Files set up by Kevin Ryan based on raw data from the Buckeye 2.0 corpus:

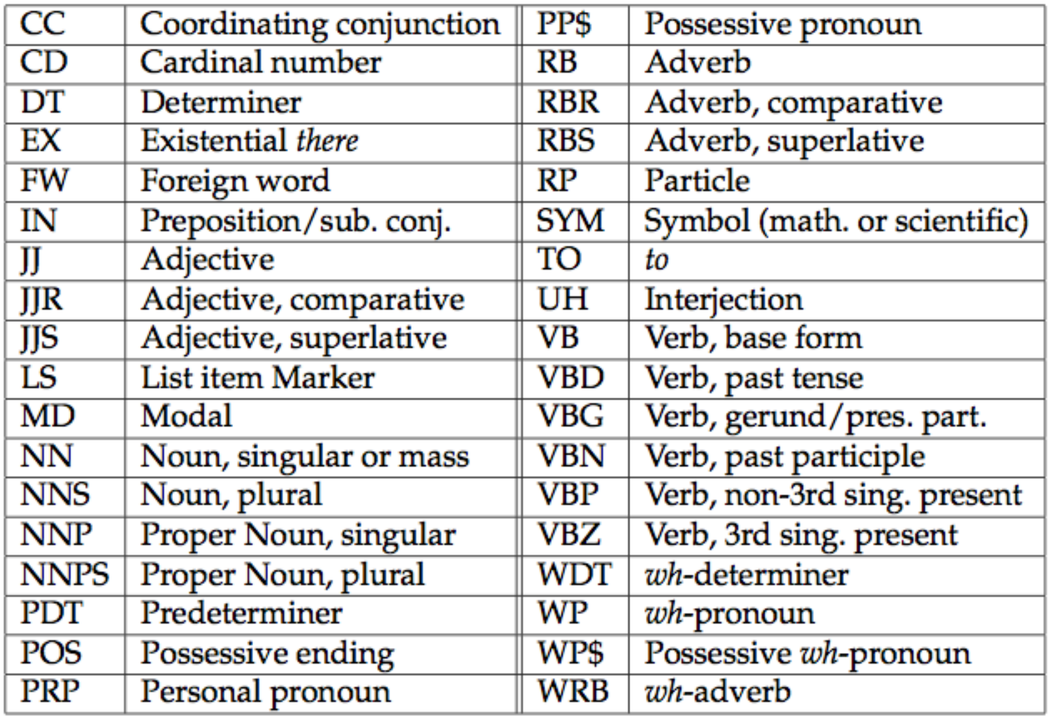
> Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

The corpus includes 40 speakers, including 20 males and 20 females. Syllable break locations follow CELEX (or estimates based on how clusters are usually divided in CELEX). Stress patterns follow CMU Pronouncing Dictionary (which includes more inflected forms than CELEX). Frequency data come from a TV transcript corpus (chosen to be representative of spoken American English), as described at en.wiktionary.org/wiki/Wiktionary:Frequency_lists. Frequencies are given for the 41,284 words in that corpus with counts of greater than 5. If the Buckeye word (as orthography) doesn't occur in that list, its frequency is given as "NA", suggesting an uncommon word or nonstandard spelling. Load the files in R as `x=read.table("segment_durations.txt", header=T, sep="\t", quote="", comment.char="")` to handle the special characters properly.

	DISC/CELEX	Buckeye	Notes	
Vowels	i	iy(n)	"beat"	$V = '[iuIUQ\{2645E13VPHF]'
	u	uw(n)	"boot"	
	I	ih(n)	"bit"	
	U	uh(n)	"book"	
	Q	aa(n)/ao(n)	"bought"	
	{	ae(n)	"bat"	
	2	ay(n)	"bite"	
	6	aw(n)	"bout"	
	4	oy(n)	"boy"	
	5	ow(n)	"bow"	
	E	eh(n)	"bet"	
	1	ey(n)	"bait"	
	3	er(n)	"bird"	
	V	ah(n)	"abbot" (schwa)	
Syllabic Cs	P	el	"bottle"	
	H	en	"button"	
	F	em	"bottom"	
Consonants	m	m		$C = '[mnNltdJ_TDSZszkgpbfvwhjrRY]'
	n	n(x)		
	N	(e)ng		
	l	l		
	t	t		
	d	d	(tap is often transcribed this way here)	
	J	ch		
	_	jh		
	T	th		
	D	dh		
	S	sh		
	Z	zh		
	s	s		
	z	z		
	k	k		
	g	g		
	p	p		
	b	b		
	f	f		
	v	v		
	w	w		
	h	h(h)	(usually "hh" in Buckeye)	
	j	y		
	r	r		
	R	dx	(not reliably transcribed)	
	Y	tq/x	glottal stop	

segment_durations.txt
(840,358 phones)

0. **Segment** unigraph DISC/CELEX transcription
1. **Duration** rounded to nearest millisecond
2. **Speaker** 1-40 consecutive
3. **File** section identifier for file containing token
4. **Word** orthography of word
5. **POS** Buckeye part of speech (see table below)
6. **Ideal** like an underlying, phonemic, or careful form of the word
7. **Surface** the (often reduced) surface transcription of the word
8. **Context** immediate segmental context (including #) with hyphen for gap
9. **Prepausal** segment immediately followed by pause (1 or 0)
10. **SylN** location of this syllable in surface word
11. **OfN** total # of surface syllables (for "ideal" syllable count see words file)
12. **SylContext** whole surface syllable with hyphen for gap (syllabification based on CELEX)
13. **SubSyl** onset, nucleus, or coda
14. **StressGuess** primary, secondary, unstressed, or unknown, based on CMU Dict
Watch out for clitics, which are often given as "primary"
(can use POS or frequency to weed these out)
"Guess" because it's based on a dictionary, not the actual token
15. **TVFreq** "NA" (<6) or a # greater than 5 (see intro). Should be logged.
Approx. <10,000 is useful for getting content words (or use POS)
16. **TimeInto** location in recording (useful for checking or acclimation correction)
17. **IsYoung** 1 if "young" (<=30), else 0 (>=40)
18. **IsMale** 1 if male (n=20), else 0
19. **IntIsMale** 1 if interviewer is male
20. **DurationNormal** duration correction for overall rate of speaker (speaker's
avg normalized segment length matches the global avg)



CC	Coordinating conjunction	PP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition/sub. conj.	SYM	Symbol (math. or scientific)
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item Marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund/pres. part.
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd sing. present
NNP	Proper Noun, singular	VBZ	Verb, 3rd sing. present
NNPS	Proper Noun, plural	WDT	*wh*-determiner
PDT	Predeterminer	WP	*wh*-pronoun
POS	Possessive ending	WP$	Possessive *wh*-pronoun
PRP	Personal pronoun	WRB	*wh*-adverb

syllable_durations.txt
(361,313 syllables)

0.	**Syllable**	syllable in DISC/CELEX transcription (boundaries follow CELEX)
1.	**Duration**	syllable length in milliseconds
2.	**Speaker**	1-40 consecutive
3.	**File**	section identifier for file containing token
4.	**Word**	orthography of word
5.	**POS**	Buckeye part of speech (see table)
6.	**Ideal**	like an underlying, phonemic, or careful form of word
7.	**Surface**	the (often reduced) surface transcription of the word
8.	**Context**	surface with this syllable hyphened out and dots between syllables
9.	**Prepausal**	syllable precedes pause (1 or 0)
10.	**SylN**	location of this syllable in surface word (see words for "ideal" count)
11.	**OfN**	total # of syllables in surface word
12.	**Onset**	onset (possibly empty)
13.	**Nucleus**	nucleus
14.	**Coda**	coda (possibly empty)
15.	**StressGuess**	primary, secondary, unstressed, or unknown
16.	**TVFreq**	frequency of word
17.	**TimeInto**	location in recording
18.	**IsYoung**	1 or 0
19.	**IsMale**	1 or 0
20.	**IntIsMale**	1 or 0
21.	**DurationNormalized**	syllable ms normalized for speaker's overall rate

word_durations.txt
(284,573 words)

0.	**Word**	orthography of word
1.	**Duration**	length of word in milliseconds
2.	**Speaker**	1-40 consecutive
3.	**File**	section identifier for file containing token
4.	**POS**	Buckeye part of speech (see table)
5.	**Ideal**	like an underlying, phonemic, or careful form of word
6.	**Surface**	the (often reduced) surface transcription of the word
7.	**ParseIdeal**	ideal form syllabified with dots
8.	**ParseSurface**	surface form syllabified with dots
9.	**Prepausal**	word precedes pause (1 or 0)
10.	**SylNIdeal**	total # of syllables in ideal form of word
11.	**SylNSurface**	total # of syllables in surface form of word
12.	**StressGuess**	compiled stress contour from CMU Dict (1 = primary, 2 = secondary, 0 = unstressed, ? = unknown)
13.	**TVFreq**	frequency of word
14.	**TimeInto**	location in recording
15.	**IsYoung**	1 or 0
16.	**IsMale**	1 or 0
17.	**IntIsMale**	1 or 0
18.	**DurationNormalized**	word ms normalized for speaker's overall rate