ADVANCING THE UTILITY OF THE TRANSCRIPT: A COMPUTER-ENHANCED METHODOLOGY

Tyler Kendall

Duke University, USA

1. INTRODUCTION

The transcript is often the primary mediating apparatus between theory and data in language research. Researchers from a wide array of linguistic disciplines and across the social sciences rely on transcripts for the analysis and presentation of their data, yet despite some important contributions to the literature (for example, Edwards 2001, Edwards and Lampert 1993, Ochs 1979) most transcripts remain text-based documents, varying in their conventions from researcher to researcher, and limited in their utility to the project-at-hand. While we know, as Jane Edwards writes, that "transcripts are invaluable [since] they provide a distillation of the fleeting events of an interaction, frozen in time, freed from extraneous detail, and expressed in categories of interest to the researcher" (2001:321), we also know that the form of and information in a given transcript will influence our interpretations of the data (Edwards 2001; Ochs 1979). Decisions as seemingly straightforward as how to layout the text to those more nuanced — like how much non-verbal information to include and how to encode minutiae such as pause-length and utterance overlap — have far reaching effects on the utility of a transcript.

This paper presents the approach to the transcript undertaken by the North Carolina Sociolinguistic Archive and Analysis Project. This approach, I argue, helps combat the confusions that arise from text-based transcripts and moves the transcript in new directions, with results that are of benefit to language researchers.

2. THE NORTH CAROLINA SOCIOLINGUISTIC ARCHIVE AND ANALYSIS PROJECT

The North Carolina Sociolinguistic Archive and Analysis Project (NC SLAAP)[1] is a research and preservation initiative being conducted jointly by the North Carolina

[*] I gratefully thank the North Carolina State University Libraries for their support of this project.

[1] For more information, see the NC SLAAP website at `ncslaap.lib.ncsu.edu/`.

Language and Life Project (NCLLP) and the North Carolina State University Libraries. The NCLLP is a sociolinguistic research initiative at North Carolina State University with one of the largest audio collections of sociolinguistic data on Southern American English in the world. The collection consists of over 1500 interviews conducted from the late 1960s to the present, most on analog cassette tape but some in formats ranging from reel-to-reel tape to digital audio and video. While one major goal of NC SLAAP is to preserve the NCLLP's recordings through digitization, it additionally provides a fertile testing ground for exploring new computer-enhanced techniques for sociolinguistic analysis and for experimenting in the storage and presentation of linguistic data, including transcript data.

3. TRANSCRIPTION AS THEORY AND DATA

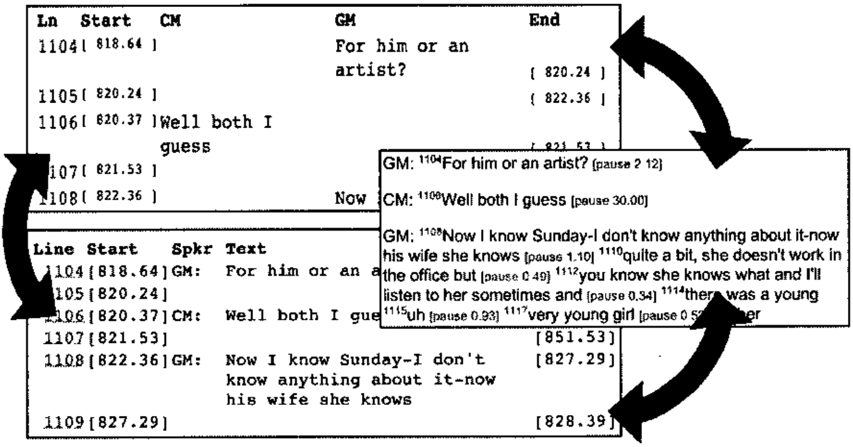
The subjective nature of transcription practice has been acknowledged in the sociolinguistic literature at least since Elinor Ochs' important (1979) paper "Transcription as theory". In her paper, and in papers that have followed (for example, Du Bois 2006, Du Bois, Schuetze-Coburn, Cumming, and Paolino 1993, Edwards 2001), scholars have worked to refine the requirements of a "basic" transcript and to identify hierarchies for the incorporation and coding of verbal and non-verbal information. These improvements are no doubt important and have led to better specificity and reliability in transcripts. However, for the most part, scholars have explored few other directions for the improvement of our transcripts. Little work has been done to move transcription away from static text-based (or document-based) representations, whether readable by human or computer.

NC SLAAP adopts the hypothesis that linguistic (naturalistic speech) data can be treated and stored as data, just as we would treat and store other types of data such as financial or customer information to use business comparisons. Similarly, NC SLAAP seeks to apply standard data management and presentation methodologies to the treatment of natural speech data. One major premise therein is the separation of content and format. Separating the data from its formatting provides a huge amount of flexibility and power, and as a result transcripts can be presented in any number of formats. For example, Figure 1 displays three different views of the same transcript data. In NC SLAAP, users can instantly switch between views.

Transcript data in NC SLAAP are stored in database tables. Each transcript is a table in the database, and each line is an entry in the database table representing a phonetic utterance by a speaker[2]. Transcripts for NC SLAAP are built using Praat[3], the open-source phonetics software, to obtain highly accurate start- and end-times for each utterance. Unlike the textual accuracy that many transcript theorists aim

[2] The determination of exactly what should constitute a transcript line is not straightforward. For NC SLAAP a line is based simply on an unbroken stretch of speech (silence-speech-silence). Other scholars (for example, Chafe 1993) focus on "intonation units" as the principal spoken utterance.

[3] Information about Praat is available at `www.fon.hum.uva.nl/praat/`.



```
Ln   Start    CM                  GM                End
1104[ 818.64 ]                    For him or an
                                  artist?           [ 820.24 ]
1105[ 820.24 ]                                      [ 822.36 ]
1106[ 820.37 ]Well both I
              guess                                 [ 821.53 ]
1107[ 821.53 ]
1108[ 822.36 ]                    Now
```

```
Line Start    Spkr Text
1104[818.64]GM:   For him or an a
1105[820.24]
1106[820.37]CM:   Well both I gue
1107[821.53]                                   [851.53]
1108[822.36]GM:   Now I know Sunday-I don't    [827.29]
                  know anything about it-now
                  his wife she knows
1109[827.29]                                   [828.39]
```

GM: ^{1104}For him or an artist? [pause 2 12]

CM: ^{1106}Well both I guess [pause 30.00]

GM: ^{1108}Now I know Sunday-I don't know anything about it-now his wife she knows [pause 1.10] ^{1110}quite a bit, she doesn't work in the office but [pause 0.49] ^{1112}you know she knows what and I'll listen to her sometimes and [pause 0.34] ^{1114}ther was a young ^{1115}uh [pause 0.93] ^{1117}very young girl [pause 0 5 er

FIGURE 1

Three screenshots illustrating different presentations of the same transcript data

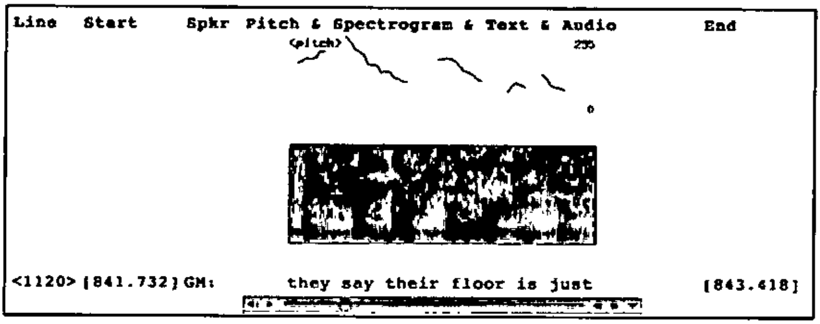
for (cf. Du Bois et al. 1993), NC SLAAP transcripts target temporal accuracy with the belief that everything else can be (re-)constructed from the audio file, either automatically by software, or manually by examining the audio for the given time range. With the start- and end-times for each utterance captured in the database and a linkage maintained with the audio, much of the other information that is often tagged or coded (for example, latching, overlap, pause length, etc.) is unnecessary.

TABLE 1

Core data elements for a data-based transcript

Speaker	Utterance Start Time	Utterance Textual Representation	Utterance End Time

In a data-based transcript model, the only data required, I propose, are those represented in Table 1. This very simple data model is actually quite powerful. Software, like NC SLAAP, can then create links between the transcript data and the audio file from which the transcript is based. Phonetic software (such as Praat) can then be integrated with the transcript to allow for real-time phonetic analysis. In other words, there is no need to code for loudness or pitch because these features can be reconstructed from the audio itself. At the same time, an approximation of standard orthography (following Chafe 1993) is sufficient for the transcript text because pronunciation features (for example, vowel qualities, *r*-vocalization, etc.) can be listened for or examined instantly via a spectrogram. Figure 2 shows a screenshot from NC SLAAP demonstrating an in-depth view of one transcript line. This example shows a pitch plot as well as a spectrogram, though other views are avail-



```
Line   Start       Spkr  Pitch & Spectrogram & Text & Audio          End
<1120> [841.732] GM:     they say their floor is just            [843.418]
```

FIGURE 2

Screenshot showing line analysis with phonetic data

able. The line can be listened to at the same time and numerical data (such as pitch measurements) can be obtained at the click of the mouse.

4. CONCLUDING REMARKS

The transcript model described here is intentionally simple and this discussion has focused on the minimum requirements for such a computer-enhanced model. Of course, it might be appropriate to include other elements such as non-verbal information (for video recordings, for instance). Nonetheless, this simple model has a number of strengths that are being made apparent through NC SLAAP.

NC SLAAP provides users with dynamic control over aspects of the transcript like formatting (whether to appear in a vertical or column-based format) and the levels of information displayed (whether to display spectrograms or other phonetic data in line with the text). It also seeks to bridge the gaps between corpus-based approaches, quantitative methods, and discourse analytic methods by providing tools for searching and querying transcripts for particular features. Overarchingly, it is hoped that the NC SLAAP software can illuminate new approaches for linguistic data management and transcription practices that can ultimately strengthen our overall linguistic research program.

REFERENCES

Chafe, W. 1993. Prosodic and functional units of language. In Edwards and Lampert, pp. 33–43.

Du Bois, J. 2006. Transcription and the delicacy hierarchy: What is to be represented? Paper given at Linguistic Society of America Annual Meeting, Albuquerque, NM.

Du Bois, J., S. Schuetze-Coburn, S. Cumming and D. Paolino. 1993. Outline of discourse transcription. In Edwards and Lampert, pp. 45–89.

Edwards, J. 2001. The transcription of discourse. In *The handbook of discourse analysis*, ed. D. Tannen, D. Schiffrin and H. Hamilton, 321–348. Oxford: Blackwell.

Edwards, J. and M. Lampert, ed. 1993. *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Lawrence Erlbaum.

Ochs, E. 1979. Transcription as theory. In *Developmental pragmatics*, ed. E. Ochs and B. Schieffelin, 43–72. New York: Academic Press.