Corpus Sociolinguistics:

Issues, Data and Tools

Stephanie Strassel (strassel@ldc.upenn.edu)
Christopher Cieri (ccieri@ldc.upenn.edu)
University of Pennsylvania
Linguistic Data Consortium and
Department of Linguistics
www.ldc.upenn.edu

NWAVE- 28, Toronto, October 1999

Definitions

Corpus

a collection of data organized for purposes of analysis

Raw Data - direct output of linguistic performance either naturally occurring or elicited.

Annotation - any process of adding value to a corpus through human judgement or automatic analysis

Transcription - a special instance of annotation. Even transcription presupposes some theory about the language

Segmentation - another special type of annotation; chopping a glob of data (virtually) into useable pieces.

LDC

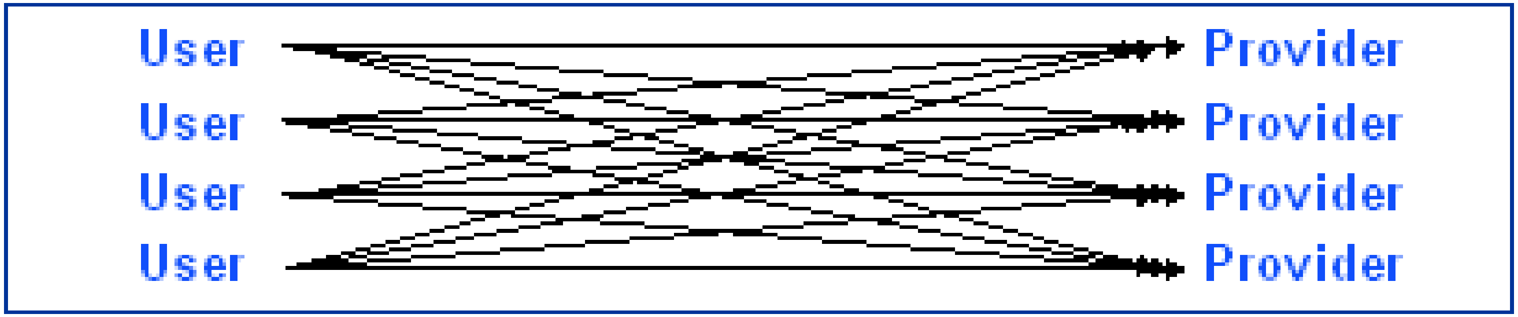
- A non-profit activity of the University of Pennsylvania
- An open consortium of Universities, Government agencies and Companies
- Founded in 1992 with DARPA/NSF support
- Now self-supporting through membership fees and corpus sales
- Mission to create, publish, promote and archive language resources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

Members

- Organizations join the consortium by the year
- Members receive all corpora published in the years in which they join (~20/year)
- Members can exercise their rights to access at any time
- Members also receive year-long access to LDC Online databases
- Some corpora may be sold but many are restricted to members.
- Archive data permanently
- Act as an intellectual property intermediary between data users and data providers

Members

- Organizations join the consortium by the year
- Members receive all corpora published in the years in which they join (~20/year)
- Members can exercise their rights to access at any time
- Members also receive year-long access to LDC Online databases
- Some corpora may be sold but many are restricted to members.
- Archive data permanently
- Act as an intellectual property intermediary between data users and data providers

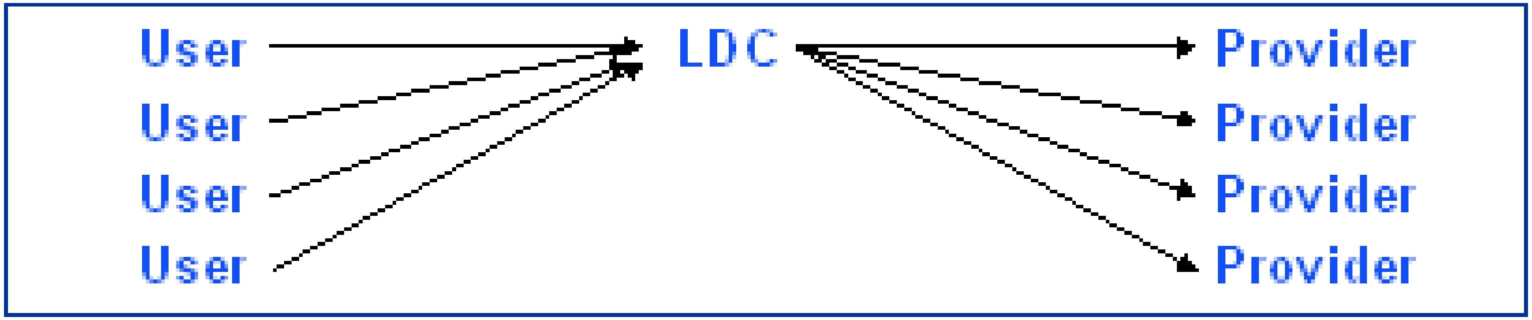


User
User
User
User

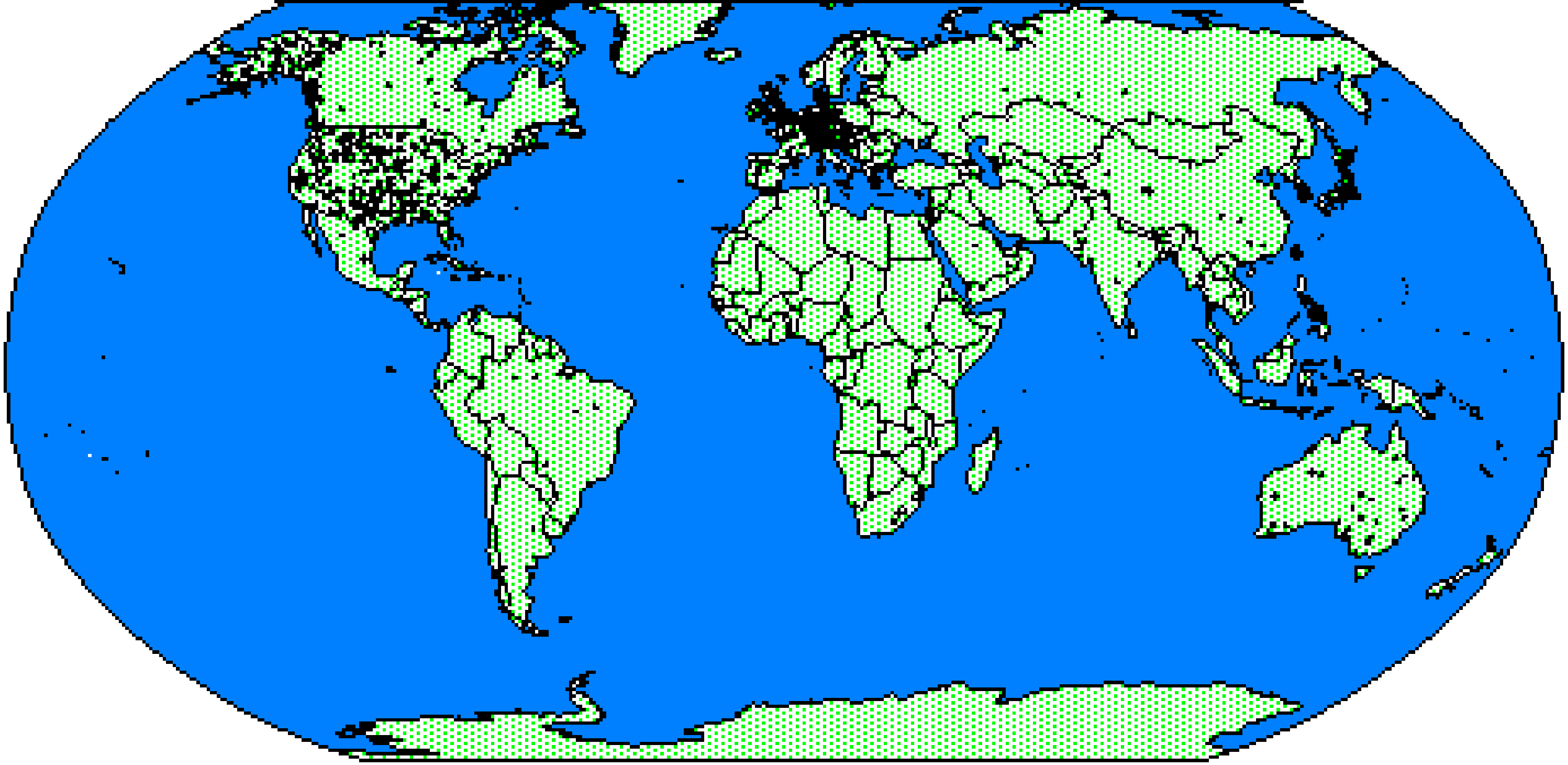
Provider
Provider
Provider
Provider

Members

- Organizations join the consortium by the year
- Members receive all corpora published in the years in which they join (~20/year)
- Members can exercise their rights to access at any time
- Members also receive year-long access to LDC Online databases
- Some corpora may be sold but many are restricted to members.
- Archive data permanently
- Act as an intellectual property intermediary between data users and data providers



LDC Members/Users



LDC as Publisher

Publish data created internally or externally for sponsored programs or community initiatives

Make data available to everyone

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

Membership fees and corpus sales support ongoing publication

Distribute

- primarily on CD, soon on DVD
- via FTP where corpus <50MBs
- where possible via LDC Online indexed in small pieces

Promote the idea of shared resources

- advise on collection, publication, IPR & Corpus Cookbook
- develop standards & tools for annotation - TalkBank
- IPR intermediary

Informed Consent

LDC subscribes to University of Pennsylvania's Human Subjects Protocol.

All subjects are notified that the data the contribute will be used for research and education.

Domains of research left underspecified since LDC serves multiple communities.

Subjects may request their data be expunged ex post facto.

LDC has done a small amount of blanking of sensitive information within interviews (expensive).

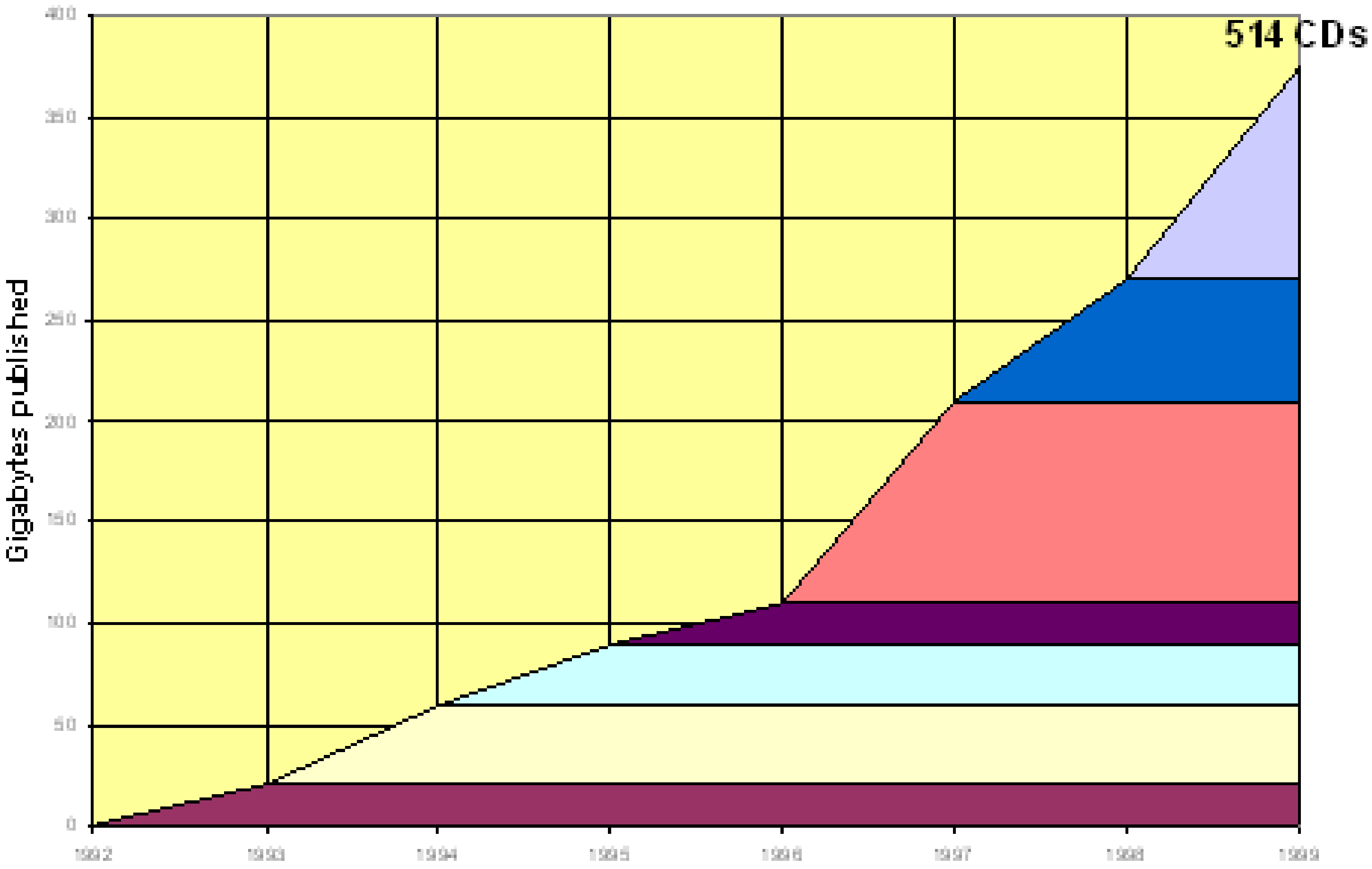
Human Subject's Protocols are updated and reviewed yearly by a University committee.

Languages

	>100				>40	8	>35
	Speech / Transcripts						
Language	*Broadcast*	*Telephone*	*WideBand*	*Parallel Text*	*Newswire/ Other Text*	*Lexicon*	*Traditional 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

How Much is Enough?



Corpus Creation

Collection

- conversational, elicited
  - telephone
  - microphone
- broadcast news
- newswire
- WWW sites

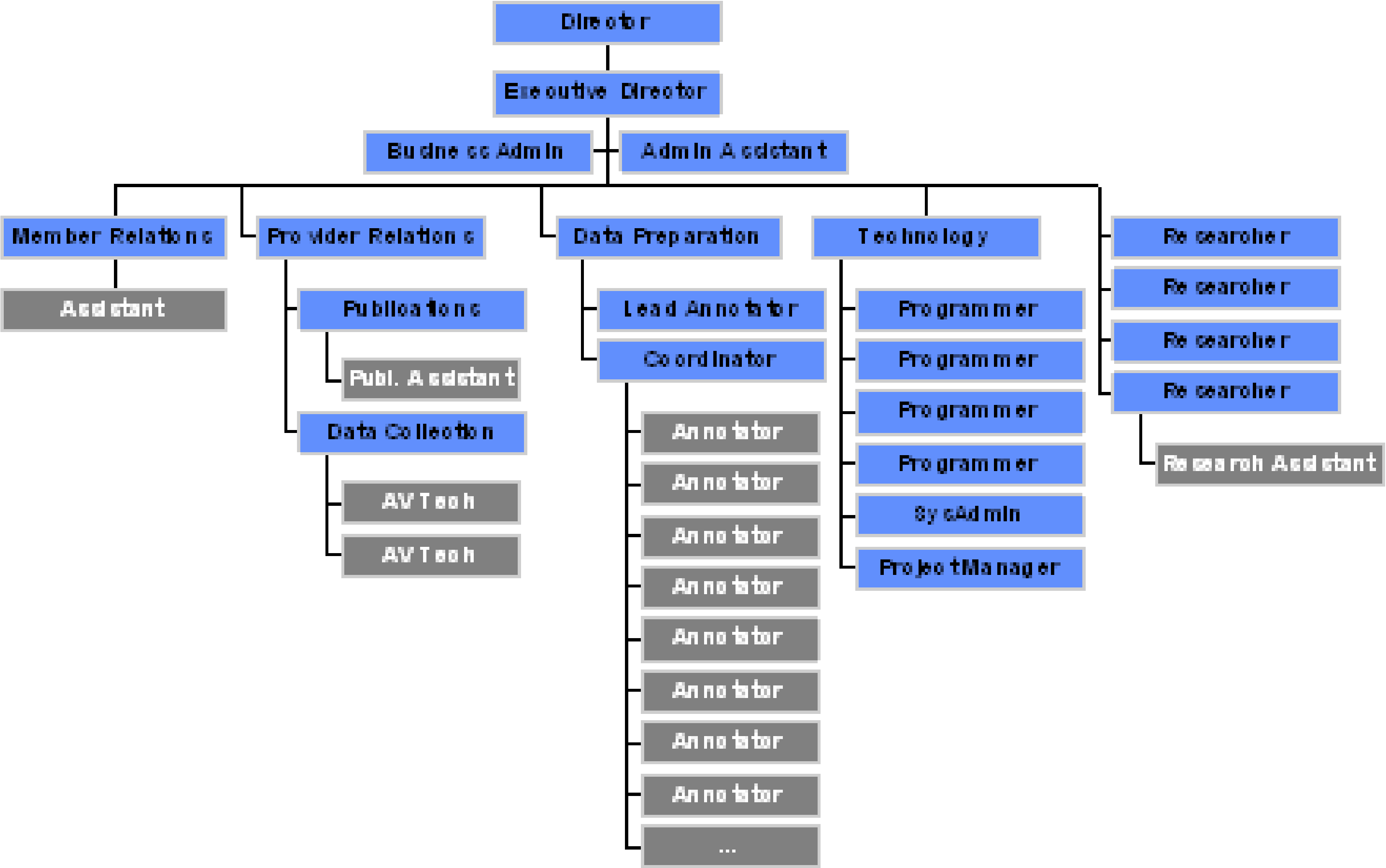
Conversion

- Audio: .sph <= .wav (PC), .aiff (Mac), .au (Sun)
- SGML for text, Unicode compliant where necessary

Annotation (with custom interfaces)

- transcription
- segmentation
- topic relevance
- novel information
- named entity

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
      - Project Manager
    - Researcher
    - Researcher
    - Researcher
    - Researcher
      - Research Assistant

Technical Resources

- 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
- Special
  - Telephone Collection - 45GB RAID disk, T1 access
  - Satellite Downlink - multifunction, receives VOA
  - Collection Workstations - newswire, WWW, broadcast audio & video
- Workstations
  - > 60 Sparcs, >20 PCs, few Macs for compatibility

Using pre-existing corpora

- Why would a sociolinguist want to use a pre-existing corpus?
  - *exploratory study* before doing individual data collection
    - » helps narrow focus to most relevant environments/speakers
    - » locate 'rare' constructions
  - to *supplement* individual data collection:
    - » lots more data, possibly greater range of data
    - » often highly searchable - get lots done quickly
- Accessing the corpora: LDC Online
  - fully available to current members
  - some corpora available with guest account
  - powerful search capabilities:
    - » search by word, phone, lemma, part-of-speech, regular expressions
    - » statistics available (frequency, mutual information, etc.)

What’s out there?

- TIMIT
  - each speaker reads 10 phonetically rich sentences
  - 630 speakers of 8 major dialects of American English
  - time-aligned orthographic and phonetic and word transcriptions and waveform file for each utterance
- HUB-4
  - Broadcast news recordings in English, Spanish, Mandarin Chinese
  - Fully transcribed, time aligned to phrasal level, speaker turns and story boundaries indicated
- CallHome
  - Unscripted telephone conversations between friends & family in English, Spanish, Japanese, Mandarin, German, Egyptian Arabic
  - Partially transcribed, time aligned to speaker turn
  - Speaker demographics (age, gender, region, education) included in documentation
  - Corresponding lexicons available

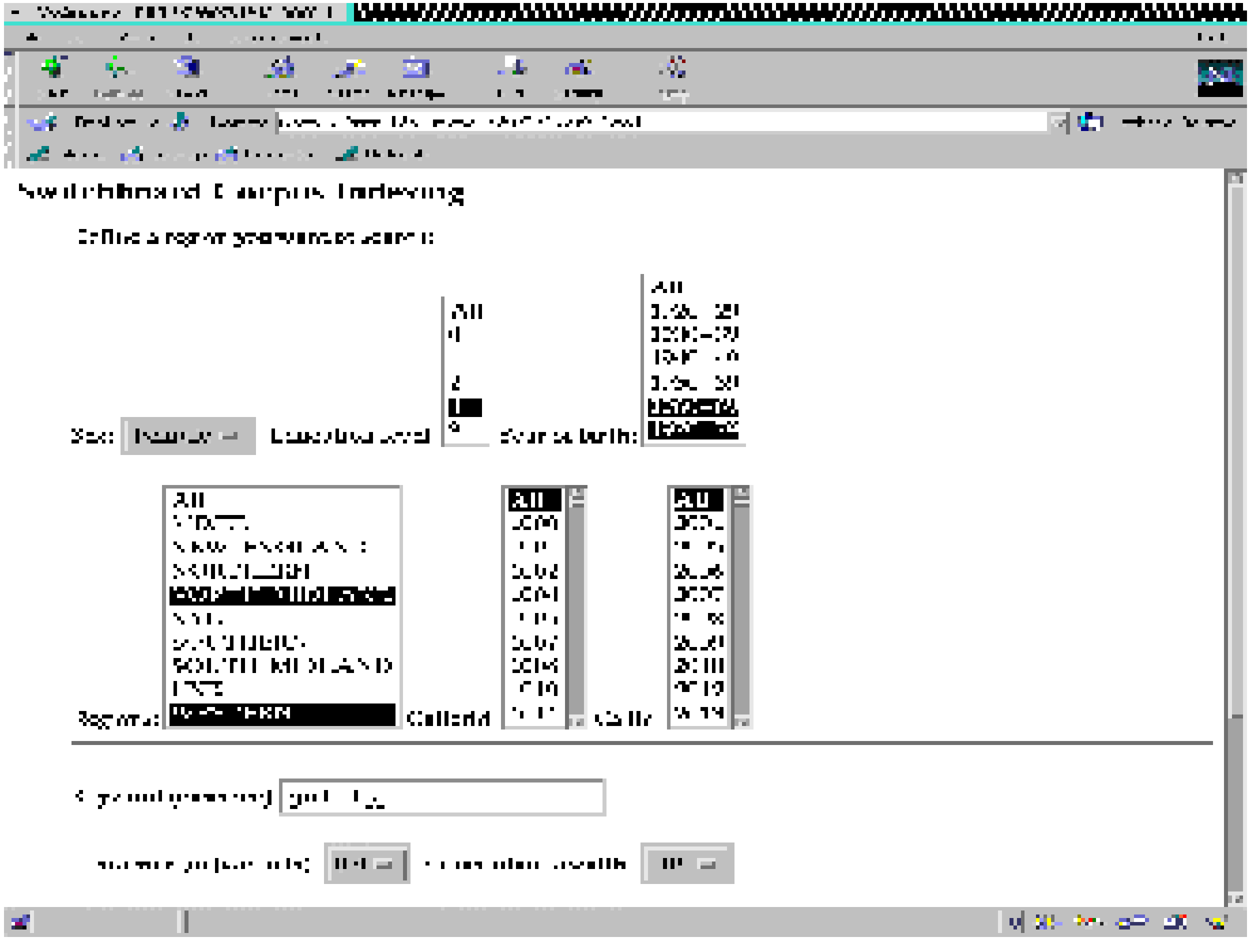
Switchboard

- Corpus contains 2430 conversations averaging 6 minutes in length; in other words, over 240 hours of recorded speech, and about 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English.
  - fully transcribed and word-aligned (timestamped to word level)
  - brand new transcripts and word segmentations

    www.isip.msstate.edu/projects/switchboard/index.html
  - rich online documentation
  - demographic information for speaker easily accessible (age, sex, region, education)
  - powerful search capabilities
    - seach by keyword, part-of-speech, pattern searching
    - search by demographic information of speaker
  - useful as speech or text corpus

Switchboard as Speech Corpus

– Search by speaker demographics, call/caller ID, keyword



Switchboard Corpus Indexing

[illegible]

Sex: Female

Education level: All, 0, 1, 2, 3, 9

Year of birth: All, [illegible]

Regions: All, MIXED, NEW ENGLAND, NORTHERN, NORTH MIDLAND, NYC, SOUTHERN, SOUTH MIDLAND, UNK, WESTERN

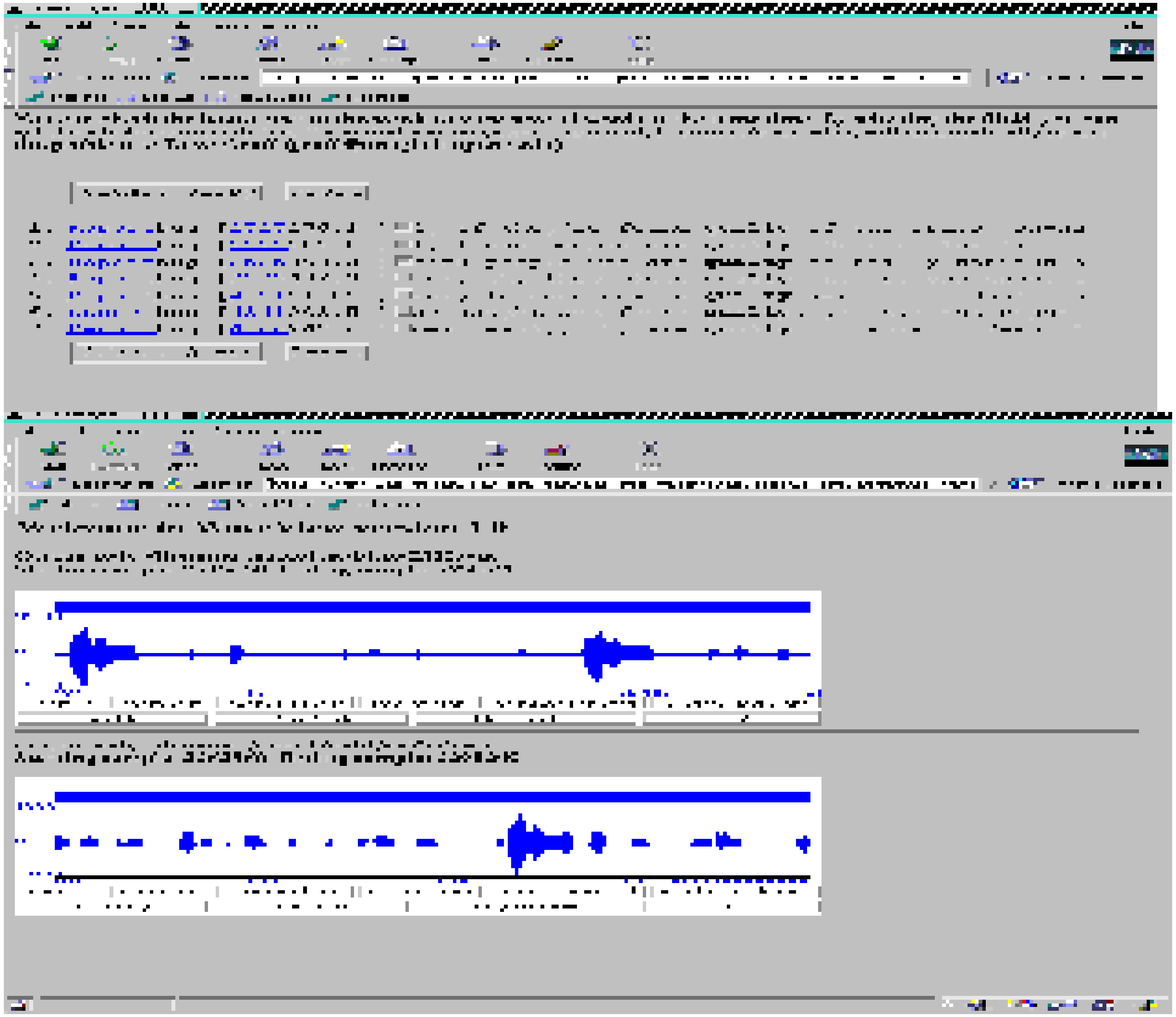
Caller Id: All, [illegible]

Call Id: All, [illegible]

Keyword (required): [illegible]

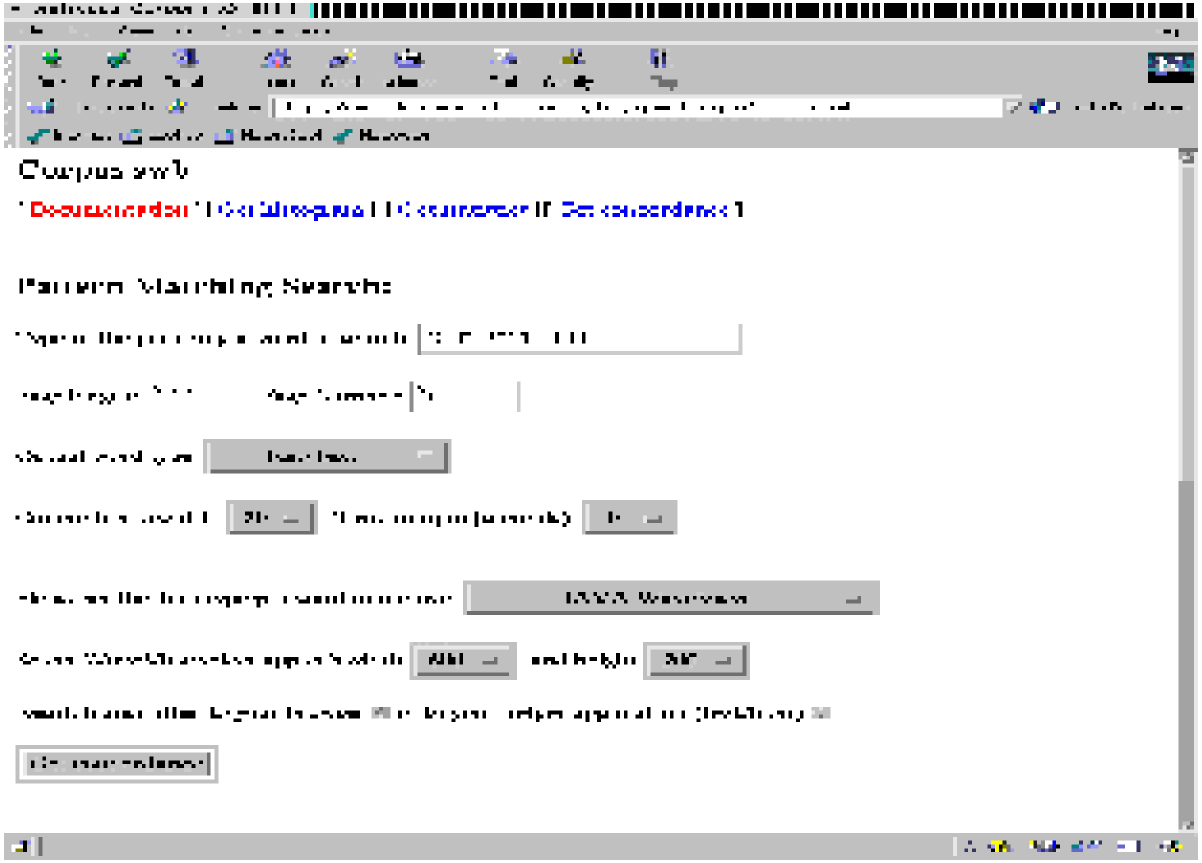
Switchboard as Speech Corpus

– Select individual items to view waveform and listen to speech sample



Switchboard as Text Corpus

- Identify pattern consisting of word(s), part-of-speech tags or combinations of these
- View concordance, histogram, or statistics for each search

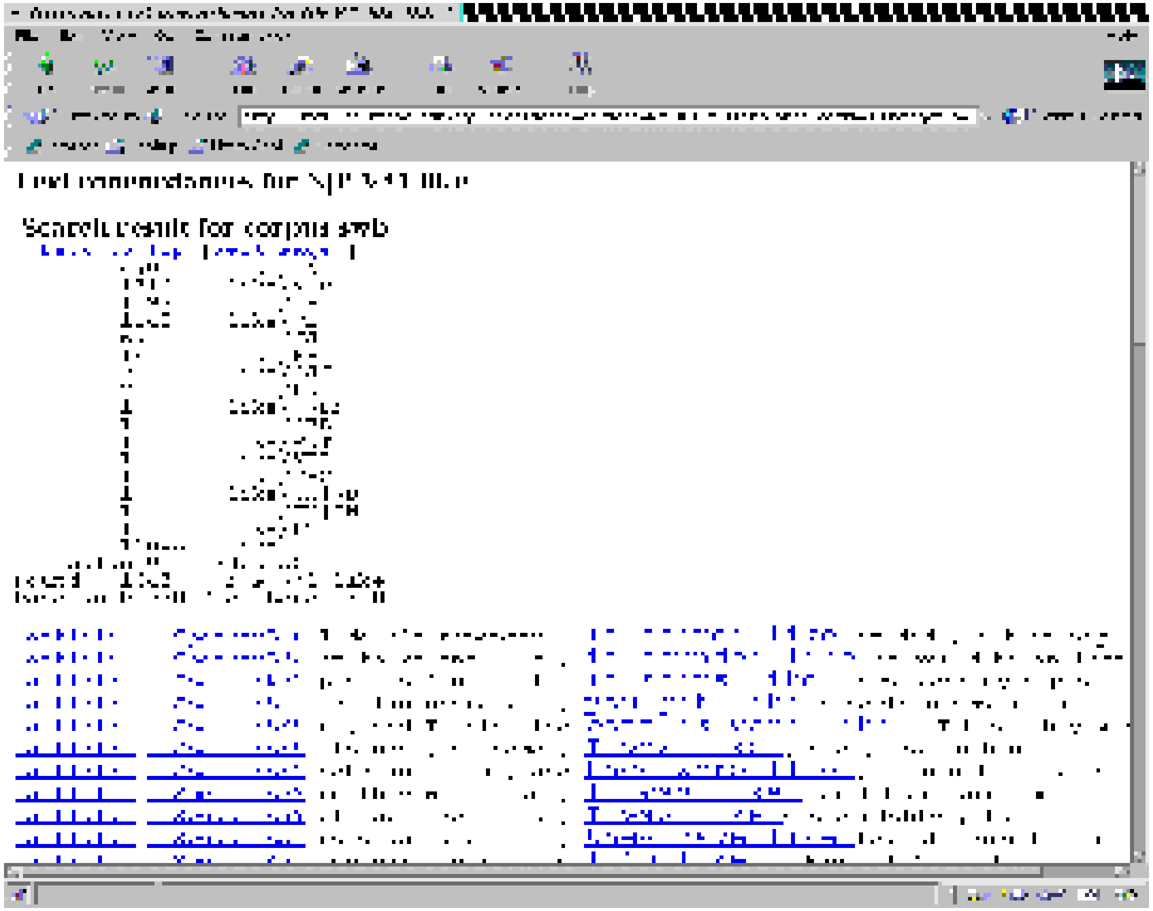


Corpus swb

Pattern Matching Search:

Switchboard as Text Corpus

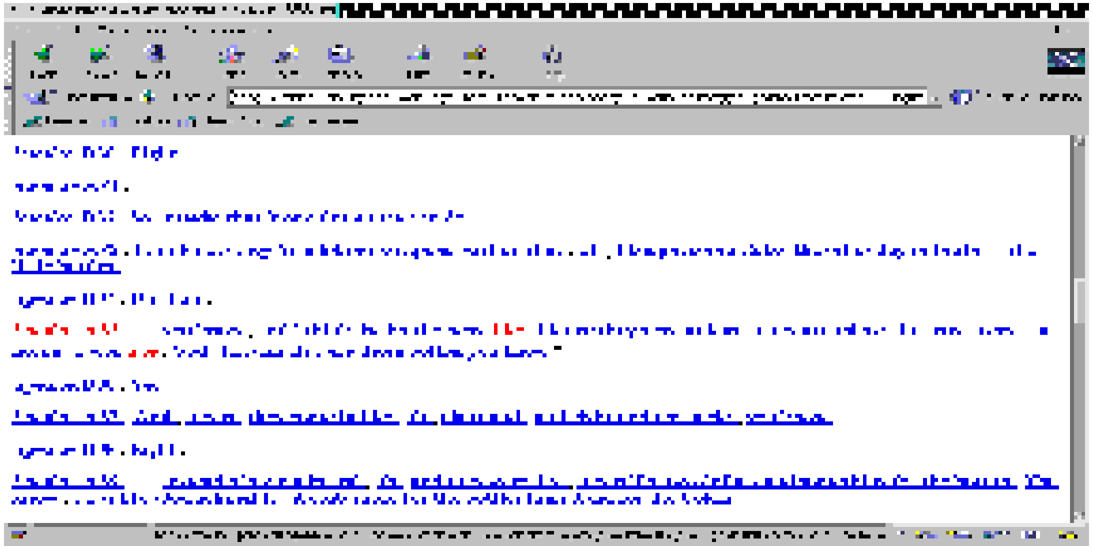
- Identifies P-O-S frequencies for keyword
- Select individual cases for listening or viewing transcripts

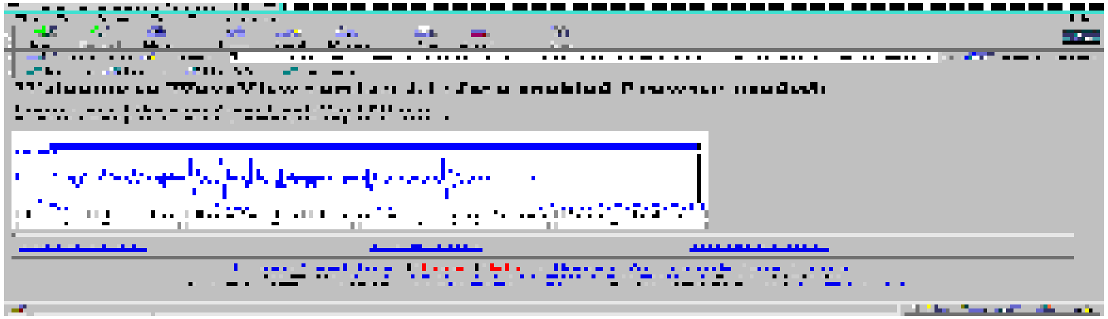


Search results for corpus swb

Switchboard as Text Corpus

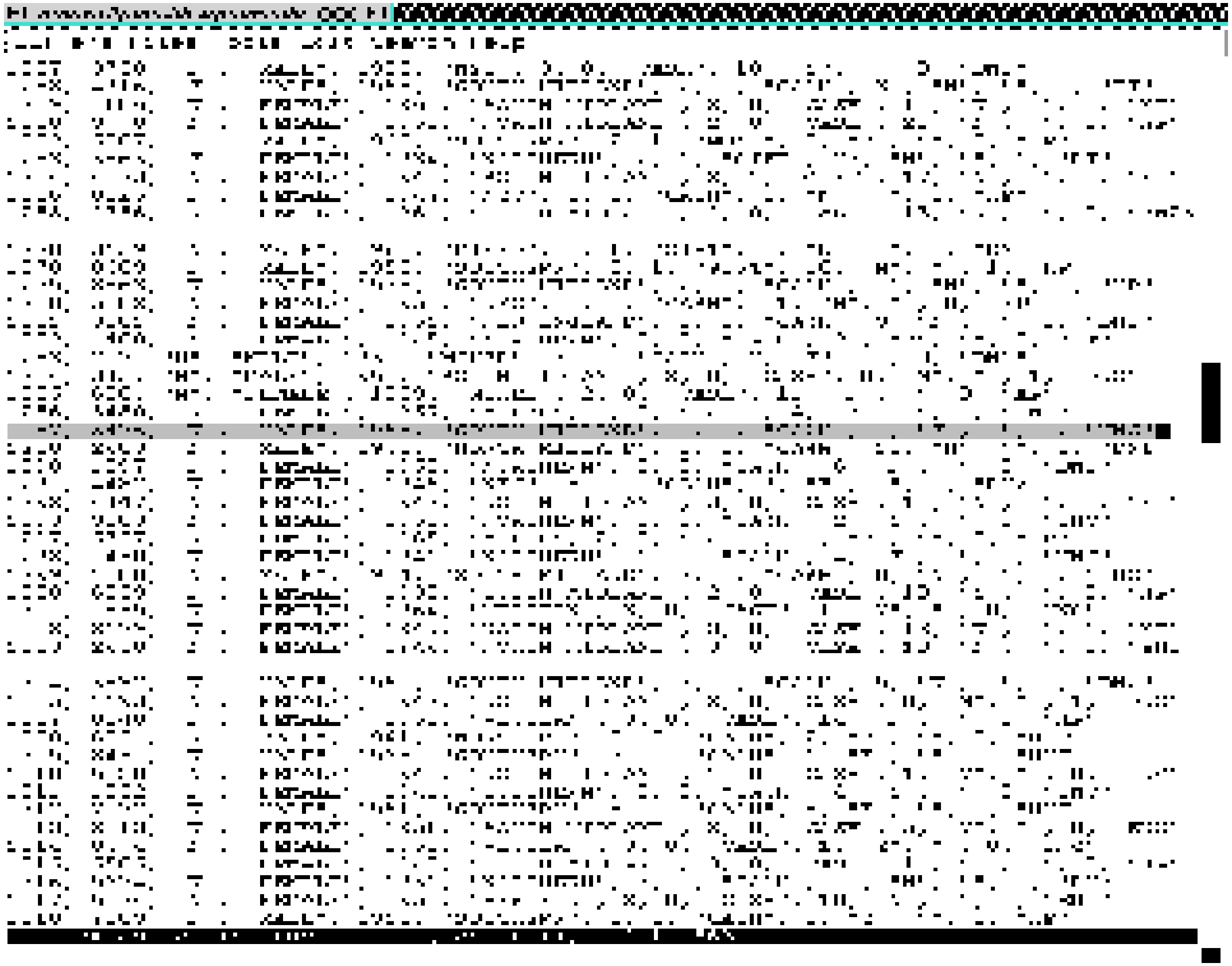
- View entire transcript
- View speaker turn
- Hear focus word & view waveform





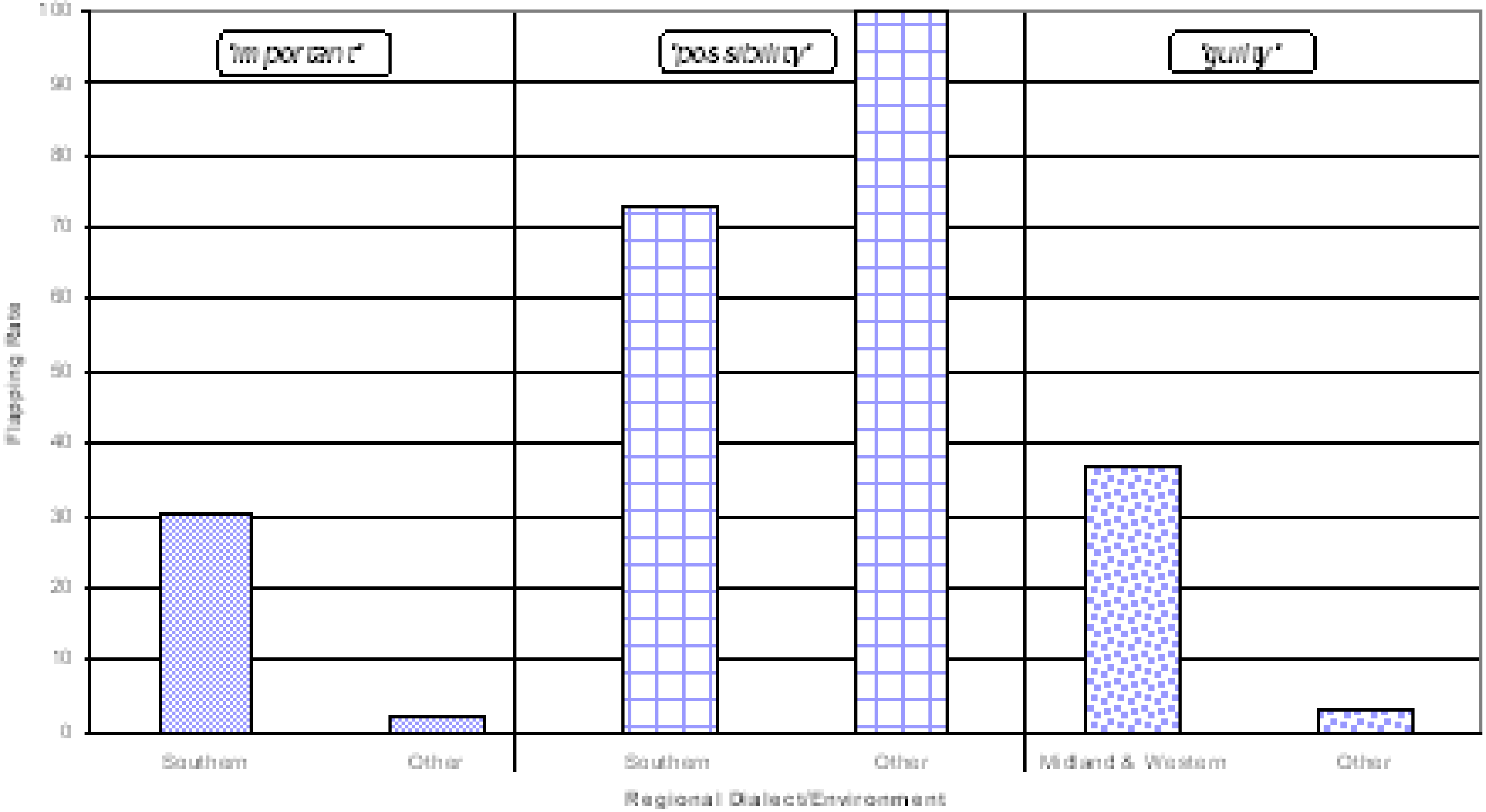
Switchboard as Text Corpus

– Caller demographics included in corpus documentation



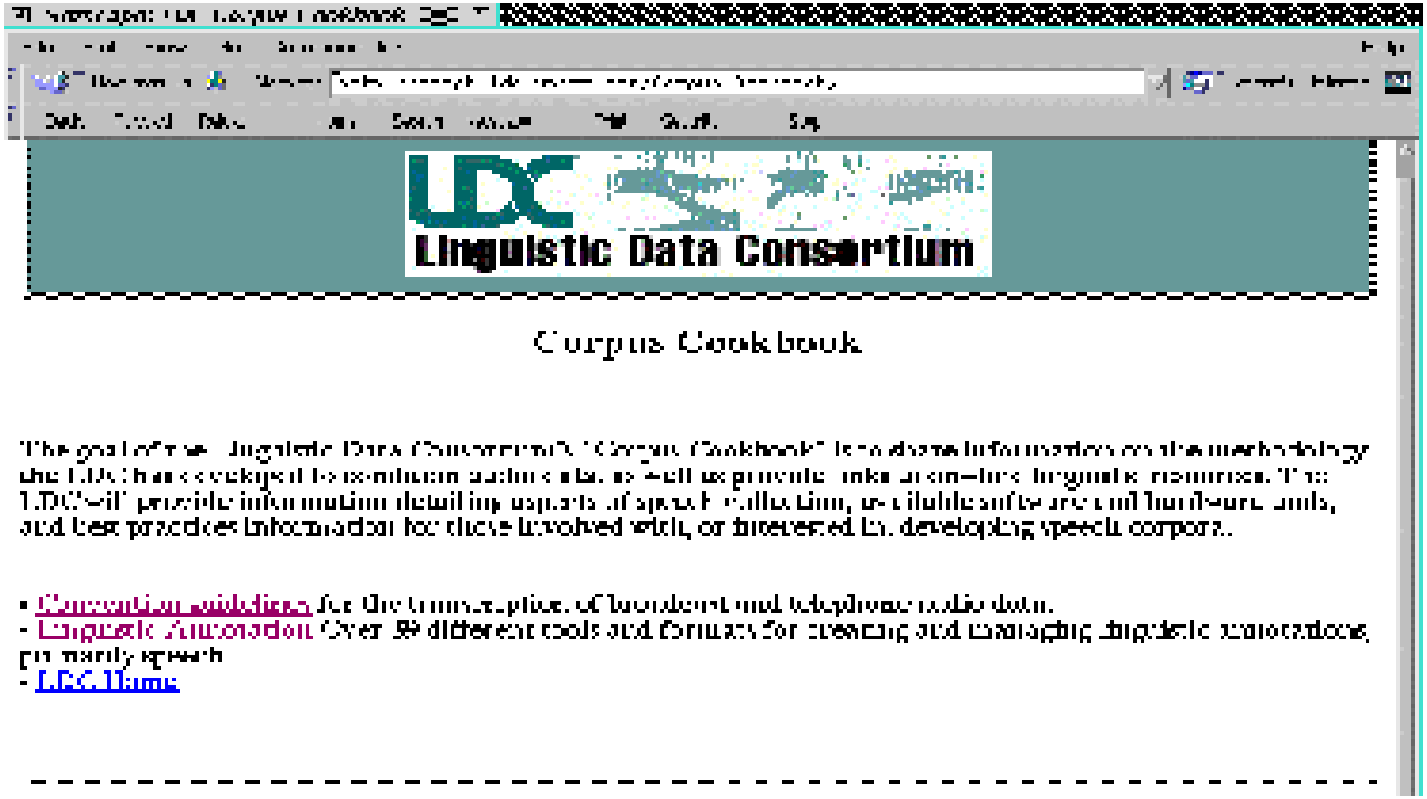
Switchboard: Getting Results

Regional Variation in Flapping: Switchboard Corpus



Tools

The LDC Corpus Cookbook:
A resource page for individual corpus creation efforts
www.ldc.upenn.edu/Corpus_Cookbook
A work in progress...


Netscape: LDC Corpus Cookbook
File Edit View Go Communicator Help
Bookmarks Netsite: http://www.ldc.upenn.edu/Corpus_Cookbook/
What's Related
Back Forward Reload Home Search Netscape Print Security Stop
LDC Linguistic Data Consortium

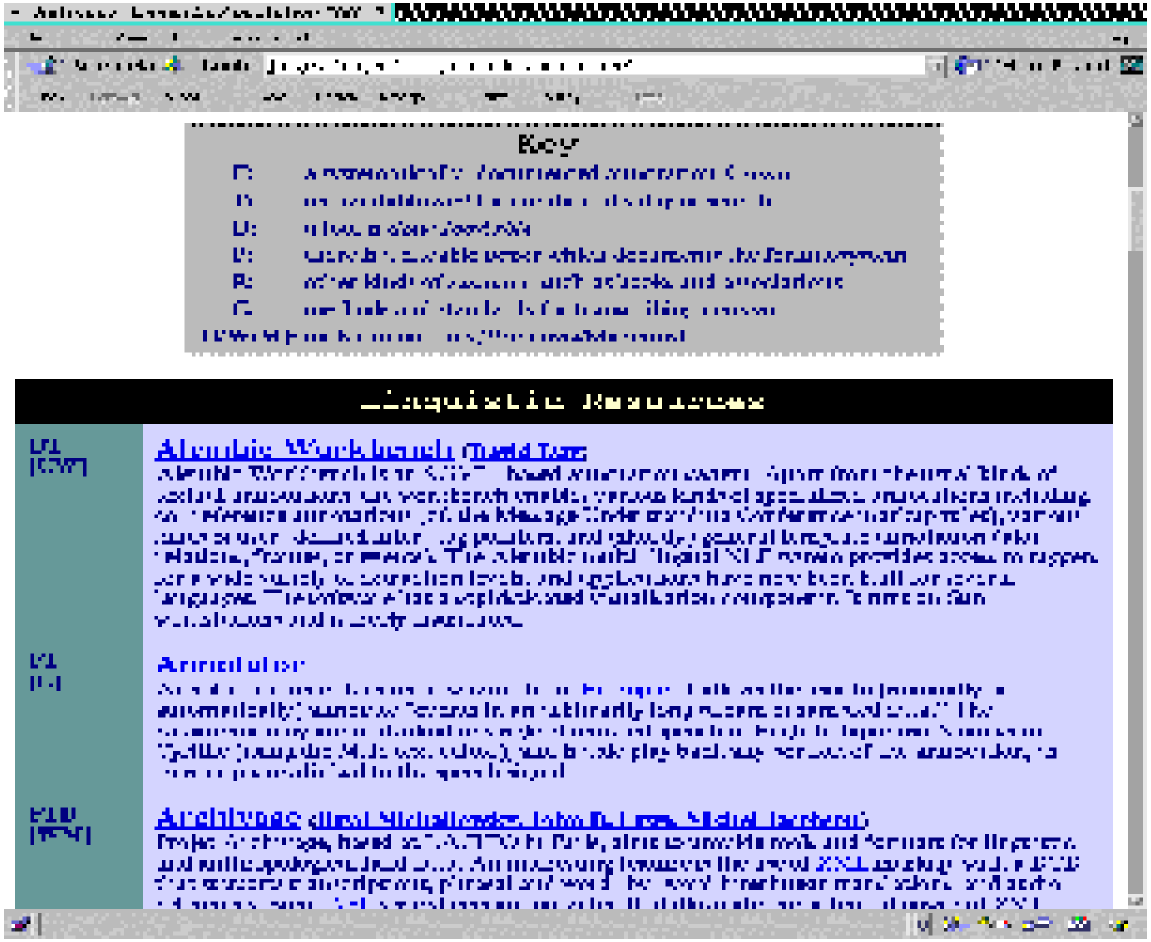
Corpus Cookbook

The goal of the Linguistic Data Consortium's "Corpus Cookbook" is to share information on the methodology the LDC has developed to produce corpora as well as provide links to on-line linguistic resources. The LDC will provide information detailing aspects of speech collection, available software and hardware tools, and best practices information for those involved with, or interested in, developing speech corpora.

- Convention guidelines for the transcription of broadcast and telephone audio data.
- Linguistic Annotation Over 30 different tools and formats for creating and managing linguistic annotations, primarily speech
- LDC Home

Tools

Linguistic Annotation Links:
Tools and formats for creating and managing linguistic annotations
www.ldc.upenn.edu/annotation



Key

[illegible]

Linguistic Resources

Alembic Workbench [illegible]

[illegible]

Annotator

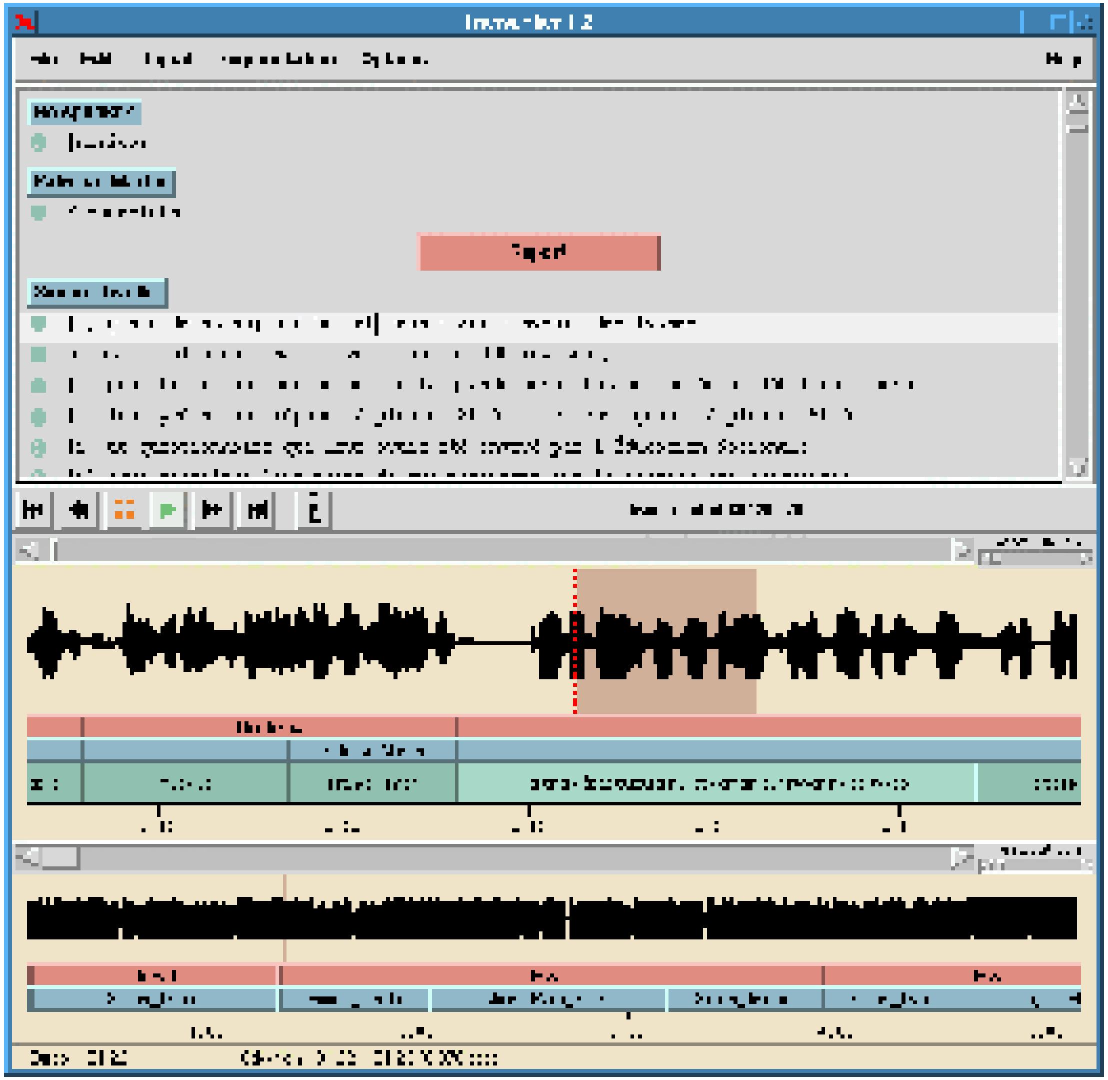
[illegible]

[illegible]

Tools

- Speech Analyzer - tool for performing various analyses of waveform and spectrograms (www.sil.org)
- TalkBank - NSF funded project at CMU and LDC to develop basic tools for annotation of data in the social sciences and humanities - 17 communities involved
- Transcriber - tool for transcribing and segmenting audio in variety of formats -- free from LDC and friends

Tools



www.ldc.upenn.edu/mirror/Transcriber

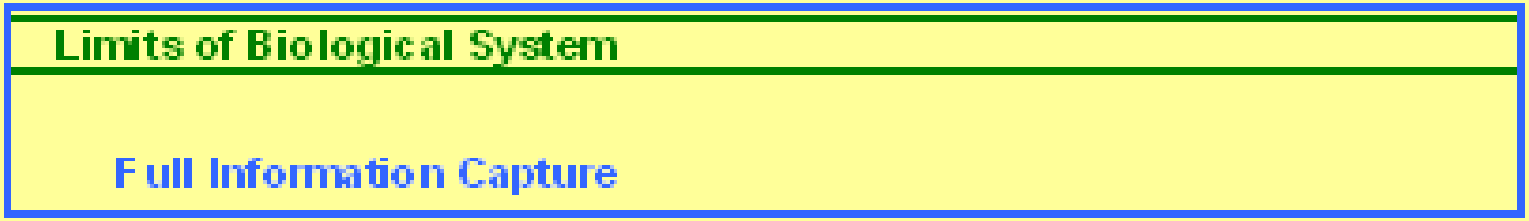
Conclusion

- Shared Data
  - supports individual research
  - supports collaborative research
  - allows comparison of competing or evolving models
  - encourages replication of results
- Data and Tools are available
- LDC wants to stimulate discussion of corpus use & creation in sociolinguistics.
  - distribute existing corpora
  - consider the community's needs as we develop corpora
  - include sociolinguists in discussions of standards & tools
  - especially, develop LDC on-line to be useful to the community

Motivation

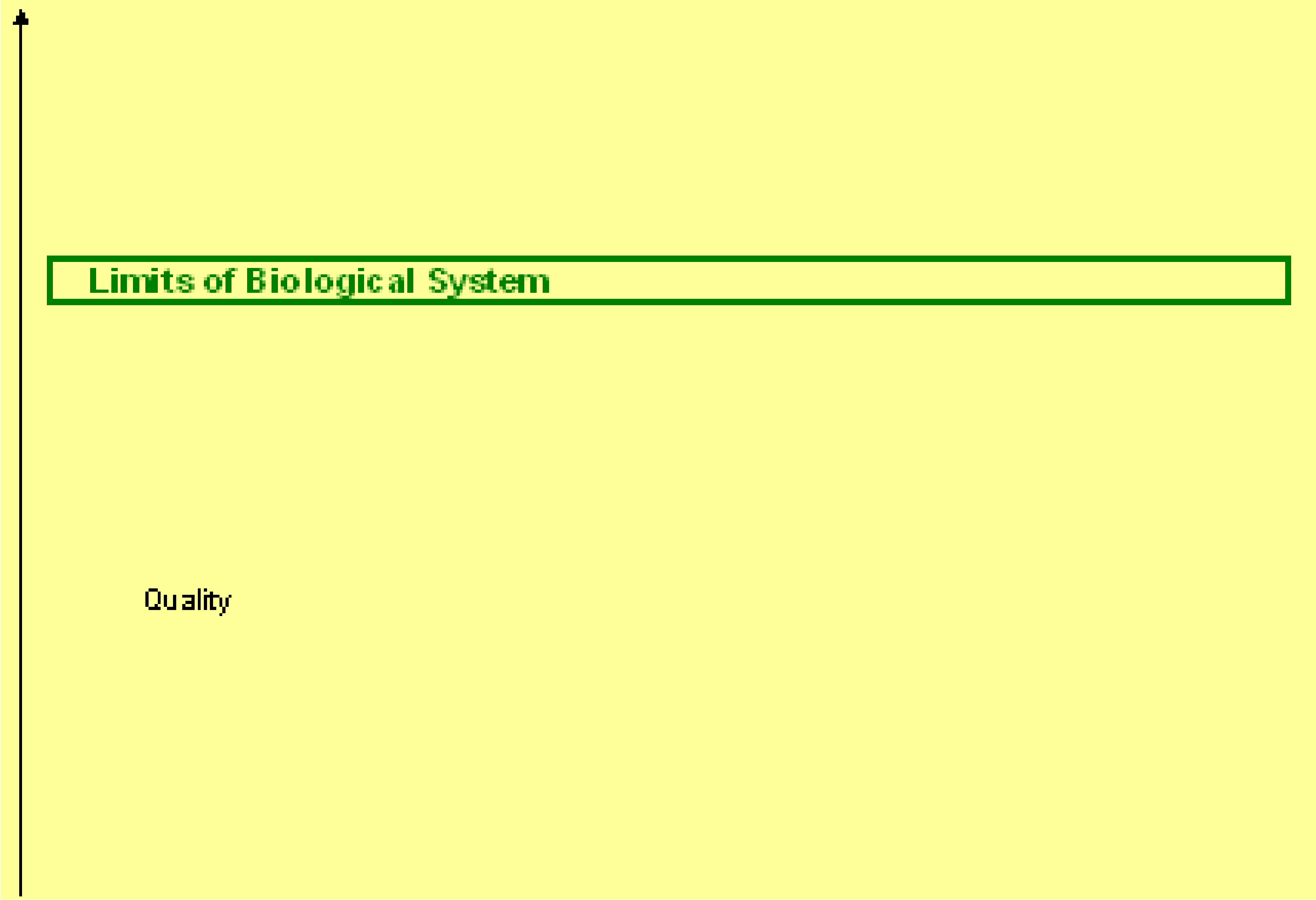
- Sociolinguistics is necessarily corpus based
  - but digital access and manipulation of corpora required specialized equipment
  - hardware needs
    - signal processing hardware
    - adequate storage (150MBs per hour of single channel 22K, 16bit audio)
    - processors to handle multi-megabyte audio files
    - current systems have all of this (40 hours on 6GB disk)
  - Software needs
    - network operating systems to handle large data
    - display and editing of sampled waveform data
    - transcription with time-alignment to audio
    - acoustic analysis
  - Data Needs
    - there's never enough

Collection Model


Limits of Biological System

Full Information Capture

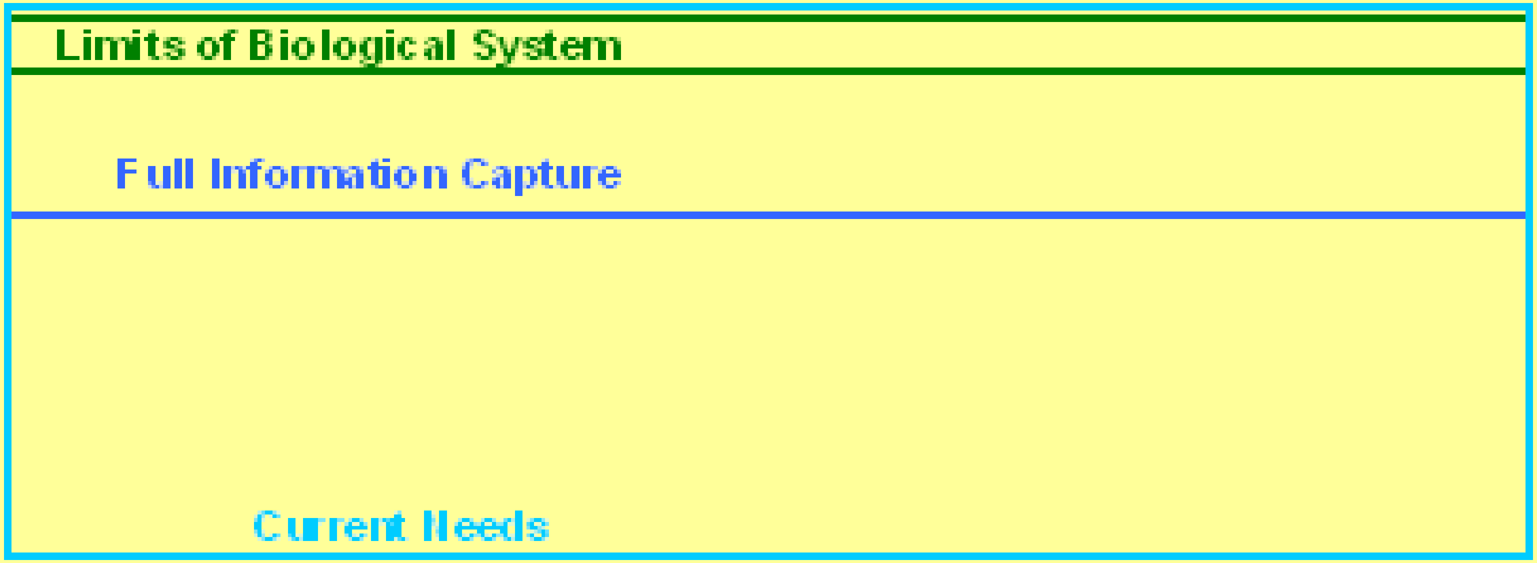
Collection Model



Limits of Biological System

Quality

Collection Model



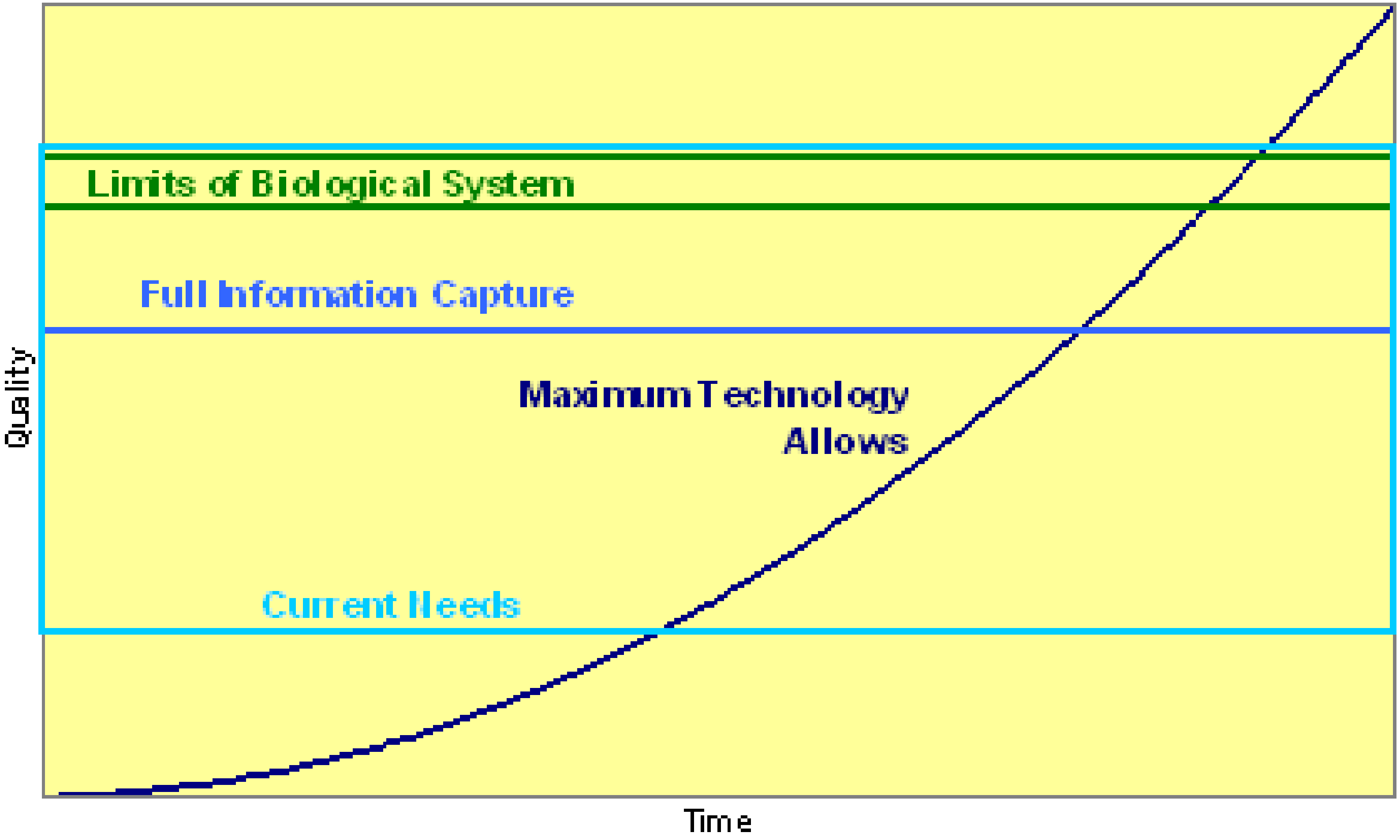
Limits of Biological System

Full Information Capture

Current Needs

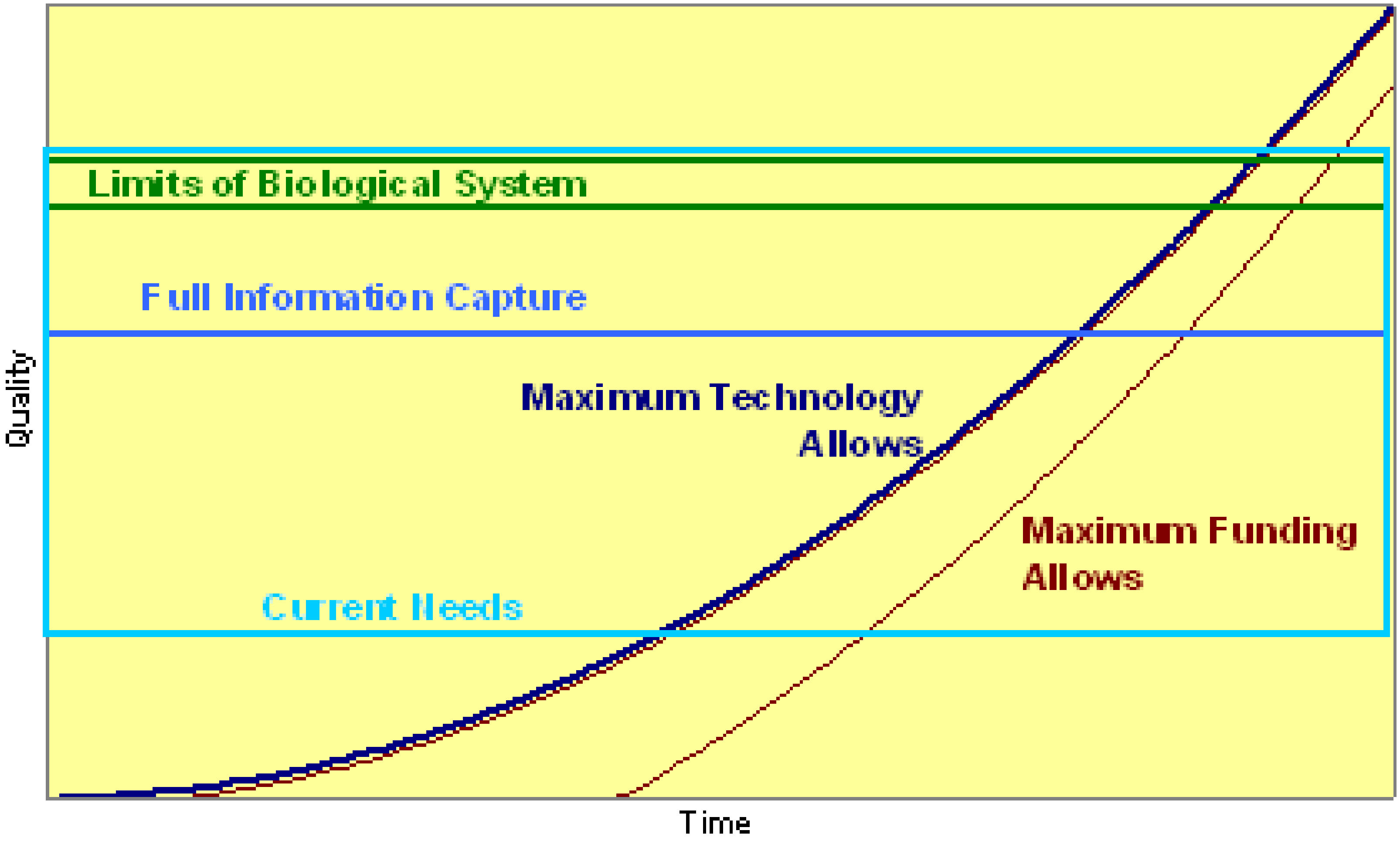
Collection Model

Options for Setting Quality



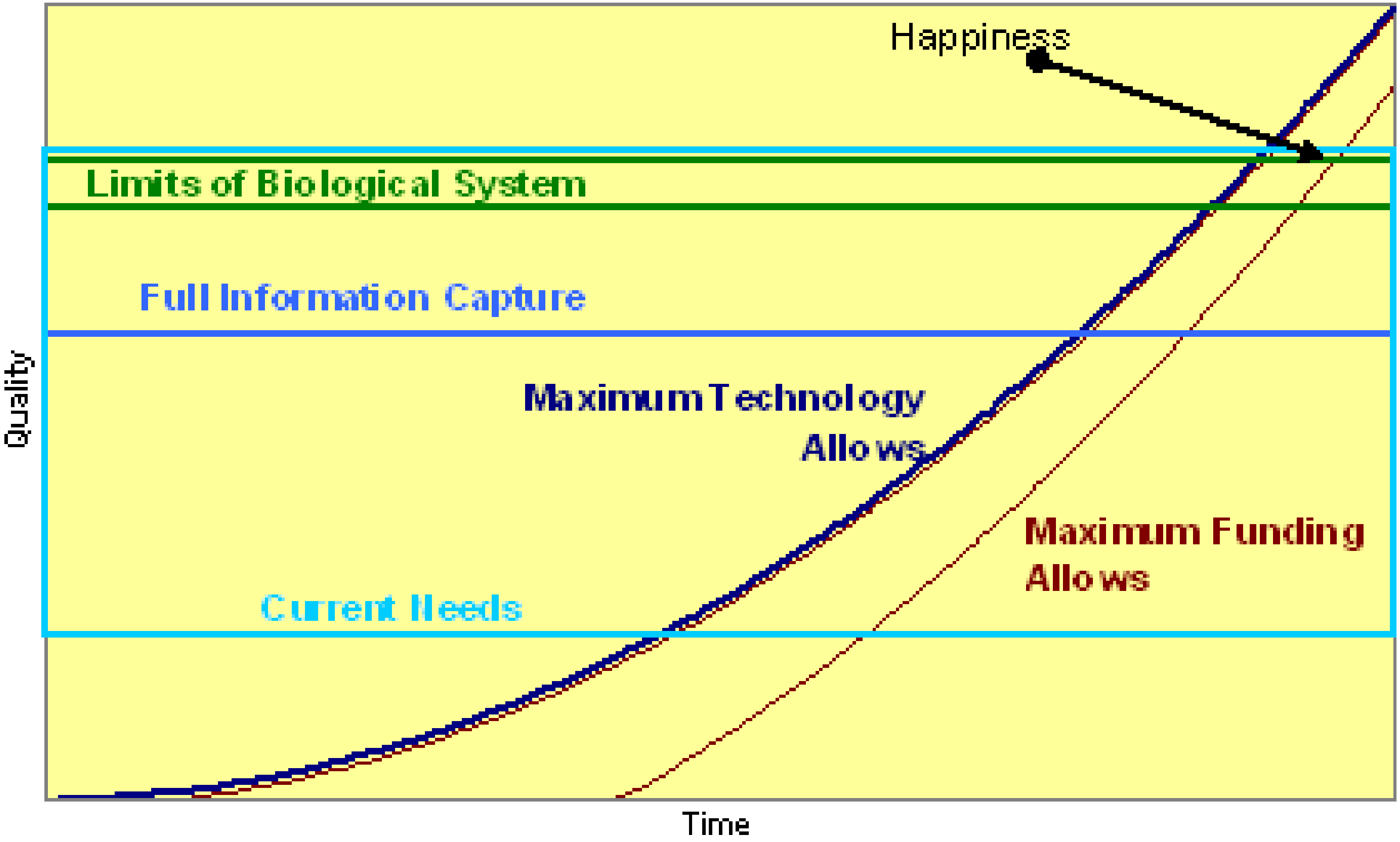
Collection Model

Options for Setting Quality



Collection Model

Options for Setting Quality


Happiness
Limits of Biological System
Full Information Capture
Maximum Technology Allows
Maximum Funding Allows
Current Needs
Quality
Time