Available Resources and type of annotation

LDC corpora

LDC96S36	Boston University Radio Speech Corpus: ➢ orthographic transcription, phonetic alignments (TIMIT phonetic labeling system; Arpabet), part-of-speech tags and prosodic markers
LDC2002S28	Emotional Prosody Speech and Transcripts: ➢ emotionally transcribed
LDC2003S06	Santa Barbara Corpus of Spoken American English Part-II: ➢ some limited prosodic annotation, lots of data on conversation type, speaker background, etc.

In addition we have tons of LDC corpora that are transcribed + speech or only speech but no additional annotation.

Other corpora

- Part of Switchboard (TIMIT phonetic labeling system; Arpabet)
  - ~9k sentences (60k words)
- Similar Chinese corpus (but rather phonemic transcription)
- TIMIT
  - 6300 sentences from 8 English dialects
  - phonetically transcribed (TIMIT phonetic labeling system; Arpabet)
- VerbMobil I and II (German, English, Japanese)
  - We don't have the sound files, but a lot annotated files
  - Superimposed Speech - SUP
  - Phonetic Segmentation PhonDat - PHO
  - Phonetic Segmentation (SAM-PA phonetic system) - SAP
  - Automatic Segmentation (SAM-PA phonetic system) - MAU
  - Word Segmentation - WOR
  - Dialogact Segmentation - DAS
  - Prosodic Segmentation - PRB
  - Symbolic prosodic Segmentation - PRS
  - Signal-based Prosodic accents labeling - LBP
  - Signal-based Prosodic boundaries labeling - LBG
  - Syntactic-prosodic labeling - PRO
  - Syntactic trees - SYN,FUN,LEX
  - Parts of Speech
  - Phonetic Segmentation
  - Segmentation in turns/sentences/chunks/etc
  - SmartKom Transliteration, Gesture Labeling, User State Labeling holistic, User State Labeling by mimic expression, User State Labeling Occlusions, Meta Linguistic Features
  - Translation - TLN

- The London-Lund Corpus of Spoken English (part of the ICAME)
  - consists of 100 texts which are ToBI-labeled. ~ 500,000 words.
  - Phonetically annotated:

	BOOK	SLIP	TAPE	
GRAPHICS	full-time	full-time	045	Hyphen[17]
	that's	that's	096	Apostrophe
	à	à	097 052	eg *à la maison*
	ä	ä	097 053	eg *Händel*
	é	é	101 049	eg *donné*
	è	è	101 050	eg *très*
	ê	ê	101 051	eg *être*
	ë	ë	101 052	eg *Citroën*
	ö	ö	111 049	eg *Höllgarten*
	ü	ü	117 049	eg *Dürer*
	ç	ç	099 053	eg *français*
PHONETICS[18]	[ə]	[ə]	064	
	[tʃ]	[tʃ]	116 115 104	
	[dʒ]	[dʒ]	100 122 104	
	[θ]	[θ]	116 104	
	[ð]	[ð]	100 104	
	[ʃ]	[ʃ]	115 104	
	[ʒ]	[ʒ]	122 104	
	[ŋ]	[ŋ]	110 103	
	[ɑ]	[ɻ], [ɹ]	126	Voiceless clicks
	[œ]	[oe]	111 101	eg in Fr. *peur*
	[ʔ]	[ʔ]	063	Glottal stop
	[ã]	[ã]	097 049	Nasal

- 9 urban dialects from Britain are collected in the IViE-corpus
  - Contains about 36 hours of recordings.
  - Very small subset is annotated for prosody and prominence.
  - F0-tier
- RNC – Corpus of German radio news
  - ~ 160 news stories
  - orthographically transliterated, words tier, morphosyntactically annotated
  - automatically word aligned
  - manually prosodically labeled, full ToBI labelling
  - phone transcription (also available as syllable-based transcription):
    - 0.570000 122 <P>
    - 0.700000 122 d
    - 0.840000 122 e:
    - 0.870000 122 R

Example annotations

Boston News Corpus

```
H#       0       4
>endsil
DH       4       5
IH+1     9      10
S       19       9
>This
HH      28       5
AA+1    33       9
L       42      12
AX      54       4
DCL     58       3
D       61       1
EY      62      16
>holiday
S       78      11
IY+1    89      14
Z      103       7
EN     110      20
…
```

XWAVES/PRAAT readable:

```
signal st43/f3ast43p1
type 1
color 76
font -*-times-medium-r-*-*-17-*-*-*-*-*-*-*
separator ;
nfields 1
#

    0.035000   76 H#
    0.085000   76 DH
    0.185000   76 IH+1
    0.275000   76 S
    0.325000   76 HH
    0.415000   76 AA+1
    0.535000   76 L
    0.575000   76 AX
    0.605000   76 DCL
    0.615000   76 D
    0.775000   76 EY
    0.885000   76 S
…
```

TIMIT

```
Word label (.wrd):
        7470 11362 she
        11362 16000 had
        15420 17503 your
        17503 23360 dark
        23360 28360 suit
        28360 30960 in
        30960 36971 greasy
```

```
Phonetic label (.phn):
(Note: beginning and ending silence regions are marked with h#)
        0 7470 h#
        7470 9840 sh
        9840 11362 iy
        11362 12908 hv
        12908 14760 ae
        14760 15420 dcl
        15420 16000 jh
        16000 17503 axr
```

Switchboard