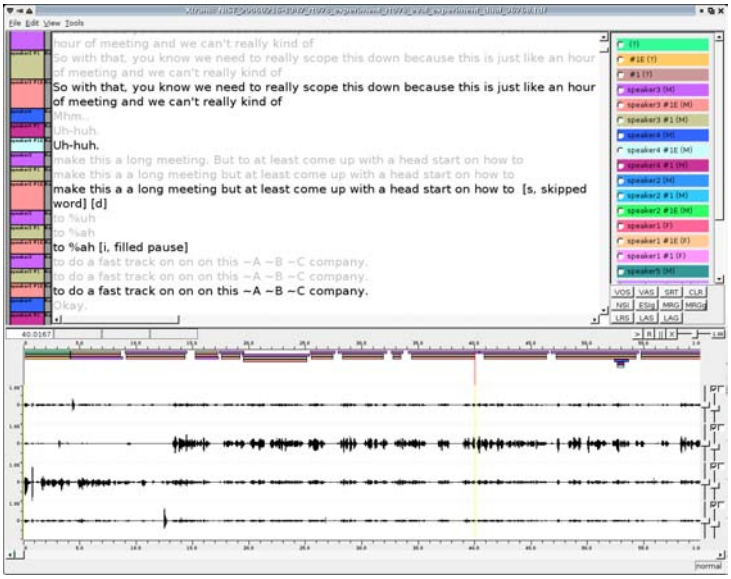


?

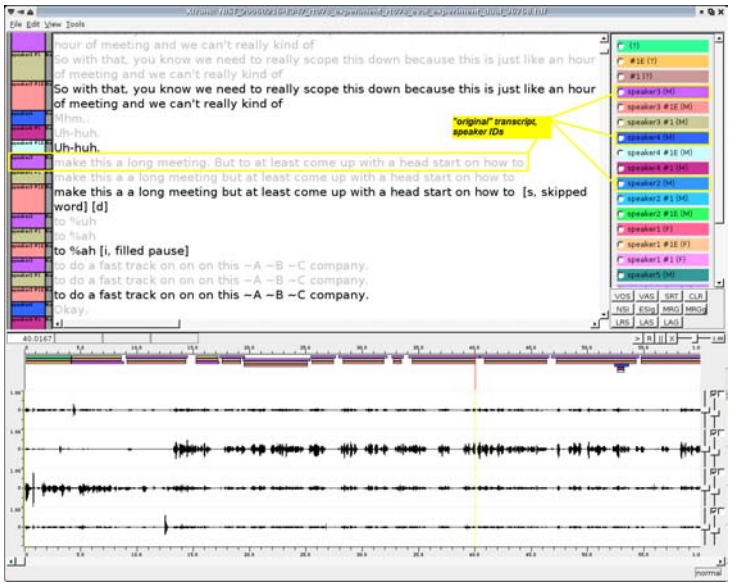
Consistency analysis

- Inter-annotator consistency study
  - First pass dual
  - Compared both files
    - Modified version of XTrans shows both files together
    - Allows space for comments
  - Labeled differences using the following distinctions
    - Significant
      - Different words, mistaken transcription
      - Missed transcription
    - Insignificant
      - Punctuation differences
      - Capitalization differences
      - Spelling of unknown proper nouns
  - If the difference is significant transcribers label it and answer:
    - Which version is correct?
      - original [o]
      - dual [d]
      - neither [n]
      - both [b]
      - unable to tell [u]
    - Was it caused by:
      - carelessness (not following guidelines)
      - misunderstanding of transcription rules
      - audio quality (static, too low, too high)
      - interactivity (overlapping speech, speaker noises)
      - speaker (speaking quickly, non-native speaker, confusing terminology)
      - other

Dual transcript comparison

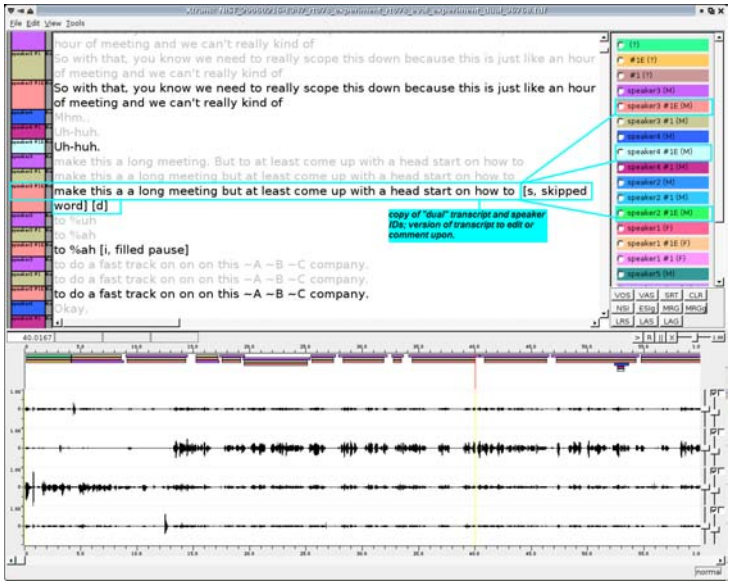


Dual transcript comparison



• RT-07 Spring Meeting Recognition Evaluation Workshop, Baltimore, MD May 10-11, 2007

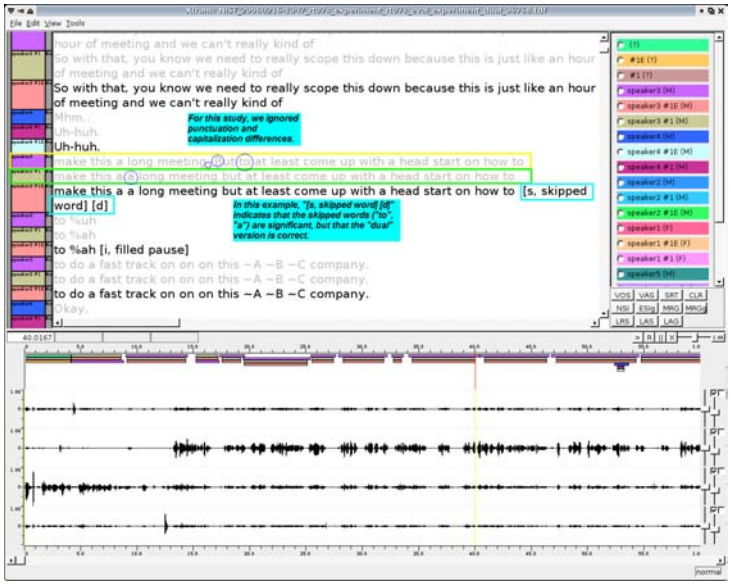
Dual transcript comparison



Dual transcript comparison

• RT-07 Spring Meeting Recognition Evaluation Workshop, Baltimore, MD May 10-11, 2007

Dual transcript comparison



Consistency results

- Qualitative results
  - Capitalization and punctuation inconsistencies are responsible for most of the errors
  - Careless errors
    - Missed partial words
    - Missed filled pauses
    - Missing markup (not a focus for 1p transcription)
  - Comprehension errors
    - Unintelligible speech (()) markers due to audio quality, speaker accent or speed
    - Misspelled proper nouns
    - Conversation context
      - For example:
        - » orig: Tha- that's true. ^Alex is kind of ((dead and air looking)).
        - » dual: That's true, Alex is kind of debonair looking.
        - » orig: and give out accu Esther reports and forecasts to the to the air crews going down range.
        - » dual: And give out +accurate weather reports and forecast to the to the air crews going down range. T-
- Conclusion
  - Quality control necessary
  - Overall quality of transcription team is very good
  - Tasks like SU annotation or topic labeling might improve consistency of sentence-final punctuation

Discussion & Issues

- Transcription issues
  - Segmentation granularity
    - Breath groups are current standard, but
    - SUs (sentence units) are used in other projects
  - Topic shift labeling
  - Speaker noise
    - Treatment of isolated speaker noises like {breath}
    - Ignore?
    - Add VSC for noises?
  - Other research tasks?
- Data issues
  - LDC has some meeting collection capability
  - Could contribute up to 10 hours
    - IHM for all speakers plus 2+ distant mics/session
    - Currently, limited video capability without new funding