SLX

The SLX Corpus of Classic Sociolinguistic Interviews

Corpus and Tools User Manual

http://www.ldc.upenn.edu/Projects/DASL/SLX

1 Introduction

Welcome to the SLX Corpus of Classic Sociolinguistic Interviews!

This document provides a brief guide to installing and using the annotation tools included in this publication. Please note that this is a pre-release version of the corpus. The tools included here are optimized for browsing and exploration of the SLX corpus data. These tools can be extended to create new time-aligned transcripts or tag additional variables within the existing corpus; however, these functions are not fully supported in the current release.

Corpus users can visit the DASL SLX Corpus site (http://www.ldc.upenn.edu/Projects/DASL/SLX) for updates, bug fixes and additional information about the SLX Corpus and accompanying tools.

The current release includes a pre-compiled package of these tools for Windows users. We plan to create a pre-compiled package for i386 Linux users in the future.

2 Installing the Corpus Tools

The Tools directory on the corpus DVD contains a file called **SLX-setup.exe**.

This is a self-installing setup file for the SLX Corpus Tools for Windows. Double-clicking on the file icon will start the installation program. The installation process should be self-explanatory: just follow the on-screen instructions. Once the installation is complete, you will find a new entry named "LDC Penn SLX Corpus Tools 1.0" in your Start Menu. Please note that the setup program installs only the tools and not the data of the SLX Corpus.

The SLX Corpus Browser is an interactive tool for exploring the SLX Corpus. The Browser works with the DVD as well as with a copy on your hard disk. The contents of the corpus can be copied to your hard disk by creating a directory on your disk, and copying all of the subdirectories on the DVD to that directory.

NOTE: the SLX Corpus Browser assumes all of the DVD's subdirectories ("speech", "transcripts", "annotations", etc.) to be under the directory specified as the SLX Corpus Location (See Section 4.1).

3 Uninstalling the Corpus Tools

Go to your Start Menu and select

```
Programs->LDC Penn SLX Corpus Tools 1.0 ->Uninstall SLX Corpus Tools
```

Note that it is possible that the old entry in the Start Menu will not be removed. This is a Microsoft Windows bug.

4 Installing IPA Fonts

The Sociolinguistic Variable Survey files contain phonetic symbols that require special IPA fonts. The original coding for the survey was done using Microsoft Excel, with the SIL font SilDoulosIPA93 for phonetic transcription.

The SIL IPA93 Fonts can be downloaded from http://www.sil.org/computing/fonts/encore-ipa.html. The font package is also included in this corpus (tools\silipa93.exe), with permission from SIL.

To install the SIL IPA93 Fonts on Windows, double click on the installation program, tools\silipa93.exe, and follow the instructions on the screen.

If you are running Windows NT, 2000 or XP, you will also need to do the following after running the installation program:

o Go to Start->Settings->Control Panel->Fonts.

o Go to File->"Install New Font" from the menu.

o Select the \WINNT\System folder, click on "Select All", and then on "OK".

If the installation is successful, you will see fonts named "SILDoulos IPA93" (among others) in your applications' (Word, etc.) font selections.

***Note**: Some of the IPA characters used in the original sociolinguistic variable survey are not displayed (or are incorrectly displayed) in DASLTrans. Our plan is for future releases of the toolkit to use a Unicode-based font that will resolve this display problem.* The corpus also includes the original Microsoft Excel versions of the variable survey forms, which display the fonts properly.

5 Exploring the Corpus

5.1 The SLX Corpus Browser

The SLX Corpus Browser is an interactive assistant that will step you through the various corpus components. To start the Browser, go to your Start menu and select

```
Programs->LDC Penn SLX Corpus Tools 1.0 ->Start SLX Corpus Browser
```

In a moment the Browser window will open, containing the list of speakers and links to the corpus documentation files plus transcripts and annotations for each speaker.

The first thing you need to do is to specify where the SLX Corpus is located on your machine. If you're working from the DVD, this location will be the DVD drive. Otherwise, it will be the directory where you placed a copy of the corpus. You can type the path to the folder directly into the SLX Corpus Location text box, or you can click on the icon next to the text box to browse through folders on your computer.

Once you have specified the corpus location, you're ready to access the data.

5.2 Documentation Files

If you click on "See Documentation", you will be shown a list of files that explain various aspects of the SLX corpus. These include:

- Overview of the Corpus
- Guide to Stylistic Annotation
- Sociolinguistic Variable Survey Guide
- Mapping of Speakers and Files
- Speaker Demographic Information

- Guide to Segmentation and Transcription
- The SLX Corpus and Tools User Manual (this document)

Select a document title and click OK to see the .pdf file.

Note: You should read through the Corpus Overview (readme.pdf) before proceeding.

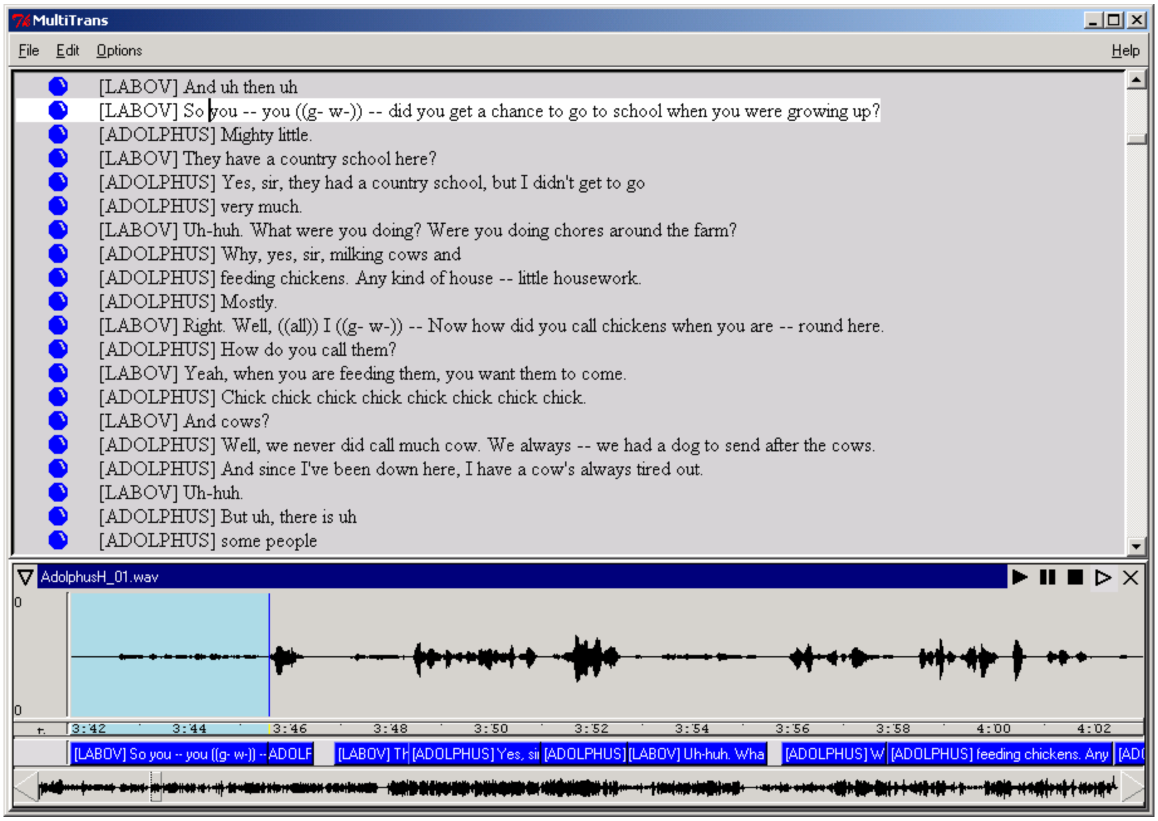
5.3 Transcript Files: The MultiTrans Tool

Overview

Back in the main SLX Corpus Browser window, select a data set to explore by clicking on a speaker's name. Now click the "See Transcription" button to see the list of transcription files associated with that speaker. The list of transcripts includes individual speaker transcripts (in the form *AdolphusH_01_ADOLPHUS.lcf,* where the name in all CAPS indicates the target speaker); it also includes merged speaker transcripts where all speakers are shown in a single transcript file (these files are named like *AdolphusH_01_ALL.lcf*). Both types of transcripts correspond to the same audio file. Select the transcript you wish you see, and click OK.

Tool Display

This will start up the MultiTrans Transcription tool. **NOTE: It may take several minutes for the audio and transcript files to load. Please be patient.** You should not click the MultiTrans or SLX Corpus Browser windows while waiting for the tool to load. When the file opens, you'll see something like this (this example shows one of the merged speaker files):



[LABOV] And uh then uh
[LABOV] So you -- you ((g- w-)) -- did you get a chance to go to school when you were growing up?
[ADOLPHUS] Mighty little.
[LABOV] They have a country school here?
[ADOLPHUS] Yes, sir, they had a country school, but I didn't get to go
[ADOLPHUS] very much.
[LABOV] Uh-huh. What were you doing? Were you doing chores around the farm?
[ADOLPHUS] Why, yes, sir, milking cows and
[ADOLPHUS] feeding chickens. Any kind of house -- little housework.
[ADOLPHUS] Mostly.
[LABOV] Right. Well, ((all)) I ((g- w-)) -- Now how did you call chickens when you are -- round here.
[ADOLPHUS] How do you call them?
[LABOV] Yeah, when you are feeding them, you want them to come.
[ADOLPHUS] Chick chick chick chick chick chick chick chick.
[LABOV] And cows?
[ADOLPHUS] Well, we never did call much cow. We always -- we had a dog to send after the cows.
[ADOLPHUS] And since I've been down here, I have a cow's always tired out.
[LABOV] Uh-huh.
[ADOLPHUS] But uh, there is uh
[ADOLPHUS] some people

The top portion of the screen displays the transcript while the bottom portion displays two views of the audio wave file plus speaker segments, and the audio playback buttons.

In the transcript display, each blue dot represents a segment boundary. The dot is followed by the speaker name and then the transcript content for that segment. You can use the scrollbar at the right of the transcript display to scroll up and down in the transcript window.

The larger of the two waveform panels corresponds to the section of the transcript currently displayed, and shows a small amount of audio at a time. If you click on a segment in the transcript file, the corresponding segment will be highlighted in the large waveform display as well.

A timeline appears beneath the large waveform display. Under this is the segment panel, which provides a graphic representation of the segments in the transcript. Each rectangular box contains a separate speaker segment, ordered in sync with the timeline and the waveform display.

Below the segment panel, the small waveform display at the very bottom of the window shows the entire audio file for the current transcript. You can use the arrow buttons at the left and right of the small waveform to scroll through it to find a particular section of the audio file, or you can drag the rectangular cursor back and forth to move through the waveform more quickly. The segment panel will move in sync with the small waveform display. If you click one of the segment boxes, the corresponding section of the transcript will then be displayed in the main transcript window.

Playback
The audio playback buttons appear beneath the transcript on the right hand side of the display. To start playback, click once with your mouse anywhere in the large waveform display. Then click the black arrow, or hit <TAB>, to begin playback from that starting point. The audio will continue playing until you stop it or until it reaches the end of the file.

During playback, the tool will move the cursor through the large waveform display while highlighting the corresponding segment boundaries, and also highlighting the corresponding segment in the transcript display.

To play an individual segment, you can click on the segment within the transcript display and hit <TAB>, or click the open arrow button in the audio playback buttons. You can also select the segment from the segment panel below, and again hit <TAB> or the open arrow to play. With either function, the tool will also highlight the corresponding segment in both the large waveform display and the transcript display.

Additional Information
To exit MultiTrans, go to the File menu and pull down Quit.

The MultiTrans tool also contains a hidden "bonus feature" that allows you to create spectrograms, pitch contours and spectrums. To access this feature, right-mouse click on the waveform display and select Create Pane -> Spectrogram/Pitch Contour/Spectrum Section. Note that this feature has not yet been fully developed or documented.

You can find complete guidelines for the segmentation and transcription of the SLX Corpus among the corpus documentation files.

The MultiTrans Tool included in this release has been optimized for browsing the SLX Corpus data. The tool also supports segmentation and transcript creation from scratch. The tool will also allow you to edit existing SLX Corpus segmentation and transcription files (although any changes must be saved to a new location.) See the User Manual on the Help menu for more information. Future tool releases will provide additional support and documentation for modifying existing transcripts and segmentation files and for creating new ones.

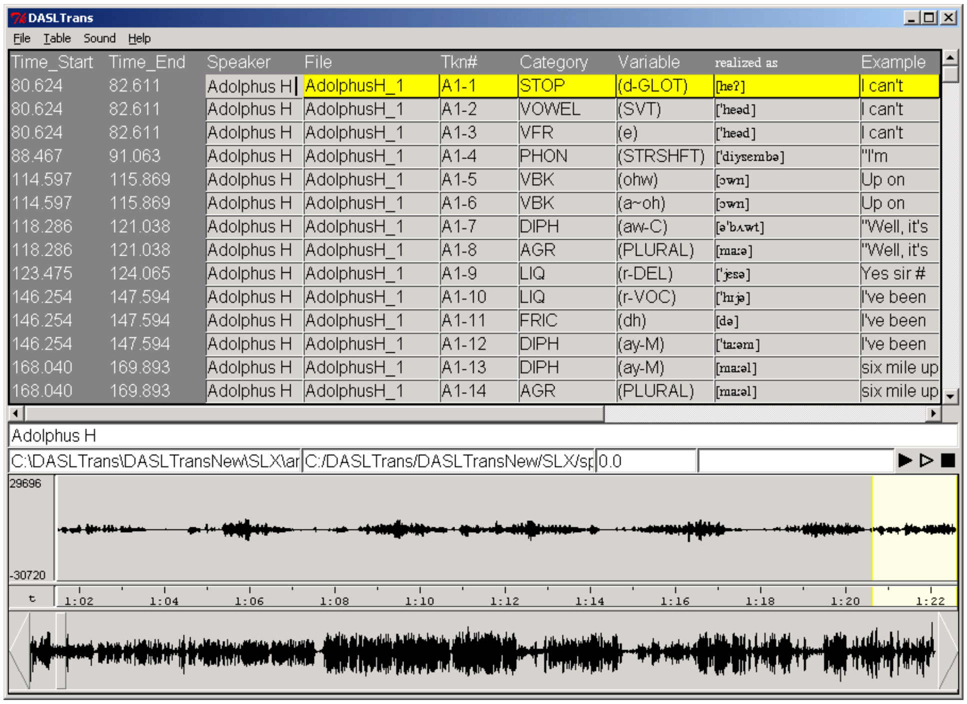
5.4 Sociolinguistic Variable Annotation Files: The DASLTrans Tool

Overview

Back in the SLX Corpus Browser window, select a speaker and click See Variables. This will start the DASLTrans tool. DASLTrans allows you to explore the sociolinguistic variable survey for each speaker. The tool also permits you to create new annotations within the existing survey.

Display

It may take several seconds for the audio and variable files to load. When the file opens, you'll see something like this:



Time_Start	Time_End	Speaker	File	Tkn#	Category	Variable	realized as	Example
80.624	82.611	Adolphus H	AdolphusH_1	A1-1	STOP	(d-GLOT)	[heʔ]	I can't
80.624	82.611	Adolphus H	AdolphusH_1	A1-2	VOWEL	(SVT)	['heəd]	I can't
80.624	82.611	Adolphus H	AdolphusH_1	A1-3	VFR	(e)	['heəd]	I can't
88.467	91.063	Adolphus H	AdolphusH_1	A1-4	PHON	(STRSHFT)	['diysembə]	"I'm
114.597	115.869	Adolphus H	AdolphusH_1	A1-5	VBK	(ohw)	[ɔwn]	Up on
114.597	115.869	Adolphus H	AdolphusH_1	A1-6	VBK	(a~oh)	[ɔwn]	Up on
118.286	121.038	Adolphus H	AdolphusH_1	A1-7	DIPH	(aw-C)	[ə'bʌwt]	"Well, it's
118.286	121.038	Adolphus H	AdolphusH_1	A1-8	AGR	(PLURAL)	[maːə]	"Well, it's
123.475	124.065	Adolphus H	AdolphusH_1	A1-9	LIQ	(r-DEL)	['jesə]	Yes sir #
146.254	147.594	Adolphus H	AdolphusH_1	A1-10	LIQ	(r-VOC)	['hɪjə]	I've been
146.254	147.594	Adolphus H	AdolphusH_1	A1-11	FRIC	(dh)	[də]	I've been
146.254	147.594	Adolphus H	AdolphusH_1	A1-12	DIPH	(ay-M)	['taːəm]	I've been
168.040	169.893	Adolphus H	AdolphusH_1	A1-13	DIPH	(ay-M)	[maːəl]	six mile up
168.040	169.893	Adolphus H	AdolphusH_1	A1-14	AGR	(PLURAL)	[maːəl]	six mile up

Adolphus H

C:\DASLTrans\DASLTransNew\SLX\ar | C:/DASLTrans/DASLTransNew/SLX/sp | 0.0

The main display window is a spreadsheet-style table view of the variables selected for each speaker plus corresponding annotations and notes. The columns shown in DASLTrans are drawn directly from the original sociolinguistic variable survey for each speaker, and include:

Time_Start	Start time of the utterance containing the variable
Time_End	End time of the utterance containing the variable
Speaker	The name of the speaker
File	The file name the data is from
Tkn#	Token number. This is a unique code that shows the speaker's initial(s), the file number of the speaker that the data is from and then a number that corresponds to the order in which the token was added. So, the 10th token of Adolphus from his file 01 has a tkn# of A1-10.
Category	This is the name for the category that the variable belongs to. (There are fewer categories than variables. See the Sociolinguistic Variable Survey Guide in the corpus documentation for the grouping of variables into categories.)
Variable	This is the name of the variable

Realized as This is the realization of the variable in IPA transcription, or standard orthography for syntactic variables. The symbol “Ø” represents the absence of a form when null is on variant of a variable (i.e., copula deletion).**

Example This is the utterance surrounding the variable – the context.

Style This it the style code for the variable in the context of the interview

Notes This includes comments about any aspect of the variable, as well as definitions for some of the British lexical items.

** ***Note: Some of the IPA characters used in the original sociolinguistic variable survey are not displayed (or are incorrectly displayed) in DASLTrans. Future releases will use a Unicode-based font and will resolve this display problem.*** The corpus also includes the original Microsoft Excel versions of the variable survey forms, which display the fonts properly.

Playback and related functions

Many of the functions of the DASLTrans tool are similar to MultiTrans. Clicking on a cell in the table will highlight the corresponding segment in the waveform view below. Look in the "example" column to see the transcription for that segment. Hit <F1> or the open arrow button to play that segment. Hitting <TAB> will not play the segment but will instead tab through the table cells for that variable. The right, left, up and down arrows move to the neighboring cells in the corresponding directions.

You may notice that several instances of the same example within a given speaker's variable table. This is due to the fact that one utterance may contain several variables of interest. Look at the "Category" and "Variable" columns plus the "Realized as" field to see which variable is currently described.

You can use DASLTrans to play the entire audio file rather than one segment/token at a time. The <F1> key will toggle playing and stopping of the audio. If an individual segment is chosen, it will play that segment only. If a single point is selected in the waveform, it will start playing from that point. As with MultiTrans, the DASLTrans tool features aligned playback and will move the cursor through the large waveform display while highlighting the corresponding tokens in the table display.

Sorting tokens

You can sort the tokens in any number of ways by double-clicking on the column headings. The default sorting is time-ordered. Double-click on the "Category" heading to sort variables by variable category. This will allow you to examine, for instance, all variables that fall into the AGR (agreement) category one after the other. You can also sort by "Variable" to see, for example, all cases of -STR-DEL (unstressed syllable deletion) for this speaker. Tokens can be sorted by any category except "Realized as" (this is due to the font display issues described above, and will be resolved in a future release). Sorting according to start and end times can also be done from the menu: Table->Sort->Sort Annotations by Start Time or Sort Annotations by End Time.

Creating new annotations

DASLTrans also allows you to create new annotations. To create a new token, select a section of audio in the waveform display by clicking and dragging your mouse over the region. Then hit the <Enter> key. This will create a blank row in the table display. The Time_Start and Time_End cells will automatically be filled in with. You can type in each cell to fill in the characteristics of the token. The Shift-right arrow and Shift-left arrow key combinations will move the insertion point within the current cell.

You can also copy an annotation from an existing row in the table. Hitting the <Shift-Enter> key combination will create and insert an annotation row in the spreadsheet. If the current region of the waveform is chosen, the start and end times of the current region are inserted. In addition, the feature values of the annotation row that was highlighted when you hit <Shift-Enter> are copied into the new annotation row.

Hit Control-d to delete the currently highlighted annotation.

If the start and end times for a token are incorrect, hit Control-g to update the start and end times of the current annotation using the current region selected from the waveform.

Additional features

The menu item Table ->Find or the key combination Control-f will bring up a dialog window to search a string in the spreadsheet. If a matching string is found in the cell, the cell will be highlighted, and the annotation row becomes the current annotation.

The Table -> View menu items allow the user to select which annotations should be displayed in the table. The Table -> Configure Columns menu item allows the user to reorder the column display and adjust column width.

The tool also contains a hidden "bonus feature" that allows you to create spectrograms, pitch contours and spectrums. To access this feature, right-mouse click on the waveform display and select Create Pane -> Spectrogram/Pitch Contour/Spectrum Section. Note that this feature has not yet been fully developed or documented.

Additional Information

To exit DASLTrans, go to the File menu and pull down Quit.

You can find complete guidelines for the Sociolinguistic Variable Survey among the corpus documentation files.

The DASLTrans Tool included in this release has been optimized for browsing the SLX Corpus data. Although the tool will also allow you to edit existing annotations and create new ones, these functions are not fully supported or documented in this release. Future tool releases will provide additional support and documentation for creating new annotations.

5.5 Further information about SLX Corpus Tools

Future tool releases will integrate the segmentation and transcription features of MultiTrans and the annotation features of DASLTrans into a single tool, allowing users to go from segmentation to transcription to coding of variables within a single system.

Both DASLTrans and MultiTrans were developed using the Annotation Graph Toolkit (AGTK - http://agtk.sf.net), also developed by LDC. Future versions of DASLTrans and MultiTrans will be available from the AGTK website as well as the SLX Corpus site:
http://www.ldc.upenn.edu/Projects/DASL/SLX

This corpus release also includes Transcriber, a multi-platform transcription tool developed jointly by LDC and DGA. LDC staff used this tool to create the initial audio segmentation and transcription for the SLX corpus. Transcriber is not currently integrated into the SLX Corpus Browser. User manuals, updates and additional information about Transcriber can be found at:
http://www.etca.fr/CTA/gip/Projets/Transcriber/