TEI and Linguistic Interviews
Bill Kretzschmar, University of Georgia

Abstract:

While P4 does offer resources for the transcription of speech (#11) and for some kinds of linguistic analysis (e.g. #15), the basic problem with linguistic interviews is that they are essentially not documents. Today they are, first of all, sound recordings, and various kinds of information and encoding can be derived from them, only one of which is a text transcription covered by TEI. A central question for use of TEI with linguistic interviews is how a text transcription is related to other kinds of digital information (e.g. sound files, acoustical plots, maps), closely followed by the question of how TEI encoding might best be implemented with other layers of text encoding (e.g. lexical, phonetic, grammatical encoding for analysis; survey-specific encoding; document structure encoding for alternate organizational units such as breath groups or prompt/response objects).

It is a pleasure to be here today, but I have to say that I feel at risk for re-enacting the old story about the emperor with no clothes. The other talks on the program are by real experts in TEI, and I cannot and will not claim such status; all of you would see through it immediately if I were to try it on. My brief this morning is Peter Robinson's injunction that " whatever seems really interesting and important to you about the TEI (where it is, what it does well, where it should go, what it could do better..) will be interesting and important to the audience." You and I have little choice now but to trust that he has collared a suitable speaker. And I am in fact right now considering whether and, if so how, to use TEI to weave together the various strands of presentation and analysis in my research project as it enters a new stage in its development. I propose, therefore, to wear my own clothes rather than new imperial ones, and so tell you about things that I know and care about as they relate to TEI.

First, a little about me and what I do. I am an empirical linguist. This means that I do not necessarily assume that each speaker of a language shares the same linguistic system, or conversely that speakers naturally possess a single linguistic system as native speakers, as structural or generative linguists might do; instead, I want to collect great quantities of real

speech from a great many speakers in order to describe what people actually say. Empirical linguists typically employ the grammatical categories postulated by structuralists and generativists, but they test each category empirically to assess its reality in use. Empirical linguists also test the distribution of words, whether as lexical units or as they embody morphological markers or pronunciations, not necessarily as elements in a contrastive system but for themselves, to observe the dynamics of real speech by real people in samples taken from whole regions or communities. When an empirical linguist makes a generalization, it boils down large quantities of speech from many sources, as opposed to the structural or generative prediction of the speech of the group on the basis of one or a few individuals. Many corpus linguists are members of the empirical group, along with survey researchers like me. Empirical linguistics is not a replacement for structural or generative linguistics; it is just different, and there is room for all sorts. The reason to talk about it here is to highlight my goals of recording speech at all, and to make clear at the outset that my goal for encoding speech is certainly not to make it fit within pre-established grammatical or other linguistic categories, but instead to open up the recorded speech to empirical analysis.

My particular interest in the field is language variation research, particularly through the American Linguistic Atlas Project. Historically, American Atlas surveys have sought to elicit about 800 targets from the everyday speech of representative speakers, in an interview of 6-8 hours' duration (see Kretzschmar et al. 1993). More recently, we have adopted a modified interview technique for work in the Western States (field interviews are in progress in California, Colorado, and West Texas) that addresses our need for specific elicitation targets in the context of a discourse-style interview (Pederson 1996a, 1996b; Pederson and Madsen 1989), this time designed to collect 360 targets in about three hours. The earliest interviews were only partially

transcribed because, before tape recorders, the interviewers had to write down the word or phrase of interest on the fly, but the new recorded interviews should be transcribed in full, with encoding in the transcription to identify elicitation targets for lexical and phonetic analysis but also full opportunity to preserve the continuous speech of the speaker for analysis of verb form frequency and other discourse features (Pederson 1996a:54-59). Our most recent interview format, now funded for testing in urban Atlanta, is the one-hour interview, in which we hope to gather fixed-format elicitation data for speech scientists as well as informal conversation that will yield tokens for up to 100 elicitation targets and also the opinions of the speaker about local speech. The whole point of the shorter interview is to speed up the time from interview to publication, so we need encoding that can be applied rapidly.

For the earlier interviews, we have used database methods to store and access the words and phrases written down during the interviews. For the three-hour Western States interviews, the first full-text transcripts from the late 1980s adopted an encoding method as shown here (Figure 1, from Pederson and Madsen 1989). Of course the encoding is primitive, but it does the same sorts of things that are enacted in TEI (principally section 11, but also sections 14 and 15): conversational turns are numbered, and curly braces identify speech by the field worker within the turn; square brackets identify grammatical features, such as the past tense verb in turn 10, and the missing preposition and determiner in turn 9; parentheses identify lexical targets; and angle brackets identify pronunciation issues, enclosing both identifying information about the target item and two different notations for non-IPA phonetics and stress marking. All data in these transcripts had to be entered with the 7-bit ASCII then available. Punctuation is included as seems best to the transcriber. The transcripts were prepared from tape recordings, but there is no indexing from the transcript back to the recording. Personally, I think that this represented a

good start, and it even had the advantage of being parsimonious with markup in a time with very limited storage resources. Not too long ago Matt Zimmerman and Betsy Barry worked through how we could convert the primitive encoding to our current needs. Figure 2 offers a non-TEI-compliant but more modern encoded portion of a similar interview. The markup identifies speakers, targets, turns, and other things of interest to us, though not yet comprehensively in this sample.

Let us now move to the present day. Fifteen years ago it was an achievement to move from separate words and phrases to full-text transcripts. Now, however, we have a more comprehensive list of demands for computer management of interviews. Within texts we want to

--display full text of transcriptions
--link sound to the text so that users can hear what they read in real time
--associate acoustical phonetic information with pronunciation targets
--associate graphical F1/F2 acoustical plots with pronunciation targets
--associate listing and tally scripts with lexical targets
--enable GIS plotting (maps) for pronunciation and lexical targets
--enable technical geography statistical functions for pronunciation and lexical targets
--enable syntactic analysis through POS tagging

Between and among texts we want to

--enable KWIC concordance displays across different texts
--extract and display individual and aggregated informant biographical information from metadata
--link informant information and text access to regional maps and lists of social variables
--link help screens and other useful information to texts and analyses

This list begins with simple display of texts, but it continues with many tasks that users now have to perform separately, including automation of a large number of tasks which heretofore have required specialized software and no small amount of sophistication on the part of the user to accomplish. There are eight different functions within texts that require encoding overhead; each of these also should be able to work across texts. Our survey interviews are also in demand for purposes other than our own, notably for inclusion in the American National

Corpus. We are also interested, of course, in keeping our data for a long time, whether in active use on a site or in an archive. The question is, can all of these demands be accommodated with TEI compliant encoding, or would it be better to implement some other strategy.

The logical design of a site based on TEI-compliant encoding would be focused on the encoded texts of the interviews (Figure 3). Information processing begins with the transcribed text, and ends with eight different kinds of output. The text may be displayed in its own right as output (1). The text will be linked to sound files, so one operation on the text will generate the sound which matches the text (2). Another operation on the text will be to call up associated acoustical phonetic data, which may either be displayed itself (3) or may be plotted on charts for display (4). Yet another pathway begins with the generation of lists and tallies from the text, which may be displayed themselves (5), routed through plotting scripts and displayed on base maps (6), or routed though statistical algorithms from technical geography. The output of the technical geography algorithms may be displayed itself (7), or displayed on base maps (8). Besides the actual transcription itself, we also need to keep metadata about the status of the interview and biographical data about the speaker. Some information about the status of the interview will be added editorially, but in the main the recorded conversations will yield this additional information–and so we also need to be careful to prevent sensitive personal information on the recording from being released with the rest of the interview. None of these operations is too problematic in itself, but I am worried about two things: the extent of and competing hierarchies within the encoding, and, perhaps more important, the fact that our research is actually based not on a text but on the audio recording, and so any text we make of it is secondary and derived.

To take the second matter first, one leading alternative to a text-centered TEI-

compliant plan is to use annotation graphs in a software package (AGTK, http://agtk.sourceforge.net/) created by Steven Bird and Mark Liberman of the Linguistic Data Consortium. The basic principle of annotation graphs is to consider that recorded speech not as a structured discourse but as a continuous stream of data. Bird and Liberman's insight after review of many linguistic annotation formats was that "all annotations of recorded linguistic signals require one unavoidable basic action: to associate a label, or an ordered sequence of labels, with a stretch of time in the recording(s)" (Bird and Liberman 1999: 1). Annotation graph software thus facilitates the association of labels with time slices. Figure 4 illustrates how a time slice at the bottom (taken from one of two data streams in a two-channel recording–a conversation) is associated with various annotations in a database structure at the top. There are eighteen fields associated with each text selection, in addition to the transcription field itself. While the figure happens to show the conversational turns in discourse, there is nothing to prevent slices from being associated with particular words or even speech sounds within a word. Bird and Liberman describe cases of multiple annotation of a single data stream, essentially the creation of annotations at different hierarchical levels, say the levels of phonetic realization, words, and syntax and discourse structures; multiple annotation of the data stream thus avoids overlapping hierarchies within a single encoding structure. Figure 5 is a modified version of Figure 3, now showing the architectural effect on our goals if we used time indexes as a central organizing object instead of a text. In short, there is not much new to see here; if anything, the architecture is cleaner because the use of time indexes allows for greater segregation of the sound and acoustical phonetic pathways. To cap it all off, the annotation graph software exports XML, so there is no reason that a clean transcription text with user-selected characteristics cannot be generated automatically from the databases in which the slices and their various annotations are

stored. The only downside that I can see here is that annotation graph structures are not yet as well-accepted as TEI, even though the LDC, the largest creator and archiver of linguistic databases, made and uses the system. For the long-haul, we may still wonder how well the system will be interpretable after the passage of time.

Now, back to question number one: can TEI-compliant encoding do the job, or as good a job as annotation graphs, particularly with respect to the amount of encoding overhead required and the likelihood of overlapping hierarchies? In a word, no, I don't believe it, not if all of the functionality required were packed into a single instantiation of the text. Some of you might convince me if you tried, because I know that you know more than I do. Segmentation below the utterance level is a key consideration. P4 offers several elements of the <seg> class for breaking out linguistically interesting parts of the text, like sentences, clauses, phrases, words, morphemes, and characters. Below this level, section 16 of P4 provides resources for description of feature structures. It appears to me, however, that this collection of elements may be difficult to apply in practice. We should keep in mind that I am interested in rapid encoding in a production setting. Can my assistants handle it all, whether quickly or not? On another front, I am sure that there are already answers in the TEI community, just unknown to me, to the old saws of linguistics that caution students about rigid segmentation. How should we mark *wanna* and *gonna*, as one word each or as two? What should we do with the fact that acoustical analysis shows that speech sounds in use do not have onsets that are clearly separate from the codas of preceding sounds? To what degree do we want to code for elements that are not there but ought to be, like those in the primitive encoding example, and should they be represented as elements in analysis or as (missing) text? Given answers to these and similar questions, how much weight of segment and feature coding will one text bear? The examples in P4, as they should, treat each

encoding practice in turn; how many of these can we pile together simultaneously? My earlier examples showed that I have to be interested in feature structures, morphology, the lexicon, and syntax, all at once, though generally not for every word in the transcript.

My inclination, and you can tell me just how wrong it is, would be to avoid the potential overload and follow the practice built into the annotation graph software to create a series of parallel texts, each with its own level of annotation. Different text files, not just a single central text file, could then be provided with a manageable degree of annotation and called separately as needed by scripts to do the work we need to do on our integrated site. Anchors and IDs could presumably be used to link the texts. The difficulty would be to keep these different text files aligned as faults were discovered in the transcript (as they always are) and editorial changes are applied. Ideally, one file would be the master version, from which updates could be shared with many specialized child copies. And this idea raises in turn the notion of standoff annotation, under which the child copies would not contain the text at all but merely some independent hierarchy of annotation with pointers back to the text in the central file.

At this point, I fear that I have outrun completely what I know of TEI. I cannot answer the questions I have asked, and I am not at all sure that I have posed them in an intelligible way. Still, these are real problems, and I need to solve them. In the early 90s I was not an early adopter of text encoding myself, preferring database methods because of the nature of my historical data and because of limitations then in mass storage–I remember too well those 10Mb and 20Mb hard drives, which were great improvements over floppies and especially over the punch cards I started with. My colleagues led the way at that time for encoding practices in my field. But now the style of my data has changed and storage and processing limitations are greatly mitigated, and I believe that TEI-compatible text encoding offers the best chance for

long-term archival preservation of data. To answer Peter Robinson's questions, that is what TEI does best right now for texts, offer the chance for long-term archival preservation. Where it should go and what it could do better, for my kind of data, is to accommodate better the notion of a continuous data stream. Bird and Liberman are right about linguistic recordings. TEI may not really have been intended for such non-texts, but I am sure that I am not the first to wish that it could be extended to streaming modes, to video sources as well as audio sources.

Another thing that I want for TEI is a tool to help my assistants apply extensive layers of encoding both rapidly and rigorously. Let me close today with an idea that may work for me, to see what you think of it. I have always thought of TEI as a kind of programming environment, a central structuring device or a project. Some LDC-connected friends from Philadelphia have encouraged me to think that encoding like TEI is just not worth thinking about since other methods fit my data better. I am wondering whether I can get the best of both worlds. What if I used annotation graphs to segment and label the streams of speech, but then exported TEI-compliant texts with the XML export utility in the software? I am not sure whether this would actually work but it seems worth a try. I would give up the text-centered model that I have always associated with TEI, but could retain the value of TEI for archival preservation. Annotation graph software would be my tool for rapid and rigorous encoding (the column/row format is particularly attractive in practical terms, because assistants would always have the possibilities for markup in front of them). AGTK would also offer the chance to make clear decisions about segmentation of the speech stream, and to link the sound and transcripts generated from the system.

My Philadelphia friends might think that I was silly to spend time creating TEI-compliant DTDs for export of texts, but I am used to wearing labels like that. Perhaps you might

be unhappy if TEI were just the clothing on a generated text and no longer the primary organizing method for the project. As I see it, better that than no clothing at all.

References

Bird, Steven, and Mark Liberman. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. Available online at http://agtk.sourceforge.net/.

Kretzschmar, William A., Jr., et al. 1993. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States.* Chicago: University of Chicago Press.

Pederson, Lee. 1995. Elements of Word Geography. *Journal of English Linguistics* 23:33-46.

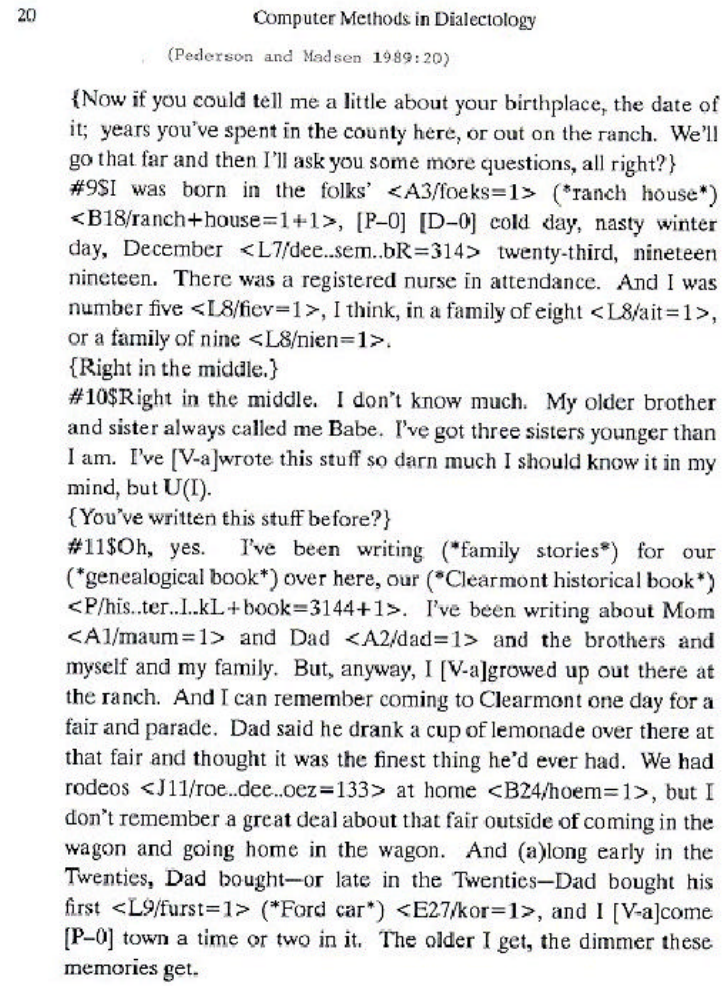
------ 1996a. LAWCU Project Worksheets. *Journal of English Linguistics* 24:52-60.

------ 1996b. LAMR/LAWS and the Main Chance. *Journal of English Linguistics* 24:234-49.

------, and Michael Madsen. 1989. Linguistic Geography in Wyoming. *Journal of English Linguistics* 22:17-24.

Pederson, Lee., and Michael Madsen. 1989. Linguistic Geography in Wyoming. *Journal of English Linguistics* 22:17-24.

Figure 1 Pederson and Madsen Early Encoded Transcript



20 Computer Methods in Dialectology

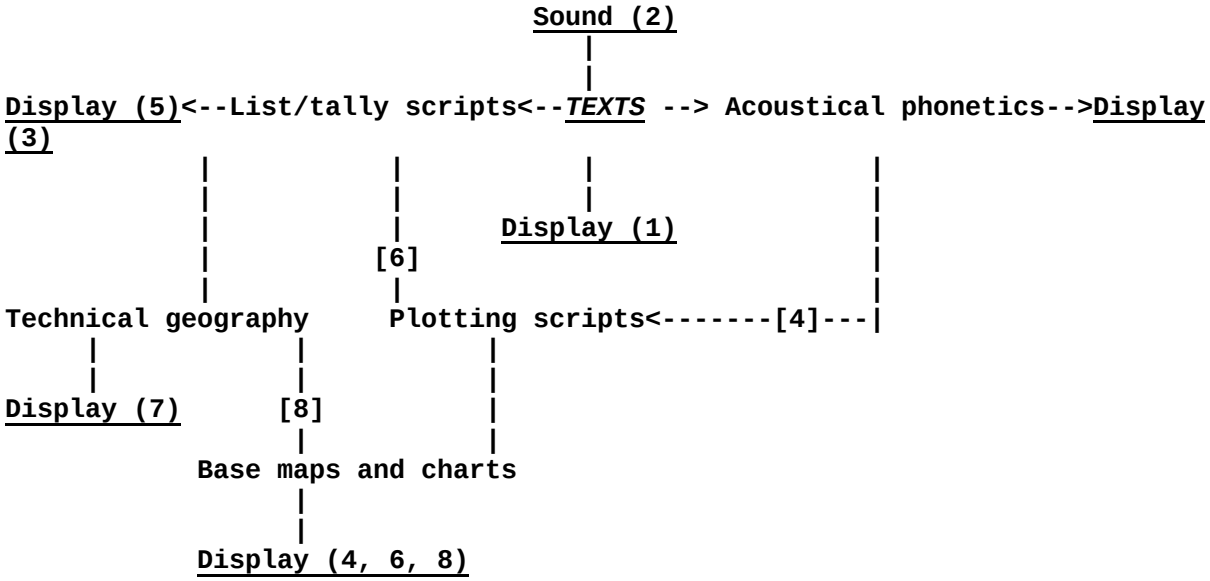
(Pederson and Madsen 1989:20)

{Now if you could tell me a little about your birthplace, the date of it; years you've spent in the county here, or out on the ranch. We'll go that far and then I'll ask you some more questions, all right?}
#9$I was born in the folks' <A3/foeks=1> (*ranch house*) <B18/ranch+house=1+1>, [P–0] [D–0] cold day, nasty winter day, December <L7/dee..sem..bR=314> twenty-third, nineteen nineteen. There was a registered nurse in attendance. And I was number five <L8/fiev=1>, I think, in a family of eight <L8/ait=1>, or a family of nine <L8/nien=1>.
{Right in the middle.}
#10$Right in the middle. I don't know much. My older brother and sister always called me Babe. I've got three sisters younger than I am. I've [V-a]wrote this stuff so darn much I should know it in my mind, but U(I).
{You've written this stuff before?}
#11$Oh, yes. I've been writing (*family stories*) for our (*genealogical book*) over here, our (*Clearmont historical book*) <P/his..ter..l..kL+book=3144+1>. I've been writing about Mom <A1/maum=1> and Dad <A2/dad=1> and the brothers and myself and my family. But, anyway, I [V-a]growed up out there at the ranch. And I can remember coming to Clearmont one day for a fair and parade. Dad said he drank a cup of lemonade over there at that fair and thought it was the finest thing he'd ever had. We had rodeos <J11/roe..dee..oez=133> at home <B24/hoem=1>, but I don't remember a great deal about that fair outside of coming in the wagon and going home in the wagon. And (a)long early in the Twenties, Dad bought—or late in the Twenties—Dad bought his first <L9/furst=1> (*Ford car*) <E27/kor=1>, and I [V-a]come [P–0] town a time or two in it. The older I get, the dimmer these memories get.

Figure 2 Sample Interview Markup

<PROMPT>I'm familiar with homestead; I'm not familiar with desert claim. What is that?</PROMPT>
<RESPONSE_NUMBER>#021</RESPONSE_NUMBER>
<RESPONSE>Well, I really, maybe Victor could tell us. I really don't know what a desert claim is. Would it be all right if he could tell us?</RESPONSE>
<PROMPT>Maybe we'll get back. Yeah, what is a desert claim?</PROMPT>
<RESPONSE_NUMBER>#022</RESPONSE_NUMBER>
<SPOUSE_RESPONSE>Desert claim is the <U(H)> tract of land that the government designated as a <U(H)> the amount of 160 acres, and people could file on that amount of land in an area where that amount of land was government land subject to those filings to the amount of 160 acres. They were entitled to one 160 acre tract</SPOUSE_RESPONSE>
<PROMPT>All right. That's helpful. That's very helpful. <U(F)> So you were born here in Durango. Were you born in a home, were you born in your house? I'm just wondering did people then.</PROMPT>
<RESPONSE_NUMBER>#023</RESPONSE_NUMBER>
 <RESPONSE> My <LEXICAL_TARGET id="A14">mother </LEXICAL_TARGET> stayed at home. </RESPONSE>
<PROMPT>G(A).</PROMPT>
<RESPONSE_NUMBER>#024</RESPONSE_NUMBER>
<RESPONSE>And had her children. She, they didn't go to the hospital. Well, neither did I when our children were born. We stayed in the home and had a country doctor.</RESPONSE>
<PROMPT>OK.</PROMPT>
<RESPONSE_NUMBER>#025</RESPONSE_NUMBER>
<RESPONSE> <U(F)> It was just an old brick house, and a <LEXICAL_TARGET id="A10">midwife</LEXICAL_TARGET> came in.</RESPONSE>
<PROMPT>That's what I was wondering.</PROMPT>
<RESPONSE_NUMBER>#026</RESPONSE_NUMBER>
<RESPONSE>Yes, my <LEXICAL_TARGET id="A14">mother </LEXICAL_TARGET> <U(F)>, when I was born, my <LEXICAL_TARGET id="A14">mother </LEXICAL_TARGET> had a doctor. It was old Doctor Oksner. He was one of the old doctors in Durango. I don't know what would happen if anything went wrong at home, if just everything didn't go bing, bing, bing. If they had problems, I suppose they took care of them right there at the home.</RESPONSE>
<PROMPT><G(A)></PROMPT>
<RESPONSE_NUMBER>#027</RESPONSE_NUMBER>
<RESPONSE>They weren't supposed to have any problems. <U(L)>.</RESPONSE>
<PROMPT>Yeah. They couldn't rush off to the hospital every time you got sick, could you?</PROMPT>
<RESPONSE_NUMBER>#028</RESPONSE_NUMBER>
<RESPONSE>No.</RESPONSE>

Figure 3. Information Flowchart for Linguistic Interviews, Text Centered



Outputs = ( ) Pathways towards outputs = [ ]

Output 1: Transcript in normal orthography
Output 2: Linked sound
Output 3: Acoustical phonetic data in lists
Output 4: Acoustical phonetic data plotted on charts or maps
Output 5: Tally of features in list format
Output 6: Tally of features in map or chart format
Output 7: Technical statistical results in list format
Output 8: Technical statistical results in map or chart format

Figure 4 Annotation Graph Tool (from PowerPoint samples, http://agtk.sourceforge.net/)

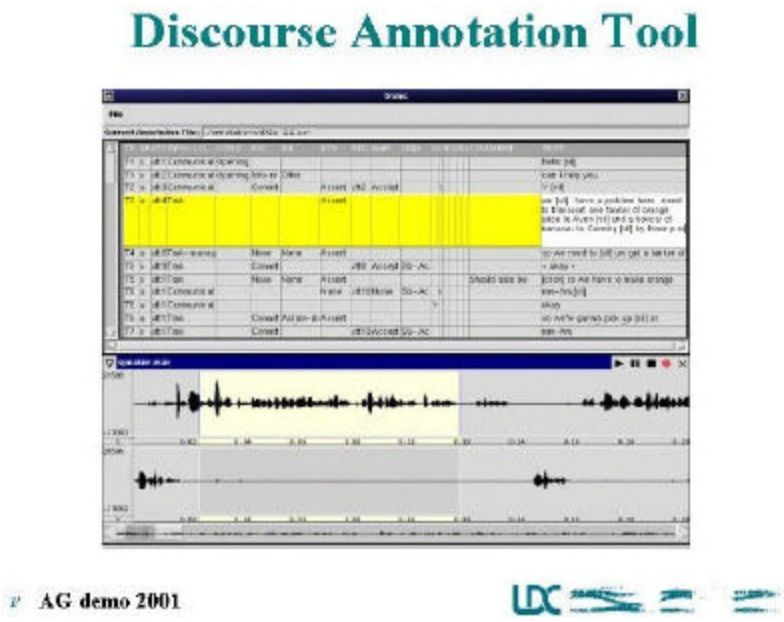
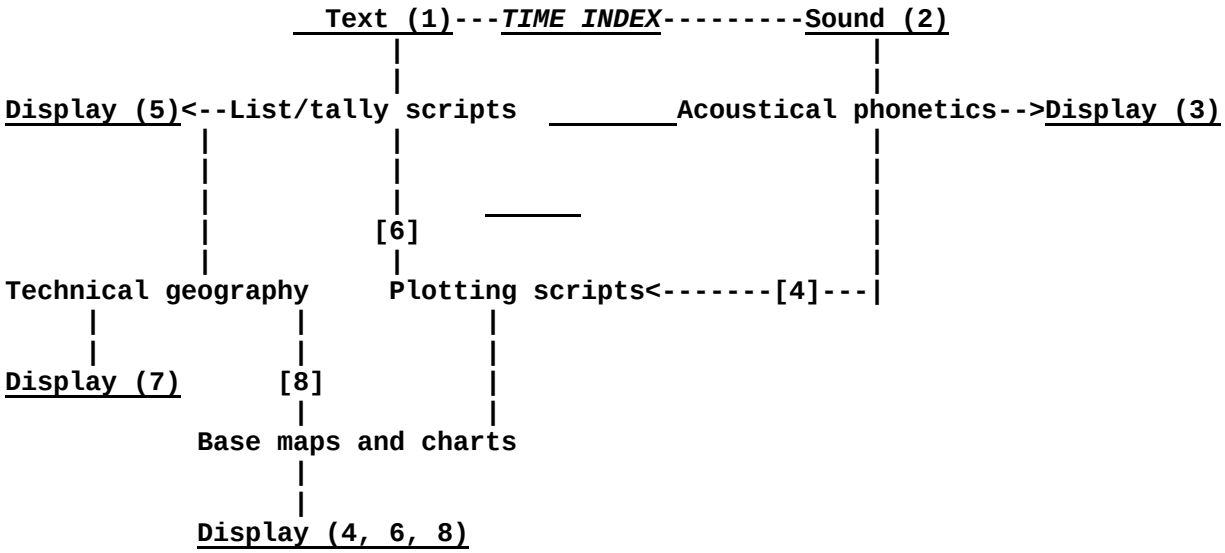


Figure 5. Information Flowchart for Linguistic Interviews, Time Index



Outputs = ( ) **Pathways towards outputs = [ ]**

Output 1: Transcript in normal orthography
Output 2: Linked sound
Output 3: Acoustical phonetic data in lists
Output 4: Acoustical phonetic data plotted on charts or maps
Output 5: Tally of features in list format
Output 6: Tally of features in map or chart format
Output 7: Technical statistical results in list format
Output 8: Technical statistical results in map or chart format