Automatic Alignment and Analysis of Linguistic Change - Transcription Guidelines*

February 2011

Contents

*This document is an adaptation of the transcription guidelines for *The SLX Corpus of Classic Sociolinguistic Interviews*, Linguistic Data Consortium, September 30, 2003. <http://projects.ldc.upenn.edu/DASL/SLX/docs/transcription.pdf>.

1 Orthography and spelling

1.1 Capitalization

Capitalization in the transcripts is used to aid human comprehension of the text. Transcribers should follow standard written capitalization patterns, and capitalize words at the beginning of a sentence, proper names, and so on.

1.2 Spelling

Transcribers use standard orthography, word segmentation and word spelling, except where explicitly specified otherwise. When in doubt about the spelling of a word or name, please consult a standard reference, like an online or paper dictionary or other reference material.

1.3 Contractions

Annotators should transcribe contractions only when a contraction is actually produced by the speaker. Annotators should take care to transcribe exactly what the speaker says, not what they expect to hear.

If a speaker uses a contraction, the word is transcribed as contracted: *they're, won't, isn't, don't* and so on. If the speaker uses a complete form, the annotator should transcribe what is heard: *they are, is not* and so on.

The table below shows some examples of how to transcribe common contractions. Please note that annotators should use the nonstandard forms *gonna, wanna, gotta, shoulda, woulda, coulda* instead of standard orthography if this is how a speaker pronounces the words in question.

Complete Form	Contracted Form
I have	I've
cannot	can't
will not	won't
you have	you've
could not	couldn't
should have	should've, shoulda
would have	would've, woulda
could have	could've, coulda
it is	it's
Marvin is	Marvin's
Marvin has	Marvin's
going to	gonna
want to	wanna
got to	gotta

Note: Please take care to avoid the common mistakes of transposing possessive *its* for the contraction *it's* (it is); possessive *your* for the contraction *you're* (you are); and *their* (possessive), *they're* (they are) and *there*.

1.4 Numbers

All numerals are written out as complete words. Hyphenation is used for numbers between twenty-one and ninety-nine only.

```
twenty-two
nineteen ninety-five
seven thousand two hundred seventy-five
nineteen oh nine
```

1.5 Hyphenated words and compounds

Annotators should use hyphens in compounds where they are required:

`anti-nuclear protests` (*not* `anti nuclear protests`)

In cases where there is a choice between writing a compound word as one word, a hyphenated word, or as two words with spaces in between, transcribers should opt for one of the latter two versions:

```
house-builder
house builder                (not  housebuilder)
```

1.6 Abbreviations

In general abbreviations should be avoided and words should be transcribed exactly as spoken. The exception is that when abbreviations are used as part of a personal title, they remain as abbreviations, as in standard writing:

```
Mr. Brown
Mrs. Jones
Dr. Spock
```

However, when they are used in any other context, they are written out in full:

```
I went to the junior league game.
I went to the doctor, and all he said was, don't worry, it's natural.
Hey mister, do you know how to get to the stadium?
```

1.7 Acronyms and spoken letters

Acronyms that are normally written as a single word but pronounced as a sequence of individual letters should be written in all caps, with each individual letter surrounded by spaces:

```
I took my G R E's.
I'll stop in to get my U P S packages.
```

Similarly, individual letters that are pronounced as such should be written in caps:

```
I got an A on the test.
How 'bout if his name was spelled M U H R?
```

1.8 Punctuation

Annotators should use standard punctuation for ease of transcription and reading comprehension. Punctuation is written as it normally appears in standard writing, with no additional spaces around the punctuation marks.

Acceptable punctuation is limited to periods, exclamation marks and question marks at the end of a sentence, and commas within a sentence. Exclamation marks are used for especially emphatic speech.

```
And it broke! Like, the bed broke.

Were there any, like, fights between different groups that you can
remember, or?
```

Quotation marks are used to indicate direct speech or thoughts within a narrative and should be used consistently for that purpose:

```
"Oh", I says, "Ain't that something", I says.

And my dad was like -- actually brought up that necklace. He's like,
"don't you have one?" I'm like, "I don't know where it is."

An' the more I thought abo- +about, about it, I thought, "Why not?"
```

2 Disfluent speech

2.1 Introduction

Regions of disfluent speech are particularly difficult to transcribe. Speakers may repeat themselves, utter partial words, restart phrases or sentences, and use numerous hesitation sounds. Annotators should take particular care in sections of disfluent speech to transcribe exactly what is spoken, including all of the partial words, repetitions and filled pauses used by the speaker.

2.2 Filled pauses and hesitation sounds

Filled pauses are non-lexemes (non-words) that speakers employ to indicate hesitation or to maintain control of a conversation while thinking of what to say next. Each language has a limited set of filled pauses that speakers can employ.

The spelling of filled pauses is restricted to these five items:[1]

	Arpabet	**Example**
ah	AA1	`Ah there we go -- water ice and snow cone, that's a good one!`
eh	EH1	`with her girlfriend eh -- well I knew, I knew her girlfriend`
er	ER1	`The kids, er, parents let kids get away with more`
uh	AH1	`'Cause I needed uh, insurance.`
um	AH1 M	`They have a little fl- um, it's like a garage but it's open.`

2.3 Partial words

When a speaker breaks off in the middle of the word, annotators transcribe as much of the word as can be made out. A single dash without preceding space - is used to indicate point at which word was broken off. If transcribers can make a reasonable guess at which word was intended by the speaker, they should include the full form of the word immediately after the truncated form, preceded by a plus sign + (without separating spaces).

```
Yes, absolu- +absolutely absolutely.
Well, I gue- +guess -- I would think this is what they intended.
```

2.4 Restarts

Speaker restarts are indicated with double dash – surrounded by spaces. Annotators use this convention for cases where a speaker stops short, cutting him/herself off before continuing with or rephrasing the utterance.

```
Did people uh -- did fights ever break out uh over hockey?
Since she -- when she died we moved from across the street.
```

2.5 Mispronounced or non-standard words

An asterisk * is used for obviously mispronounced words (*not* regional or non-standard dialect pronunciation), or for words that are made up on the spot by the speaker or idiosyncratic to that speaker's usage. Annotators should transcribe using the standard spelling and should not try to represent the pronunciation.

```
They have as much *knowledgement about things as we've got.

He insisted that we ((*teak)) -- talk to him in Italian.
```

[1] Please refer to Appendix C for details on the ARPAbet phonetic transcriptions.

```
An' I said, "What *picture?" An' they looked at me an' they said,
"Oh, I don't know. We seen some *picture an' we thought it looked
like you" because they knew I didn't know (()).
```

3 Additional markup

3.1 Unclear or unintelligible speech

Sometimes an audio file will contain a section of speech that is difficult or impossible to understand. In these cases, annotators use double parentheses (( )) to mark the region of difficulty.

If it is possible to make a guess about the speaker's words, annotators should transcribe what they think they hear and surround the stretch of uncertain transcription with double parentheses:

```
And she told me that ((I should just leave.))

Four blocks in another direction is ((Epiphany)).  It's a Roman
Catholic church.

Like, maybe, once every two years I go to ((Wawood)).
```

If an annotator is truly mystified and can't make out at all what the speaker is saying, empty double parentheses should be used to surround the untranscribed region.

```
And she came down with the big (( )) and start whackin' us both
over the head.
```

3.2 Interjections

The following standardized spellings are used to transcribe interjections. Interjections do not require any special symbol.

Interjection	**Arpabet transcription**
`duh`	D AH1
`eee`	IY1
`ew`	UW1
`ha`	HH AA1
`hee`	HH IY1
`huh`	HH AH1
`huh-uh` *(neg.)*	AH1 AH0
`hm`	HH M
`jeepers`	JH IY1 P ER0 Z

Interjection	Arpabet transcription
jeez	JH IY1 Z
mm	M
mhm *(pos.)*	AH0 HH AH0 M, AH0 M HH AH0 M
nah	N AA1
oh	OW1
ooh	UW1
uh-huh *(pos.)*	AH0 HH AH1
uh-oh	AH1 OW2
whoa	(HH) W OW1, HH OW1
whew	(HH) W UW1, HH Y UW1
whoops	(HH) W UW1 P S
yay	Y EY1
yeah	Y AE1, Y EH1 AH0
yep	Y EH1 P
yup	Y AH1 P

3.3 Other transcription symbols

In addition to the transcription conventions outlined above, the following symbols are used to for the transcription of other kinds of noises made by either the main speaker or one of the other participants in the interviews:

{BR}	breath	(The speaker takes an audible breath.)
{CG}	cough	(The speaker coughs, or clears his/her throat.)
{LS}	lip smack	(The speaker smacks his/her lips.)
{LG}	laughter	(The speaker laughs.)
{NS}	noise	(Loud background noise, e.g. a door slamming, cars honking etc.)

```
What'd I do when I was young, well uh {BR}
No, well {CG} -- well I worked for an insurance company.
We probably were too stupid to think that. {LG}
```

3.4 Formal methods and style coding

If a recording contains a section on formal methods (i.e. a part of the interview where people are explicitly asked to read out texts or word lists, or to make judgments about language use), the appropriate two-letter style codes should be indicated on a separate style tier.

For the reading passage and the word list, the whole section containing the formal method in question should be marked by one annotation unit containing the two-letter style code. For the semantic differential and the minimal pairs sections, only the words specifically targeted should be annotated with the respective two-letter codes; the rest of the section should be coded as "L" ("Language"):

Code	**Formal method**
SD	*Semantic differential:* The speaker is asked to describe the difference in meaning between two closely related words.
RP	*Reading passage:* The speaker is asked to read out a text passage/list of sentences.
WL	*Word list:* The speaker is asked to read out a list of individual words.
MP	*Minimal pairs:* The speaker is asked to read aloud pairs of words and is asked whether they differ in pronunciation.

If style coding is done in the conversational part of the recording, the following one-letter codes from the Style Decision Tree (Fig. 1) should be entered on the style tier:

Careful Speech		**Casual Speech**	
R	Response	N	Narrative
L	Language	G	Group
S	Soapbox	K	Kids
C	Careful (Residual)	T	Tangent

4 Some general considerations

Annotators should not try to correct non-standard grammatical features; e.g. "I seen him" for "I saw him" should be transcribed as spoken. The same goes for words that are used in a non-standard way: annotators should transcribe what is spoken, not what they expect to hear.

```
These kids come up, says, "Give me a couple bucks." I says, "No I
can't."

And my brother was messin' around:  "He ain't gonna come. He ain't
gonna come." I says, "He don't come, I'm gonna kill you."

She was telling her that, y'know "Youse are going together but
youse ain't saying youse are gonna get married," y'know.
```

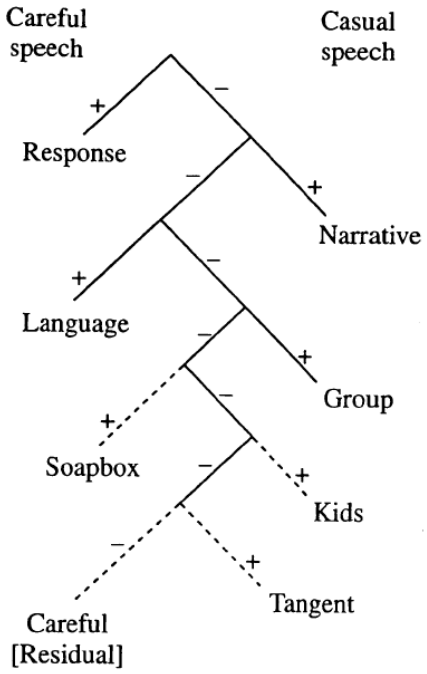

Figure 1: **Style Decision Tree** for stylistic analysis of spontaneous speech in sociolinguistic interviews.

However, annotators should not try to imitate a speaker's non-standard pronunciation. Standard spelling should be adopted even for non-standard pronunciations. Obviously mispronounced words (as opposed to non-standard pronunciations) should be marked with the asterisk * symbol.

Exceptions to this are the coding of final *-ing*, which should be transcribed as either *-in'* or *-ing*, depending on the speaker's pronunciation, or other elided sounds, which should be marked by an apostrophe.

```
I stopped hanging out with 'em.
So pretty much he's runnin' around free.
But he's nothin'. An' he's stuck in bein' nothin'.
```

A Example transcript

The following lines show an example transcription exported from the original ELAN file:

```
BC   I don't know. {LG} He was plannin'. We'll go to see the one
     show. We were gonna double date with his brother and his
     girlfriend.

BC         We'll see the one show and then we'll go out.
IV   Mhm.

BC   Oh man. This k- -- he was cute. (( )) He's got my lunch
NS                                                          {NS}

BC          at school.
IV   (( ))

BC   And I says -- 'cause I didn't go to school last week. I was
     gonna go.

BC   I says, "I'll *murderize him. I'll kill him."

BC   My mom says, "No you know what you do?

BC   You go into school, and you act as if nothin' ever happened. You
     forgot all about it.

BC   And wait till it starts buggin' him. He'll say to himself, "She
     doesn't care. Let's try again."

BC   And then she says, "when he says let's, you know, go out, you
     know.

BC   Say no. {LG} And then start on him." {LG}
IV                                        {LG}

IV   {LG} Did you follow her advice?

BC   Oh I didn't go to school last week. I was over the hospital.

BC   Yeah.

IV   Yeah. Have you talked to your mom about problems with boys?

BC   Yeah.                                                 Yeah.
```

IV Do you think she's really (( )) on it? Maybe?

IV Did she date a lot when she was younger?

BC My grandfather didn't believe in it, she said. {LG}
IV Oh yeah?

BC Yeah.

BC She says, like you know one time she went out with father. Twenty-one.

BC He made her be in at eleven o'clock. Mhm.

BC {LG} It was somethin'. 'Cause they were rollerskatin'. She said {BR} they went

BC ((and)) it was five after.

BC And it took about ten minutes to get home from the L stop.

BC She said, "We put our rollerskates on, me and your aunt, and rollerskated down the street.

IV {LG}

BC Met your grandfather at the corner, and he beated me so bad. It wasn't civil for bein' ten minutes late."

BC And he made her be in at eleven o'clock (( )).
IV {LG}

BC And I says, "that's like how you're so strict."
IV (( ))

B Files to turn in

The following files should be turned in via email when annotators reach the end of a sound file to be transcribed:

- **ELAN transcription file (*.eaf*)**
- **list of names and addresses** (see below)

If annotators are aware that they are prone to misspelling words, they should check the final transcription for spelling errors. This can be done by e.g. exporting the transcript as a tab-delimited text file in *ELAN*, and spell-checking the exported text file in a word processor such as *Word*.

Due to the confidential nature of the interviews transcribed, all material will have to be anonymized at a future point in time by removing all identifying names and addresses from the recordings. To facilitate this process, transcribers are asked to keep a list of names and addresses that occur in the sound file while they are transcribing. This list can be a simple spread sheet with the name/address in one column, and the approximate time stamp of the occurrence in the sound file in a second column (see Fig. 2). The following should be listed:

- full names (first name *and* last name - just first names are okay)
- specific addresses (many people talk about where they live or grew up)

```
Uh, my name is Jim Haney.
I live at twenty-five thirty-eight South ((Hicks)) Street.
We went to grade school together. His name's Morgan Zantori.
```

General references to locations in the city (e.g. Broad Street, City Hall) or to well-known public figures do not need to be listed.

```
Yeah. I walk to Broad Street then I take the Broad Street Subway
down to Center City.
```

	A	B	C
1	**Time**	**File: PH06-1.Jenny.1Mono**	
2	0:00:47	((Stanny)) Keller	
3	0:01:37	Mifflin Street	
4	0:01:45	Buffalo	
5	0:03:15	Fitzgerald	
6	0:05:28	Lisa	
7	00:06:10-11	Lisa McCafferty and Stephen	
8	0:08:27	Jimmy	
9	0:08:28	Casey	
10	0:08:35	Casey	
11	0:11:21	Lisa McCafferty	
12	0:11:34	Casey	
13	0:12:09	Mrs. Mack	
14	0:13:01	Uncle Fred	
15	0:13:25	Mrs. Stutsky	
16	0:13:58	Michael Zancolli	
17	0:16:15	Barbara	
18	0:20:05	Stephen	
19	0:20:06	Lisa	
20	0:21:30	Lisa	
21	0:22:19	Diane	
22	0:27:58	Fitzgerald Street	
23	0:30:13	Jimmy	
24	0:38:46	Casey	
25	0:38:55	Jimmy	
26	0:40:11	Our Lady of Mount Carmel	
27	0:40:17	Saint Marie Goretti	
28	0:40:20	Saint John Newman	
29	0:41:35	Casey	
30	0:41:38	Jamie	
31	0:41:55	Casey	

Figure 2: **List of names and addresses.**

C The ARPAbet

Vowels			Consonants		
Phoneme	**Example**	**Transcription**	**Phoneme**	**Example**	**Transcription**
AA	bot	B AA T	B	be	B IY
AE	bat	B AE T	CH	cheese	CH IY Z
AH	but	B AH T	D	day	D EY
AO	bought	B AO T	DH	that	TH AE T
AW	bout	B AW T	F	fee	F IY
AY	bite	B AY T	G	go	G OW
EH	bet	B EH T	HH	he	HH IY
ER	bird	B ER D	JH	just	JH AH S T
EY	bait	B EY T	K	key	K IY
IH	bit	B IH T	L	late	L EY T
IY	beat	B IY T	M	me	M IY
OW	boat	B OW T	N	knee	N IY
OY	boy	B OY	NG	sing	S IH NG
UH	put	P UH T	P	pay	P EY
UW	boot	B UW T	R	read	R IY D
			S	sea	S IY
			SH	she	SH IY
			T	tea	T IY
			TH	thanks	TH AE NG K S
			V	vain	V EY N
			W	we	W IY
			Y	yes	Y EH S
			Z	zoo	Z UW
			ZH	pleasure	P L EH ZH ER

Stress is indicated by digits following the stressed vowels. There are three levels of stress:

Value	Level of stress
0	no stress
1	primary stress
2	secondary stress

Example:

`in`	AH0 N, IH1 N
`the`	DH AH0, DH AH1, DH IY0
`dictionary`	D IH1 K SH AH0 N EH2 R IY0
`stress`	S T R EH1 S
`is`	IH1 Z, AH0 Z
`indicated`	IH1 N D AH0 K EY2 T AH0 D
`by`	B AY1
`digits`	D IH1 JH AH0 T S
`following`	F AA1 L OW0 IH0 NG
`stressed`	S T R EH1 S T
`vowels`	V AW1 AH0 L Z

D Summary of transcription conventions

Category	Condition	Markup	Example	Explanation
Orthography and spelling	**Numbers**	spelled out	twenty-five, one oh nine, one hundred thirty-seven	Write out in full; dashes for twenty-one through ninety-nine.
	Contractions	transcribe as spoken	can't, I'm, gonna	If you hear a contraction used, write it as a contracted form.
	Punctuation	comma, question mark, exclamation mark, period, quotation marks	, ? ! . "	Limited to these symbols.
	Pronounced acronyms	no special markup	NAFTA	Write letters with all caps, no space between letters.
	Individual letters	surrounded by spaces	I before E, Y M C A	Individual letters spelled out, with spaces in between.
Disfluent speech	**Filled pauses**	no special markup	ah, eh, er, oh, uh	Limited to this list.
	Partial words	`-, (+)`	absolu- +absolutely	Speaker-produced partial words are indicated with a dash. Transcribe as much of the word as you hear. Indicate intended word immediately afterwards, preceded by a plus sign.
	Speaker restart	`--`	I thought he – I thought he was there.	Used when the speaker stops short and then repeats him/herself, or abandons the utterance completely, restarting with a new sentence.
	Mispronounced or non-standard words	`*`	*knowledgement	Speech errors or idiosyncratic vocabulary. **NOTE:** Do not use this symbol to indicate non-standard but common regional/social dialect pronunciations. Transcribe non-standard pronunciation variants or mispronounced words using standard orthography.
Other markup	**Unclear or unintelligible speech**	(( ))	They lived ((next door to us)).	Parentheses indicate a transcriber's best attempt at transcribing a difficult passage, or, if left empty, an entirely unintelligible passage.
	Interjections	no special markup	uh-huh, yeah, mhm	Use standardized spellings.