Sharing, Structuring and Processing Data: Part 1: Advantages and Challenges

Christopher Cieri
University of Pennsylvania, Linguistic Data Consortium
ccieri AT ldc.upenn.edu

This work was supported in part by NSF Grant BCS #1144480 with supplemental funding from LDC and continues from the resulting LSA 2012 workshop. Thanks are due to all workshop participants as well as Christine Massey, Laurel MacKenzie, Brittany McLaughlin and Marian Reed for their unflagging assistance developing and organizing the LSA workshop.

Linguistic Data Consortium

The Problem

- Data is critically important in the quantitative analysis of linguistic variation
- However, data methods, especially sharing, are inadequate to need and lag behind other language related fields where
  - sharing is the default
  - studies based on data not publicly available are criticized or ignored
  - entire multi-year, multi-site programs rely on common data
- Zinsmeister & Breckle 2013:
  - “*The transfer of information structure between two verb-second languages and the filling of the Vorfeld is contrastively investigated by Bohnacker and Rosen (2008). However, their analysed data is not published as a reusable annotated corpus.*”
- Habash et al 2013:
  - “*Al-Sabbagh and Girju (2012) describe a supervised tagger for Egyptian Arabic social networking corpora […] They report 94.5% F-measure on tokenization and 87.6% on POS tagging. […] We do not compare to them since their data sets are not public.*”
- Przybocki 2007:
  - “*NIST has coordinated annual evaluations of text-independent speaker recognition from 1996 to 2006. This paper discusses the last three of these, which utilized conversational speech data from the Mixer Corpora recently collected by the Linguistic Data Consortium*”

Why is Shared Data not the Default?

- Early sociolinguistic works defined a research program based on a
  - new domain: speech community
  - new data type: sociolinguistic interview

Need for new data collection extreme; utility of data exchange marginal

- Sharing difficult
  - copying audio tapes
  - suffering quality degradation with each copy
  - grappling with a multitude of tape formats.
- Lack of tools for indexing audio even if speech were transcribed made analysis difficult.
- In the field, the researcher could
  - interact directly with informants
  - adapt elicitation practice as needed
  - identify linguistic variables and the factors that may co-vary with them.

- Notwithstanding benefits of original field data collection, as in all sciences, there are equally valid needs to
  - build upon prior work
  - compare individual studies
  - track phenomena through different communities, communicative situations
  - hypothesize and evaluate hypotheses about general processes and
  - analyze more data than any single fieldworker can accumulate.
- One may exploit published accounts but is then limited to data, conclusions reported in comparable form.
- Need for replications of prior work, re-analyses of existing data becomes inevitable as field matures, number of new concepts and analytical tools grows.
- Impediments to effective sociolinguistic data sharing: not willingness but impartial/new technical support, methodology.

- Today the potential for shared data is much greater because:
  - We identify and compare different groupings of speaker.
  - We recognize that other communicative situations are interesting.
  - ~50 years of sociolinguistics = a lot of field data
  - Even more data available from other sources (HLT research).
  - Data is digital, sharing is easy, common audio formats are ~universal, copying is lossless.
  - Tools exist to support transcription and find audio based on transcript.
  - Forced alignment technologies provide even finer alignment at the word and phone level.
- And we have the following addition motivations
  - Funding agencies increasingly demand plans for sharing data long-term
  - US OSTP directed agencies to make data, publications freely available

Possible Futures

- Forced to share data, we do:
- data sets scattered
- transcripts partial or absent
- coding<->source links ambiguous
- coding practice
  - differs by site
  - acquired through apprenticeship
- essential terms assumed same
- only required data shared

- only data, publications shared

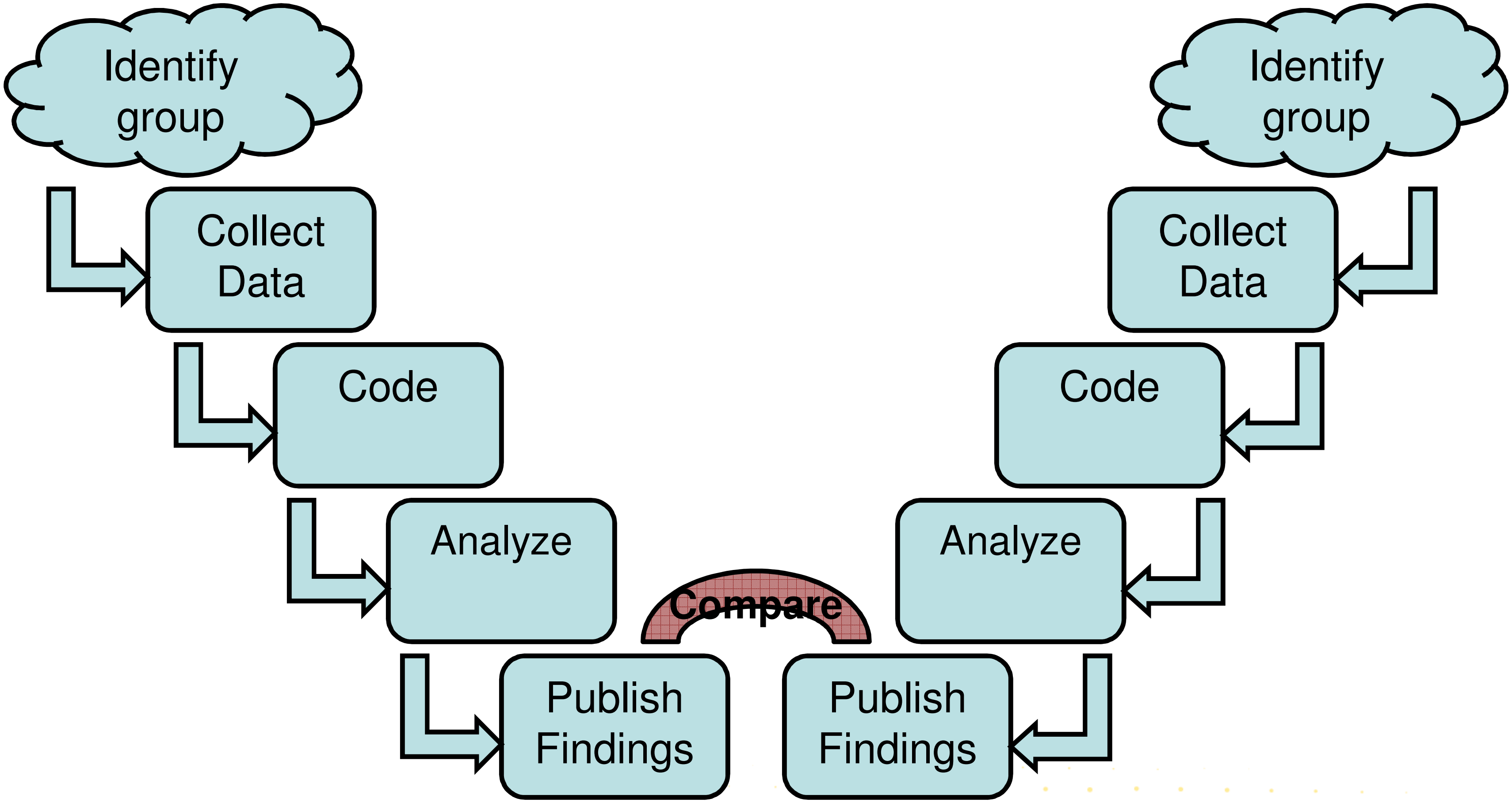
- current domains dominate

Possible Futures

- Forced to share data, we do:
- data sets scattered
- transcripts partial or absent
- coding<->source links ambiguous
- coding practice
  - differs by site
  - acquired through apprenticeship
- essential terms assumed same
- only required data shared
- only data, publications shared
- current domains dominate

- We seize the opportunity:
- data sets collected, indexed
- transcripts complete
- coding<->source links exact
- coding practice
  - unified, where possible
  - formally defined
- essential terms defined
- all data shared
  - source, transcription
- all resources shared
  - specifications, coding, analytic procedures (Praat/R scripts), tools
- new domains studied in shared data

Comparison via Papers



Case Study: t/d deletion

- loss of coronal stops in word final consonant clusters
- one of earliest, most frequently studied of sociolinguistic variables, "*a showcase for variationist sociolinguists*" (Patrick 1992)
- figures into many issues: development of variable rules, positivism vs. empiricism, constraint ordering, age grading, functionalism, lexical phonology, exponential hypothesis, language transfer, dialectology
- Incidence ranges about as much as possible (3-97%)
- "*The inherent difficulty in sociolinguistic analysis, the question of what to count, is highlighted by the wide range of difference in coding practices by researchers.*" (Santa Ana 1991)
  - nature of variable: continuous/discrete/categorical, specific/general reduction
  - linguistic and non-linguistic factors considered
  - factor values
  - (features of) tokens considered or excluded

Comparing via Papers: Coding Differences

<table>
<tr><td>C</td><td>C</td><td rowspan="3">~V</td></tr>
<tr><td>L</td><td rowspan="3">~C</td></tr>
<tr><td>G</td></tr>
<tr><td>V</td><td>V</td></tr>
</table>

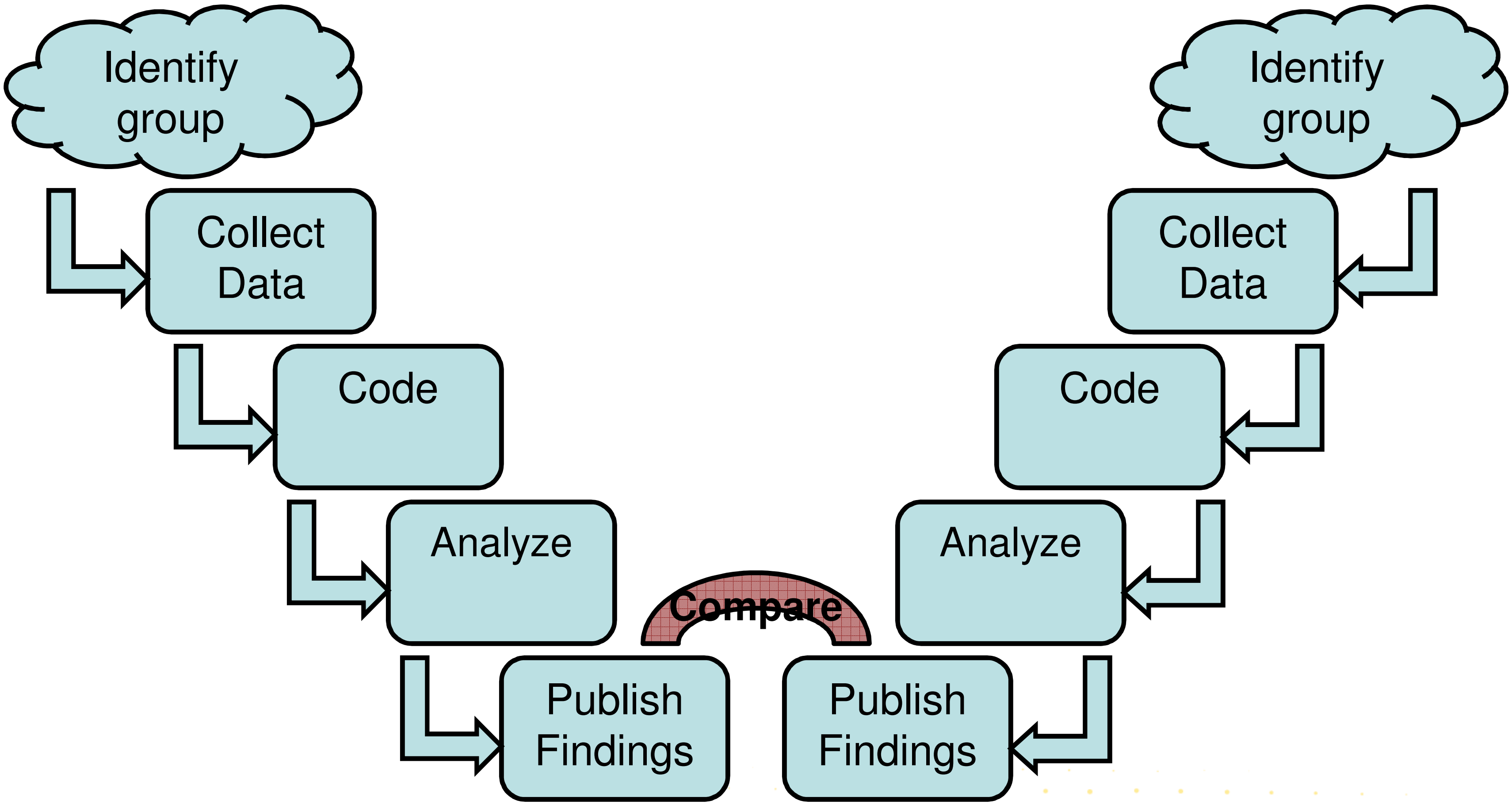
- Following environment affects t/d deletion: generally C>L>G>V
- Difference between consonants, liquids and glides sometimes not significant
- Studies variably coded following segment as vowel (V) versus non-vowel (~V), consonant (C) versus non-consonant (~C)
- When reported, pause sometimes disfavors (V), sometimes favors (C)
- Issues:
  - When differences create non-/partial overlaps, comparison impossible or dubious.
  - How long does a pause have to be to be a pause?

Comparing via Papers: More Coding Differences

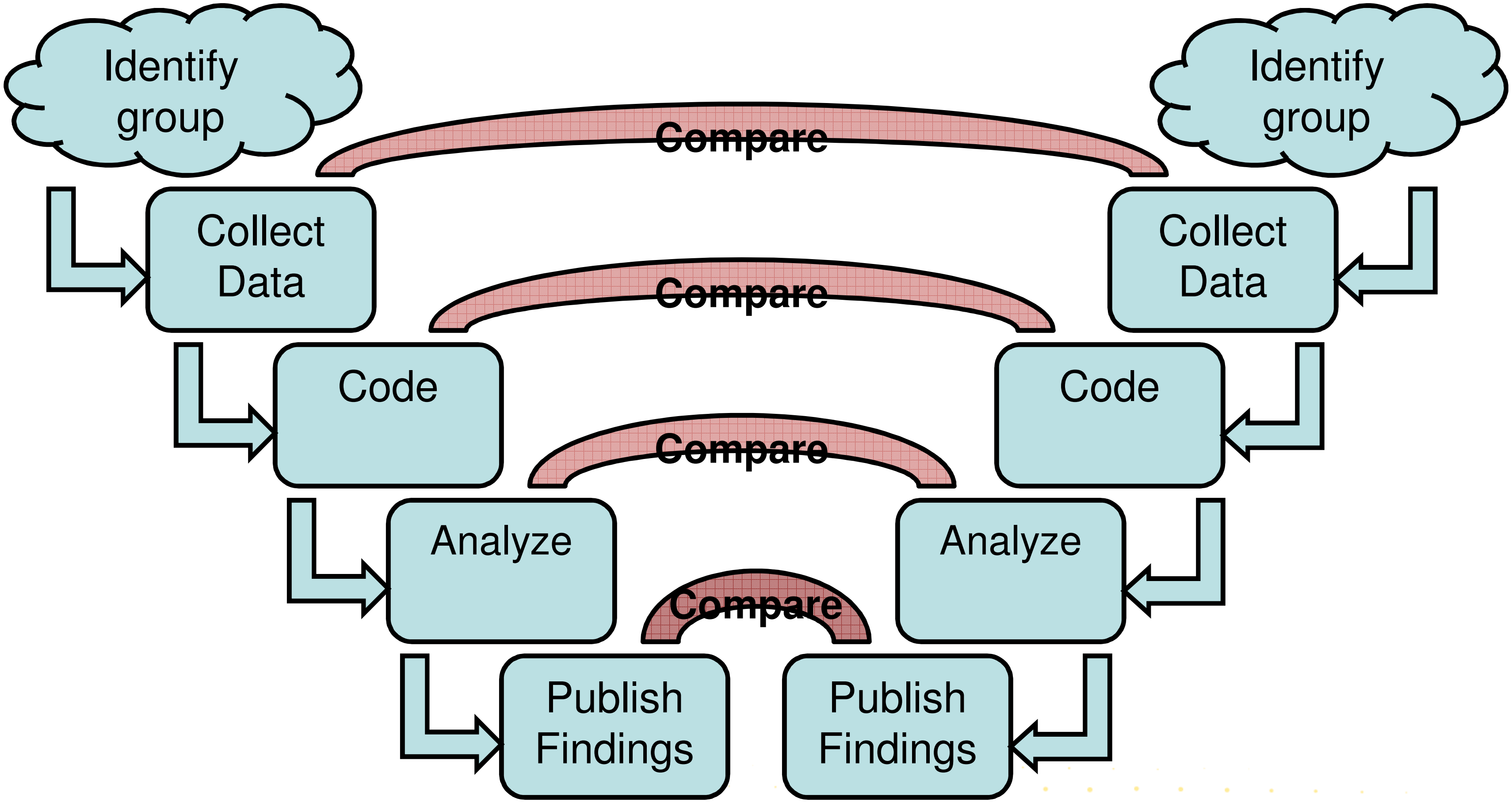
pb td kg ʔ	■	stop	stop	stop	stop	■	stop	obstruent
fv θð h		fricative	fricative	fricative	fricative		spirant	
sz ʃʒ	sibilant	sibilant	sibilant	sibilant	sibilant	sibilant		
m(n)ŋ	nasal	nasal	nasal	nasal	nasal	nasal	sonorant	sonorant
l	■	l	l	l	l	l		

- Preceding environment (manner of articulation) affects –t/d deletion
- Santa Ana (1991:51) reviews 7 studies: coding, order & significance of effect differ. Figure 2 schematizes
  - Column 1: lists a subset of English consonants (my guess as to what the categories mean)
  - Columns 2-9: show different treatments, outcomes
  - Black cells: study did not report on that preceding environment.
  - Gray cells: differences were not reported as statistically significant
  - Missing cells, mismatches inhibit comparison.
- Did 4 of these 7 studies really need to differ on this dimension?
  - What would happen if the journal editors rejected papers with unjustified differences?

Comparison via Papers



Other Comparisons



Using Shared Data: Domains

- Speech Community original, commonest sociolinguistic domain, but Handbook of LVC (now 10 yrs. old!) reviews work in social network (Milroy), community of practice (Meyerhoff), family (Hazen)
- Is it OK to have different groupings?
  - LCRL 1968 deals with the NYC speech community and the youth groups: Aces, Cobras, Jets and T-birds
- Can we compare speech communities with other groups? We do.
  - Coding speakers for age, sex, SEC, education and then correlating with linguistic variables is comparing a speech community with something else (a sub-sample)
  - "*Another aspect of the regularity of t,d simplification is the fact that the basic relationships are repeated in group after group across neighborhoods and age levels.*" (LCRL 1968)
  - "*Data recorded in the Belfast study were examined to compare the language patterns of 46 speakers from three low status urban working-class communities – Ballymacarrett, Hammer, Clonard. Eight phonological variables, all of which were clearly indexical of the Belfast urban speech community, were analyzed in relation to the network structure of individual speakers.*" (Milroy 2002)

Using Shared Data: Elicitation Methods

- Sociolinguistic interview original, commonest elicitation technique
- But
    - Labov (1966) reports on rapid and anonymous interviews
    - Bell (1977, 1982, 1984, 1988, 1991) reports on on broadcast news
    - Fox (2001), Yaeger-Dror (et. al. 2011), Jurafsky (et. al. 2000), Mackenzie (2012) studied t/d deletion, prosody and syllable final s lenition using conversational telephone speech
    - Strassel & Cieri (2002) showed t/d deletion in conversational telephone speech similar to sociolinguistic interviews

Case Studies: Advantages

- Hedberg, et. al. 2010
  - TOBI annotated and categorized discourse function of 200 wh-questions in CallHome (Canavan et. al. 1997) and Fisher (Cieri et al. 2004)
  - confirm predominance of falling intonation in English wh-questions
  - suggests that wh-questions ending in a high boundary tone "query information that does not bear directly on the current topic" without actually changing the topic; that is they pose side-issues.
  - Fisher alone includes 11,699 ~10 minute transcribed conversations, required many person years of effort at a cost of >$1.2M
- Fox (2000)
  - used CALLHOME Spanish Transcripts and Lexicon to index CALLHOME Spanish Speech corpus for all instances of syllable final (s)
  - coded >24,000 tokens for lenition
  - coded or calculated confidence level, syllabification and following pause duration
  - inherited information about token, context, speaker dialect, sex, age
  - >20% of tokens annotated multiple times (intra-coder=88%, inter-coder=76%)
  - published all annotations (Fox 2001)

Case Studies: Advantages

- Yaeger-Dror, et. al., (2011)
  - "*The LDC sound archives provide a plethora of corpora for comparative analysis of speakers from different regions and different cultures. … Such a study would not have been feasible at all before the recent advances in technology which have made it possible to store large corpora and carry out acoustic and statistical analysis of such large corpora. Only these advances have made it possible to supersede the analysis made in the 1980s based on smaller corpora …*" (p. 161)
  - analyzed variation in pitch prominence of negatives as function of demographics, situations
  - upon 9 publically available corpora, >175 hours of speech (134 transcribed), 3 languages, 2 situations
  - exploit relatively unmonitored conversation among intimates and specific interaction types, disagreements, difficult to elicit otherwise
  - corpora large, high quality, some fully transcribed to permit reliable sampling
  - beyond analysis, published findings, additional transcripts now freely available.

Case Studies: Challenges

- Language Corpus: collection of recorded observations of linguistic behavior selected and annotated for a specific purpose
- Optimization toward specific purpose maximizes utility for intended users, creates challenges for others. However corpora can be re-used as above and some are created for purposes very close to our own hearts
  - SLAAP: http://ncslaap.lib.ncsu.edu/
  - OSCAR: http://oscaar.ling.northwestern.edu/
  - Buckeye Corpus: http://vic.psy.ohio-state.edu/
  - Digital Archive of Southern Speech: http://catalog.ldc.upenn.edu/LDC2012S03
  - Malto Speech and Transcripts: http://catalog.ldc.upenn.edu/LDC2012S04
  - Nationwide Speech Project: http://catalog.ldc.upenn.edu/LDC2007S15
  - SLX Corpus: http://catalog.ldc.upenn.edu/LDC2003T15
  - MALACH Interviews, Transcripts: http://catalog.ldc.upenn.edu/LDC2012S05

Case Studies: Challenges

- Fox's (2000) work with CALLHOME Spanish corpora required adaptation. Changes to the transcripts and the limitations of available forced-alignment lead to loss of 3000-4000 (s) tokens, she estimates.
- Yaeger-Dror, et al, 2011: precautions for working with published data
  - CALLFRIEND corpora support language identification
    - short phone calls, male & female, multiple ages, definitive language labels
    - included additional speech, subject metadata (for some future beneficiary)
    - Not completely transcribed
  - Broadcast News support ASR
    - audio, characteristic channels, time aligned transcription, little speaker metadata
    - no speaker age, expensive to provide given many unknown speakers
    - dialect divisions inappropriate for new re-use, had to be recoded
    - removed calls with evident power asymmetries
    - noted limited social class variation; most college educated, computer literate
    - Japanese corpus (not LDC) contained interaction among demographic features; males from one region and females another

Linguistic Data Consortium

Creating Sharable Data within Anyone's Means

- **<u>Media</u>** format incompatibility issues largely a thing of the past.
- Acquire and save metadata that would be impossible to recover
- Transcribe Audio (Elan, Xtrans. Transcriber, **not** Word)
- Use transcripts to index audio
- Know current best practice, think twice before diverging
- Specify coding practice so that unaffiliated researchers can replicate
- During coding save pointers to original data
- Code with fine-grain, collapse later
- Maintain audit trail: number & save version before major changes
- Malcah's question: "*What do we not yet code for appropriately?*"

Motivation

- Stoker ’97 provides early justification for transcription in a related field:

Motivation

- Stoker '97 provides early justification for transcription in a related field:

He accordingly set the phonograph at a slow pace, and I began to typewrite from the beginning of the seventeenth cylinder.

…

He thinks that in the meantime I should see Renfield, as hitherto he has been a sort of index to the coming and going of the Count. I hardly see this yet, but when I get at the dates I suppose I shall. What a good thing that Mrs. Harker put my cylinders into type! We never could have found the dates otherwise.

- Stoker, Bram (1897) Dracula

Motivation

Linguistic Data Consortium

Motivation

- efficient coding allows researcher to
  - consider greater percentage of tokens per variable
    - Strassel & Cieri 2001: 28,311 tokens annotated for -t/d deletion
  - investigate more variables
  - minimize misses
  - improve accuracy and balance
  - improve consistency
- retains
  - accurate time, sequence information
  - mapping among sound, transcript, tokens, coding
- encourages re-use of data
  - each additional pass requires less effort than original
  - re-use & reanalysis profits from previous preparation
- with emerging tools, effort required for first pass also decreasing
  - "*It saves money because it saves money.*"