Corpus Creation and Quality Control at the LDC

Stephanie Strassel

strassel@ldc.upenn.edu

Linguistic Data Consortium

University of Pennsylvania

Philadelphia, PA 19104

www.ldc.upenn.edu

LDC Background

- A non-profit activity of the University of Pennsylvania
- An open consortium of universities, government agencies and companies
- Founded in 1992 with DARPA/NSF support
- Now self-supporting through membership fees and corpus sales
- Mission to create, publish, promote and archive language resources
- for education, research, clinical practice and technology development related to language

LDC Principles

Publish the data that researchers need

- data for sponsored programs (TDT, Hub-4, OLEADA,...)
- data from community initiatives (ACL/DCI, Unipen..)
- data from non-LDC projects (CSAE, CELEX, Trains...)
- LDC-funded data (Treebank, COMLEX, WordNet...)

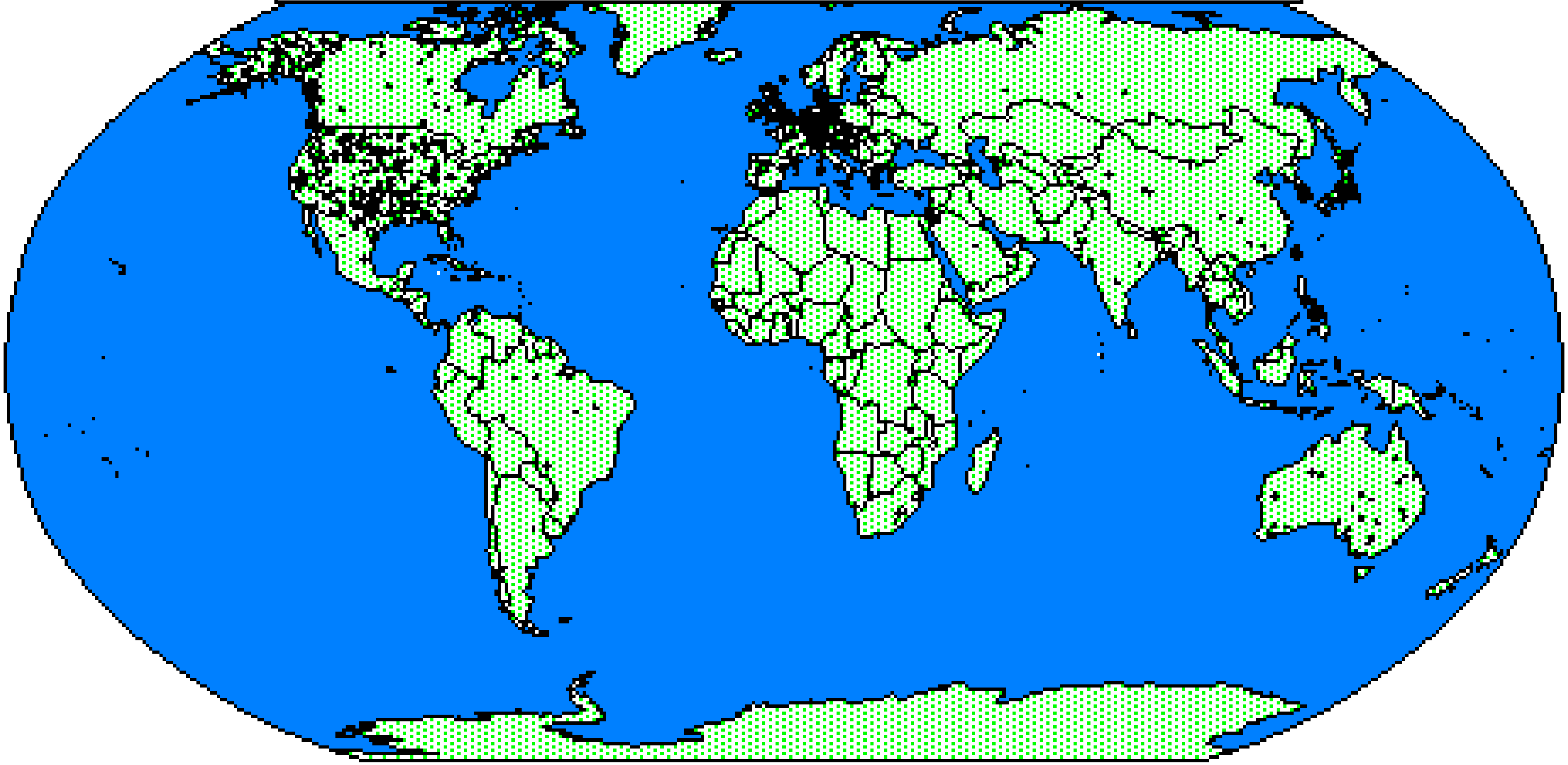
Make data available to everyone

- consortium membership is open to all
- most databases available to non-members

Promote the idea of shared resources

- IPR intermediary
- advice on collection, publication and IPR issues
- development of standards & tools for more useful publication

LDC Members/Users



Languages

	>100			>40		8	>35
	Speech / Transcripts						
Language	*Broadcast*	*Telephone*	*WideBand*	*Parallel Text*	*Newswire/ Other Text*	*Lexicon*	*Traditionl. Dictionary*
Arabic (Egyptian)		🟦			🟦	🟦	
Czech	🟦						🟦
Dutch					🟦	🟦	🟦
English	🟦	🟦	🟦	🟦	🟦	🟦	🟦
French		▨		🟦	🟦		🟦
German		🟦		🟦	🟦	🟦	🟦
Hindi		▨					
Japanese		🟦	🟦		🟦	🟦	
Korean		🟦				🟦	
Mandarin	🟦	🟦		🟦	🟦	🟦	
Persian		🟦			🟦		
Portuguese					🟦		🟦
Russian		🟦			🟦		🟦
Serbo-Croatian					🟦		🟦
Spanish	🟦	🟦	🟦	🟦	🟦	🟦	🟦
T. Putonghua			🟦				
Tamil		▨			🟦		
Thai					🟦		
Turkish					🟦		
Vietnamese		▨					

- **Afrikaans, Bamileke, Basque, Estonian, Hungarian, Italian, Kazakh, Kurdish, Latvian, Manding, Polish, Slovene, Ukrainian, Uzbek, Xhosa, Yoruba**

Uses of LDC Data

Linguistic technologies

- Topic Detection & Tracking, Information Retrieval, Message Understanding
- Speech Recognition and Speech Synthesis
- Machine Translation
- Language and Speaker Identification
- Language Teaching

Linguistic research topics

- Parsing
- Sense Disambiguation
- Discourse Modeling
- Prosody
- Language Acquisition & Language Teaching
- Sociolinguistic Variation Studies

LDC Resources: Specialized Equipment

Servers

- Unagi/Morph/X - research computing, LDC Online
  - 2 Sun E4000 multi-processors with >1GB RAM
  - >1TB disk shared
  - Two 3.5TB tape robot for backup and near-line storage
- Easter - separate administrative server, RAID, tape robot
- Dedicated fiber-optic network

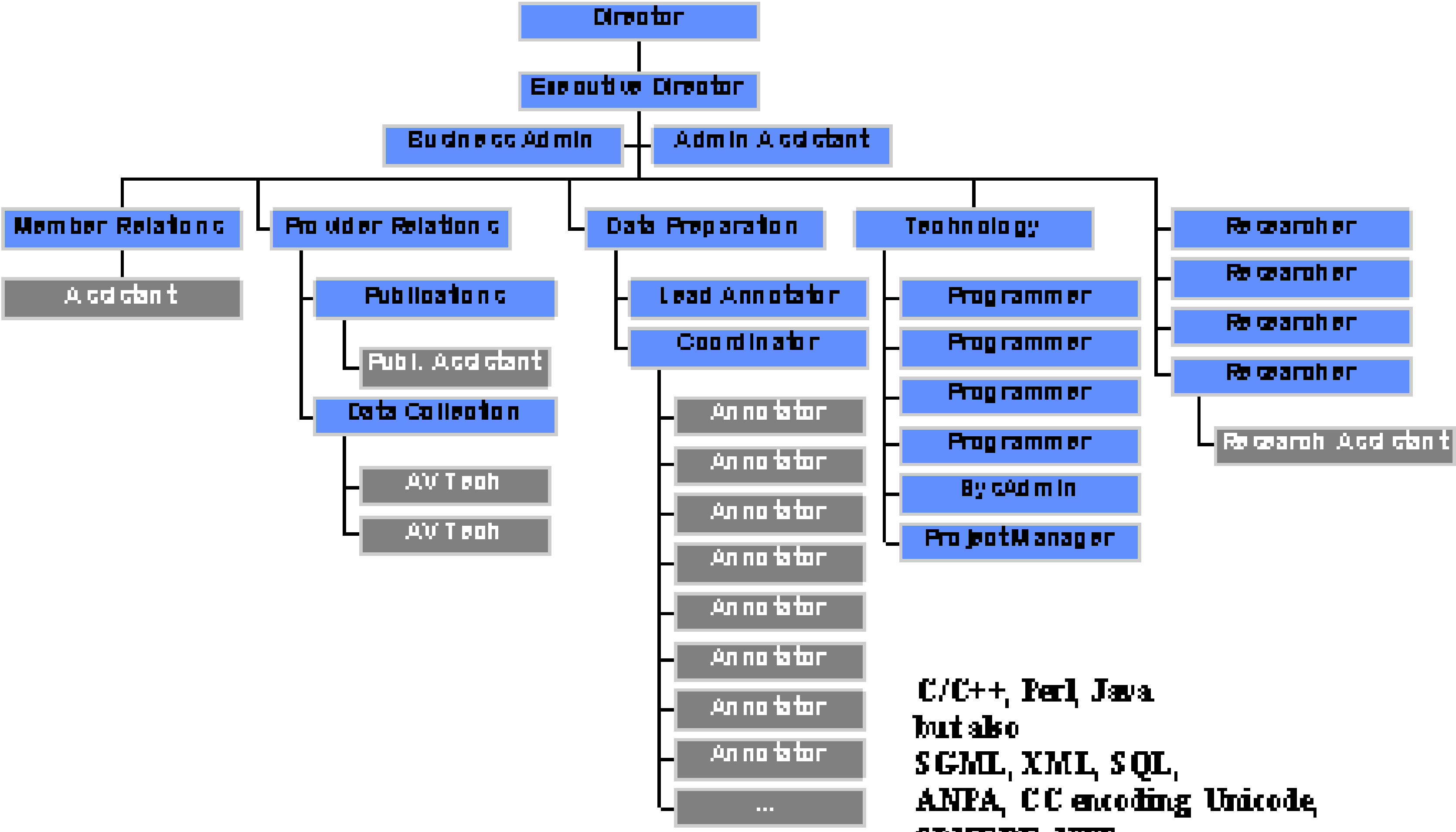
Collection resources

- Telephone Collection - 45GB RAID disk, T1 access
- Satellite Downlink - multifunction, receives VOA
- Collection Workstations - newswire, WWW, broadcast audio & video
- A/V receivers & recorders, CC decoders, DATs, etc.

Workstations

- > 60 Sparcs, >20 PCs, few Macs for compatibility

LDC Resources: Staff



- Director
  - Executive Director
    - Business Admin
    - Admin Assistant
    - Member Relations
      - Assistant
    - Provider Relations
      - Publications
        - Publ. Assistant
      - Data Collection
        - AV Tech
        - AV Tech
    - Data Preparation
      - Lead Annotator
      - Coordinator
        - Annotator
        - Annotator
        - Annotator
        - Annotator
        - Annotator
        - Annotator
        - Annotator
        - Annotator
        - ...
    - Technology
      - Programmer
      - Programmer
      - Programmer
      - Programmer
      - SysAdmin
      - ProjectManager
    - Researcher
    - Researcher
    - Researcher
    - Researcher
      - Research Assistant

C/C++, Perl, Java
but also
SGML, XML, SQL,
ANPA, CC encoding, Unicode,
SPHERE, UTF
PerlTK, PerlDbi, Emacs-Lisp

Corpus Planning

Sponsors (or other initiators) with specific needs

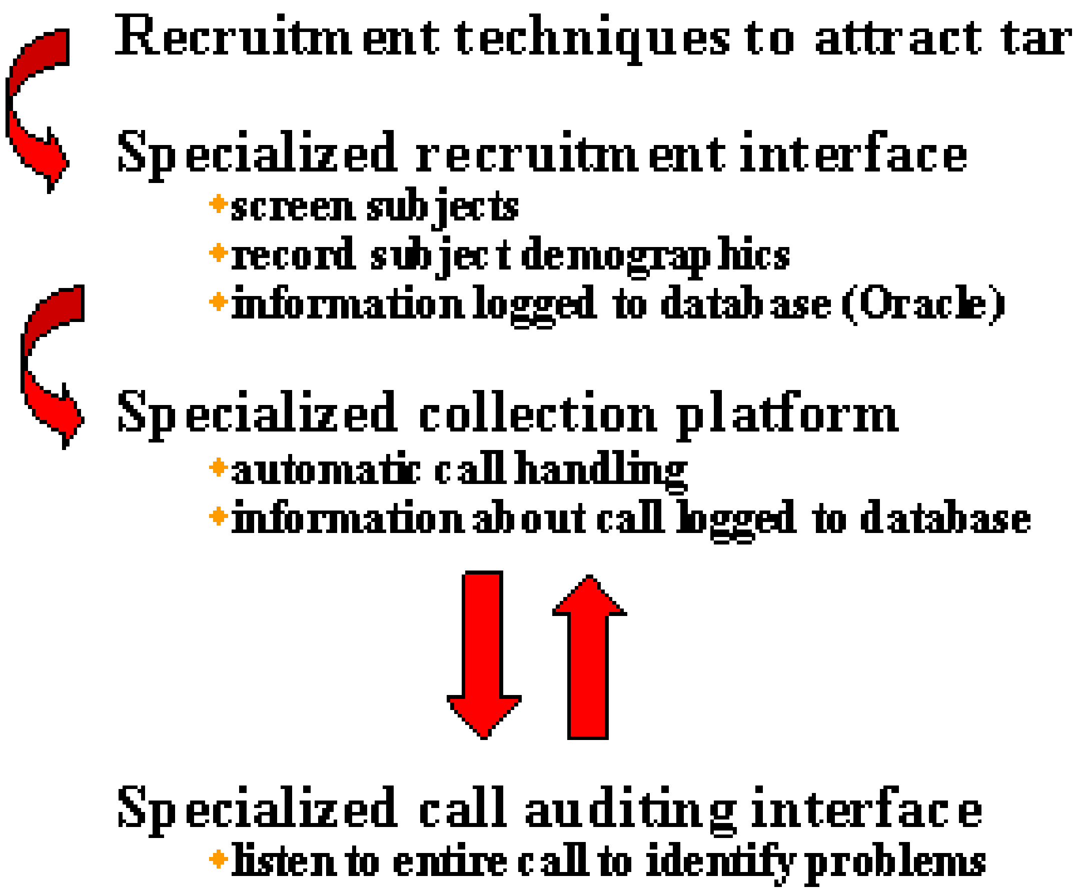
Final data structure (the deliverable)

Design Specifications

- data format (speech & text)
- types of annotation
- annotation specifications
- method of distribution
- etc.

Structure of the final product informs all aspects of corpus creation, including QC

Data Collection: Telephone Speech



Recruitment techniques to attract target subjects

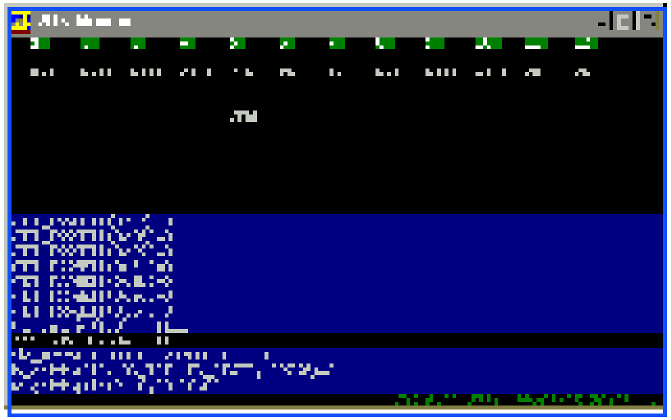
Specialized recruitment interface
- screen subjects
- record subject demographics
- information logged to database (Oracle)

Specialized collection platform
- automatic call handling
- information about call logged to database

Specialized call auditing interface
- listen to entire call to identify problems

Platform Overview

- Off the shelf equipment (Dialogic , Dell)
- Written in Perl and VOS, a structured, telephony enabled, interpreted programming language
- Easy to program/maintain and very reliable
- Applications are flexible - collections in new languages can be started by simply recording new prompts and making a few minor adjustments to the collection environment.
- Adequate resources to run multiple collections simultaneously (>740 hours of digitized speech versus 40 hours on old platform).



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

Annotation: Staff & Training

Staff

- Large (30+), transient annotation team
- 3 fulltime managers

Training

- Ongoing individual & group training

(~15% of annotation budget - time & financial)

- general orientation to LDC and task
- specialized tool training (interfaces, etc.)
- application of annotation spec
- practice files, "quizzes"
- regular feedback: weekly meetings, email lists, etc.

All resources put into Annotation Guide available on web and hardcopy

Emphasis on documentation and communication

Transcription QC

Multiple, complete passes over the data

Specialized tools for each pass

- Audio segmentation
- Background tagging
- First pass transcription
  - apply transcription specifications (verbatim transcription, additional markup)
- Second pass transcription
  - file checked for common segmentation & transcription errors
  - additional automatic checks performed (spelling, syntax, etc.)
- etc.

Additional QC measures

- ~5% of data at each pass is "spot checked" by team leader
- individual annotator performance monitored daily
- regular feedback

Additional Annotation QC

Dual annotation for 5-10% of data

- double-blind assignment: separate individuals annotate same file
- part of regular work assignment

Discrepancy

- team leader resolves results of dual annotation
- these results reported as Kappa score
- results fed back into training

Precision

- team leaders search for miscategorization of data
  - incorrect timestamps, event identification, speaker turns, lexical tags, etc.

Recall

- team leaders search for uncategorized data
  - missing timestamps, speaker identifications, etc.

Each QC task involves specialized tools, interface

Data Formatting QC: Types of Corpora

Low Maintenance

- collections of speech only

Medium Maintenance

- newswire text collections
- existing text archives

High Maintenance

- transcripts of speech
- lexicons
- manually annotated text

Data Formatting QC: Considerations

Keep speech data isolated from higher maintenance derivatives

Large text collections may need cosmetic "retagging"

- markup needs can evolve over time
- a given corpus may serve multiple tasks needing different markup
- raw material may change format

Multitask usage may require alternate filtering

Prior to publication, multiple "sanity checks" to locate errors

- Check speech & text file headers
- Markup meets expected format
- Character filtering
- No missing attributes or tokens
- Cksum, check file size
- Plausible word/second rates
- Multiple annotations refer to identical source data

...etc.

Publication

(Moving toward) in-house replication on CD-ROM

Eventual move to DVD-ROM

- and internet distribution

LDC-Online

- Available from LDC's web page
- All LDC data online
  - exceptions: IPR issues
- Sophisticated search and retrieval via standard web browsers (audio as well as text retrieval)
- Some materials available to the public
- Everything available to current members
- Potential to expand & improve this

FTP delivery of corpora <50 MB

QC - data maintenance

Conclusion

- planning
  - corpus design, final data structure inform all other stages
- specialization
  - staff
  - skills
  - tools
  - tasks
- reiteration
  - multiple passes over the data at every level of corpus creation
  - multiple individuals involved in each stage of creation
- communication
  - critical for staff at every stage to know big picture