# Telephone Speech Corpora:

## New Needs, Languages, Methods and Technology

(http://www.ldc.upenn.edu/)

*Alexandra Canavan, Kevin Walker,*
*Christopher Cieri, David Graff*

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104

# Introduction

- Current Projects
  - CallFriend Russian: Collection, Transcription, Lexicon
  - CallFriend Korean: Transcription and Lexicon
  - Switchboard Cellular (SB-2 Phase IV): Collection
- Platform
- Process
  - auditing
  - transcription
  - lexicon development
- Materials for Language Modeling
- Timeline
- Summary of Available Resources
- Possible Future Developments

# Call Collection

- CallFriend Russian
  - 140 unique speakers
  - minimally 20-minute, maximally 30-minute domestic calls
  - demographic info obtained at registration (age, gender, education, country of birth, city where raised)
- CallFriend Korean
  - 120 unique, domestic calls collected in 1995-1996
  - between 5 and 30 minutes each (average 25 minutes)
  - nominally 15 minutes will be transcribed of each of 100 calls
- Switchboard Cellular
  - 190 unique speakers participating in ~10 calls each
  - 5-minute recorded calls with focus on GSM users
  - topics will be given to participants (similar to Phase III topics)
  - demographic info obtained at registration (age, gender, education, country of birth, city where raised)

# Features

- **Software**
  - Currently running Windows NT 4.0 Server w/SP4
  - Perl: database interaction (DBI, DBD), process management
  - VOS: call flow, T-1 signaling, call recording
  - Oracle: PIN verification, call statistics, call activity
- **Hardware**
  - Dialogic PCI based T1 Interface card
  - Dual Processor Dell PowerEdge 4300, w/2 P II 450Mhz processors
  - 48 GB capacity PERC RAID Level 5, 32 MB cache
  - 6 UltraWide SCSI, 10000 RPM Seagate Cheetah Harddisks
  - 256 MB RAM
  - APC 1400 UPS
- **Features**
  - Cost-effective and easy to program and administer.
  - High Capacity: T-1, up to 12 simultaneous 2 channel conversations could be recorded. Up to 760 hours of digitized speech can be stored on the local RAID.
  - Secure: Redundant drives, power supplies, UPS, tape robot does daily backup.
  - Good support/good relationship with Parity Software (suppliers of VOS)
  - Exploits strong knowledge base at LDC in employed technologies (Oracle, Perl)

# VOS Language

The VOS programming language (Parity Software) is easy to learn, but allows for a great deal of control over telephony resources. VOS scripts are structured and resemble languages like Perl or C, allowing LDC to draw on the programming expertise of existing staff. Familiarity with the programming model also make maintenance and development of new applications much easier.

**VOS provides specialized functions which serve as a front-end to the Dialogic API. Here is an example of the VOS code needed to answer an incoming call on a T-1 channel using E&M Wink signaling.**

This loop waits for the wink generated by the telco to signal an incoming call.

Sets the A and B signaling bits for the given timeslot.

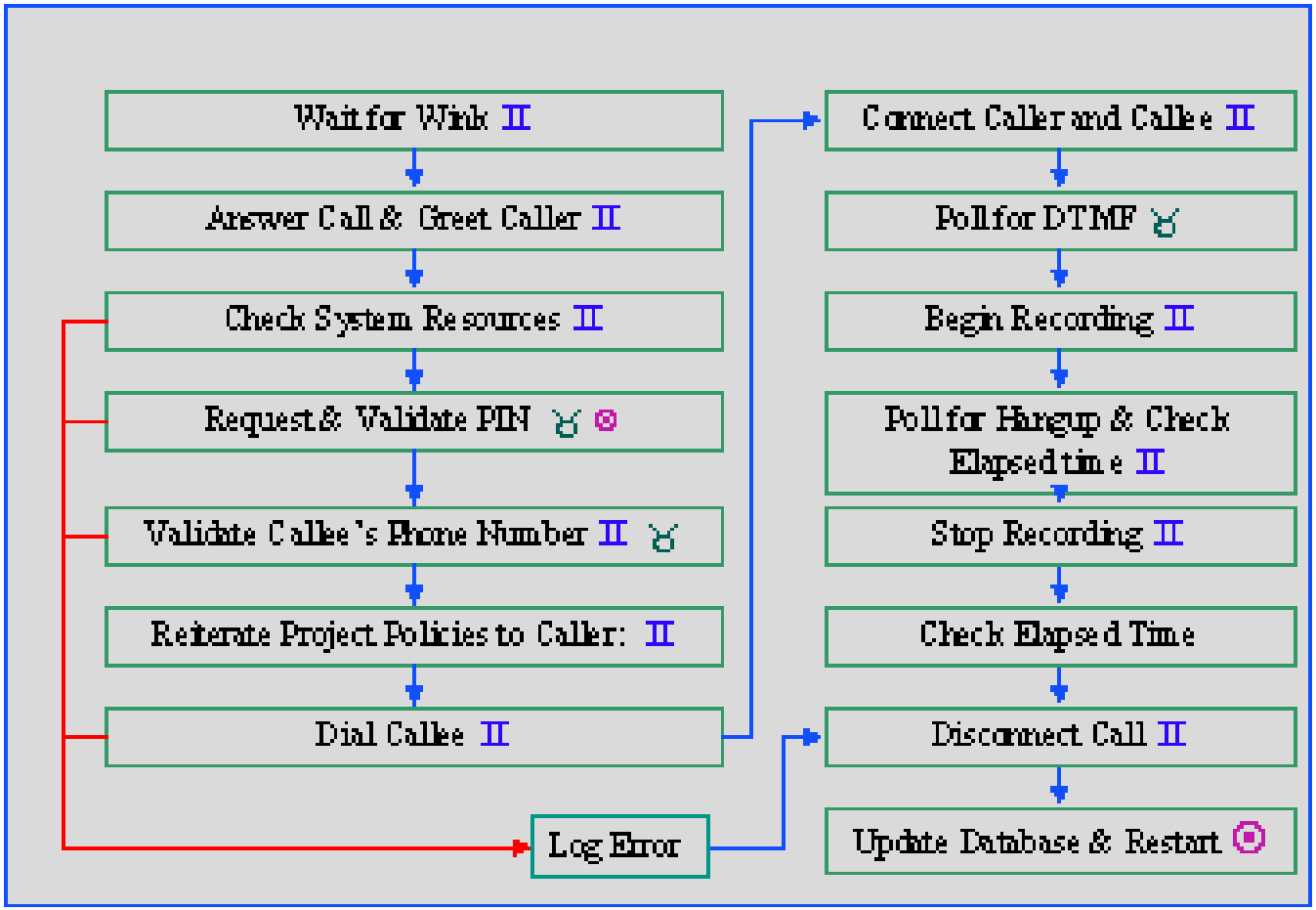
```
func answer_incoming_call()

     DTI_clrsig(DTIboard, DTIchan, 3);
     DTI_clrtrans(DTIboard, DTIchan);
     DTI_watch(DTIboard, DTIchan, "Aa");
     DTI_use(DTIboard, DTIchan , "a");
     DTI_clrtrans(DTIboard, DTIchan);
     do
          DTI_wait(DTIboard, DTIchan);
     until (DTI_trans(DTIboard, DTIchan, "A"));
     DTI_wink(DTIboard, DTIchan);
     DTI_clrtrans(DTIboard, DTIchan);
     sc_getdigits(VOIchan, 17, 10, 10);
     ADinfo = sc_digits(VOIchan);
     sc_clrdigits(VOIchan);
     ani = substr(ADinfo, 2, 10);
     dnis = substr(ADinfo, 13, 4);
     DTI_setsig(DTIboard, DTIchan, 3);
     return(ani,dnis);

endfunc
```

# Call Flow

## High Level Overview of CallFriend Russian Call Flow



# Auditing

- CallFriend Korean and Russian
  - listen to entire call prior to transcription
  - mark gender information of caller and callee
  - identify dialect for caller and callee when confident
  - make judgements on quality of call (echo, bg noise, distortion)
- Switchboard-2 Cellular
  - listen to three of five minutes
  - verify speaker identification across calls with same PIN
  - make judgements on quality of call (echo, bg noise, distortion)
  - remark on known disruptions (call waiting, traffic, static)
- Rejection
  - non-native speaker of target language
  - repeat speaker
  - non-target language > % 5 of call

# Transcription

- Audio segmentation
  - identify turn boundaries
  - place timestamps at turns and additional intervals as needed
- First pass transcription
  - follows HUB-5 conventions
    - » verbatim transcription of all speech on each channel
    - » disfluencies, overlapping speech fully transcribed
    - » standardized representation of acronyms, non-lexemes, interjections, noises
- Second pass transcription
  - file checked for common segmentation & transcription errors
  - spell-check performed (including proper names)
- Additional QC measures
  - regular spot-checking by "language leader"
  - weekly meetings, email list to discuss issues & give feedback

# Lexicon

- Sources
  - transcripts, reference, morphological expansion
- Content
  - Orthographic form
    - » native character set
    - » Romanized form where appropriate
  - Pronunciation(s)
    - » citation type, for standard reference dialect
    - » variants, predictable variants derived by rule
  - Morphosyntactic features
    - » morphological analysis where necessary
    - » names, foreign words, etc.
  - Frequency counts
    - » in transcripts and other corpora

# Methods

- Words from
  - transcripts
  - available dictionaries
  - other corpora
  - hand-entered basic vocabulary classes
- Minimal subset selected for coverage of conversational speech
- Pronunciations
  - rule-generated and hand checked if possible
  - entered entirely by hand if necessary
- Morphosyntactic features
  - generated by FST if possible
  - entered by hand if necessary

# Supplemental Resources

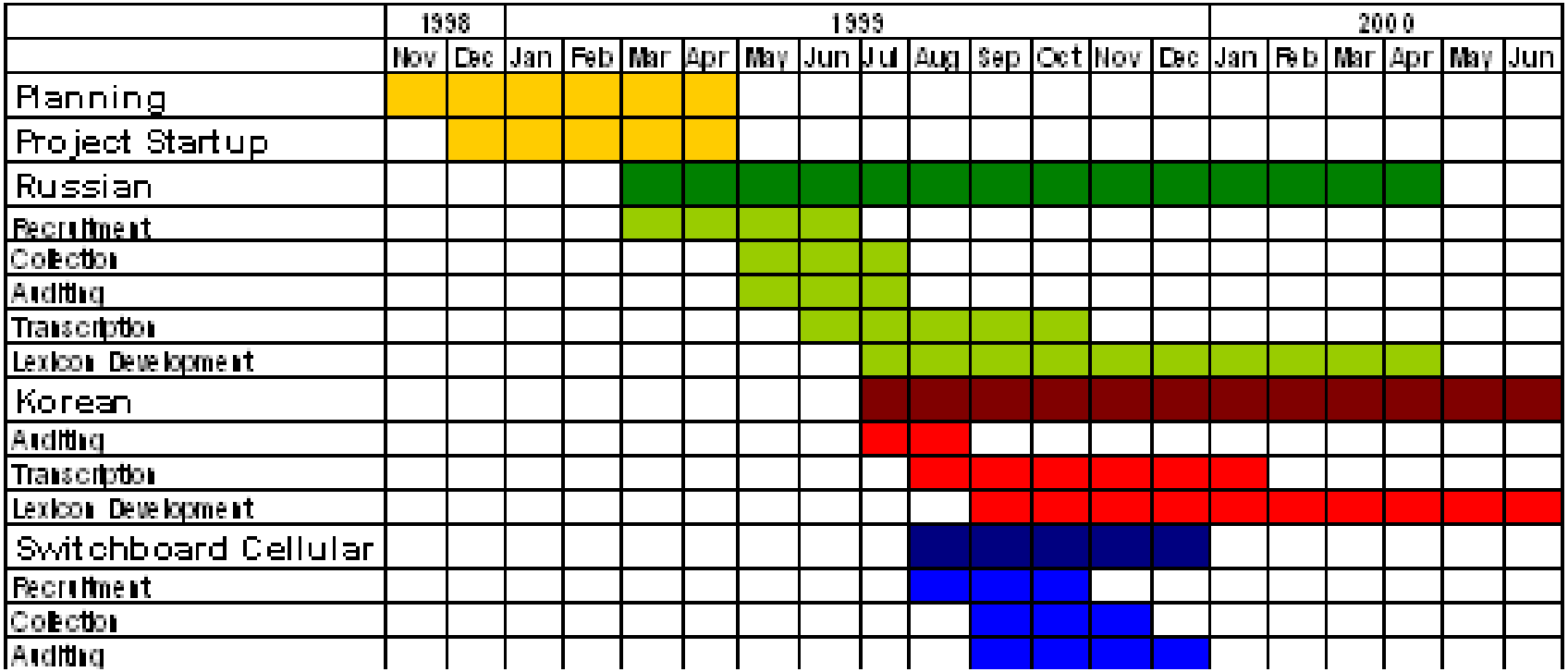
## Lexicons

| Language | Original | Supplement | 1999 Release |
|---|---|---|---|
| Mandarin | 44,405 | 5,190 | 5,405 |
| Spanish | 45,582 | 5,235 | 5,960 |
| Japanese | 80,688 | | |
| German | 318,809 | | |
| English | 90,987 | | 3,902 |
| Arabic | 16,641 | | 23,724 |

## Text for Language Modeling

| | | |
|---|---|---|
| Russian | Russica-Izvestia Information | 13,718,860 |
| Korean | Korean Press Agency | 16,378,651 |

# Timelines


1998
1999
2000
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Planning
Project Startup
Russian
Recruitment
Collection
Auditing
Transcription
Lexicon Development
Korean
Auditing
Transcription
Lexicon Development
Switchboard Cellular
Recruitment
Collection
Auditing

# CallFriend

| Language | Collected | Trans. | #Trans. | Released | N.Released | Prob. | Lex. | From-To (Dates) |
|---|---|---|---|---|---|---|---|---|
| Egyptian Arabic | 177 | No | 0 | 60 | 96 | 21 | No | 06/22/95-11/03/96 |
| Canadian French | 179 | No | 0 | 60 | 100 | 19 | No | 09/29/95-11/22/95 |
| English | 340 | No | 0 | 120 | 217 | 3 | No | 06/18/95-04/01/96 |
| Farsi | 120 | No | 0 | 60 | 49 | 11 | No | 06/27/95-10/26/96 |
| German | 106 | No | 0 | 60 | 37 | 9 | No | 06/18/95-03/31/96 |
| Hindi | 116 | No | 0 | 60 | 51 | 5 | No | 06/23/95-03/17/96 |
| Japanese | 101 | No | 0 | 60 | 34 | 7 | No | 06/26/95-03/11/96 |
| Korean | 128 | No | 0 | 60 | 60 | 8 | No | 06/20/95-02/23/96 |
| Mandarin | 271 | Yes | 42/4hr | 120/42* | 144 | 7 | Yes | 06/24/95-01/21/96 |
| Spanish | 203 | Yes | 106/30hr | 120/106** | 77 | 6 | Yes | 06/20/95-11/02/96 |
| Tamil | 109 | No | 0 | 60 | 38 | 11 | No | 06/30/95-02/11/96 |
| Vietnamese | 112 | No | 0 | 60 | 46 | 7 | No | 07/17/95-10/21/96 |

* Mandarin 120 calls released as CallFriend, 42 released as Hub-5

**Spanish 120 calls released as CallFriend, 106 released as Hub-5

# CallHome

| Language | Collected | Trans. | #Trans. | Rel. | Not Rel. | Prob. | Lex. | From-To (Dates) |
|---|---|---|---|---|---|---|---|---|
| Egyptian Arabic | 291 | Yes | 200, 10 Min. | 120 | 80 | 158 | Yes | 06/21/95-02/02/96 |
| English** | 371 | Yes | 200, 10 Min. | 120 | 80 | 148 | Yes | 06/13/95-03/24/96 |
| German | 235 | Yes | 200, 10 Min. | 100 | 100 | 29 | Yes | 06/25/95-03/31/96 |
| Japanese | 95* | Yes | 200, 10 Min. | 120 | 80 | 0 | Yes | 06/17/95-11/15/95 |
| Mandarin | 40* | Yes | 200, 10 Min. | 120 | 80 | 6 | Yes | 06/20/95-07/22/95 |
| Spanish | 15* | Yes | 200, 10 Min. | 120 | 80 | 18 | Yes | 06/13/95-08/13/95 |

*Number Collected at LDC, original collection done at Texas Instruments

** 30 domestic calls included in publication

# Switchboard

| Switchboard | | | | |
|---|---|---|---|---|
| Collected | Female | Male | Total | Trans. |
| 2400 | 241 | 302 | 543 | Yes/All |

| Switchboard-2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Collected | Female | Male | Total | Avg. Calls | Trans. | #Trans. |
| Phase I | 3638 | 358 | 299 | 657 | 11.07 | Yes | 20 |
| Phase II | 4472 | 352 | 327 | 679 | 13.17 | Yes | 20 |
| Phase III | 2728 | 348 | 292 | 640 | 8.53 | Pending | Pending |

# Future Developments

- Transcribe Current CallFriend Collections
  - Egyptian Arabic, Canadian French, English, Farsi, German, Hindi, Japanese, Tamil, Vietnamese
- Additional CallFriend Collections in New Languages
  - which ones?
- Continued Cellular Collection
- Enhancements to Lexicons
  - English, Spanish, Arabic updates to be released this summer
  - Japanese updates in progress