

TRANSCRIPTION AS SPEECH-TO-TEXT DATA TRANSFORMATION.

by

BETSY L. BARRY

(Under the direction of William Kretzschmar)

ABSTRACT

This dissertation proposes a method of preparing natural language data for analysis, focusing on the transformation of spoken language into written form in a Language Variation framework. The protocol devised here is one that strives to maintain the integrity of the primary linguistic data under investigation, while providing a theoretically straightforward method of data preparation that places a reasonable time and resource constraint on the analyst. This project also seeks to establish practical guidelines for the transformation linguistic data from sound to transcript that will encourage and support data-sharing amongst the linguistic community and greater research community at large.

INDEX WORDS: Linguistics, Corpus Linguistics, Empirical Linguistics,  
Data Transformation, Transcription Theory, Dissertation (academic)

TRANSCRIPTION AS SPEECH-TO-TEXT DATA TRANSFORMATION.

by

BETSY L. BARRY

B.A., University of Mississippi, 1990

M.A., University of Memphis, 1998

A Dissertation Submitted to the Graduate Faculty  
of The University of Georgia in Partial Fulfillment  
of the  
Requirements for the Degree  
DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

© 2008

Betsy L. Barry

All Rights Reserved

TRANSCRIPTION AS SPEECH-TO-TEXT DATA TRANSFORMATION.

by

BETSY L. BARRY

Approved:

Major Professor: William Kretzschmar

Committee: Marlyse Baptista  
Diane Ranson  
Lewis C. Howe

Electronic Version Approved:

Maureen Grasso  
Dean of the Graduate School  
The University of Georgia  
December 2008

## DEDICATION

For Stella who inspired me to finish what I started and Kevin who showed me that sometimes the journey is more important than the destination. I love you both.

## ACKNOWLEDGMENTS

First and foremost I would like to thank Dr. Kretzschmar for his unwavering support and guidance over the past decade. Dr. K., you are an extraordinary mentor and I could not have done this without you. I am also grateful to my committee members, Diane Ranson, for her great insight and feedback and Chad Howe, for coming through for me in the 11th hour. A special thanks to Marlyse Baptista who has been a steady source of inspiration to me both personally and professionally over the years. Last but not least, I want to thank Clayton for indulging me with far too many conversations about data transformation and other equally fascinating topics, and for helping me wrangle  $\LaTeX$ . Loyal friend indeed.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS . . . . .	v
LIST OF FIGURES . . . . .	vii
LIST OF TABLES . . . . .	viii
CHAPTER	
1 INTRODUCTION . . . . .	1
2 LITERATURE REVIEW . . . . .	13
3 THE ATLANTA SURVEY PILOT . . . . .	43
4 THE INTERSECTION OF THEORY AND PRACTICE . . . . .	59
5 ANATOMY OF A TRANSCRIPT PART 1: METADATA HEADER . . . . .	82
5.1 METADATA . . . . .	83
6 ANATOMY OF A TRANSCRIPT PART 2: TEXT BODY . . . . .	126
7 CONCLUSION . . . . .	160
BIBLIOGRAPHY . . . . .	169
APPENDIX	
A APPENDIX 1: ORIGINAL ATLANTA PILOT TRANSCRIPT . . . . .	174
B APPENDIX 2: LV TRANSCRIPT . . . . .	205
C APPENDIX 3: LV DTD . . . . .	234

## LIST OF FIGURES

4.1 JW wave files . . . . .	59
-----------------------------	----



## LIST OF TABLES

3.1	Population Demographics. . . . .	45
6.1	Positive Contractions. . . . .	134
6.2	Negative Contractions. . . . .	135
6.3	Miscellaneous Contractions. . . . .	136
6.4	Informal Contractions. . . . .	137

## CHAPTER 1

### INTRODUCTION

Empirical linguistic research, or research based on actual language usage, is the cornerstone of Language Variation. In Variationist research, there are many fundamental components of designing and carrying out a successful study of spoken language, from sampling procedures to statistical analyses of linguistic features. Although field-wide attention has been given to different aspects of Variationist methods, the interval between data collection and analysis, the space where linguists routinely transform their language data from speech to a written medium and prepare it for various analytical endeavors, has been largely ignored on both a practical and theoretical level. In an effort to remedy the lack of attention given to data preparation practices in the field, this dissertation is focused on developing a protocol to assist Language Variationists in navigating the nebulous area of data transformation in their research. Although this task is a practical one, it is necessary to address data transformation processes at the conceptual level before tackling the subject in an applied framework. Therefore, at the beginning stage in this endeavor, I am going to start at square one by defining data transformation processes and discussing their role in the context of Language Variation research. Next, I am going to develop overall, field-wide objectives for these processes, followed with a discussion that focuses on how our theoretical assumptions about language intersect with our methods of preparing spoken language data for analytical and investigative endeavors. Clarifying and refining some basic premises in this manner will lay the foundation for the practical exercise that lies ahead.

The first thing I want to clarify is: What is Language Variation as a field of study? In this dissertation, I use the general term Language Variation to refer to a field of empirical linguistics that focuses on patterns of social variation and regional variation, that is, linguistic patterns of variation associated with speakers of different ages, genders, socioeconomic classes and geographical locale. Specifically, it encompasses both Dialectology, which has historically been associated with regional variation and change in spoken language, and Sociolinguistics, which has focused mainly on how socioeconomic class and other socially constructed categories influence linguistic patterns of variation.

The second important question that needs to be addressed is: What are data transformation practices and how do they fit within Variationist methodology? As with anything relating to language, the answer is not entirely simple. As Saussure discussed at length, spoken language as an object of study is a particularly challenging subject because in order to study it, linguists have to record it. In Saussure's time, this meant creating a written record, which he pointed out is a necessary, yet complicated task as it calls on using one distinct linguistic system (writing) to study another (speech) (see Saussure 1915/1959). In the one hundred years since Saussure wrote about the subject, complicated issues involved in creating written representations of spoken language have not changed. While our methods of recording have radically changed, the relationship between sound and text is still vastly complex. In Language Variation research, spoken language data usually consists of a recording of spoken language from an informant or informants engaging in a personal narrative or in an interview. Even when captured on audio tape the details of a speech event are lost almost as soon as they materialize. Facial expressions, gesticulations, the relationship between interlocutors, and other paralinguistic elements that fully contextualize a speech event are as transient as sound waves themselves. This is a fact of gathering data in the field. The audio recording of spoken language is already once removed from the primary data source, the spoken linguistic event itself. Once back in the laboratory, the empirical linguist takes further measures of preserving spoken language data and preparing it in a manner that lends itself

to investigation. The most prevalent approach to data preparation is to transform sound to text, by indexing an audio recording of a speech event to a text-based representation record of that event, a practice most commonly referred to as transcription. In Language Variation in particular, audio recordings of spoken language are routinely transformed into a transcript to expedite frequent referencing of linguistic data throughout the pre-analytical and analytical stages of research. With the creation of a transcript of a spoken linguistic event, the resulting body of transcribed language is now twice removed from its primary data source. The spoken language has literally been transformed from one medium to another on two counts, first by capturing it in an audio recording, and second by taking the audio recording and creating a graphical representation via transcription. Both instances are transformative processes because they involve alterations of spoken language; however, it is the latter data transformation process, the initial rendering of a graphical representation from a spoken linguistic event, or the creation of a transcript, that I want to focus on primarily in this dissertation.

Transcribing spoken data is a customary data transformation practice in Language Variation studies and transcripts play an important role in the analytical phases of our research. In the most general sense, a transcript is a graphic representation of a recorded speech event. On a more intricate level, transcripts are a complex link between fleeting sound and enduring text that creates a “road map” of a body of spoken data that enables a researcher to navigate and locate linguistic information without having to access sound files repeatedly in order to do so. Ideally, it is one aspect of a comprehensive body of information that will become the object of a linguistic investigation. As prevalent as transcribing spoken data is in Language Variation research, understanding the field-wide relationship between a transcript and the speech event that it represents is an essential aspect fully understanding this data transformation process. In the context of Language Variation in particular, a historical account of the role of transcripts in research helps elucidate the existing relationship between the

practice of transcribing spoken language data and the Variationist framework in which it is routinely employed.

Before Language Variation was a discernible field in empirical linguistics, empirical language research that focused on patterns of variation did so in the context of regional dialects. Linguistic Atlas research, rooted in Dialect Geography, dates back nearly a century and is a precursor to modern Language Variation research and an important contributor to modern Variationist methodology. In this historical context, transcripts were the first and only step in preparing data for linguistic analysis. In fact, they were the data. This model of research, in place long before audio recording technology, involved conducting lengthy, extensive interviews with a number of targets decided upon in advance, and subsequently recorded by hand if and when they occurred during the course of the interview. Interviews called for topic-directed conversations in order to insure that the most number of targets would occur during the course of the spoken linguistic event. Linguistic Atlas field workers like Guy Lowman were extensively trained to be able to quickly write down lexical, phonetic, and grammatical targets while not interrupting the flow of the exchange between participant and researcher. The transcripts they created, which consisted of notations of target linguistic features, were the only linguistic data available to them for subsequent analyses. The transcripts were all they had to go on. The historical model of “transcript as empirical language data” persisted until the introduction of the audio recorder in the 1960s, an event that roughly corresponded with the emergence of Sociolinguistics.

Once audio recordings became standard features of field research, the role of the transcript shifted somewhat. Transcripts were no longer the one and only method of creating and preserving a record of a spoken linguistic event, nor were they the only data informing the analytical stages of empirical research. As such, it stands to reason that the old model of a transcript would change, as our focus on observing linguistic phenomena was now enhanced by another principal data source: sound. Although entire conversations were captured in recordings, early audio recording technology was bulky, cumbersome, analog reel-to-reel

equipment that did not easily lend itself to continual and repeated referencing, especially at different intervals throughout the recording. A transcript was created to ease this burden on linguists as they were not able to go back and forth between one medium to another without a degree of difficulty. One big difference in the role of the transcript during this particular phase of empirical language research is that instead of having to adhere to a set of decided upon linguistic targets to focus on during the transcription process, researchers were able to transcribe entire spoken events, including as much linguistic information as possible by employing various systems of notation, highlighting various kinds of linguistic and paralinguistic information. As a result, the task of identifying potentially linguistically interesting features became an inextricable aspect of creating a transcript, and transcription systems were employed as a means of rendering entire recordings of spoken language into a linguistically “authentic” textual representation. Consequently, transcripts became surrogates for primary recorded data sources. The need for creating a transcribed version of a speech event did not diminish with the introduction of recording technology, but rather the reasons for creating a transcript and the role of the transcript in general changed considerably. The model of “transcript as empirical language data,” however, still prevailed in the field.

Fast forward several decades and recording technology has evolved significantly. Analog recording technology has given way to digital recording technology, and sound recordings created in the field are now housed on computers as audio files. In the past decade especially, it is now possible to easily access and reference an audio file of principal spoken data in ways that analog technology could not afford. Digital audio files can be accessed at any point during the recording, thus easily accommodating continued and repeated referencing. Researchers no longer have to spend time creating a transcript that showcases every linguistic feature of a spoken linguistic event, as the audio recording can now be easily referred to for this information.

At present, the role of the transcript in empirical linguistic research is poised to shift yet again due to the fact that linguists are now able to access primary spoken data recordings and

go seamlessly back and forth between one medium and another. As a field, we are on the verge of entering into another phase with respect to data transformation. In light of the evolution of recording technology and its impact on the nature of data transformation processes in the field, there is one important question that needs to be raised and properly addressed in Language Variation scholarship: Are transcripts still considered principal language data? In theory, the answer to this question is no. They do not have to be. In today's research environment, the audio recording is considered the principal language data and the transcript is a research aid used to supplement the aural spoken event with a graphical representation. It is an index between sound and text, not a substitute of text for sound. This is a direct result of the data transformation phase we find ourselves in today as an empirical linguistic research community, brought on by technological advances—arguably the most influential forces in Variationist research over the past decade. As a field, we have not properly considered the manner in which technology informs our methods of studying spoken language. Without question, digital technology and general advances in computer-based technology are having a substantial impact on how we handle and prepare our language data.

For a vast majority of researchers, creating the transcribed representation of their language data is still an exercise in creating *the* body of data to analyze, a body of text highlighting a multitude of linguistic features, despite the fact that technology is affording more methodologically sound alternatives to this practice. To complicate matters, technology-centered methodologies that are being incorporated into Variationist research are being done so in an ambiguous ways, as evident in Tagliamonte (2006), where the creation of a transcript is synonymous with the creation of an electronic-based corpus of analyzable data, a topic discussed at length in Chapter 2. Transcripts are being created not only as bodies of linguistic data for qualitatively-based human investigation, but also as prepared electronic texts to be subjected to various sorts of computer-based, automated analytical processes. In fact, these two data transformation processes are quite distinct, as they represent two very different data transformation processes. The first process is the initial transformation from

audio recording to text-based transcription. The second process is a transformation from text-based transcription to analyzable corpus. The former being a sound-to-text transformation and the latter being a text-to-text transformation. Again, these are two distinct processes; however, the confusion surrounding these different data transformation processes should not come as any surprise. The field of Language Variation has not properly addressed how recording and computer-based technology has impacted data transformation processes because data transformation processes have never been properly addressed in the context of Variationist methodology to begin with. In the past few years especially, we as a field have moved from one ill defined concept of data transformation to another, ill defined concepts that have encouraged inconsistent and imprecise practices with respect to the treatment of our language data. At the very least, Language Variationists need to move beyond the old model of having a transcript as the one and only referenced body of data during linguistic analyses. We need to take care that the audio recording of spoken language remains the focus of empirical investigations—not the written representation of spoken language. Linguists must not make the mistake of relying solely on a transcript to answer any and all inquiries and to guide all lines of linguistic investigation, despite the historical precedent that gives credence to this practice. Rather, Language Variationists should acknowledge and fully incorporate the current state of technology, which accommodates better data transformation practices in the field. Furthermore, data transformation processes should be firmly situated in Variationist methodology, clearly defined at the conceptual level, carefully considered at the theoretical level, beginning with an updated definition of a transcript and a clear set of objectives concerning the role of a transcript and subsequent data transformation processes in Variationist research. Only then can these data transformation processes be carried out in a principled manner at the practical level.

To begin with, it is time to reconsider the role of the transcript in Variationist methodology. In this initial data transformation process of creating a written text from sound, it is necessary to leave behind the idea of a transcript as a faithful and exhaustive linguistic



rendering of an audio recording. It places far too much responsibility on the transcriber to catalog important linguistic information during the act of transforming sound to text. This practice can have a negative impact on the transcription process, as trying to create a consistent representation is a difficult task in and of itself, but to add to this the onus of identifying and cataloging even a limited set of linguistic features introduces another layer of complexity to an already difficult and time consuming process. It can inhibit the creation of a text-based representation of the primary speech data as the transcriber's attention is divided between the transformation of speech into text and categorizing or tagging linguistic information. These are different processes and should be recognized as such. Transcripts as catalogs of linguistic information create a situation ripe for inconsistency as the urge to highlight as much linguistic detail as possible is not always compatible with clear, concise and most importantly, regular treatment of the data. When transcribers are charged with too many responsibilities, they will produce erratic results. This is always a drawback to human attempts at mechanical tasks. Consistency and accuracy aside, it also runs the risk of having a transcript so haphazardly marked with information that it becomes virtually unreadable to the human eye, producing something that only specialists, usually the researchers themselves, can read and make sense of. This can also compromise basic computer-based, automated methods of handling information as computers can only reliably process data that is consistently defined and identified across the board. It also renders the transcript useless for future research by other linguists or another audience altogether. It is impossible to create a transcript that presents every possible linguistic and paralinguistic detail. Optimally, a transcript should accommodate all empirical linguistic research. It should be a snapshot of language at a moment in time that anyone can examine for content as well as form.

The role of a transcript needs to be primarily archival in nature, having the potential to be both linguistically and culturally meaningful, and existing in a form that is amenable to both researchers and some limited computationally-based methods of processing. The idea that a transcript must do everything, from archiving to highlighting all potentially

interesting linguistic and paralinguistic information, is one that empirical linguists need to rethink. It encourages the practice of viewing the transcript as a surrogate for primary linguistic data during the pre-analysis and analysis stage of research, a throwback to the old model of transcript as principal language data. The more saturated a transcript becomes with specifically marked linguistic or paralinguistic details, the less a researcher feels the need to refer back to the primary linguistic source for reexamination to confirm or challenge initial observations. Technology now makes it possible to keep audio recordings as primary data sources, available for researchers to access, reference and evaluate. If audio recordings are viewed as the primary data source, the point of reference in language research, the role of the transcript shifts considerably. In this model, transcripts are simply graphical representations of a spoken language event, creating a visual layer to an otherwise aural experience. They are simply indexes used to complement the primary speech data, rather than surrogates for the primary data altogether. This model also eases the weighty task of creating a transcript of a speech overfull with linguistic features and observations, saving such practices for another data transformation process, rather than the initial data transformation stage. In this model, the transcript is a text based archive that offers a visual association with an audio based archive, the former created largely to give support to the latter.

When properly viewed as indexes serving in an archival capacity, transcripts become the initial data transformation process, not the only one. As such, the objectives for creating a transcript are at least twofold: On one end of the spectrum, transcripts must accommodate human eyes and be a useful source of orientation for referring back to the primary audio recording. Transcripts are created by human researchers for human researchers. Regardless of the fact that they exist entirely in a digital and electronic medium and support some types of machine automated processing capabilities, transcripts still must be suited for human readers and public consumption. On the other end of the spectrum, transcripts must be readied in a manner that does in fact take into consideration the electronic, computer-based environment in which they will reside. When a transcript is an archival text and an index

in a digital and electronic medium, created to accommodate researchers in other fields as well as empirical linguists, it is practical to assume that a transcript will be transformed into a body of analyzable text, via a secondary text-to-text transformation, that will be the subject of automated, computational processing at some point in the future. This fact should be considered at the outset of the creation of a transcribed body of spoken discourse as it has practical ramifications. Despite various levels of computational savvy on the part of researchers, it is reasonable to assume that the majority of all transcriptions are created and housed in a digital and electronic format in a computer-based environment. The degree of automated processing will range greatly for each body of research, but in anticipation of facilitating data sharing in a technological age, a transcript as an archival record should be readied for basic electronic processing, tailored for locating material and for discerning basic information about the body of spoken language data. As the universe of natural language data grows, so does the need for automated methods of locating and investigating collections of language data. Computers are better at processing large amounts of data than humans. At present, transcripts must be prepared to accommodate this reality by the use of practical, consistent and transparent methods of preparation, conditions that are an absolute necessity for successful machine readable material.

Another aspect of the updated model of a transcript proposed in this dissertation is establishing who the intended audience of a transcript is in the context of empirical research and how to accommodate this audience via this initial data transformation process. Even in very specialized areas of Language Variation research, the best approach for determining who to accommodate when preparing language data is to presume the broadest possible audience. Otherwise, a transcript runs the risk of being suitable for some types of research and some audiences at the exclusion of others, a practice which would also be contrary to creating an archival text suited for public consumption. If the transcribed data reflects a series of interviews from a particular region of a country for example, the archive may be of interest to local historians, families of the interview subjects, and other laymen researchers

whose inquiries and insights are as valuable as the veteran empirical linguist. In this respect, the transcript as archive is a part of a collection of language data not so unlike a collection at a library. A library is organized in a way that anyone can successfully navigate. It reflects a system of organization that allows both grade school students and the most seasoned intellectual access to information in a clear, predictable, and consistent manner. While the books themselves may reflect a wide range of subject matter, genres, and content, they are all systematically organized on the principle that everyone should have equal access to them. Similarly, transcripts as digital and electronically based archives should be organized on the theory that they cater to all audiences in order to accommodate a wide range of research and interests. With these objectives of data transformation in mind, as well as this proposed model of a transcript in place, it is now possible to go forth and tackle the task of developing general guidelines to facilitate the process of preparing spoken language data for analytical and investigative endeavors in the context of Language Variation methodology.

This dissertation culminates with a protocol for transforming an audio recording of primary linguistic data into a graphical representation, or transcript. It provides step by step instructions for creating a theoretically and methodologically straightforward sound-to-text data transformation in a Variationist framework, also taking into consideration data-sharing across various empirical fields. The destination, however, is only one small part of this dissertation. The journey of the decision making processes involved in general data preparation practices and the insight into the overall impact it has on our research is just as important as the resulting protocol. As such, this dissertation begins by reviewing the body of scholarship that addresses the practice of transcription and other methods of readying natural language for the analytical stages of research. As there is no cohesive body of literature that addresses data transformation processes, Chapter 2 draws on scholarship from various fields in empirical linguistics in order to give some needed context to the ideas presented here. After drawing on many different perspectives from different empirical fields, Chapter 3 offers a glimpse into a Linguistic Atlas pilot project conducted in a Language Variationist frame-

work, focusing on data collection, data transformation and basic data preparation practices. Moving from a broad survey to a first-hand account of data handling and preparation and the decision-making processes involved provides a useful point of reference for initiating the process of developing a both a theory and a method of transcription qua data transformation. Chapter 4 looks closely at the intersection of theory and practice in transforming sound into a transcript, further developing the basic concepts of data transformation and the role of the transcript introduced in the preceding pages. Specifically, this chapter begins to flesh out several important theoretical issues that impact practical methodology, thus requiring careful thought and deliberation at the outset of the data transformation process. Chapters 5 and 6 focus on practical applications by going step by step through the actual stages of transcribing and preparing data for both archival purposes and future text-to-text data transformation processes. Chapter 5 begins by dissecting the component parts of a transcribed body of spoken language, focusing on readying a transcript with contextualizing information, known as metadata. Chapter 6 tackles the body of the transcript proper, addressing issues that arise when transforming speech into writing and offering guidelines to negotiate these issues. This dissertation concludes with Chapter 7, which offers the complete protocol consisting of a set of guidelines and descriptions that clarify each respective guideline. In addition, a fully prepared transcript is included in the Appendices which demonstrates the data transformation protocol in its entirety.

## CHAPTER 2

### LITERATURE REVIEW

In the field of Language Variation, there is no cohesive body of methods-based scholarship that specifically addresses transforming and preparing spoken language data. In fact, as far as the treatment of language data in the post-data collecting, pre-analytical stages of linguistic research is concerned, there is very little Language Variation scholarship that has even placed data transformation processes, such as transcription for example, in the context of research methodology at all. Over the last several decades, the treatment of research methods in the field has focused mainly on the following areas: sampling procedures or locating and selecting research subjects, collecting data or fieldwork strategies, and types of analyses and interpretation (e.g., Milroy 1987, Milroy and Gordon 2003, Johnstone 2000, Tagliamonte 2007). There has also been some scholarship devoted to the technical aspect of capturing spoken linguistic events in the field (e.g. Ives, 1995, Everett 2004, Newman and Ratliff 2001). There is a conspicuous “gray area” in the literature, however, with respect to how Language Variationists approach data transformation processes, how these processes influence the analytical stages in research, and importantly, how these processes reflect and reinforce our assumptions about language. In order to fill in the gaps and piece together a composite picture of scholarship relevant to the ideas developed here, it is necessary to look to other empirical linguistic fields and historically significant works that traverse both theory and methods, and that offer valid points of orientation for going forth.

I am going to begin with scholarship that addresses the theoretical issues surrounding speech and the graphical representation of language with Ferdinand de Saussure’s *Course in General Linguistics*. In the first chapter of this historical work, Saussure discusses the

relationship between speech and the graphical representation of language, offering one of the most relevant observations with respect to the ideas developed in this dissertation. When talking about the study of language, Saussure states:

But we generally learn about languages only through writing. Even in studying our native language, we constantly make use of written texts— Writing, though unrelated to its inner system, is used continually to represent language. We cannot simply disregard it. We must be acquainted with its usefulness, shortcomings and dangers (23).

In this one paragraph Saussure identifies perhaps the most serious issue facing empiricists studying the spoken form of language, the fact that in order to do so we rely on the written form. Saussure also goes on to clarify that Language and writing represent two distinct systems of signs, the latter existing solely for representing the former. Importantly, as an object of study, linguists are interested in spoken forms of Language alone; however, the spoken word, he argues, is so inextricably linked to its graphical representation that the latter “usurps” the former (23–24). He offers several insightful reasons behind our tendency to privilege the written word over the spoken. First, the “preserved” aspect of the graphic form gives us a false sense of stability. This is compounded by notion that writing is much easier to grasp than sound. Along this line, he argues that people tend focus more on visual impressions as they are more lasting than aural ones. Saussure also points out that literary language and codified forms of language such as dictionaries and grammars, which is comprised of “rules” of usage, affords the written form a certain air of importance. Finally, Saussure states that when a disagreement between language and orthography arises, for the vast majority of people the written form wins out (24–25). Systems of writing, even systems which are phonetically-based, are characterized by a discrepancy between the written and the spoken form of language. He concludes that in reality, “writing obscures language” (30).

After this lengthy discussion on the “tyranny of writing” as he calls it, Saussure explores how to diminish the obstacles of empirical observation of the spoken form. He states that

linguists need a means of “transcribing articulated sounds that will rule out ambiguity” (33). He acknowledges that “countless systems” have been introduced but then he ponders the justification for substituting a phonologically-based system for one that is already in use. He then concedes such a system should be for the linguist only: “The advantages would not be sufficient to compensate for the inconveniences. Phonological exactitude is not very desirable outside of science” (34).

Saussure also has an interesting take on the importance of identifying a discernible unit of organization in Language before tackling the actual study of it. He argues that language is made up of real objects that must be identified and treated as the focus of synchronic Linguistic studies, or studies of the state of language as opposed to an era, which is diachronic or historical in nature. We must endeavor at the outset to delimit or define these real objects, or as he calls them “concrete entities” even though as practitioners of a language we scarcely even consciously perceive such units, although we do not doubt their existence. The concept of “real” objects or “concrete entities” must not be confused with “natural” objects or entities, as Saussure stresses that these objects must be identified and defined by the linguist. The fact that we define these objects for investigation means that they are arbitrary, not natural (106–108). Even though Saussure was addressing conventions of spoken language in a synchronic framework, because the creation of a transcript seeks to create a view of the state of an individual’s language at a moment in time and because a transcript also must have a discernible entity or unit of organization at the outset, his theoretical stance on concrete entities is relevant in the context of representing spoken language.

Saussure identifies the basic, organizing units of a language, concrete entities, as existing by way of associating an idea or “sign” with a cultural phenomenon or the “signified.” If practitioners of a language do not have both elements, the sign and the signifying phenomenon in which to assign total meaning, then we are dealing with mere abstractions or theoretical musings only, the stuff of psychology rather than linguistics (103–104). Concrete entities must be discernible, but they also must be able to form an association with abstract



ideas in order to convey meaning. Thus if a stream of speech is broken up into a succession of phones, then each phone is relegated to having phonological import only as they must be delimited as a concrete entity in order to sustain an idea in order to convey meaning. Saussure describes the entity or unit of a language as a string of sound “to the exclusion of everything that precedes and follows it in the spoken chain is the signifier of a certain concept” (104). Practitioners of a language are able to identify its organizational units by indexing concepts with utterances, and according to Saussure, this process actually involves visualizing a concept with sound-images, by the process of delimitation (104–105). When discerning a concrete entity, Saussure claims that although parsing units at the level of the phone is possible, it is not realistic as individual sounds do not always contain a corresponding sound to the unit meaning. On the other end of the spectrum, he states that trying to identify the sentence a basic organizational or concrete unit in Language invariably falls back on the word with its “grammatical characteristics” and thus the sentence cannot be considered a concrete entity (106). Despite the fact that there is not a tidy one to one correspondence between the Saussurean concrete entity and a word, the idea of a word as an organizational unit when representing spoken language in a transcript exhibits the same relationship as a sign (a word) and the act of signifying (indexing words with representative spoken forms). In spoken language as well as written language, words or combinations of words are the organizing units in which we drape meaning and/or function over. Although there are traditional units linguists seek to identify and represent in a stream of speech, such as phones, words, syntactic units, these units are firmly established in our academic linguistic repertoire. We refer to these units out of convenience, not because they are inherently the most desirable units of organization.

I now want to turn to a more recent treatment of the intersection of theory and practice by looking at the most cited article on the subject of theoretical issues surrounding transforming speech into a graphical representation. In Elinor Ochs’s seminal 1979 article, “Transcription as Theory,” the linguistic anthropologist and discourse analyst penned one of the first schol-

arly works that puts forth the idea that the act of transcribing natural language data, or “performance data” as she refers to it, into a text-based medium should be recognized as a wholly theoretical exercise in the context of language research. Specifically, Ochs contends that a researcher experiences “selective observation” during the transcription process which is a reflection of their own personal “theoretical goals and definitions” (44). Ochs goes on to make several insightful points with respect to the practical issues surrounding data transcription. She contends that selectivity is not only inevitable, but it is encouraged, as the more aware the transcriber is of his or her “filtering process,” the more likely they will strive to be clear in their decisions. Ochs asserts that by ignoring the transcription process, or by considering it to be a rote exercise devoid of thoughtful contribution, the outcome will be a text-based representation of a linguistic event that is of no value in a research capacity (44–45). Another benefit of selectivity she claims is that it will ensure that a transcript does not contain an overwhelming amount of linguistic information, as this can cause it to be “difficult to follow and assess,” again rendering a transcription useless for empirical research (44). Thus, striking a balance between inclusion and exclusion with respect to representing linguistic information is paramount in Ochs’s model of transforming sound to text. The priority of the transcription process is to be selective and carefully thought out, but it must also be theoretically transparent so that the underlying assumptions that drive the process can be easily distinguished within the realm of research goals and objectives.

The points Ochs raises in her pioneering article bolster an argument towards field-wide conventions in the area of transcription for psycholinguistics, specifically as it pertains to children’s language behavior. She is a proponent of developing and adhering to standards and guidelines in order to create more comparable research. In her words: “On a very practical level, we have not developed a set of conventions for representing the verbal and nonverbal actions of young children. We have no meta-language, no easy way to identify and compare actions and interactions” (44). The idea that a lack of shared conventions in the fields of empirical language studies hinders the ability to compare data, and thus the outcome of the

analyses performed on data, is one that echoes throughout Ochs's article and is one that is still relevant almost 30 years later.

Another influential article on the topic of transcription in the field of Psycholinguistics, published in 1999 by Daniel O'Connell and Sabine Kowal is "Transcription and the Issue of Standardization," which appeared in the *Journal of Psycholinguistic Research*. Aside from Ochs, O'Connell and Kowal's article is one the most cited on the topic. O'Connell and Kowal begin their look into transcription by placing the act of rendering sound into text as an exercise in preservation in order to transform the otherwise ephemeral occurrence of spoken language into an enduring, accessible form. The act of preserving spoken language by the creation of a transcript is motivated by the objectives of facilitating linguistic analysis, objectives which no doubt impact both the function and form of the transcript (105). As such, O'Connell and Kowal acknowledge the idea put forth by Ochs that act of transcription is a theoretical exercise rather than rote, a-theoretical undertaking. To illustrate this point, they begin their discussion on transcription practices by referring to a pioneering study by Maclay and Osgood (1959) in field of psycholinguistics that exposed the difficulty of process. Upon hiring secretaries as scribes to transform spoken discourse into text with the seemingly simple instructions to use normal English orthography and create as literal a transcription as possible, while avoiding any subjective normalizing of the data, the researchers learned that the transcription experience proved extremely difficult and tedious for the hired scribes. For example, when transcribing utterances such as "I dunno," scribes were torn between creating a literal translation and using standard orthographic conventions. After all, representing such an utterance as "I don't know" could be construed as normalizing the data. Additionally, the habit of leaving out certain vocalizations such as "uh" or "um" was problematic. In fact, despite the instructions given to the secretaries, the analytical nature and intense decision making process involved in creating a transcript completely trumped what seemed like straightforward instructions and made for a process riddled with difficulty and inconsistency (106). One interesting result of this study was that many empirical researchers concluded

that they should invest in their own data and undertake the task of transcription themselves, and rightfully so. Moreover, this study was also the first of its kind to demonstrate that there is no straightforward, simple way to transform sound recordings into text-based representations. It demonstrated that creating a transcript is a process characterized by tremendous decision making and critical thinking, and the outcome is a subjective representation that varies greatly from scribe to scribe. Again, this pioneering study resulted in convincing empirical researchers that they should invest in creating transcriptions themselves; however, it did not result in any kind of collective field-wide attempts to create general guidelines for the process of transforming sound to text (105–106).

O’Connell and Kowal refer to this ground-breaking study not only to highlight the complexity in data transformation processes, but to initiate an in-depth discussion on issues of standardization and guidelines when it comes to preparing data for analysis. The authors set out to survey a number of transcription systems from different areas of research in an attempt to explore how the notion of standardization with the context of encoding spoken discourse into a text-based form, with all of its various features both linguistic and extralinguistic, manifests in different fields that rely on empirical research. Just as Maclay and Osgood showed high variability between scribes, predictably, O’Connell and Kowal’s survey yields vast differences in how different fields react to transforming spoken language into a transcript: Some protocols specified a nuanced encoding scheme that relied on recording as much linguistic detail as possible, while others focused on extralinguistic features in order to contextualize as much information about an utterance as possible. All protocols, however, represented an arbitrary set of rules and measures that may have differed in form and function, but were developed in order to offer a complete solution to creating an accurate written representation of a spoken event. The point that O’Connell and Kowal deftly make is that no foundation or even loose guidelines for transcription stand a chance as long as data preparation practices, such as transcription, are not even considered in a methodological context, even in the very fields that rely on this kind of data transformation for their

research. The issue that has prevented any unifying ideas surrounding the transformation of sound to text, as O'Connell and Kowal point out, is that until transcription is deemed a significant methodological step in research, and until this intersects in a meaningful way with goals and objectives that drives its creation in the first place, no guidelines or means of standardizing will inspire wide acceptance in any field (112).

Another valuable point that O'Connell and Kowal make in their article is to stress the importance of not considering a transcript to be an original data base. As a derivative body of information, they contend that because a transcript is created to expedite an analysis, its function is only aligned with a particular investigative angle being explored in a particular research context (112). They contend that a transcription system can only be properly evaluated in the context of how it is to be used in a research endeavor. When you take this point into consideration, coupled with the reality that no transcription can replicate the entirety of the spoken linguistic experience, it appears as if standard guidelines are an unattainable goal in empirical research. O'Connell and Kowal argue that any transcription protocol comprised of strict encoding conventions risks undermining a scientific approach to the study of language. It places unproductive restrictions on language data (112). Moreover, such restrictive systems inhibit sharing resources amongst researchers.

Aside from Psycholinguistics, another field of study that has contributed significantly to the scholarship on transcription systems is Conversation Analysis. Conversation Analysis is often viewed as a methodological extension of Sociolinguistics, focusing on the social implications of spoken exchanges between interlocutors (Markee 2004, 3–5). While traditional Sociolinguistics focuses on patterns of variation in language usage attributable to extralinguistic social characteristics, Conversation Analysis does not focus on speech itself, as it is considered an artifact of the interaction between two folks communicating. In the Conversation Analysis primer, “Everyday Conversation,” Robert Nofsinger states that conversations are “social actions” towards one another. He goes on to say that linguists in this field must develop and hone their intuition and reasoning with respect to these actions in

order to “focus on what people *do* in conversation (their “moves”) rather than merely what they *say*” (emphasis his) (13). This field-wide objective necessitates an extremely detailed look at conversation in a qualitative framework. Furthermore, paralinguistic information and pragmatic information are essential components in this framework and must be studied as such (13–15).

The most significant contributions in the area of transcription in this field were made by Gail Jefferson and in fact the standard for transforming sound to text in the field is known as the Jefferson System. The theoretical assumptions that support Conversation Analysis rest on the social interaction aspect of human communication and rely not only on audio recordings of an interaction, but on video recordings as well. Thus, the data transformation systems employed by this area of study are unique. In the Conversation Analysis model, the paralinguistic and extralinguistic nuances of the social interaction during the course of an exchange are considered centerpieces of empirical investigation, connected to a larger social phenomenon that is human communication (Markee 2004, 5–7). As such, data transformation is more about representing nonverbal information than verbal information, a fact that is reflected in highly elaborate transcription systems that allow a scribe to add layers upon layers of paralinguistic and other contextualizing information to the graphical representation of a spoken exchange. A system of transcription such as Jefferson’s is geared toward quality rather than quantity, as it is a method of transforming a small stretch of conversation into a body of analyzable empirical data. A general reference guide to the Jefferson system is found in several methods-based texts (e.g. Nofsinger 1999, Jefferson 2004).

I now want turn to methodology-centered scholarship that deals with the treatment of spoken language data in a Language Variation framework. In the field of Language Variation, there have been a few seminal methods-related works published in the past decade that address data transformation processes to varying degrees. In Milroy and Gordon’s 2003 text *Sociolinguistics: Method and Interpretation*, one of the more comprehensive treatments

of language variation methodology, the authors devote one paragraph to the subject of transcription:

One methodological point of considerable importance in pilot work concerns transcription. Transcription of any kinds is invariably a selective process, reflecting underlying theoretical goals and assumptions (Ochs 1979: 44). It is therefore unwise at the pilot stage, when these goals and assumptions are still being formulated, to idealize too far from the data. Moreover, an over-abstract representation can conceal important information. When the objectives of the analysis are clearer, a selective transcription will be more useful than a detailed one, which at that later stage of the research is likely to contain much unwanted information (143).

Although Milroy and Gordon do not address transcription processes in detail, they do at least identify transcription as an important aspect of variationists' methodology. This paragraph manages to acknowledge the importance of transcription practices and it also raises some valuable points for the researcher's consideration, even if those points are not thoroughly elaborated elsewhere the text. The authors do caution about the selective nature of producing a text-based representation of language data. They suggest that early stages of data preparation should avoid creating what they refer to as an "overly abstract" transcript that can potentially obfuscate important details before research goals and assumptions are entirely formulated, thereby hindering a comprehensive analysis later on. Milroy and Gordon also suggest that creating a selective transcript, or one that is tailored to the particular research objectives, should be undertaken later on in the investigation when it is likely that these research objectives are more clearly articulated in the research.

While not explicitly stating it, Milroy and Gordon's paragraph on transcription hints that the authors consider the creation of text-based representations of spoken linguistic events as a two-tiered process. The first process occurs in the early research stages and consists of a general representation of the primary language data, reserving production of a second, more

selective representation that caters to specific research questions and goals for the analytical stages of research.

Another important Variation methods–centered text to address transcription is Barbara Johnstone’s 1999 text *Qualitative Methods in Sociolinguistics*. The author brings up several important matters a researcher must consider when creating a written representation of language. She highlights the decision making process as paramount in producing transcripts. For example, she points out the fact that conventions such as punctuation and sentential units are not products of natural speech, but rather standard orthography. Punctuation marks can “signal facts about syntax as often as they signal facts about what speech sounds like” (118). A period organizes a chunk of speech into a sentential unit we typically think of as a type of declarative statement, as opposed to a question mark that indicates a question. But as McLemore (1991) points out, speakers often have rising intonation at the end of a declarative sentence, which could be appropriately indicated by a question mark. Thus the question mark would serve to approximate an acoustic element in the discourse rather than adhere to an orthographic convention. This example illustrates a decision wholly typical of the transcription process, and one that often does not present itself until the transcription is underway. As Johnstone points out, it is very difficult to exhaustively predict the types of decisions that one will face when creating a transcript. Thus, Johnstone concludes her discussion of transcription by addressing several points that a transcriber should consider before deciding how to create a text–based representation of language data. First, one should address the theoretical biases inherent in the transcription system. It is the responsibility of the researcher to address potential biases head on, making them explicit and making an effort to recognize the manner in which biases potentially affect the outcome of the analysis (120–124). Secondly, one should consider carefully how the transcription system represents the speakers who contributed the data. How is the speech of these informants represented in the transcription and how does this reflect on the informants? For example, is a rural speaker’s dialect emphasized by making use of “eye dialect?” Preston (1996) defines “eye dialect” as



any orthographic liberty taken to approximate pronunciation of a nonstandard dialect. As both Johnstone and Preston point out, these kinds of liberties in spelling can only serve to color a transcript with condescension and disrespect and transcribers have a responsibility to their informants that warrants taking this into consideration. Johnstone's final point is that a transcription system has to fit its purpose within the context of the research endeavor (118–124). Transcripts are almost always created with a purpose in mind. A good transcription system has clearly defined parameters, explicitly laid out in a theoretical framework, as well a commitment to a systematic and consistent representation of the primary language data it is designed to represent (124). The idea that a transcribed body of data be created for a specific research purpose ensures that a transcription system is more transparent and less convoluted for both the researcher and the research consumer. This topic is explored at length in forthcoming chapters of this dissertation.

A recent text in Language Variation methodology, published in 2006 by linguist Sali Tagliamonte entitled *Analyzing Sociolinguistic Variation*, offers the most comprehensive account of Language Variation methodology in contemporary scholarship to date. While the majority of the text covers different aspects of linguistic research, including devising a sampling procedure, carrying out field methods, and developing various quantitative and qualitative procedures for the analytical stages of linguistic research, Tagliamonte does devote an entire chapter to language data, including transformation processes such as digitizing language corpora and transcription protocols. She takes a very hands-on approach in discussing data preparation issues in empirical linguistic research and a condensed and practical discussion of the various aspects of Language Variation research methods.

In her chapter entitled “Data, data, and more data,” Tagliamonte begins to fill in the gap between gathering data and analyzing data, the stages repeatedly referred to in this dissertation as data transformation. Her discussions stem from what she refers to as the “tried-and-true procedures” of her own experiences, using her own data to illustrate various techniques in preparing data that have worked for her over years of conducting research

in the field. The author's approach to handling and preparing language data is somewhat different from typical Language Variation methods in that she refers to her collection of data as a corpus, and her to protocols as centering on "corpus-building" techniques (50). Her terminology suggests that she takes her data handling and preparation cues from the field of Corpus Linguistics, which relies on principled protocols for creating collections of accessible, comparable texts prepared for automated processing to serve a multitude of analytical purposes. Likewise, the focus of Tagliamonte's approach to preparing language data in her Data chapter centers on creating a body of transcribed language data that is organized, linked and accessible in a variety of meaningful ways so as to be maximally functional in an automated processing capacity, ready for all manner of linguistic analyses.

It is useful to make a small clarification regarding the Tagliamonte text. When the author discusses the components of her "corpus" of linguistic data, she is talking about her research and how the concept relates to her own research (50). It is worth taking a moment to explain how corpus is used in the field of Corpus Linguistics versus Language Variation, as it is representative of a concept promoted throughout this dissertation. Generally speaking, corpus and corpora refer to a body of language information made up of a variety or sample of different materials. The material that comprises individual corpora is prepared according to some protocol or specified convention to create comparable texts. In creating a corpus of natural language, there is an underlying assumption that computer-assisted methods will facilitate collecting, archiving, and retrieving information from the body of language data (Biber et al 1998). There is also an underlying assumption that a language corpus should be a reusable linguistic resource, available to any researcher with a need for natural language data (Leech et al 1995). Thus computer-mediated processing and data accessibility and sharing are at the foundation of Corpus Linguistics, and likewise are implied when the term corpus and corpora are used in this dissertation. Tagliamonte does explicitly refer to the former, but the idea of data-sharing as an objective amongst the Language Variation community is not a theme explicitly discussed in her scholarship. Thus it is unclear as to whether this is a factor

in data preparation and transformation practices. On the contrary, other linguistic research endeavors, such as Kretzschmar et al 2006, do explicitly state a policy of public accessibility with respect to empirical data. Tagliamonte is, however, obviously an advocate of creating general methodological guidelines in the field, which is a running theme throughout this text.

In her *Data* chapter, Tagliamonte begins by looking at transcription practices, including raising the issue of clarity of purpose with respect to creating a body of transcribed language data to accommodate particular research goal and objectives. She encourages striking a balance between creating a representation that has enough linguistic information as to be useful in an investigative setting but straightforward enough to facilitate readability and not put an undue burden on the transcriber. Overall, her goal is to create a machine-readable corpus that adheres to some discernible transcription protocol and presumably, one that can be adopted and emulated by other Language Variationists (54–55). Tagliamonte describes a “first rule” of transcribing as creating an authentic representation of the data, and then offers an example of representing syntactic and morpho-syntactic variants that diverge from prescriptive grammar. She says that such variations should to be represented as authentically as possible and not normalized to adhere to the prescriptive standard.

Tagliamonte advocates using standard orthography and standard punctuation during the transcription process, and although she does not fully elaborate on either suggestion, she does demonstrate several instances from her own data to offer clarification. Likewise, she surveys a range of spoken language phenomena and then offers examples how each phenomenon would be represented with the use of standard orthographic and punctuation practices. Her examples include such occurrences as false starts and pauses, transcribed with use of hyphens and ellipsis, respectively (58). She also offers examples for coding inaudible utterances, speech sounds and extralinguistic sounds that contextualize an utterance such as laughing or knocking, as well as overlapping speech by multiple informants (58–59). Tagliamonte thoroughly surveys the typical range of speech phenomena that one will encounter while transcribing spoken discourse by presenting several useful examples that demonstrate the depth

of the decision making processes that characterize creating a transcript. She also advocates developing a good, working protocol for transforming sound to text, specifically, analyzable text suited for linguistic research. The author stresses the importance of keeping track of all of the conventions employed during the transcription process in an official capacity. In fact, Tagliamonte includes her own personal protocol in the appendix that can be used as a model.

Tagliamonte devotes several pages in her Data chapter addressing what she calls labeling, or making all of the aspects of data from the recording, to the transcript, and to the extralinguistic information that contextualizes the data, link together in a meaningful way. Her methods suggest creating categories with an eye towards supporting information retrieval and data manipulation that will aid in linguistic investigation. She spends the bulk of the discussion about preparing data on how to go about identifying and representing linguistic phenomena in the transcribed text-based representation of a speech event. As her overall goals of the book are to give linguists in the Language Variation tradition a comprehensive reference manual to guide them through the entire research process, her discussions of various data transformation issues revolve around presenting workable solutions to handling data. Throughout her explanations of handling data and preparing it for analyses, there are implicit references to theoretical issues surrounding the decision making processes, but Tagliamonte does not plainly state the relationship between issues of data handling and preparation processes to the theoretical assumptions that support them. Overall, Tagliamonte's treatment of language data initiates a discussion not often seen in the context of Variationist methodology. She highlights the fact that transforming data into analyzable text is a formidable exercise in decision making, and then goes on to offer clear and reasonable solutions to aid in the process. She advocates creating a protocol, an official record or log, that validates the various decisions made during the transcription process, as well as documenting the encoding schema overall.

At the end of Tagliamonte Data chapter she devotes a paragraph to addressing the practice of digitizing of data (67–68). As it stands, Tagliamonte’s mention of digitization is mainly in reference to the idea of the preservation of language data; however, she does not fully elaborate the role of transforming data into a digital medium or otherwise within the realm of archiving data and the impact that this currently has and will continue to have in Language Variation research. Her acknowledgment of the practice of digitization establishes a potential connection between creating a transcript and sharing data with other researchers in the field looking for comparable language data for their own linguistic research.

As far as contemporary Language Variation methodology is concerned, the Johnstone and Milroy and Gordon, and Tagliamonte texts represent the preeminent “methods focused” scholarship in the field. They are essentially the three primary reference texts in contemporary Language Variation studies that incorporate transcription practices, and data handling and preparation in general, into a methodological discussion. Aside from Tagliamonte’s chapter on data, absent from these Variationists’ methodological discussions is the idea that various means of preparing language data in a research environment are transformative processes, by way of converting spoken language data into another medium for the purpose of analysis you are fundamentally changing the language data. Thus, as far as field-wide scholarship is concerned, data transformation processes remain largely invisible. As far as other individual accounts or research projects go in current Language Variation research and scholarship, at best you can get a very limited idea of what a researcher’s data transformation practices look like if they include transcribed or text-based data in the analytical discussion as an example of some linguistic feature. To illustrate this point, consider the following example included as one of several in William Labov’s seminal work in the field, *Principles of Linguistic Change: Social Factors*, Vol. 2 (2001):

So Margaret said, “I’m not talking to you,” she said, “Celeste S., I’m talking to John Santorini– So uh– she comes down walkin’ with the angel. So I says,

“Listen, you want the angel? She says, “Yeah.” I said, “Well you stand and hold it for Christmas. ‘Cause there’s no angel gonna be put on my property” (390).

Nowhere in this canonical series does Labov address his methods for preparing his spoken data for analysis, or any transcription conventions in general, in any way that would give insight into processes used for identifying and encoding linguistic features or normalizing language data. One can assume that recorded data was transcribed according to some conventions, but because data preparation and transformation were not included in any methodological discussion it is reasonable to assume that this example of data mirrors the transcription of the primary linguistic data. This example illustrates the manner in which language data is typically presented in the context of Variationist scholarship. Labov’s linguistic studies are so widely referenced in the field of Language Variation that despite a lack of methodological transparency, or any explanation of data handling and preparation practices at all, they are considered paradigms in the field and they are widely emulated in Variationist research.

To risk sounding repetitive, in the context of Variationist scholarship, the potential issues surrounding data transformation have been all but neglected on a theoretical level and given only a cursory treatment at the practical level. Still, there are other fields and subfields in empirical language research that do incorporate data handling and preparation into a methodological framework and can inform a survey of the topic in current literature. Taking a cue from Tagliamonte 2007, one of the logical “first stops” when looking to other empirical linguistic fields for support in the area of preparing natural language data is Corpus Linguistics. One important distinction with respect to data handling and preparation in a Corpus Linguistics framework, as with other fields in Humanities Computing, is that researchers in this tradition rely on the consistent and rigorous treatment of data in order to facilitate automated means of processing. In other words, Corpus Linguistics operates on a basic assumption that language data must accommodate computer assisted methods of processing. Thus computer based methods are at the foundation of the field and any data transforma-

tion processes are considered an extension of this fact. Thus Corpus Linguistics concentrates on preparing textual data in a computer environment to contribute to large repositories of natural language created for various kinds of empirical linguistic research.

While most corpus data preparation begins with a text based representation, there are in fact a few language corpora in existence that are transcriptions of spoken discourse (e.g. London–Lund Corpus, British Academic Spoken English (BASE) Corpus, Michigan Corpus of Academic Spoken English (MICASE), Switchboard Corpus). The area of Corpus Linguistics does not necessarily advocate a particular transformation protocol for rendering sound to text, but it does require a consistent and principled treatment of language data, including transcription, in order to create comparable bodies of information that are representative of the genre from which they originate (Biber et al 1998, Kennedy 1999). Scholars who create natural language corpora must spend a good deal of time thinking about design and organization of data, which entails principled decision making about preparing data, as well as meticulous record keeping detailing the decision making process. It is a widely held assumption in Corpus Linguistics that corpora are repositories of language data that are meant to be shared, accessible to any empirical researcher that is interested in natural language text. This fact, coupled with the underlying objective of preparing language data for automated processing, lends itself to a certain amount of rigor when dealing with natural language texts. Although Language Variation incorporates computer–mediated methods and automated processing into its research in varying degrees, it is not a driving force in the field like it is in more technology–centered fields such as Corpus Linguistics; however, as Tagliamonte’s text suggested, the role of technology and computer–assisted methods of analyzing language data are now very influential forces in Language Variation methodology. As such, Language Variationists must look to fields like Corpus Linguistics in order to gather valuable insight necessary for handling and preparing data within this context, as Tagliamonte implies, in light of the continued prominence that computers and computer–mediated processing is playing in linguistic research.

In *Corpus, Concordance, Collocation: Describing English Language*, John Sinclair states that creating a corpus of analyzable data must take into account issues of archiving, processing, and basic information retrieval. As such, contextualizing information must be recorded in order to create a comprehensive bibliographic record of the language data. Additionally, the actual language text must be subjected to some standard conventions in a widely recognizable format to facilitate automated processing, which relies on consistency and accessibility. He also recommends that codes meant for computer and human mediated processing be identified and classified in an official protocol or record of some sort, as documentation is key in creating transparent methods (20–21). Sinclair goes on to recommend keeping text as “clean,” or as free from extraneous encoding as possible, for reasons that need to be at least considered by researchers in order to strengthen their relationship with their language data. First, every research endeavor, every linguistic investigation is going to prioritize different types of information. Thus it is advantageous to create a corpus of natural language that is accommodating rather than exclusive (20–21). Secondly, as a corpus is a body of data that should not only accommodate different types of research, but different research from different fields each with their own sets of priorities and theoretical frameworks, one cannot assume in organizing a body of analyzable data that all assumptions that drive linguistics, such as grammatical categorization for example, are shared assumptions that cut across fields that rely on empirical language data for research (21). Keeping text as free of extensive amounts of encoded material, from narrowly defined to very general, ensures that the potential of a body of language data is maximally useful in a variety of research settings, not simply one particular, and possibly quite narrow, research endeavor.

The idea of creating clean text is a particularly important one in Sinclair’s model of language transformation. As a transcript is a text-based rendering of an audio recording, and one half of a preserved record of a spoken linguistic event, Sinclair points out there is methodological precedence for creating an initial text-based representation free from the analytical processes involved in encoding linguistic and extralinguistic information, typical of



transcription in Language Variation research. Theoretically, this is also a desirable position. Sinclair states: The computer is incurably and pedantically data oriented. For a linguist, there are no short cuts to his or her abstractions. The first stage of adaptation is to resign oneself to plodding through the detail. Attention to detail makes one rapidly aware of oversimplification, and some of the accepted generalizations seem to be in need of a tighter specification (29).

The careful consideration and the attention to detail making decisions about preparing language data can be obfuscated when a researcher focuses only on “accurately” representing or highlighting a certain linguistic feature, which is no doubt based on some preconceived assumption, or “abstraction” as Sinclair puts it, about Language rather than the reality of preparing language data, especially spoken language data (22). Again, in the most general sense, this suggests a principled approach to creating a text-based representation of spoken language that should be a cornerstone in all empirical linguistic fields that rely on sound-to-text transformation.

Aside from Corpus Linguistics, there are other fields in the Humanities Computing arena that also offer useful perspectives when dealing with natural language data in this context. In *Spoken English on Computer* (1995), a group of linguists and technologists came together in an edited collection of essays that offers a body of scholarship that specifically addresses the intersection of computer technology and data transformation practices. The collection traverses the range of transformation practices, from recording language data, to transcribing it in an electronic format, to enriching a transcript with linguistic information, to organizing data for specific areas of research. The authors situate all of these practices in a historical context as well as placing them in a contemporary research setting as all parts of a comprehensive, collaborative effort aimed at creating and maintaining spoken language corpora for preservation and investigation (3–7).

The introduction of the text begins by outlining the type of research needs met by developing spoken corpora. These needs range from researchers who need guidelines for means

of transforming spoken data for personal research endeavors to include basic preservation and information retrieval in an electronic format, to linguists who intend to develop applications for specialized areas of research, such as language recognition technology (1–2). As all levels of empirical research can benefit from computer–assisted methods, a fact that becomes even more apparent as technology evolves and becomes more widely available, so researchers would do well to work towards embracing computer–based methodology with respect to their data handling and preparation practices. Perhaps the most important aspect of this collection of essays is that it manages to cover all of the various aspects of data transformation processes, the procedures and preparation that occurs after data have been gathered in the field, leading up to the actual analytical methods of a research endeavor. It addresses the theoretical motivations for readying language data for automated processing, and thus starts at square one in its treatment of how language data intersects with computational technology on its journey to become an analyzable body of information. While this text does not specifically address how field specific areas of empirical research, such as Language Variation for example, incorporate principles of creating and maintaining spoken corpora in an electronic environment, there are a couple of essays in particular that are good points of reference for fields that are currently poised to meet the challenges of incorporating more rigorous data transformation procedures into its methodological framework, while taking into consideration computer–mediated approaches to preserving and analyzing natural language data.

The text itself is divided into three parts, the first is a collection of essays dealing with the theoretical aspects of creating of transcribed text and how this intersects with incorporating computer–based methods into language research. The second part of the book features essays that focus on creating applications for more specialized language research and the third part surveys a number of existing transcription systems and mark–up protocols. Generally speaking, the text offers an technology–oriented view of language transformation and preparation that moves from the theoretical to the practical, while appealing to a non–specific

audience of language researchers. The first section of the text, the theoretical issues section, is perhaps the most relevant section to this dissertation. Thus, all of the essays deserve review and consideration in their own right. The first essay in the text, “Principles and alternative systems in the transcription, coding and mark-up of spoken discourse,” is authored by Jane Edwards and addresses transcription practices within the context of transforming sound to text in an effort to ready data for automated processing. Edwards begins by making some useful distinction in transformation practices that are worth noting here. First, the author defines transcription as the creation of a text-based representation of a spoken linguistic event that records “who says what, in what manner (e.g. prosody, pause, voice quality), to whom and under what circumstances (e.g. setting, activity, participant characteristics and relationships to one another)” (20). She also differentiates between broad and narrow transcriptions. She then distinguishes the practice of “coding” text as focusing on organizing text according to some abstract categorization, such as syntactic parsing for example (20). Finally, she defines mark-up as the process by which structural units of organization are imposed on a text, and which in turn are used to symbolize information (20–21).

Once Edwards defines transcription, coding and mark-up, she goes on to discuss the underlying principles involved in encoding systems. As a result, she develops a theory of these transformation processes, including principles of category design, readability, and computational tractability (21). As far as category design is concerned, this can be summed up as approaching the decision making process in identifying “dimensions of interest” with extreme rigor (22). As such, in creating categories that will be imposed onto language data you must be able to discern in the data the instance that requires a system of categorization and the applicable category. Categorization must be consistently applied. Edwards insists that in creating categories, it is necessary to “exploit the contrastiveness” of each, in order to avoid confusion in the boundaries of prescribed categories (22–23). Next, Edwards discusses the issue of readability, which she interprets primarily as machine readability. Finally, transformation processes must take into consideration the intersection of computational processing

and practices of representing language data in order to expedite automation. For example, she mentions the need to curb what she refers to as “non-meaningful variability,” an example being representing abbreviations via acronyms and spelling them out over the course of a text. In short, computers rely on consistent, predictable means of representation and all of her theoretical stances supporting transformation processes reflect this fact (25).

One final point that Edwards brings up is relevant to the idea of sharing resources and standard methods, an important theme throughout this dissertation. In creating corpora of spoken language data to be accessible to any researcher, it is important to work towards creating comparable data across language corpora. When working with bodies of transcribed text for example, coding elements must be easily interpretable in the event that they will undergo further transformations in order to converge toward a standard. In order for this to happen an “exhaustive” list of coding elements, or a detailed protocol, must be produced (22–24). Edward gets her point across that one should not underestimate the importance of documenting the analytical procedures of all data transformation processes. It will enable other researchers to decide whether or not your data is comparable to their own, or another accessible corpus of language data, as well go on to make further data transformations for analytical purposes (31–32). Again, Edwards expresses a standard that would benefit any empirical linguistic field, especially Language Variation.

Another relevant article in *Spoken English on Computer* is authored by Guy Cook and entitled “Theoretical Issues: Transcribing the Untranscribable.” In this essay the author advocates a critical assessment at the relationship between an original audio recording of a speech event and its transcribed counterpart with respect to computer based methodology and language research (35). He begins by arguing that there are two aspects to automated processing concerning transcriptions, the use of computer mediated processes to retrieve and present transcribed language data, and the use of computers to aid in quantitative analysis of transcribed language data. The two distinctions are related to the idea of human readability and machine readability, distinctions that are characterized by their own set of issues, the cen-

tral one being that human readability requires qualitatively meaningful information, which is in turn encoded to facilitate machine readability, which is then used to make quantitative assessments about spoken discourse (35–36). These issues regularly collide when deciding on how to proceed in transcribing spoken language, a fleeting aural phenomenon that does not lend itself to writing easily, in a manner that provides the consistency needed for deriving exact measurements. For example, a transcriber might decide to highlight an interruption as a linguistic feature, but concedes that there are two types of interruptions, one that is meant as obtrusive speech and is offensive to an interlocutor and one that is simply beginning to speak out of turn, not mean or interpreted as intrusive, and in turn decides to differentiate between the two types. The distinction that is made between obtrusive and non-obtrusive are intuitive and speculative assessments about a linguistic phenomenon, and hence not necessarily amenable to an exact measure. Cook concedes that this is not necessarily a reason to forego identifying such things, but rather to understand the underlying assumptions that accompany these types of decisions (36–37).

Cook also discusses the intersection between human readability and machine readability by examining the notion of words and text as phenomena that have currency in both the linguistic and transcription realm equally. While some may interpret words as linguistic phenomena that link the aural with graphology and text with the collection of words that comprise a body of transcribed speech, Cook argues that words have a psychological existence apart from speech and writing, making them the subject of as much interpretation as discerning between an obtrusive interruption versus a non obtrusive one. This assertion logically leads to the idea that “all transcription is interpretation” or that “the text is already mark-up,” which creates a theoretical gulf in the practice of transforming sound into a text-based form (38). The key, Cook insists is acknowledging the difference between these two forms of language, that creating a text based representation of a speech event does not make that writing more like its spoken counterpart than other types of written language, but rather that the distinctions lie in the context of the knowledge of the participants that

is necessary in speech to add the psychological layer to an interaction, the paralinguistic and the situation that surrounds communication amongst interlocutors (41–42). In other words, in Cook’s model of Language, the extralinguistic characteristics of the interlocutors are essential in approaching speech as a linguistic phenomenon. Cook then goes on to present possible solutions in transcribing paralinguistic, in which he focuses on creating two encoding categories that correspond to “situation,” “participant knowledge,” and “participant attitudes.” He advocates establishing boundaries in order to not be bogged down in the “infinity of detail” that recording such information can offer. He also readily acknowledges that there is no theoretically rigorous basis for the selection of elements of situation, for example, but that transcribers do as they have always done in the past and make ad hoc decisions felt to be important to fully characterize a speech event. He refers to it as a matter of expediency, rather than a theoretically defensible decision (48–49). In short, Cook asserts, a researcher must recognize the paradox in the task of transcribing spoken discourse that it is an exercise in transcribing that which does not lend itself to transcription, to straddle to scientific and the nonscientific as it were, and to avoid the pretense that transcribed speech is somehow an accurate depiction of the aural phenomenon in which it represents (51–52).

In the third article, “Adequacy, user–friendliness, and practicality in transcribing,” Wallace Chafe argues that as transcriptions must be accompanied with sound files, as language researchers must guard against “accepting a transcription as given” (61). As he is coming from the position of Corpus Linguistics, where language corpora are created and shared amongst the research community at large, he is a proponent of creating accessible sound and text, so that researchers can have access to the primary data form, not just the transformed and transcribed body of data. He states that taking any given transcription and moving forward with multiple, various analyses spares the researcher from engaging in the time consuming practice of transforming sound to text, but also the researcher loses out on the valuable insight gained from have a relationship with the data from the beginning, which

in most cases will include accessing the sound file to create a more comprehensive picture of the language data (61).

The next several chapters of the book's first section begin to move away from focusing on transcription and into the actual preparation of a body of language data via encoding guidelines for electronic text analysis. James Monaghan's "Whole-text analyses in computerized spoken discourse," addresses some issues concerning representing spoken English with respect to the design of the corpus, or creating a corpus that accommodates various analytical endeavors, and encoding features in a machine readable capacity. He argues that spoken language corpora should be created and maintained in a manner than lends itself to adhering to standard methods and guidelines, should one have the desire to transform an entire corpus and make it available according to a particular standard (63–65).

John Sinclair focuses on the intersection of theory and practice with his essay entitled "From theory to practice." As Sinclair is one of the most prominent Corpus Linguists and one of the first to produce a computer-based corpus of spoken language, his insight into the connection between technology and preparing spoken language for research is invaluable. In this essay, Sinclair points to the importance of pursuing the relationship between the sound wave and the written record (99–100). He draws attention to economy of time and creating maximally functional bodies of spoken data and stresses the importance of guidelines to navigate the creation of extensive bodies of natural language poised to meet a variety of research needs. Sinclair also stresses the importance of both human readability of corpora as well as machine readable data, both of which are accompanied by their own set of issues and concerns. As a means to both of these two objectives, Sinclair advocates the use of standardized methods of data preparation to facilitate readability on both fronts, adding that any standard must remain user friendly, not be burdensome in both time and economy, and avoiding "interfering" with the text as much as possible (108–109). This last point sets the precedent for his idea of a "clean" text-based body of transcribed language data developed in his later scholarship.

The fifth and sixth chapters of *Spoken English on the Computer* introduce an idea that is central to this dissertation, that of encoding standards and guidelines. The fifth chapter by Lou Burnard is introduction to one such standard entitled “The Text Encoding Initiative: an overview.” The Text–Encoding Initiative, hereafter TEI, is one of the more prolific set of encoding guidelines available to both academia and industry, which has developed standards for electronic text mark–up. At the time of the publication of this article, TEI was in its infancy but had already published a list of recommendations for encoding language corpora that would facilitate the exchange of bodies of data, or data–sharing, by way of incorporating like mark–up to produce like corpora for a machine readable environment (69). TEI focuses on two main areas of encoding electronic texts: First, it looks at what information should be encoded and secondly, it addresses how the encoding should be represented in a text (71). Over a decade after this essay was written, the same basic tenets frame TEI’s guidelines. TEI has since published five volumes outlining its encoding protocol. Today, TEI is a world–wide, interdisciplinary set of extensive encoding guidelines used by a range of institutions to accommodate all kinds of text, from literary to linguistic in nature. It offers a multitude of annotative choices for marking up all manner information in a body of text. Currently, TEI is one of the most prolific, comprehensive standards of data preparation in areas of the Humanities that utilize text–based computational processing. TEI guided text–encoding must incorporate several required elements in order to render an electronic text TEI compliant. The required elements pertain to designating metadata information in a header pertaining to a description of the body of text as a whole, including bibliographic and editing information as well as contextualizing extralinguistic information pertaining to the text, and the incorporation base elements for encoding the body proper of electronic data, including a set of hierarchical elements designed to standardize the organization of a text. Historically, TEI was one of the first initiatives to propose the standard treatment of metadata, prefixing it to the body of a text in the form of a header, a protocol now preferred in most encoding protocols (see <http://www.tei-c.org/Guidelines/index.xml>).



On the topic of standards and guidelines for electronic text–encoding, there are a few more prominent ones that are relevant to the ideas developed in this dissertation. The Encoded Archive Description, hereafter EAD, began at the University of California, Berkeley library in 1993 as a venture into developing nonproprietary protocols for automated methods of searching and retrieving archival texts. From its inception, the project organizers recognized the need of accessing information beyond what was included in conventional machine–readable cataloging, or MARC records.<sup>1</sup> The requirements for the encoding protocol include the ability to present extensive and interrelated descriptive information found in archival finding aids, thus it mainly pertains to metadata. The protocol also had to be able to preserve the hierarchical relationships existing between levels of description, to represent descriptive information that is inherited by one hierarchical level from another, and to move within a hierarchical informational structure. And importantly, the protocol had to support element–specific indexing and retrieval. Today, EAD standards focus on bibliographic information to facilitate finding and retrieving archived electronic texts. In other words, the standard hinges on metadata information. EAD is the standard that the Library of Congress adheres to and it has become a protocol highly recognized and regarded in the field of library science (see <http://www.loc.gov/ead/eaddev.html>).

Another highly regarded set of standards in text encoding circles is put forth by the Open Language Archives Community, or OLAC, is an international community of scholars dedicated to cataloging language resources for the purpose of making linguistic data accessible to the linguistic research community. The OLAC guidelines are primarily metadata oriented. The metadata set itself is based on the Dublin Core metadata set, a metadata initiative that prescribes an encoding scheme designed for locating resources in an inter and intranet environment. The objectives of this metadata initiative are simply to provide a protocol that facilitates locating resources, sharing resources and managing information. The Dublin Core

---

<sup>1</sup>MARC records are the most widely recognized standard for creating machine readable bibliographic records for texts. MARC is used by the Library of Congress (see <http://www.loc.gov/marc/umb/>).

traces its roots to Chicago at the 2nd International World Wide Web Conference, October 1994, when a group of technical experts from a range of fields began discussing the difficulty of locating resources on the Web. Like EAD, Dublin Core and subsequently OLAC are rooted in archival practices of location and retrieval. Thus, its set of standards is applicable mainly to metadata, rather than offering an exhaustive list of encoding methods for a multitude of text types, such as TEI (<http://www.language-archives.org/>). All of these standards, TEI, EAD, and OLAC will be revisited at length in Chapter 5.

While each of these sets of standards and guidelines are systems that offer specific ways of dealing with preparing the body of data that comprises electronic texts, many of the most stringent aspects of each deal with encoding metadata. As metadata is a subject that is specifically connected with preparing bodies of text in an electronic and digital environment, it is a topic that is most likely foreign to many empirical linguists but one that should be relevant to any researcher creating and storing language data in a computer based environment. Perhaps the most comprehensive treatment of metadata standards and the topic in general, is Susan Lazinger's 2001 text "Digital Preservation and Metadata." This piece of scholarship offers one of the most comprehensive texts on the subject to date. Lazinger's text addresses the historical importance, theoretical significance and practical considerations of metadata as it pertains to electronic text encoding and digital information. Lazinger specifically addresses the role of computer-mediated methodology in accumulating and disseminating information and the importance metadata plays in this capacity. She considers electronic texts, from transcribed speech to literature, as archival by their very nature and thus in need of proper accompanying documentation that will contextualize these preserved texts (8–10). She goes on to give relevant categories of metadata as they apply to all electronic texts, which in turns sheds light on encoding protocols such as TEI and OLAC, and how these sets of standards incorporate the idea of including contextualizing information in order to describe any kind of electronic data (142–143).

Before ending this survey into the scholarship of empirical linguistic methods, I would like to revisit a research project briefly mentioned in the preceding paragraphs that deserves some additional attention within the context of a methodological discussion, especially as it pertains to Language Variation. The Linguistic Atlas of the United States and Canada comprises a body of research projects that have spanned the past century and represent a massive collection of natural language data of spoken North American English. Due to the sheer size of spoken language data produced for these projects, the Linguistic Atlas has relied on principled methods for handling and managing these data, which have spawned countless studies and scholarship in the field of Language Variation, Dialectology in particular. Furthermore, under the direction of William Kretzschmar the Linguistic Atlas projects have lead the way in digital archiving endeavors of large corpora of comparable language data, employing computer-mediated applications for analyzing and manipulating large quantities of spoken language data on the Web. Today, these projects continue to incorporate traditional Language Variation methodology into a modern computational environment with a commitment to facilitating scientifically sound research practices. Data transformation processes are at the center of Linguistic Atlas research. Even so, the complexities of transforming sound to analyzable linguistic text in the context of a large-scale research endeavor were the impetus for the ideas developed in this dissertation. In the next chapter I will describe in detail a pilot research project conducted for the Linguistic Atlas in the fall of 2003. The Atlanta Survey pilot, was carried out according to a rigorous sampling procedure, organized data gathering endeavors, as well as carefully thought out transcription protocols, all developed and executed to prepare the spoken data in a manner comparable to spoken data gathered and prepared for other regional Atlas projects. As a body of research scholarship, the Atlas projects have wholly informed the ideas in this dissertation; however it is the Atlanta pilot project in particular that is the foundation, the impetus for the theoretical discussion and subsequent practices developed here.

## CHAPTER 3

### THE ATLANTA SURVEY PILOT

The Atlanta Survey pilot was a Linguistic Atlas research project conducted under the direction of William Kretzschmar.<sup>1</sup> The goals of this pilot project were threefold: First, we planned to survey the speech of residents of Atlanta Georgia. Secondly, we were testing new methods of data collection and preparation for the Atlas projects. Thirdly, we were committed to offering all of the data, to include full text transcripts and sound files, on the Linguistic Atlas website (<http://www.lap.uga.edu>) for review and study by the research community at large. Essentially, developing modern research methodology and sharing language resources were as important in this pilot as the actual investigation of the speech of individuals in Atlanta, Georgia.

Issues of economy and practicality, or reasonable resource expenditure while collecting a useful body of linguistic data, were central aspects of the interview design for the Atlanta Survey project. One of the biggest problems in linguistic survey research is to find the right balance between the nature and extent of the interview, and the costs of carrying out the linguistic survey. The methods originally used for the American Linguistic Atlas surveys, questionnaires of nearly 1000 target items which required six to eight hours for the field worker and informant to complete, were very costly. The editorial labor of processing field records into lists for analysis was also costly and difficult, even when the editorial staff was recording only single words or short phrases from field notes. After tape recorders came into use, which made it possible to carry out full transcriptions of the interviews, such transcriptions were not made because of the enormous amount of time required: 5200 hours

---

<sup>1</sup>This pilot project was funded by an NSF grant number SBR-0233448.

of taped interviews were collected for the Linguistic Atlas of the Gulf States (LAGS; Pederson 1986–92) alone, which would have required approximately 52000 hours for creation of full transcripts—26 years. The interview format for the Western States devised by Lee Pederson (Pederson 1996) restricted the number of elicitation targets and cut the interview time to three hours, but even this reduction requires nearly a week of skilled labor to create a full transcription of approximately 300 pages. The alternative at the other end of the spectrum, the telephone interview, is much more rapid and cost effective, but it has its own problems. Guy Bailey pioneered the use of the telephone for a planned linguistic survey by adding a few questions to the Texas Poll, an existing survey instrument for opinion polling in the state (Bailey and Bernstein 1989; see p. 8 for discussion of problems). Subsequent use of the telephone for linguistic survey research, notably in Labov’s Telsur Project, has refined the method, but it remains the case that sound transmission over telephone connections reduces the frequency response of recordings (see Labov 2001: 156–157; this is an important issue when acoustical measurements are part of the research protocol, as they are for Telsur). Such interviews typically limit the types and extent of elicitation targets, and telephone interviews are unavoidably unnatural as the researcher trades reduced travel expenses for the distortions introduced by a remote conversation. Another cost effective survey method, postal questionnaires, has been used effectively by Chambers in his Canadian Golden Horseshoe study (Chambers 1998). This method, however, relies on reported speech, and so the data that can be gathered by it are somewhat limited.

Taking these data-gathering techniques into consideration, the Atlanta Survey set out to test a cost-effective interview format for linguistic survey research: the one-hour interview. The interview was designed to yield five different kinds of linguistic information: 1) acoustical phonetic measurements in the manner preferred by speech scientists, fixed-format elicitation; 2) acoustical phonetic measurements from free conversation, such as those typically collected by sociolinguists; 3) grammatical information from free conversation, such as that typically collected by sociolinguists; 4) lexical tokens, a subset of those collected in Western

States Linguistic Atlas interviews; and 5) perceptual linguistic information, such as that collected by Preston and others. Interviews were conducted in person by Linguistic Atlas field workers Iyabo Osiapem and the author. The first forty minutes of the interview consisted of conversation regarding the local way of life, at which time the field worker attempted to draw out information from the speaker about family and household topics like those in the first portion of the Western States Linguistic Atlas interview. The next ten minutes of the interview consisted of fixed-format elicitation of stressed vowels. The last ten minutes consisted of a series of questions designed to determine the speaker's perceptions of local speech.

Fulton County and DeKalb County are the area where field work was conducted. They are the two most populous counties of the Atlanta metropolitan area, and had a total population of approximately 1.5 million people according to the 2000 census (<http://www.census.gov>). The following table offers the 2000 population, the percentage of African Americans, and the number of Linguistic Atlas interviews that were executed previously in the county:

Table 3.1: Population Demographics.

County	2000 Population	%Af-Am	LAMSAS-LAGS interviews
DeKalb	665,865	54.2	6
Fulton	816,006	44.6	16
Totals	1,481,871	48.9	22

Parallel samples of ten African American speakers and ten non African-American speakers were planned; however, only eight non African-American speakers located and interviewed. For analysis purposes, the two counties were considered as one geographic unit, and the sample was drawn without attempting to balance Fulton against DeKalb. Overall, the sample was to achieve a sampling ratio of approximately 1:75000, but the reality of the sampling ratio was closer to 1:85000.

In order to qualify for participation, all the subjects must have had English as their primary language, must have been adults (age 18 or older, because of the contract and

payment for the interview), and, while ideally lifelong residents, must have lived in the area for at least half of their lives. These qualifications meant that we were actually sampling a target population less than the full population of the region. We excluded children, the most transient residents, and residents for whom English was not the primary language. These exclusions were necessary concessions in order to accommodate the contemporary mobility of the population and to the fact of multilingualism in the community.

Ideally, the sample called for 20 interviewees. Half of the interviewees were male and half female. We did not attempt any finer notion of gender than biological sex, although we readily admit that a more sensitive classification might yield important information. Subjects were grouped into binary classifications for occupation based on the concept of the “linguistic marketplace.” According to this idea, jobs are ranked based on the necessity for using the standard variety at work. Census categories where it is important for employees to use Standard English include executive, administrative, managerial, professional specialties, technicians and related support, sales, and administrative support. Occupations that require less use of the standard are precision production, craft and repair, machine operators, assemblers, inspectors, transportation, material moving, handlers, equipment cleaners, helpers, laborers, and service. Fifty-six percent of all employed persons in the state of Georgia fall into the Standard English categories and forty-four percent work in the other categories. Speaker biographies routinely revealed specific occupations, for potential analysis in more detail. We used the terms “white collar” and “blue collar” as convenient designations for these occupations, but we understand that the ranking of jobs implies different levels of education, economic status, and life views. Since our study was based on 20 speakers, we simply could not expect to create too many potential categories for analysis for fear of overburdening our sample. We were, therefore, controlling for three binary variables, race, sex, and occupation. However, it is possible to conduct analysis with age as a scalar variable, even though it is not one of the quota criteria. The social characteristics chosen as aspects of

the quota sample were done so as to allow comparisons with other historical sociolinguistic and dialectology surveys.

Households were selected randomly. Subjects were compensated for their time, typically about one hour. Interviewees were solicited from a phone list purchased from the University of Georgia's Institute for Behavioral Research. The phone list contained numbers of approximately 3000 households in Dekalb county and Fulton county. Fieldworkers randomly called the list of numbers of Atlanta residents and followed a scripted questionnaire in order to ascertain whether or not someone at the household was interested in participating in the research and whether or not someone in the house hold fit our research criteria and quota sample. The variables will be ranked in the following order: race, occupation, and sex. If the planned quota for one race has been filled, we will continue the process of randomly selecting locations until a person of the other race can be located. The same is true of the other variables: we will continue the process of randomly selecting locations until a person who can satisfy the quota requirements can be located. This procedure corresponds to the procedure of sampling "with replacement," which is appropriate for any population for which no complete list of the members is available.

Each speaker selected for the sample provided detailed biographical information as part of the interview, including places of residence, family history, education, specific occupation, and social contacts. It may thus be possible to carry out analysis of other variables, or more detailed analysis of planned variables, than the sampling methodology explicitly allowed. Our sampling methods are designed to satisfy quotas by race, sex, and occupational status so that we may execute valid statistical tests; the methods do not thereby exclude additional analysis, but there may be skewed subsamples for additional variables. If we attempted to guarantee the distribution of speakers according to additional variables, or with more than a binary distribution for our planned variables, we would have been required to interview a great many more speakers for the sample, and thus put at risk our ability to carry out the study.



Each speaker signed a contract which explained that the interview was to be used for research purposes, that the entire interview will be published on the Internet and potentially in other media (except for information that could identify the speaker or information which needs to remain personal to the speaker), and that the confidentiality of the speaker will be protected to the extent permitted by law. In exchange for signing the contract and completing the interview, the speaker was paid fifty dollars.

All in all, 18 interviews were conducted by the author and Iyabo Osiapem. As mentioned, the goal was 20 interviews, however, finding participants that filled the sampling quota proved difficult as the project neared completion and the quota fields were more narrowly defined. In general, the “cold call” approach to soliciting interview participants became increasingly difficult towards the final interviews. In the end, the ability to locate two male, non African–American blue collar workers proved too demanding a task given our time constraints and research considerations for this pilot study.

The first section of the interview plan focused in part on collecting the speaker’s demographic information. Interviewee’s were asked to report their ethnicity, occupation and other personal information such as place of birth and religious affiliation and basic social network information. All identifying personal information was available only to the interviewer, the project director and those collaborating on preparing the language data. Family information, such as parents’ occupations and birthplaces, information about siblings and extended family, was also included in the demographic portion of the interview. After a nominal amount of personal and family information was collected, the interview moved into discussing general household items such as furniture, food items, and the like. The interviewer had a number of household–related topics in which to guide the conversation in order to produce certain lexical targets and pronunciation targets, which corresponded to linguistic targets of past Atlas studies (see Kurath and McDavid 1961, Pederson 1986–1992). This being the most substantial part of the interview, ideally producing around 30 to 40 minutes of speech, it

was designed ideally to elicit a maximum amount of language data from the interviewee based on a minimum of prompts from the interviewer.

The second part of the interview, referred to as the fixed format elicitation, was designed to elicit stressed vowel forms for acoustical phonetic measurement. This portion of the interview typically lasted around 10 minutes. Thirty cues (words) were written on index cards and shown to the interviewee and in turn the interviewee was to pronounce each cue. The index cards were then shuffled and the task was repeated two more times for a total of three readings of each cue.

The last part of the interview was a discussion of the interviewee's perceptions of local speech. Interviewees were encouraged to comment on whether they thought there were differences in speech according to any of the variables of the sampling plan, such as African-American versus non African-American, blue collar versus white collar, etc. All in all, the Atlanta Survey Project resulted in a fair and complete test of the one-hour interview format, based on its application in an urban setting with speakers of different social characteristics. The social characteristics chosen as aspects of the quota sample were done so to allow comparisons with other historical sociolinguistic surveys. All interviews were recorded on a digital recorder onto a compact-disc using a cda sound file format. The interviews were then rerecorded directly onto a standard PC using standard sound editing software. Interviews were saved into .wav files for both archiving and editing purposes. Once sound files were stored on a local PC the transcription process got underway.

The procedure for transforming the Atlanta data from sound into text was as follows: First, each one-hour interview was allotted approximately 10 hours for transcription, plus an equal amount of time for proof-reading. The encoding schema was divided into three parts: metadata, DTD, and text. Metadata appeared at the top of the file and described speaker variables, particulars about the interview, and particulars about the project itself. Metadata elements followed the emerging Dublin Core and OLAC international standards, incorporated the following:

- I:** interviewer's prompt to label a turn, in the format I: (name) ... \I. The name attribute is optional, if only one interviewer conducted the interview.
- R:** response by respondent to label a turn, in the format R: ... \R
- A:** response by auxiliary speaker to label a turn, in the format A: (name) ... \A The name attribute is optional, if only one auxiliary speaker was heard.
- V:** vocabulary item. Must be labeled with a descriptor from the work sheets, or with a separate gloss, in the format V:(E9) ... \V, V:(ghost) ... \V. An empty attribute in the format V:( ) ... \V, may be used to identify interesting forms without a descriptor or gloss, e.g. V:( ) brother \V.
- G:** grammatical item. must be labeled with a descriptor from the work sheets, or with a separate gloss, in the format G:(E9) ... \ G, G:(verbal particle) ... \G ... An empty attribute in the format G:( ) ... \G, may be used to identify interesting forms without a descriptor or gloss, e.g. G:( ) fixing to \G.
- P:** pronunciation item. must be labeled with a descriptor from the work sheets, or with a separate gloss, or both, with the = sign preceding any gloss, in the format P:(A12) wife \P, P:(A12=waIf) wife \P, P: (=br6) brother \P. Initially, we will use broad IPA for glosses, in the OED low-ASCII transliteration ... An empty attribute in the format P:( ) ... \P, may be used to identify interesting forms without a descriptor or gloss, e.g. P:( ) brother \P. Phonetic glosses need not be provided for every pronunciation item, only items of interest, because we intend to provide acoustical phonetic plots for pronunciation items.
- Q:** editorial query/question, in the format Q:( ) ... \Q. Queries should be used by coders whenever there is a question about the status of the transcript or coding. Each query should describe the nature of the question within ( ), with the complete section of the

text or code being questioned to follow before the ending code. Query codes will all have to be removed before publication.

The following codes do not require an ending code:

**T:** time stamp (refers to elapsed time from the beginning of the recording, not to any particular span of time within the interview), in the format T:(mm.ss). Time stamps should be entered at the beginning of the speakers' conversational turn, and optionally more often as appropriate.

**U:** non-word utterance or feature. must be labeled with one of the following descriptors: (c) cough, (f) false start, (h) hesitation, (i) interruption, (l) laughter, (p) pause, (e) external stimulus. May optionally use an = sign plus an explanation, e.g, U:(p=2seconds),U:(e=door slams).

**D:** deletion of an expected feature. must be labeled with a descriptor for what has been deleted, optionally with an = sign plus what has been deleted, e.g. D:(copula), D:(pronoun=it)

**X:** unintelligible section of the tape. must be labeled with a descriptor for what is unintelligible, such as a word or period of time, such as X:(word), X:(10s).

The Document Type Definition (hereafter DTD) portion of the encoding schema was the actual list of codes used to mark up the body of the transcript. The following list below was used as the beginning set, but the option for adding codes as circumstances demand was and is available. The DTD was intended as a separate file alongside the transcription file containing the metadata and transcription proper. Most core codes were labeled with beginning and ending tags in order to facilitate converting them to XML later on in the data transformation process.<sup>2</sup> Basic codes are meant to translate into XML tags, which consist of an element and which may or may not include an attribute. Elements are the part of the tag

---

<sup>2</sup>There are several excellent XML reference texts. The one consulted in Linguistic Atlas research is Rusty Harold's text *The XML Bible* 2nd ed.

that acts like a keyword and is meant to semantically define the information being encoded in the text. Attributes are secondary pieces of identifying information that complements an element, more precisely defining the text that is being encoded. Thus, the transcription DTD was considered a blueprint for one that was to be produced for an XML document or electronic text. The following excerpt demonstrates the list of codes used in the transcription process:

**Title** Atlanta Survey Pilot

**Creator** interviewer(s), transcriber(s)

**Subject–Keywords** language variation; speech science; perceptual variation; Atlanta speech

**Description** 60–minute interview in three parts: conversation, direct elicitation, perceptual elicitation; interviewer’s or transcriber’s description of “quality” or characteristics of the interview

**Publisher** William A. Kretzschmar Jr.–Sonja Lanehart

**Contributor** Speaker name, address, phone; also identify any other voices of auxiliary speakers heard on the tape

**Date** date of interview, dates of transcription

**Resource Type** transcribed audio recording

**Format** reference to current encoding document

**Resource ID** a seven–character code, in which the first two characters are digits 01–20 for a serial ordering of interviews; one character for ethnicity, A (African American) or N (non AfAm); one character for occupational level, W (white collar) or B (blue collar), one character for sex, F (female) or M (male); and two characters representing the age of the speaker in years at the time of the interview. E.g., 01AWF42.

**Source** Linguistic Atlas Project

**Language** English

**Relation** interview related to early section of Western States work sheets; NSF BCS–0233448 “SGER: Atlanta Speech Sample”)

**Coverage** Atlanta, Fulton and DeKalb counties

**Rights** copyright UGA, Linguistic Atlas Project

Standard orthography was the norm, using “dictionary spelling” (not literary dialect spelling—the test is whether the spelling can be found in a dictionary or the LAGS indexes). “Sentence units,” which are not necessarily grammatical sentences but word groups thought by the transcriber to cohere in some way, were to be marked with a capital letter at the beginning, and a period and one space at the end. Other use of caps and periods were limited, and periods were not to be followed by spaces unless at the end of a sentence unit. Proper names were to be capitalized, but no caps or periods were to be used with titles like mr or mrs or dr, and other abbreviations were to be avoided as much as possible. The ? symbol was used to mark questions. Colons and semicolons were not to be used and comma use was to be limited as much as possible. Other punctuation or symbols such as ’ “ \* @ # % & ( ) ! ~ [ ] \ / | < > were not to be used because they were likely to be confused with symbols in use in XML or other text encoding. Readability was facilitated by following a common practice in formatting: the T tag was to be put on its own line; a “hard return” was to be used after the end tags of a conversational turn and one blank line was to be left between conversational turns; the start tags for speakers were to be put at the left margin, and one Tab after the colon following a code was to be used before starting to type in the transcribed language data. The transcription process was carried out by Linguistic Atlas research assistants varying degrees of linguistic experiences. The bulk of the transcriptions were carried out by Nikki Kong, a former student of linguistics and Speech and Language

Pathology graduate student. During the course of the encoding process, Dr. Kretzschmar and I were on hand to answer and questions that arose but the majority the decision making process was left up to the discretion of the transcriber with the stipulation that consistency be of paramount importance after a decision was reached.

Before discussing the transcription process, it is useful to understand some technical aspects of the digital sound recordings. Each interview sound file was approximately 700 MB due to the fact that a high-quality format was preferred for optimal sound presentation. This is significant for two reasons: First, because this file format preserves the natural language data in the most preferable manner as it privileges quality; however, it does produce large files that can be unwieldy during the editing process as they take longer to load into memory and subsequently manipulate. Secondly, in terms of making sound files available on the Web, it is not realistic to offer large, high-quality .wav files, due to global and local networking feasibility with respect to access/download time, except perhaps in T-2, 3 or 4 environments or exceptional computing environments atypical of a general research laboratory. Thus, if sound is to be made accessible on the Web it must be either through a succession of small .wav files or through compressed sound files such as MP3s. The .wav format is more desirable as it guarantees the inclusion of as much of the nuanced phonetic data as possible; however, for those who do not wish to use the language data in this capacity, MP3s can be made readily available.

A primary objective in making the Atlanta speech corpus accessible to the research community was to facilitate the types of data-sharing endeavors that would stimulate advances in both presentation and analysis of research data, essentially furthering our understanding of the complexities of patterns of variation in American English. This pioneering effort in data-sharing has yet to be fully realized due to the time and resource-intensive nature of the undertaking. In addition to matters of resource expenditures, many issues arose in our attempts to prepare interviews fully and fairly for analysis. It was always our intention to apply basic methods of transcription, encoding and presentation, which would become

a model for future applications of interactive linguistic analysis of spoken corpora. The encoding issues that cropped up are ones that came to light once we became immersed in actually transforming the language data from sound into text. In other words, the nature of these problems could not necessarily have been anticipated until a certain familiarity with the language data was reached; however, it is also assumed that these issues are not unique to this particular research.

The transcribers were provided with the same worksheet used by interviewers that detailed numerous linguistic targets that could be expected to surface in interviews. Additionally, examples of linguistic targets were provided to further aid in correctly identifying and labeling pronunciation, grammatical and lexical targets. Despite this, confusion arose on a regular basis about how to go about encoding a particular feature. Encoding confusion often arose when an unexpected variant appeared in an interview, or when some linguistic information could potentially be more than one target. For example, in an interview the informant produced a variant of the lexical target “grandmother” by referring to someone’s “mother’s mother.” At first glance, this information might be recognized as a realization of the pronunciation target “mother” and only labeled as such. This type of issue is fairly common, especially in language variation research and it is one that is somewhat easy to resolve by multiple passes through the data. Plus, the transcriber always had recourse to another opinion when questioning whether or not a feature was in fact a token of a particular type. Here are portions of a two different transcripts to illustrate the point. A full transcript from the Atlanta Pilot is available in Appendix 1: Original Atlanta Pilot Transcript:

R: Oh absolutely. Oh, I go to-I can spend a week in Greenwood and come back to work and they’ll probably look at me and say, "WHAT? What are you talking about?" cause it, cause, you know they already, they already know at work that I have (clear throat) I have probably a more southern draw than anybody in the office because most of the people in the office are co-ed-- they went to college so they’re, they are college educated and uh, and, and then we uh, uh have pretty much of a younger crowd in out office some of the older people um, but like our owner went to Kentucky and he’d, he was, uh, he’s from New Orleans but you cannot even tell there’s no way (clear throat).



I: Cause New Orleans-

R: New Orleans is, is very distinct.

I: What did it sound like to you? What does New Orleans speech sound like?

R: Oh my gosh it's a , it's a cross between uh southern, uh southern white, black, and French and-- (laugh) and God knows else was thrown in there but uh, it's, it's uh, uh, half the time I can't-- a real New Orleans person I can't understand a word they're saying uh or if you watch a movie on tv and, and it's you know feelin' the New Orleans and your-- it's suppose to be New Orleans people and can't understand them and, and I don't know where they-- it's just such a mixture they whole Cajun thing but I think it's fascinating. I just love to hear them talk.

In addition to the set list of linguistic targets, transcribers were also told to label any “unique” feature encountered in the data. The DTD portion of the transcription protocol was flexible in that it made allowances for encoding linguistic targets not included in the interview protocol itself but that emerged during the course of transcribing things of potential interest that might warrant some investigation down the road. The “uniqueness” of a particular linguistic feature was left up to the individual transcribing the data, or the researcher during an additional pass through the data. A seemingly simple task, this allowance in the protocol proved to be very problematic because it placed too much responsibility on the transcriber. Basically, the idea of coming up with a set transcription method is to try and reduce the guesswork when encoding data. When dealing with pronunciation in particular, you risk the idea of “unique” or “interesting” being equated to “anything that is markedly different from my own speech,” “my own” referring to the speech of the transcriber. This simply is not a reliable method of encoding data as it allows for too much inconsistency across transcriptions. Consider the following example from the Atlanta data:

R: Um, and they actually live out of L.A. now. So we were able to spend some time with them when we lived just north of L.A.

In the first pass, the primary transcriber did not tag this chunk of discourse for any unique feature; however, upon another review by a different research assistant, the phrase “live out of” was tagged as a unique expression for stating where an individual lived. So, what was not necessarily “unique” to the primary transcriber seemed fairly salient and worth highlighting to another, thus illustrating the dilemma surrounding what constitutes uniqueness and the extremely subjective nature of a linguistic feature that stands out as such. It is not uncommon for transcribers to diverge from a standard protocol during the course of the transcription process as interesting or unique information emerges. This practice, however, can drastically undercut consistency and uniformity, which are important elements when basing analyses on text-based transcripts. The process of producing a text-based body of data that can reliably represent a speech event is a time-intensive, difficult process at best and while the transcriber will always be faced with decisions it is best to try and make the decisions as objective as possible.

Addressing these issues and others, along with a commitment to high standards in handling linguistic data, bore out a principled examination of the practices in the area of data preparation. The fact that sampling methodology and interview protocols were so meticulously planned out and executed, coupled with the methods of analysis that were planned for the linguistic data, this carefully outlined enterprise presented a good opportunity to notice the more vulnerable aspects of the entire research endeavor. The careful attention to methods underscored the complicated nature of the data transformation processes, even though a solid working protocol had been devised to address these aspects of research. The issues experienced in this research project were the same kinds that characterize empirical linguistic research in general and Language Variation in particular: Data transformation processes, such as the creation of a transcript and the subsequent creation of an electronic text poised for analysis, have simply not received the attention that other aspects of empirical methods have, such as sampling for example. Transcription is rarely treated as the convergence of theory and practice, which it is. The decision making process that defines

transforming sound to text has not been thoroughly addressed in either a theoretical or a methodological context and so a standing protocol is little more than a laundry list of suggestions that may or may not give a scribe the guidance they need in order to work through some of the complexities that they will be faced with during the transcription process. A set of guidelines must be preceded by theoretical deliberation, including the development of a working definition of a transcript and the examination of the assumptions that we make about language that ultimately shape any practical means for transforming data from sound to text. This is the crucial first step in the data transformation processes developed here.

In the next chapter I want to discuss the practice of taking spoken language data and transforming it into a graphical representation from a theoretical perspective. In essence, I want to look at where theory and practice intersect with respect to data transformation. Taking the lessons learned from the Atlanta Pilot study, I want to thoroughly consider data transformation processes at the theoretical level, in a Variationist framework, and in a manner that informs the practical endeavor of developing a set of guidelines for transforming speech into a transcript.

## CHAPTER 4

### THE INTERSECTION OF THEORY AND PRACTICE

As this dissertation seeks to develop general guidelines for data transformation and data preparation practices in a Language Variationist framework, it is about methods; however, the methods we employ as Language Variationists are wholly informed by our theoretical positions about Language. There is no such thing as atheoretical methodology. Thus in the context of the ideas developed here, we must take into consideration not only practical issues surrounding spoken language data, but also the theoretical issues that inform them. With this, I want to begin by stating a fact about audio recordings of spoken discourse, a fact that Saussure wrote about, which is central to our concept of transcription as an initial data transformation process: There are no natural units in a stream of speech. To demonstrate, consider the following images:

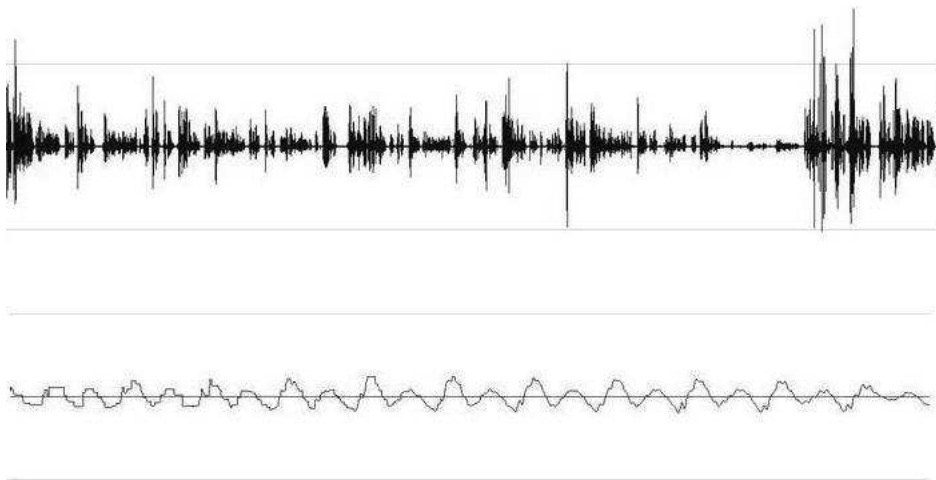


Figure 4.1: JW wave files

Both of these figures represent a snippet of a stream of speech from one of the informants in the Atlanta Survey. The first image represents less than two minutes of steady discourse and the second one represents around one tenth of a second of speech. If you look carefully at these screen shots one thing becomes clear: There are no inherent divisions, innate components, or units of organization in these speech signals that correspond with any graphical method of representation. While there may be perceptible visual elements that imply a sound is occurring, such as a spike in the wave versus an absence of the wave entirely, there is no intrinsic property of these wave forms that translates into discernible, meaningful units of language, especially at a text-based level of representation. As such, there is no “natural” way to create a transcript. This is a simple fact about transforming speech as a purely aural phenomenon into organized, readable units on a page or a computer screen. Likewise, there are no “accurate” means to render a stream of speech into some graphical representation that exists on a page, as no graphics can adequately overcome the problem of scale in representing the level of detail characteristic of spoken language. There are physical properties of a speech signal that anchor it to the medium in which it exists, which in turn renders the notion of accuracy in transforming that speech signal into something else, like a transcript, an unattainable goal. Accuracy is in the eye, and ear, of the beholder and not a purely objective, quantifiable or measurable property of a transcription of a stream of speech.

In the absence of any natural or accurate method of graphically representing a speech signal, we must turn to practical solutions for doing so. Keeping this in mind, I want to now consider several examples of transcribed speech, briefly discussing the qualities of each with respect to their relationship to the primary spoken language from which they were derived. The first two are referenced elsewhere in this dissertation.

So Margaret said, “I’m not talking to you,” she said, “Celeste S., I’m talking to John Santorini— So uh— she comes down walkin’ with the angel. So I says, “Listen, you want the angel? She says, “Yeah.” I said, “Well you stand and hold

it for Christmas. ‘Cause there’s no angel gonna be put on my property” (Labov 2001 389–390).

This example is one from Labov’s *Principles of Linguistic Change: Social Factors*, referenced in Chapter 2. As previously discussed, this example of sound-to-text transformation presents a very particular representation of spoken discourse. It indicates a personal narrative on the part of the speaker, as such it uses appropriate narrative style, as well as quotative punctuation conventions. It incorporates alternative spellings for some word forms. The example of transcribed speech that Labov includes is meant to illustrate a distinctive vernacular in its representation of an informant’s speech, as it incorporates various means both inside and outside the boundaries of standard orthography in which to do so. The use of alternative spellings, the narrative style and punctuation, seem to reinforce the impression that Celeste S. is distinctive and “nonstandard” in her manner as well as her linguistic patterns of language usage. The use of these conventions highlights these particular aspects of Celeste’s speech, demonstrating a specific manner of representation.

Now, I want to look at a very different example of transcribed speech from one of the informants of the Atlanta Pilot Study. This snippet of transcribed speech is a representation of the first speech signal from the audio wave file image featured in this chapter<sup>1</sup>:

its really funny he started out in theological school and started out as a preacher and then for some unknown reason to all of us he decided he didnt want to do that anymore and never talked about it again um (clear throat) started work here at riches and worked for riches for oh i dont know maybe ten or twelve years and he worked in their carpet department at downtown riches that was the only riches there was and uh they had a really big carpet department and he worked there

In this example, there is no punctuation, use of capitalization, or any other conventions of standard orthography outside of spelling conventions and the maintenance of white spaces to indicate word boundaries. There are representations of um and uh vocalizations, as well as a representation of the informant clearing her throat at one point. In this string

---

<sup>1</sup>The following 3 examples are from an interview with JW, an informant in the Atlanta Pilot study, interviewed in 2003 by the author and transcribed by Nikki Kong.

of transcribed speech there is no attempt at indicating particular pronunciations of different forms, as with the example from Labov. Thus, this example does not necessarily lend itself to making any inferences about the speaker's linguistic patterns of variation. In fact, this level of representation more resembles a normalized chunk of discourse, readied for machine readable processing at the tokenized level.<sup>2</sup> All in all, this chunk of discourse is not the most reader friendly, as it flouts expectations of written language with its lack of punctuation and accompanying capitalization conventions; however, it is more amenable to human eyes than the following:

itsreallyfunnyhestartedoutintheologicalschoolandstartedoutasa  
preacherandthenforsomeunknownreasontoallofushedecidedhedidnt  
wanttodothatanymoreandnevertalkedaboutitagainumclear throat  
startedworkhereatrichesandworkedforrichesforohidontknowmaybe  
tenortwelveyearsandheworkedintheircarpetdepartmentatdowntown  
richesthatwastheonlyrichestherewasanduhtheyhadareallybig  
carpetdepartmentandheworkedthere

This example completely defies expectations in terms of written language and it is all but unreadable. It would also confuse any English language based computer processing application, so it is useless in terms of machine readability as well.

The very same passage could also be represented like this:

**its rɪli fəni hi stɑrəd awt ɪn θiələdʒɪkəl skul ænd stɑrdəd awt æz ə prɪtʃr ænd ðen fɔr  
səm ənəʊn rɪzən tu ɔl əv əs hiʃ dəsɑrdəd hi dɪrɪn want tu du ðæt ɛnɪmɔr æn nəvər tɔkt  
əbawt ɪt**

This very broad phonetic transcription privileges a sound-based level of representation using IPA, a system of notation that Language Variationists, and most empirical linguists are well versed in and use routinely in their linguistic research when demonstrating variation

---

<sup>2</sup>Automated processing of a body of text is often preceded by tokenization. Tokenization breaks text down into its constituent parts, or tokens. For computers, a token can be any organizational unit. Tokenization can occur at a number of different levels: a text could be broken up into paragraphs, sentences, words, syllables, or phonemes. And for any given level of tokenization, there are many different algorithms for breaking up the text. For example, at the word level, it is not immediately clear how to treat such strings as “can’t,” “22.50,” “N.Y.” and “so-called.” Each tokenization algorithm will have processing consequences down the line. For a complete treatment of tokenization in the context of automated processing, see Jurafsky and Martin 1999

in pronunciation. This transcription is based on word level transcription. It should be noted that this level of representation does not capture the effects of continuous speech, but rather offers a very general sound-based representation at the word level. While it is an acceptable level of representation to many, it certainly does not accommodate a general audience and it is not necessarily a convention that lends itself to basic computer-mediated processing. This offers a more narrow method of representation, but a fit one nonetheless.

Moving on to a final example of sound-to-text transformation, I want to look at speech transcribed using the Jefferson transcription system, the most widely acknowledged standard in the field of Conversational Analysis <sup>3</sup>:

1. **CPO:** Is that o[ka:y.]
2. **Caller:** [ Fine. ] =yes.
3. [°that's fine.°]
4. **CPO:** [ ↓Brilliant ] okay,
5. **Caller:** °.Hh° (0.2) u:m (0.1) >I'm sorry
6. I'm a little bit< emo:~ tional
7. tod[ay~ .hih]
8. **CPO:** [Tch Qh::] go:sh I'm so:rry,
9. **Caller:** ~I've got a little four year old grandson,~
10. [huh]
11. **CPO:** [Yea]h:,
12. (0.3)
13. **Caller:** ~My son w(h)as s(h)ixtee:n~ (0.5)
14. er fifteen when he was bor:n.
15. (0.3)
16. **Caller:** .Hhh [And um (.)] he and his er (0.2)
17. **CPO:** [ °Mm::° ]
18. **Caller:** girlfriend split up.

---

<sup>3</sup>The reference for this example is found here: <http://www-staff.lboro.ac.uk/~ssjap/transcription/transcription.htm>.



19. (0.9)
20. **Caller:** ((swallows)) ~and since then um:~ (0.2)
21. she's had (0.4) several boyfriends, (0.6) .hh
22. but since the baby was bor:n
23. I've had him (0.3) every week
24. (0.5)
25. **CPO:** [°Ri:ght° ]
26. **Caller:** [I have him] from em (0.4) ((swallows))
27. Thursday through to Sundays.
28. (0.4)
29. **CPO:** Ri:ght.
30. **Caller:** Erm (0.1) she doesn't come from a very
31. good family,
32. **CPO:** [°M m : : .° ]
33. **Caller:** [[[sniffs]]] Her (0.4) step-dad (0.2)
34. abused her (0.4) sister.
35. (0.8)
36. **CPO:** Ri:ght=
37. **Caller:** =And er (0.6) I just don't feel my grandson's
38. being looked after properly
39. **CPO:** Tch °oh: [ g o : : s h° ]
40. **Caller:** [An he's had a] black eye:
41. la:st weekh,
42. **CPO:** Did he:?
43. **Caller:** An a cigarette bur:n .hh hh
44. **CPO:** Oh my g[ o : : s h : : ]
45. **Caller:** [She's now got a n- ] a new boyfriend
46. ((sniffs)) (1.1) and er hh .hh they live in Sawley
47. **CPO:** Yea:=

48. **Caller:** =which is like (0.3) three quarters of a  
 49. mi- e- three quarters of an hour away from  
 50. where we live

This transcribed speech has very different appearance than others. First, it was created to represent high levels of linguistic and paralinguistic information central to empirical language research in the tradition of Conversation Analysis. It employs unique punctuation and symbols, all which represent a specific linguistic or paralinguistic feature, typically characteristic of a spoken exchange. For example, underlining indicates vocal stress or emphasis. Colons indicate elongated sounds or syllables, the more used the longer the sound or syllable. The numbers in parenthesis represent a pause rounded up to the nearest tenth of a second (see Jefferson 2004, Nofsinger 1999). This system is one that is employed by a small segment of the empirical linguistic research community, and thus is interpretable to the select few who are trained to use it. It is not a standard level of representation as far as machine readability is concerned either. In fact, this example exhibits a level of representation not found in previous examples, as it goes beyond the level of standard orthographic conventions of spelling and punctuation. It goes beyond representing pronunciation at the word level. It offers a novel organization of stretches of speech organized around pauses and speaker turns, as well information about linguistic patterns of pronunciation at the syllable level. It depicts a more nuanced level of representation and in doing so this transcribed bit of speech demonstrates a very particular relationship with the stream of audio on which it is based.

Despite the vast differences in all of these examples and despite the fact that they are designed to highlight different aspects of a stream of speech, they all have one thing in common: They all reject conventional systems of representation and instead embrace very particular ways of representing speech. Generally speaking, they are all unconventional attempts to wrangle a speech signal into a written form presumably to accommodate, expedite, and demonstrate the investigation of language, in varying capacities. Is any one of these examples a more natural means of representation than the other? Absolutely not. Is any one

more accurate than the other? Only if you equate accuracy with level of detail. Accuracy is a matter of opinion and not an objective quality that is demonstrably measured or quantified in a scientific manner. These systems of representation are employed by human scribes, thus guided by impressionistic means and deduction, not by some mathematically based equation or application.

These examples demonstrate the range of representation that is available to empirical researchers who transform their spoken data into a graphical form. While each level of representation privileges certain information and favors certain levels of organization, all of these text-based representations are derived from a series of decisions on the part of the scribe who employed them. These decisions range from when to use non standard spelling conventions to convey certain pronunciations, what vocalizations to include or leave out, whether or not that particular sound was an alveolar stop or a glottal stop, to what constitutes elongated syllabic formation. Each level of representation was the result of subjective decision making processes driven by the need to transform sound to text in order to facilitate the study of language in some particular manner, coupled with the researcher's assumptions about language in general.

In empirical linguistics and in Language Variation in particular, we have not properly embraced the fact that there is no natural way to get from sound to text. We have not rightly acknowledged that there are just different methods of representation motivated by different factors, which end up catering to different audiences. As a field, Language Variation has not addressed these issues, nor have we addressed the complexities of the decision making process that characterize data transformation in general. We certainly have not acknowledged the manner in which these things potentially bias our transcribed spoken data and ultimately shape our analytical endeavors and inform our theories of language usage. None of these factors have been formally considered and discussed in the context of Variationist theory and methods. Saussure described it best when he deliberated spoken language as the object of study and written representation as the means of studying it:

Language and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first. The linguistic object is not both the written and the spoken form of words; the spoken forms alone constitute the object. But the spoken word is so intimately bound to its written image that the latter manages to usurp the main role. People attach even more importance to the written image of a vocal sign than to the sign itself (23–24).

In almost 100 years, very little has changed despite the fact that we can now physically record audio and have it available to us throughout the duration of our analytical endeavors. We still rely on preparing graphical representations of speech in order to study it and we still view these graphical representations as empirical data. As a field, we need to reevaluate this position. We also need to renegotiate our relationship with our spoken data. In today's technology driven research environment, we must do better in our approach to data transformation processes.

First, we must go forth armed with the knowledge that data transformation processes require taking a steady audio stream of a spoken language and imposing on it a method of representation or organizational unit that is not inherent in the data itself, but rather decided upon and implemented by the researcher, transcriber, or analyst in order to prepare the data for subsequent use. We must also readily acknowledge that while there is no natural means of representation when transforming sound into a transcript, that accuracy or authenticity are also unrealistic goals, as the divide between a purely aural phenomenon and an imposed graphical representation make the production of an “accurate” or “authentic” rendering an impossibility. And finally, when creating a graphical representation of a speech signal, or transcript, it is important to recognize at the outset that the process is a result of arbitrary decisions made by the researcher or transcriber, arbitrary in that these decisions are indeed subjective and selective in the context of the act of representation. This fact is not dependent on the role of a transcript or a set of objectives that steer its creation although it is intimately connected to it, but rather it is a simple reality of the act of transforming sound to text,

and one that has implications for future use and research. Perhaps the most important thing to keep in mind is in the initial transformation process of creating a transcript is that we are creating a representation of the data, not the body of primary data itself. An audio recording of spoken language is our primary data, the centerpiece of our investigations, and the transcript is a visual aid that gives us a useful way in which to access and navigate our primary research data. In short, we are now officially evolving beyond the model of “transcript as empirical data.”

With that statement, it is now a good time to revisit the role of the transcript outlined in Chapter 1, organizing the general objectives of this data transformation in a way that lays the proper foundation for the practical task of developing a set of guidelines for transforming sound to text, as well as informing the decision making process therein. First and foremost, transcripts are archival in nature. Transcripts are a graphical representation meant to complement an audio recording of primary data, not represent primary data as lists of linguistic features. Next, transcripts should accommodate wide audiences, not just linguists and other language experts. Finally, transcripts must not only be easily read and interpreted by human eyes, but they must also be machine-readable, operating on the assumption that they will be created and housed in a computer-based environment as a text file or some other text-based document. All of these aspects of a transcript are inextricably linked. All inform one another in significant ways and will heavily impact initial data transformation processes. Therefore, it is necessary to delve into each tenet a little further.

First and perhaps most importantly, a transcript is an archival record, one which creates an index between an audio file of a spoken linguistic event and the graphical representation of that spoken linguistic event. In the same way that an audio recording preserves a spoken linguistic event at a moment in time, a transcript as an initial rendering of sound to text creates a complementary artifact to this preserved record. Viewing a transcript in an archival capacity has significant implications, effecting the nature of a transcript in specific ways. As a text-based archive, a transcript is perfectly suited to exist as a part of a growing

repository of natural language. Transcripts as archives are ideal as a general resource, as a link to a primary data source that encourages and accommodates a wide audience with varying interests. Language initiatives such as OLAC and the Linguistic Data Consortium rely on text-based transcripts in order to create massive online libraries of natural language resources, which are easy to locate and access.<sup>4</sup> Thus transcripts as archives serve a vital function with respect to the ever expanding universe of language data, paving the way for productive avenues of investigation and future research endeavors in the Language Variation community and beyond.

Importantly, viewing the transcript as an archive, rather than *the* object of a particular linguistic analysis, allows the researcher to focus on the initial rendering of sound to text as just that, an initial rendering. It allows the transcriber to focus on creating a transparent representation of the audio recording in a balanced capacity, rather than focusing on a linguistically-slanted interpretation of the audio recording. When a transcript moves beyond existing as a linguistically-centered body of data and into the realm of being an indexed-archival record, it has the ability to be developed in a consistent, clear manner that creates a discernible relationship between audio and text. In this way, a transcript is a transparent representation of recorded speech.

Transcripts created in a transparent manner are able to accommodate a general audience and a wide range of research interests. It allows anyone the opportunity to consider the initial transcript as a part of a non-biased body of information, one that is a reasonable representation of a body of primary language data. In this role, transcripts become a general resource, and one that permits any researcher to include the body of information in their own research endeavors, prompting future data transformations that would cater to their own research needs. In the research climate we find ourselves in today, natural language data can and should be shared amongst anyone who has an interest in it. As OLAC and the LDC

---

<sup>4</sup>The Linguistic Data Consortium, or LDC, is a research group at the University of Pennsylvania dedicated to archiving and sharing linguistic data, as well as developing various technologies, analytical tools and methods for aiding empirical linguistic research. They are located on the web at <http://www ldc.upenn.edu/>.

demonstrate, we are in a position to not only archive our data, housing large repositories of it for future use, but we are in a unique position to make sound and text-based data completely available to others via the internet or intranet, and through various media. Transcripts as archival records, prepared to accommodate a general audience, are an integral part of this equation.

As an archive of natural language, a transcript must be readable to human eyes, but also readied to facilitate basic machine readability, because in today's empirical research community transcripts are created and stored in an electronic and digital medium in a computer-based environment, one of many results of the influence of technology that must be taken into consideration when initially transforming spoken language data. Facilitating human and machine readability are two key components of initial data transformation processes and the role of a transcript, as both aspects have implications for data-sharing amongst the research community at large. First, human readability for a general audience operates on the assumption that a transcript will be available for public consumption and accessible to anyone. This in turn necessitates a unifying means of representation, one that relies on traditional practices of representing information and not another novel standard or unique system that caters only to a narrow audience of experts in the field. Novel or exclusive systems of representation by design do not accommodate a general audience, but rather alienate anyone not trained to employ and interpret the system. Furthermore, more exclusive methods of representation should be reserved for further data transformation processes, employed when a body of natural language is being prepared for analysis. Again, the idea of transparency is key in this respect. Transcripts must exhibit a clear, reasonable relationship to primary data if they are to be accessible to a general audience. Inclusive, human readability depends on it.

While human readability is important, machine-readability is also crucial in outlining the general objectives of a transcript. Transcripts are no longer bodies of hand-written notations. Transcripts are routinely created in a computer-based environment and exist as text-based files, a fact which necessitates taking machine-readability into consideration in the

initial data transformation process. Just as human readability and comprehension hinges on transparency in representation, machine-based processing relies on consistency when parsing natural language. Computers require consistent, uniform, and principled treatment of information. Likewise, text-based applications and other automated processes require consistency. Although transcripts are not meant to be analyzable bodies of data, they must be created with an eye toward future text-to-text transformations that will be undertaken in order to create a text-based body of analyzable data. Basic search and retrieval functions are predicated on uniform methods of representation with respect to both humans and machines. The success of basic keyword in context applications, such as Wordsmith tools, is also conditional on the consistent treatment of information.<sup>5</sup> In all of these situations, inconsistent, erratic treatment of information does not lend itself to reliable processing results. On a general level, standards of consistency only stands to improve both our methods and practices of handling and preparing spoken data and our attention to data transformation processes in general.

These basic tenets will provide a meaningful framework for the decision making process that guides creation of the transcript proper. Now that we understand what a transcript is, what it is not, and what its purpose is in the context of our empirical research, we can now move toward the complex decision making process that characterizes this aspect of data transformation with direction and clarity of purpose. One of the first important decisions in transforming a speech signal into text is settling on a unit of organization that will provide the means of graphical representation. An organizing unit is essentially the smallest material unit of representation used to index a speech signal with its graphical counterpart. In order to accommodate our definition of a transcript, a unit of organization has to be one that bridges the gap between sound and text in a unifying manner in order to create a practical index between the two mediums. A unit of organization must accommodate human eyes and should also be a machine readable a means of archiving and mapping information in a manner that accommodates a general audience. In the Language Variation research community, the

---

<sup>5</sup>Wordsmith tools is a package of lexical analysis software and a staple application amongst Variationists and Corpus linguists. It can be found at <http://www.lexically.net/wordsmith/>



idea of a general audience spans a range of linguistic interests, from phonetics to syntax, to studies in discourse and pragmatics; however, it is important to bear in mind that using purely linguistically determined units of organization in a transcription would run counter to the new model of a transcript developed here, as it runs the risk of creating a narrowly defined audience, potentially distancing a general audience outside the Language Variation community that is interested in the study of language. Assuming a general research audience, as well as various specialists in the field, an organizational unit should not prejudice the body of data for folks on either end of the continuum. The most desirable unit of organization is one that accommodates everyone, one that adheres to theoretical as well as practical constraints, a culturally recognized entity with linguistic import rather than a purely linguistic one, and one that is functional in a computer-based environment.

Based on the role of the transcript as defined here, the unit of organization that best lends itself to this approach is the word. In short, the word as the basic organizational unit of the text-based representation of spoken language is desirable from a linguistic, cultural and computational standpoint. Linguistically speaking, the word has prominent status in all areas of the grammar: In phonetics and phonology, morphology, syntax and pragmatics, the word is a traditional unit of analysis and study. Although the notion of a word in the field is not without debate in theoretical circles, the idea of a word as a linguistic unit has long theoretical and empirical tradition. Considering the word from a linguistic standpoint is important when justifying its use as an organizational centerpiece of a transcript; however, justifying the word as an organizational unit does not require satisfying any or all of the theoretical debates surrounding the concept. Words are desirable from a practical standpoint, despite their potentially awkward theoretical nature. The notion of a word is not scientific but rather based on people's intuition about language habits. If you take the word out of the realm of linguistics it ceases to seem controversial. In cultural contexts for example, the word exists outside of realm of controversy regarding its existence or properties. Culturally, the word has the distinction of being one of the only universally applicable abstractions that

the majority of folks make about language (Sinclair 2004, 24–25). Ask the average person what a language consists of and you are most likely to get “words” as a response.

Computationally, words can be discrete processing entities that are parsed by even the simplest of text editors as well as the most complex corpus-based analysis tools. Large corpora of language data are routinely tokenized, or stripped down to a word level existence, to normalize data and expedite computer processing.<sup>6</sup> Most Natural Language Processing parsing applications are written to expect input to be a sequence of words rather than a sequence of individual characters or sentence-based units. For example, parts of speech, morphological, word class and lexical taggers depend on words as basic processing units. More advanced expert systems, such as augmentative communication systems, which help the disabled to communicate, are also predicated on word level units of organization (see Jurafsky and Martin 1999). In terms of computer-mediated processing, words are the most traditional and practical units of organization.

From all three standpoints, linguistically, culturally and computationally, the word embodies the best choice for a unit of organization in which to systematically index speech signals to text-based data. But as this research seeks to establish guidelines for Language Variation research in particular, it may be beneficial to briefly focus on the word as organizational unit and its linguistic credentials in order to better understand its role in the study of both spoken and written language. Deciding on the use the word as organizational centerpiece may seem like a menial exercise, but it is in a significant decision with important ramifications. I would like to refer back to the discussion of Saussure (1915/1959) in Chapter 2, about the importance of focusing on a discernible unit of organization in Language before tackling the actual study of it. Saussure helps put into the perspective the theoretical and practical endeavor involved in simply deciding what unit of organization is the centerpiece in a model of language. Even decisions about settling on a basic unit of organization have

---

<sup>6</sup>Normalization of data refers to ways of making text conform to a norm or standard, introducing a level of consistency which is maximally interpretable by computers. Tokenization is a method of normalization.

theoretical implications, which in turn affect how we go about studying language. Crucially, as Saussure writes, these decisions must be considered before undertaking the transcription and not during or after (106–108). Researchers and transcribers, especially in the Language Variation community, often take this aspect of creating a transcript for granted and assume that such decisions are foregone conclusions, rather than issues that need developing, examining, and explaining. Deciding on the word as the unit of organization has theoretical and practical consequences; however, resolving this issue up front clears the way for consistent and systematic transcription practices. Theoretical stances must be made explicit not only through overtly stated them, but by consistent research practices that demonstrate them in a transparent way, especially within the context of preparing data as this crucial research step is the most overlooked aspect of methodological explanations in linguistic research.

Notably, John Sinclair (2004) says this on the idea of words as units of organization:

“The starting point of the description of meaning in language is the word. There are two primitives in language form, the other being the sentence. The sentence is the unit that aligns grammar and discourse, and the word is the unit that aligns grammar and vocabulary” (24–25).

Sinclair’s use of “align” is particularly interesting. It privileges the function of the word as a building block of language that establishes the relationship between the structure of a language, or grammar, and the lexis of a language, or vocabulary. In other words, it suggests that the importance of words lies in the practical manner in which they establish and maintain the connection between grammar and vocabulary in a language. Sinclair also suggests the sovereignty of words as units is firmly ingrained into our notion of language, regardless of possible ambiguity or instances where words rely on collocates to create meaning. He goes on to point out:

“the majority of writing and printing conventions, words are separated by spaces, and thus have physical recognition of words as units, and this is built firmly into

our general model of language. A text is therefore seen as a succession of discrete items, those items being words (24–25).”

Uneasiness surrounding the theoretical “nature of the word” may be justified in some respects and require disambiguation, but in practical terms this does not interfere with the decision to rely on words as organizational entities, as any debate surrounding the word is one that exists primarily for linguists, not for the vast majority of practitioners of a language (Sinclair 29). Ultimately, a transcript is a matter of public record, an archive accessible to all and as such, it should not favor information for one group at the exclusion of another. Thus the word as the unit of organization is the best starting point for the text-based representation of a speech event.

After the initial process of establishing the word as the unit of organization in the creation of a transcript, it is necessary to consider how to consistently represent words in the transcript as these two decisions are not one in the same. This decision requires thoughtful examination into the relationship between spoken words and the written words meant to represent them. If words are to be the organizational units of a transcript, there must be an obvious association between words uttered in the context of a spoken exchange and their written counterparts. It has been proposed elsewhere that attempting to put speech events into written form is purely an exercise in interpretation, rather than an objective task (see Ochs 1979). In fact, the truth is probably somewhere in between these two positions:

Is there a reliable relationship between the words of a transcript and the words of a spoken speech event? An exchange between individual speakers of the same language entails the ability to perceive and reproduce sameness. The commonality of perception by users of a given language means that on one level the transcription of words in a speech event is not an interpretation or systematic translation from one mode to another, but in fact a repetition (Cook 39).

Cook then elaborates:

This view does not imply, however, that words reside in the physical form of the message independent of interpretation. Words are not the sound waves of speech, nor the marks of writing, but the interaction of those physical phenomena with the competence of a receiver. Among literate expert users this is uniform enough to create a core of agreement compromising both the linguistic forms and their denotations. There is interpretation involved, but of so particular and uniform a kind, that it may in practice be discounted, and the word “interpretation” reserved for those further aspects of meaning (such as connotations, illocutions, perlocutions and schematic associations) on which expert users disagree (Cook 39).

In Cook’s estimation, participants in a speech event rely on a degree of consensus in which to base a relative agreement that the intended message is in fact being delivered and received at the very least on a word-level basis. The habit of language amongst practitioners creates an expectation that negotiates form and meaning in an exchange. Likewise, this expectation of agreement is the foundation of the transcriber transforming spoken speech data from an audio recording into a written counterpart. The transcriber’s linguistic competence facilitates a word-level understanding of the recorded speech signal that is then transformed into writing. If we can go forth with the idea that a written transcript is at least a practical means in which to document a speech event, at least at the word level, we can place the exercise in the realm of interlocutors who rely on conventions of language usage, or habits of usage, in order to parse the message. Rather than viewing the transcript as an interpretation of a speech event, we can consider it a transformation from one unique medium to another: A transformation from spoken language into a written representation. A written representation of a word as a unit of organization could possibly take on many forms, as there are several transcription systems that have conventions of notation for all the various levels of representation available to them. I am, however, proposing the use of standard orthography as a means of representing words in this data transformation process. In the model of a transcript developed here, practical

considerations pave the way for adhering to traditional methods of representation, rather than novel methods that reject convention. Again, this is inextricably tied to the notion that a transcript as an archive be accessible by a general audience, available to the research community at large as matter of public record. As such, it then becomes necessary to address the use of standard writing conventions as a means of facilitating this transformation.

In Language Variation in particular, the use of standard orthographic conventions may seem counterintuitive. In a field that seeks to highlight variation, the very idea of standard orthography, which seeks to minimize deviation and normalize language, may seem like an unnatural means of representation. It is also true that writing conventions leave out a sizable amount of information important in spoken language, such as rhythm, intonation, loudness or softness, voice quality, and pauses, to name several (Halliday 1989 30). Such information has no effective means of representation in English writing systems. English writing conventions focus on the representation of words, ignoring prosodic information as well as paralinguistic features, both of which are important in imparting meaning and interpreting speech events. Over the years, a variety of systems of notation have been invented for transcribing speech events to remedy these limitations in standard orthography. In Language Variation in particular, researchers often employ a mixture of both standard orthography and specialized systems of notating linguistic information. These specialized systems range from the inclusion of IPA to indicate phonetic information, part of speech tags to highlight grammatical categories, and other techniques unique to the research project at hand. Empirical researchers in subfields such as Conversational Analysts, Discourse Analysis and Psycholinguistics to name a few, employ various methods of annotating prosodic and paralinguistic information such as Jeffersonian transcription, CHAT, TEI, COBUILD, and others (see MacWhinney 2000, Bunard 1995 and Payne 1995, respectively).<sup>7</sup> Therein lies a dilemma:

---

<sup>7</sup>Other empirical fields' transcription systems are designed to represent a multitude of linguistic and paralinguistic information and support the "transcript as empirical data" model. While surveying such systems lies outside of the scope of this dissertation, it is a worthwhile venture to look at each system as a window into field-wide assumptions about language and empirical linguistic research.

either adhere to standard orthography and its limitations, create a novel system of notation suited specifically to your research, or employ an existing alternative system that incorporates notational counterparts for prosodic and paralinguistic information. As we have seen these systems only serve to highlight certain types of linguistic information and encourage the practice of creating a linguistically-centered representation of stream of speech. This puts us at risk of falling back into old habits of viewing transcripts as empirical data, a tradition we are trying to get away from.

Highlighting this type of linguistic information should be reserved for the second tier of data transformation, the creation of a more narrowly defined electronic text made specifically for analytical purposes. This describes a text-to-text transformation process, not the initial sound-to-text creation of a transcript. Milroy and Gordon 2003 can be credited with putting forth the idea of developing a more narrowly focused representation of the spoken data, after an initial transcription of the data is in place. The secondary text-to-text transformation that I am proposing here, one that focuses on creating a body of analyzable linguistic data, can be viewed as roughly equivalent to the process of “selective transcription” described by Milroy and Gordon. They acknowledge the importance of creating a selective transcription as a second level of representation in order to accommodate particular analytical objectives or linguistic avenues of investigation (143). Along these same lines, a text-to-text transformation will allow for a transcript, which is primarily an archive and not a body of analyzable data, to become an electronic text, a body of information created to support various analytical endeavors.

Turning our attention back to the initial speech to text data transformation and a method of representation therein, it necessary to keep in mind the definition of a transcript developed here and how best to accommodate this definition. In keeping with our model that a transcript needs to be primarily an archive of public record, readable to human eyes and computers, not tilted too far towards the linguistic side of presenting language data to accommodate a general audience, and importantly, created in an accessible and accommodating

manner, the most practical choice is to employ standard orthographic conventions. Standard orthography is functional choice for several reasons: First, it is an established convention. There is no learning curve involved in employing it and in understanding it. It is a means of representation completely accessible to a general audience as well as the scientific research community. The union between text-based archives and standard orthography is both historically and culturally relevant, partly because it represents a moment in time in the history of a language, which includes culturally relevant information that contextualizes the moment. Moreover, the use of an enduring method of representation makes a transcript more archival in nature, an artifact that speaks to many rather than being one individual's measurement of the linguistic nature of the text, sacrificing the content of the body of information by propelling the linguistic form into the limelight. To illustrate this point one needs only to refer back to the speech transcribed via Jefferson system in the Conversation Analysis framework earlier in this chapter. This example of transcribed speech offers a very particular representation of a conversation whose narrative is overshadowed by layers of encoded paralinguistic information. As such, this particular type of representation is not at all conducive to creating an archival record.

Also, as previously stated, the use of standard orthographic conventions, as opposed to some novel method of annotation, discourages the creation of a transcript as empirical language data, encouraging rather the repeated and continual reference of the audio file as primary data. Another benefit of standard orthography is that when it is employed, the transcriber must understand and confront the apparent restrictions involved in transforming spoken language into a text-based form, and in doing so they can appropriately abandon any ideas of "naturalness" or "accuracy" they may harbor with respect to transforming sound to a graphical representation. Otherwise, the results will be akin to the example of transcribed speech from Labov 2003 referenced earlier in this chapter and in Chapter 2. Labov's use of eye dialect and creative punctuation interspersed with standard spelling and punctuation conventions to approximate an informant's pronunciation in a few choice instances, in what can



only be assumed as an attempt to render a more “accurate” representation, in fact has quite the exact opposite effect. Because standard orthography clearly does not accommodate the nuances of spoken forms, an attempt to bypass this limitation creates an uncertain relationship between primary data and transcript, which calls into question consistency and creates a stilted representation. If standard orthography is the preferred means of representation, then a researcher must adhere to its conventions. As for the limitations of using standard orthography to represent speech, the key is to face these limitations and understand them in order to make decisions of representing information in a consistent, transparent manner. For the vast majority of spoken utterances, there is no counterpart in standard orthography that can adequately represent them. Features of spoken language such as stress, or a unique pronunciation of a certain form, these features of spoken language that make it linguistically interesting simply do not lend themselves to representation in standard orthography. However, when the primary goal of the transcript is taken to be to index information rather than to try and accurately replicate information, then this ceases to be a troublesome issue. The stress of a particular form, or its pronunciation, is easily recovered from the indexed audio recording via the transcript, as this is a primary objective of creating the transcript in the first place: To serve as a useful directory to the analyzable body of recorded data. The use of standard orthographic methods supports the model of a transcript as an archival record, as an index between a primary data source and a secondary one, and a collective record of information available for public consumption. Any specialized representation of speech, such as one that focuses on linguistic features for example, is a secondary step in data transformation undertaken once the primary index is in place, and according to the needs of a particular research endeavor or analysis.

As standard orthography is an entire system of conventions for representing written language, it is an altogether different aspect of language than spoken language.<sup>8</sup> In reality, writing conventions were not necessarily meant to reproduce speech but are another facet

---

<sup>8</sup>Standard orthographic conventions are discussed in practical terms at length in Chapter 6.

of language altogether. As Saussure states, you can no more assume that writing is a full depiction of speech as you can assume that a photograph is as accurate as looking at someone in person (Saussure 1915,1969, 24). Just like a photograph, writing is meant to be an enduring record, a likeness that immortalizes a moment in time. Each word represents a moment on a timeline in which a speech event occurred. A transcript is a visual aspect to that timeline, creating a bridge between a recording and a text-based body of data created for the purpose of analyzing language. Standard orthography is a means to an end: A systematic representation and indexation, one that is best suited to accommodating a variety of potential research endeavors and a variety of fields. Furthermore, standard orthography is completely suited to computer mediated processing, as it is easily dealt with by the most straightforward word processing application to the most complicated natural language processing program. It is also a step in the right direction with respect to encouraging standard guidelines and practices in the treatment of natural language data.

Now, in the interest of moving from the realm of abstraction to practical application, it is an appropriate time to turn to the transcript as a physical record, a material object that exemplifies the ideas discussed in this chapter in order to facilitate the process of transforming an audio recording into a text-based representation. With words as our organizing units and standard orthography as our means of representation, we can now work through the practical endeavor of transforming sound to text.<sup>9</sup>

---

<sup>9</sup>This dissertation focuses on the manual transformation of speech-to-text conducted in an empirical, Variationist research context; however, there are applications that make use of speech recognition technology, which automate the conversion of speech-to-text, popular within the NLP community. These applications, such as Nuance Communication's Dragon's Naturally Speaking, would not alleviate the need for human intervention and proof-reading, as their indexing accuracy rates vary greatly, a fact that is not as big of an issue in NPL research as it is in Language Variation research. In fact, speech recognition technology does not handle dialectal variation well and these types of applications rely on acclimating to one voice over time to improve accuracy levels, which would make them less effective for a variety of speech types from many different individuals. As such, the use of speech recognition transcription technology would require thorough proof-reading and revision of initial renderings. The level of proof-reading and revision required would diminish any time saved by employing such technology on the front end. Therefore, it is not recommended to use these types of applications to automate initial data transformation of speech-to-text in an empirical linguistic research setting.

## CHAPTER 5

### ANATOMY OF A TRANSCRIPT PART 1: METADATA HEADER

So far, we have developed a working definition for a transcript. In doing so we created a framework in which to approach the transformation of an audio recording into a transcript which takes into consideration the role of a transcript as a cultural artifact, a body of data for all manner of linguistic inquiries, and a cohesive text that accommodates automated, computationally based processing. We have justified words as the basic unit of organization for a transcript and argued that the best means of transforming spoken words into their written counterparts is with standard orthographic conventions. All in all, we have generated a discussion with respect to the underlying theoretical assumptions that inform the process of transforming spoken language into a graphical representation, paving the way for the creation of a consistent and systematic written representation of a speech event. Now it is necessary to develop more specific procedures that will contribute to the overall process of creating a cohesive body of transcribed spoken language data. This entails looking at the transcript as both a complete body of data, with all of its various components, as well as conceiving of it as a part of a larger, ever expanding corpus or repository of natural language data.<sup>1</sup>

A transcript exists in two basic parts: One part is the body of transcribed data itself and the other is a part that contains information about the body of transcribed data, otherwise

---

<sup>1</sup>At this point in the process, it is worth recommending that you use a basic text editor for transcription. Transcriptions as text files will allow for more flexibility in later data transformation processes. There are a number of great text editors that are either free or very reasonable. My personal recommendation is editpad lite, a free download available here: <http://www.editpadlite.com/editpadlite.html>. If you prefer to use Word or another word processing program, I recommend that you create a parallel text file as well.

known as metadata.<sup>2</sup> Both parts play an important role in situating the words of the text into a larger context, a context required to fully convey and interpret the speech event as a whole. In this chapter I will first focus on the part of the transcript known as metadata, describing what it is and why it is important to language research. Then I will look at some existing metadata standards, discussing the strengths and shortcomings of each. Finally, I will present an overview of a metadata protocol developed specifically for kind of language data common to Language Variation research, prepared in a manner that accommodates collaboration in other empirical fields. This task entails outlining the basic components of the metadata, as well as rationalizing the method used to in representing metadata in the protocol.

## 5.1 METADATA

The development of metadata has its roots in the history of electronic text encoding, but the concept has been around much longer, perhaps as long as text archiving practices themselves. In the early 1990s, however, different groups of text–encoding experts interested in text–based disciplines in the Humanities, such as history and literature, decided to develop guidelines for encoding documents in a digital environment for archival and processing purposes. Such efforts were specifically developed to support and utilize the state of the art technological advances of the time. With the creation of electronic text repositories and collections of data, there was an immediate need for “structured” ways in which to organize and classify data in order to expedite automated processing for locating, navigating and investigating electronic data stored in electronic repositories (Lazinger 140–141). Metadata was developed to satisfy these basic necessities while sustaining a research environment that supports large and varied collections of text–based data as a primary object of study. Thus as text based disciplines

---

<sup>2</sup>This is a level of organization has been informed by electronic text markup protocols. This practice is important for 2 reasons: 1) it adds contextualizing information to the transcript proper and 2) it prepares the transcript for future data transformation processes that will result in the creation of an electronic text.

made the move to digital media, early pioneers in the movement were greeted with emerging standards and guidelines to referee the inter and intranet-wide effort in preparing texts for efficient identification and handling in order to accommodate cross-disciplinary inquiries (Lazinger 2001, 142–143).

Essentially, metadata is information that characterizes a body of data as a whole, providing a range of details about it. In other words, metadata is data about language data.<sup>3</sup> Metadata serves many important functions: First, it is a method of cataloging information in order to provide a reasonable way in which to locate and assess that information on the Internet, or on any computing environment for that matter. Metadata should provide language databases, and any knowledge based information system, with an easy and effective and systematic method of navigation and preservation by the inclusion of bibliographic information that uniquely identifies it. Secondly, metadata describes the content of the body of data, essentially providing a detailed descriptive component providing both linguistic and extralinguistic context essential for evaluating the content. Finally, metadata provides information about the encoding protocol of a body of data as well as administrative details that identify things like who created and edited the data and when. A wide variety of data collections, from digital archiving projects, to electronic dictionaries and encyclopedias, as well as online libraries, all manner of text repositories, rely on metadata for these purposes (see Burnard <http://users.ox.ac.uk/~lou/wip/metadata.html>, accessed July 11, 2008).

As electronic text repositories have continued to evolve in content and in form, standards for both encoding electronic texts and metadata alike have also evolved to accommodate maximal computational processing and encourage data sharing in the research community. The inclusion of metadata is viewed as a fundamental aspect of the anatomy of electronic text collections, in particular those that adhere to the importance of practical encoding guidelines like those suggested here; however, any collection of texts can benefit from the practice of including metadata, especially texts representing natural language or some body

---

<sup>3</sup>This definition of metadata so ubiquitous in computing-ese that I have no idea how to cite it. I would have to cite every reference text on the subject.

of natural language intended for linguistic analyses. For metadata not only normalizes the practice of including essential information that contextualizes text-based language data, but it does so in a way that readies the data for sharing and other collaborative research efforts.

Early text encoding standards and practices such as TEI favored metadata as an introductory element, or a collection of information located at the beginning of electronic texts and collections of texts (see Burnard 1995). Today in the field of electronic text creation, the standard practice for including metadata is to do so in a header or an introductory structure that comes before the body of the text both to systematically catalog information and to set the stage for the language data that follows. In many respects, metadata can be likened to the front matter of a written text which contains information about the author and general publishing information, a table of contents or chapter outline, and other pertinent introductory material, from acknowledgments to an abstract. Likewise, metadata is front matter that provides information such as when the data was gathered, how it was recorded, who recorded it, and who the intended audience is, as well as extralinguistic characteristics of the informant and any other information that is necessary to fully contextualize the speech event. Both physical and electronic media rely on this front matter as an essential aspect of introducing a text. For this reason, metadata is considered a required element in the creation of electronic texts, or texts that are specifically prepared for digital archiving, information retrieval, or general display and manipulation in a web-based environment. And perhaps most importantly, metadata is a key element of data sharing as it not only contains details that are essential in fully describing and characterizing a body of data for other researchers, but it renders these details as easily searchable and retrievable.

In the field of Language Variation, the kind of information found in metadata is not something that linguists typically include in the transcript proper. Field-wide standard methods of electronic data preparation are not yet considered to be objectives in Language Variation research community and the idea of metadata as a necessary component of language data is an altogether foreign notion to most linguists in this subfield. The type of informa-

tion that comprises metadata is usually only available to the researcher and not considered an extension of the language data, but rather extraneous information to be outlined in general methodologies describing the research. But even in general methodological discussions in *Language Variation*, the systematic disclosure of how data was prepared, who prepared data, how and why the data was normalized in a particular manner and other significant details is not customary. The protocol developed in this dissertation, however, includes metadata in the transcribed body of language data. This is an important exercise for several reasons: First, metadata is an imperative aspect of archiving collections of information and as collections of natural language data, transcripts are essentially archival in nature. In *Language Variation* and other empirically driven linguistic fields, language data, ranging from audio files to electronic texts, are often created and stored in an electronic representation. In particular, transcribed data regularly resides on a computer or some other accessible, portable medium. This common practice automatically adds an archival element to the data. It is customary for researchers to want to add an element of posterity to their data, as it is often the case the same body of data will be referred to for a variety of investigations by researchers; however, it should also be customary to consider one's transcription as primarily an archive in a collective system of historically relevant information. Metadata has developed as a method to place electronically stored texts in a wider, collaborative system while individually maintaining the information about the texts for easy access and processing.

Secondly, metadata provides information that situates the body of data in a broader context within the research community, a vital aspect of data sharing and making data available to a general audience. Certain types of bibliographic information that place the transcribed data in a historical context, such as when and where the data were collected, are best disclosed in the transcript itself, specifically in the transcript header. Certain types of technical information, such as how data was normalized and what types of written conventions characterize the transformation from sound to text also are best disclosed in the transcript header. Including metadata prepares a transcript in a methodologically transparent manner, readying

the data for a variety of inquiries and types of automated processing by a wider audience and adding to the growing body of linguistic research made accessible to the research community at large.

Finally, including metadata, in particular including metadata in a header of a transcript, is a sound practice because it initiates the data transformation process in a systematic, uniform style providing relevant extralinguistic and methodological information to all who access it. It is also a good way to set the stage for the information that follows. In basic terms, including metadata as an integrated, necessary part of the transcription process is a sound methodological practice for the Language Variation, Sociolinguistics, and any other field whose research focuses on transcribed spoken language data because it significantly increases the overall quality of the data. Subsequently, an increase in the quality of data will increase the overall quality of analyses carried out on the data.

Preparing quality data in a systematic manner is an imperative in research communities that rely on data sharing. For example, researchers in applied fields like Corpus Linguistics, which center on developing and analyzing large collections of language data, do so with the assumption that a variety of investigations and research interests be accommodated by the data. In the field of Language Variation, however, the idea of sharing one's linguistic data with the research community at large is a novel concept, as is the concept of preparing one's data with well thought out, defensible, and standard means. It is fair to say that research in the field is more results oriented, while the series of processes that lead to the results oftentimes get overlooked. Language Variation is poised to benefit from following the example of applied methods like those found in Corpus Linguistics. First, the Language Variation research community represents over 40 years of efforts in the field of documenting natural language usage. Certainly the body of data that exists would be an enormous resource for other investigations in linguistics and beyond. More importantly, contributing to a collective knowledge base not only increases the quantity of accessible data, but also the quality. Data sharing relies on gathering and preparing by others necessitates principled, detailed treatment



in order to give layers of context to the data, which in turn enables you to understand any observations made with a degree of reliability and validity. In this way, sharing data depends on making information about your data completely available. The inclusion of metadata alone would be an important first step for Language Variationists in improving field-wide methodology.

Lou Burnard of the Text-Encoding Initiative argues that metadata is vital in restoring context to naturally occurring language that has been transformed into a body of analyzable linguistic data:

Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language. A typical corpus analysis will therefore gather together many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata can restore that context by supplying information about it, thus enabling us to relate the specimen to its original habitat. Furthermore, since language corpora are constructed from pre-existing pieces of language, questions of accuracy and authenticity are all but inevitable when using them: without metadata, the investigator has no way of answering such questions. Without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity (<http://users.ox.ac.uk/~lou/wip/metadata.html>).

Including metadata is a straightforward, practical decision, but deciding what specific types of information to include is a much more difficult task. There are a number of existing guidelines and standards for representing metadata in electronic texts and language corpora that can be referenced when making these decisions. Many of these guidelines were developed as a means of describing written texts for specific research projects and then broadened to accommodate other types of data and more general purposes. Others were created for the sole purpose of proposing an archive-based system of identification to facilitate locating

where collections reside. Existing guidelines span a range of data preparation techniques, from the categories and structural organization of metadata, to specific encoding elements or categories intended to organize the content of the data. In short, different encoding standards handle metadata in unique ways. As encoding protocols develop and evolve, and as empirical fields like Language Variation embrace the use of standard practices in preparing data, it is necessary to consider what types of information are most essential and useful when describing a body of natural language data such as a speech event transformed into a transcription, which is most often the impetus for investigations in the field. A practical place to begin formulating and devising metadata is surveying some prevalent standards in Humanities oriented fields to compare and contrast what metadata categories encoding experts consider important.

To begin this survey process, I have chosen to review some of the more established metadata guidelines in the area of the text encoding, particularly those affiliated with Humanities in the academic community: TEI, Encoded Archive Description (EAD) and MARC, OLAC, and the Dublin Core. Chapter 2 included a brief overview of each of the guidelines, as well as an indication of the type of research each were developed to support. Here I will offer a summary of the metadata categories followed by a discussion of their relevance to empirical linguistics, specifically Language Variation data transformation. All of these standards are XML-based; however, for the purposes of focusing on metadata content rather than encoding organizational structure, this exercise will list the encoded element followed by a description of the category. Each explanation will conclude with a brief illustration of the encoded metadata structure. A full demonstration of XML encoded language metadata and language data will be presented in Chapter 7.

Before initiating the metadata comparison, it is worth noting some major differences in the encoding protocols under review. TEI represents a comprehensive encoding standard that recommends guidelines not only for recording metadata, but for encoding the entire body of a text with the option of several potential methods of organization. TEI offers an expansive

set of standards for imposing an encoded structure onto text in order to ready that text for all manner of investigation. By comparison, EAD and OLAC are focused on metadata standards, mainly to ensure a prevailing standard method for tracking down electronic texts in an internet and intranet environment. Thus while the motivation behind all of the encoding protocols is to move towards set guidelines of standards and practices to facilitate data sharing amongst the research community at large, the individual goals of these standards are distinctive, and the effects apparent in the varying levels of complexity exhibited by each standard. Guidelines focusing on archiving and retrieval offer standards that are very minimal and straightforward, while guidelines proposing a data structure for all manner of electronic texts present multi-layered, exhaustive means of categorizing data. Again, however, it is important to note that all guidelines are driven by the notion of implementing a standard that will encourage and facilitate making data accessible to the research community at large. The most practical set of standards will fall in the middle ground, creating a set of useful categories for metadata that are not over simplified, as this will not adequately reflect the richness of the data nor meet the requirements of the research community. It is also important not to overcomplicate matters by attempting to be too all-encompassing and trying to anticipate the spectrum of investigations, creating a system that too complex for most researchers to employ. To get a better idea of the need for a protocol that bridges the gap between overly complex and simplified, I will start at the “complex” end of the spectrum, as demonstrated by the first guideline under consideration, TEI. This overview represents the most basic levels of metadata representation recommended by TEI. Although it is quite lengthy, the complexity of TEI standards and the level of granularity in TEI’s metadata categories dictate this protracted discussion.

### 5.1.1 TEI

The TEI header was designated to house metadata that provided bibliographic, descriptive, encoding, and editorial information about texts that adhered to its standard

In order to focus on describing spoken language data, the TEI header that is offered here is one developed specifically for the text of spoken language transformed into a written transcript. Likewise, the sub-elements related to each higher level element will be explained only if they are candidates for metadata of transcript of spoken language. A full description and explanation of the TEI header can be found here:

Traditionally, XML tags consist of an element and may or may not include one or more attributes. Elements are the portion of tag that semantically define the encoded information. Attributes complement elements by further defining the encoded information. Attributes are assigned values. The basic structure of an XML tag looks like this: `<element attribute = "value">`. Subelements refer to elements that must be nested within another specified element. This relationship is often referred to as a parent-child relationship.

The elements and subelements of the TEI header are not always transparent and easily understood by name, thus warranting a closer look at each. I will first describe the four major parts of the header and then look at each part in more detail, including both required and optional elements of each part. Then I will look at each element individually, and each of the respective sub-elements therein.

TEI header itself, tagged `<teiHeader>`, contains a wide variety of key information about a text. It is appended to every TEI-conformant text as a title page. The header has four major parts:

`<fileDesc>` describes a complete bibliographical description of the text. This element should ultimately serve as a catalogue entry in an archival-like capacity. The file description also includes information about the source or sources from which the file was obtained.

`<encodingDesc>` describes the intersection between a text and the encoding protocol used to describe it. This element may contain detailed descriptions of the nuances of the transcription and how these nuances are realized in various levels of encoding.

<**profileDesc**> contains classification categories and contextual information crucial in describing the text. This includes information such as subject matter, the individuals involved in contributing to the language data, and the like. This element may contain a restricted vocabulary to aid in automated search and retrieval of information.

<**revisionDesc**> allows the researcher to document the history of changes and the editing process during the development of the text.

In addition to these elements, there are other required elements and supplementary elements that can be used as needed. None of these elements have attributes except <**fileDesc**>, which can utilize a type attribute if the body of the text is specified as a linguistic corpora. Of these, only the <**fileDesc**> element is required in all TEI headers; the others are optional but highly recommended nonetheless. It is important to note that the elements occurring within the TEI header may contain a variety of content including free prose, grouping elements, declarations and descriptions. The first and only required element in the TEI header, the <**fileDesc**> element, contains three required elements and four optional elements:

<**titleStmt**> contains information about the title of a work and its contributors.

<**editionStmt**> contains information concerning the edition of a text.

<**extent**> describes the size of the text.

<**publicationStmt**> contains information about the publication or distribution of a text.

<**seriesStmt**> if a text belongs to a series, this element contains information about said series.

<**notesStmt**> assembles notes providing supplementary information about a text to the bibliographic description.

<**sourceDesc**> supplies a bibliographic description of the original text or texts from which the encoded text was developed

The <**titleStmt**> element is the first component of the <**fileDesc**> element, and is required. It contains the title given to the electronic work, together with information having to do with the researcher or other parties responsible for developing the text, known as the “statement of responsibility.” It may contain the following elements:

<**title**> contains the text’s title.

<**author**> contains the author’s name.

<**sponsor**> gives the name of a sponsoring institution.

<**funder**> specifies the name or organization responsible for a research project’s funding.

<**principal**> gives the name of the principal researcher responsible for the development of a text.

<**respStmt**> supplies a “statement of responsibility” the person or persons responsible for the intellectual property of a text when other elements specifying authorship are not sufficient.

<**resp**> gives a description detailing a person’s intellectual property with respect to text.

<**name**> contains a name with a type attribute to identify who or what is being named.

The <**editionStmt**> element is the second component of the <**fileDesc**> element. It is not required but it is recommended. This element groups information relating to a single edition of a text. It contains either phrases or more specialized elements identifying the edition and those responsible for it:

<**edition**> describes the specifics of one edition of a text.

<**respStmt**> supplies a statement of responsibility for the text when other elements are not sufficient.

<**name**> contains the name.

<**resp**> contains a phrase describing the extent and nature of a person's intellectual responsibility.

The <**extent**> element is the third component of the <**fileDesc**> element. It is optional. <**extent**> describes the approximate size of the text as archived on some digital medium. The <**publicationStmt**> element is the fourth component of the <**fileDesc**> element and is required. It groups together information about the publication of a text. It may contain either a description or groups of the following elements:

<**publisher**> the name of the organization or institution responsible for publication.

<**distributor**> the name of a person or institution responsible for distributing a text.

<**authority**> the name of a person or institution responsible for making an electronic file available, if different than publisher or distributor.

The <**seriesStmt**> element is the fifth component of the <**fileDesc**> element and is optional. It groups information about the series, if any, to which a publication belongs.

The <**notesStmt**> element is the sixth component of the <**fileDesc**> element and is optional. It collects together any notes providing information about a text needed to supplement other bibliographic information.

The <**sourceDesc**> element is the seventh and final component of the <**fileDesc**> element. It is required. It is used for providing bibliographic details about the source or sources from which the text was created. If the text is derived from a spoken source rather than a written one, it is a good practice to record additional information about the recording that represents its source by using the following elements:

<**scriptStmt**> gives a citation of details for the script used for a spoken transcript.

<**recordingStmt**> gives description of the recordings used in transcription.

<**recording**> gives details of an audio or video recording event used as the source of a spoken text.

<**equipment**> provides technical details of the equipment used for an audio or video recording of a spoken text.

<**broadcast**> describes a broadcast as the source of a spoken text.

It is important to note that detailed information about the informants or interview setting should be included in the appropriate division of the profile description, rather than as part of the source description. The source description is reserved for details about the technical aspects of the source itself.

The <**encodingDesc**> element is the second major subdivision of the TEI header. It details the methods and data preparation protocol used to develop the transcript. It may also include sets of coded definitions used by other components of the header. Though not required, it is highly recommended. The content of the encoding description may be a prose description, or it may contain elements from the following list, in this order:

<**projectDesc**> describes in detail the function of the manner in which an the text was encoded, along with any other relevant information about the process of creating the text

<**samplingDecl**> contains a prose description of the sampling methods used in the development of a corpus or collection.

<**editorialDecl**> gives details of editorial standards and practices that guided the encoding process.

<**tagsDecl**> provides detailed information about the XML tagging used in the text.

<**refsDecl**> used to demonstrate how canonical references are made for the text.



<**classDecl**> contains one or more taxonomies defining any classification codes used in the text.

<**fsdDecl**> defines the feature system declaration which contains descriptions for types of feature structure.

<**metDecl**> identifies the manner metrical patterns are represented.

<**variantEncoding**> declares the method used to encode important variation.

The <**projectDesc**> element is the first of the nine optional subdivisions of the <**encodingDesc**> element. It may be used to describe, in prose, the rationale behind the encoded work together with any other pertinent information about the development procedure. This information is highly recommended for linguistic corpora as it may be used for a detailed justification of why one encoding protocol was chosen over another.

The <**samplingDecl**> element is the second of the nine optional subdivisions of the <**encodingDesc**> element. It contains a prose description of the rationale behind the sampling used to create a representative corpus.

The <**editorialDecl**> element is the third of the nine optional subdivisions of the <**encodingDesc**> element. It is used to give editorial practice specifics used during the encoding of a text. It may contain one or more of the following specialized elements:

<**correction**> This is used to identify whether the text was amended during or after data capture. It is used to identify what procedures have been implemented with respect to omissions, truncations, etc.

<**normalization**> This gives details of what normalization processes were used and if they were identified by use of tags.

<**quotation**> This details how quotation marks are handled.

<**hyphenation**> This details how hyphens are handled.

**<segmentation>** This details how the text is segmented.

**<stdVals>** This details the standardization methods behind any standardized values supplied for numeric values or dates.

**<interpretation>** This identifies where any “interpretive” information has been provided during the encoding process.

The **<tagsDecl>** element is the fourth of the nine optional subdivisions of the **<encodingDesc>** element. It is used to record the how often specific elements appear within the text, any relevant information relating to the usage of specific elements not found elsewhere in the header, and it provides a definition for the default mode of an element. This information is conveyed by the following elements:

**<rendition>** gives information about the proposed version of one or more elements.

**<tagUsage>** gives information about the usage of a specific element within a text.

The **<refsDecl>** element is the fifth of the nine optional subdivisions of the **<encodingDesc>** element. It is used to document the way in which any standard referencing scheme built into the encoding works, either by a prose description or the use of the following elements:

**<refsDecl>** specifies how canonical references are made for the text. This can have a doc-type attribute.

**<step>** specifies one component of a canonical reference defined by the stepwise method.

**<state>** specifies one component of a canonical reference defined by the milestone method.

The **<classDecl>** element is the sixth of the nine optional subdivisions of the **<encodingDesc>** element. It is used to group together definitions for any descriptive

classification protocol used by other parts of the header. Each part of the protocol is identified by a **<taxonomy>** element, which may contain a bibliographic citation or a descriptive typology. It may contain the following elements:

**<taxonomy>** defines a typology used to classify texts either by a bibliographic citation or by a taxonomy.

**<category>** contains an individual descriptive category within a user-defined taxonomy.

**<catDesc>** describes some category within a taxonomy or text typology.

The core **<profileDesc>** element has three optional components, represented by the following elements:

**<creation>** contains information about the creation of a text.

**<langUsage>** describes the languages, registers, dialects etc. represented within a text.

**<textClass>** groups information which describes the nature or topic of a text in by way of a standard classification procedure.

The **<langUsage>** element may contain one or more **<language>** elements, each of which takes attributes specifying the orthographic conventions used for the text. In addition, a prose description may also be added give further relevant information. **<langUsage>** describes the languages, sublanguages, registers, dialects etc. represented within a text. **<language>** characterizes a single language or dialect used within a text. The attribute *wsd* is used to specify the orthographic convention declaration and the *usage* attribute gives a percentage of how much of the text uses the specified language or dialect.

The second component of the core **<profileDesc>** element is the **<textClass>** element. This element is used to classify a text according to a variety of methods.

The final subelement of the TEI header, the **<revisionDesc>** element, provides a register following the changes made to a text. It is an optional element, but its usage is highly

recommended, especially where a text changes hands amongst a number of researchers. Ultimately, any TEI-compliant text should require a log during the development and editing process. `<revisionDesc>` provides a summary of the revision history for the text. Within the `<revisionDesc>` element, the required subelement `<change>` summarizes changes made to a particular version of a text which is being consulted by several researchers. The log consists of a list of entries, one for each change using either the `<list>` element or as a series of special purpose `<change>` elements, each of which has the following parts:

`<date>` identifies a date. May take a value attribute which requires a standard format, usually `yyyy-mm-dd` and a certainty attribute indicating the exactitude attached to the date.

`<respStmt>` gives a statement of responsibility for someone responsible for the intellectual content of a text, edition, recording, or series.

`<item>` contains one component of a list.

In addition to the elements and subelements listed here, all tags used in the TEI header can take the following global attributes: *n*, *id*, *rend*, *lang*, *type*, *creator*, *date.created*, *date.updated*.

The simplest form of a basic TEI header looks this this:

```
<teiHeader>
  <fileDesc type=corpora> [ ... ] </fileDesc>
  <encodingDesc> [ ... ] </encodingDesc>
  <profileDesc> [ ... ] </profileDesc>
  <revisionDesc> [ ... ] </revisionDesc>
</teiHeader>
```

The following example header is available at a website dedicated to TEI Guidelines (<http://tei.oucs.ox.ac.uk/P5/Guidelines-web/en/html/HD.html>). It is an expanded header, including both mandatory and recommended information to satisfy bibliographic, information about content, information about an encoding protocol, and a revision history of a fictitious text found in the Oxford Text Archive. This example header is of an average size,

larger than one that included only the minimal requirements yet smaller than one that incorporated most of the elements, their subelements and attributes.

```

<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Common sense, a machine-readable transcript</title>
      <author>Paine, Thomas (1737--1809)</author>
      <respStmt>
        <resp>compiled by</resp>
        <name>Jon K Adams</name>
      </respStmt>
    </titleStmt>
    <editionStmt>
      <edition>
        <date>1986</date>
      </edition>
    </editionStmt>
    <publicationStmt>
      <distributor>Oxford Text Archive.</distributor>
      <address>
        <addrLine>Oxford University Computing Services,</addrLine>
        <addrLine>13 Banbury Road,</addrLine>
        <addrLine>Oxford OX2 6RB,</addrLine>
        <addrLine>UK</addrLine>
      </address>
    </publicationStmt>
    <notesStmt>
      <note>Brief notes on the text are in a
        supplementary file.</note>
    </notesStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <editor>Foner, Philip S.</editor>
          <title>The collected writings of Thomas Paine</title>
          <imprint>
            <pubPlace>New York</pubPlace>
            <publisher>Citadel Press</publisher>
            <date>1945</date>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>

```

```

<encodingDesc>
  <samplingDecl>
    <p>Editorial notes in the Foner edition have not
      been reproduced. </p>
    <p>Blank lines and multiple blank spaces, including paragraph
      indents, have not been preserved. </p>
  </samplingDecl>
  <editorialDecl>
    <correction status="high" method="silent">
      <p>The following errors
        in the Foner edition have been corrected:
      <list>
        <item>p. 13 l. 7 cotemporaries contemporaries </item>
        <item>p. 28 l. 26 [comma] [period] </item>
        <item>p. 84 l. 4 kin kind </item>
        <item>p. 95 l. 1 stuggle struggle </item>
        <item>p. 101 l. 4 certainy certainty </item>
        <item>p. 167 l. 6 than that </item>
        <item>p. 209 l. 24 published published </item>
      </list>
    </p>
    </correction>
    <normalization>
      <p>No normalization beyond that performed
        by Foner, if any. </p>
    </normalization>
    <quotation marks="all" form="std">
      <p>All double quotation marks
        rendered with ", all single quotation marks with
        apostrophe. </p>
    </quotation>
    <hyphenation eol="none">
      <p>Hyphenated words that appear at the
        end of the line in the Foner edition have been reformed.</p>
    </hyphenation>
    <stdVals>
      <p>The values of <att>when-iso</att> on the <gi>time</gi>
        element always end in the format <val>HH:MM</val> or
        <val>HH</val>; i.e., seconds, fractions thereof, and time
        zone designators are not present.</p>
    </stdVals>
    <interpretation>
      <p>Compound proper names are marked. </p>
      <p>Dates are marked. </p>
      <p>Italics are recorded without interpretation. </p>
    </interpretation>
  </editorialDecl>

```

```

<classDecl>
  <taxonomy xml:id="lcsh">
    <bibl>Library of Congress Subject Headings</bibl>
  </taxonomy>
  <taxonomy xml:id="lc">
    <bibl>Library of Congress Classification</bibl>
  </taxonomy>
</classDecl>
</encodingDesc>
<profileDesc>
  <creation>
    <date>1774</date>
  </creation>
  <langUsage>
    <language ident="en" usage="100">English.</language>
  </langUsage>
  <textClass>
    <keywords scheme="#lcsh">
      <list>
        <item>Political science</item>
        <item>United States - Politics and government -
          Revolution, 1775--1783</item>
      </list>
    </keywords>
    <classCode scheme="#lc">JC 177</classCode>
  </textClass>
</profileDesc>
<revisionDesc>
  <change when="1996-01-22">
    <name>CMSMcQ</name> finished proofreading
  </change>
  <change when="1995-10-30">
    <name>L.B. </name> finished proofreading
  </change>
  <change when="1995-07-20">
    <name>R.G. </name> finished proofreading
  </change>
  <change when="1995-07-04">
    <name>R.G. </name> finished data entry
  </change>
  <change when="1995-01-15">
    <name>R.G. </name> began data entry
  </change>
</revisionDesc>
</teiHeader>

```

This metadata header is an example of an exhaustive protocol that attempts to account for any and every manner of classification with respect to describing a body of language data. A considerable amount of information including linguistic, extralinguistic, bibliographic, and editorial data is accommodated through a far-reaching list of elements, subelements and attributes. Even in its most rudimentary form, the TEI metadata standard is an unwieldy addition to a body of text-based data. Because TEI offers a comprehensive encoding protocol for every genre of text, ranging from literary works to spoken language corpora, and because it is a complete encoding package for both metadata and for marking up a body of text, its encoding requirements exhibit a high degree of granularity in every respect. TEI metadata categories must anticipate the expansive encoding requisites for the body of the text, and thus these categories exhibit a multitude of choices in anticipation of an array of encoding possibilities that could be used to organize and semantically define the body of the text. For example, the second major metadata element in the TEI header, the <encodingDesc>, is devoted to delimiting a multitude of categories that exhaustively describe how the content in the body of text is tagged. This entire section of the metadata is useful mainly if you are using TEI as a encoding system to transcribe the entire body of spoken data, as it focuses on declaring how references are made, how taxonomies of codes are employed, how linguistic features are encoded, including how "variation" is represented. There are nine recommended elements that related to the body of the text in this capacity, only one of which is potentially useful to data transformation in a Language Variation framework as defined in this dissertation: <samplingDecl>. This element allows you to described sampling methods used in the development of the corpus, but other than this one category, the entire section is relevant for facilitating the transformation of transcript to electronic text for example, but not necessarily an initial rendering of sound to text. Specifically, this aspect of the metadata offers a level of detail that assumes a body of spoken data is being readied as a corpus of empirical data in content and in structure. If the sheer number of possible categories was not daunting enough, the fact that the vast majority of categories do not "fit" the task of



describing spoken language data makes TEI an impractical model for a metadata standard. In short, more does not necessarily equal superior.

### 5.1.2 EAD

EAD is a practical approach to metadata and can be implemented easily with very little training. There is extensive support documentation available for reference. Generally speaking, EAD offers a protocol for describing a text on the metadata level as well as the data level; however, as this encoding initiative focuses on archival practices, the protocol is biased towards organization that facilitates methods of searching and retrieval of bodies of data, or bibliographic information (<http://www.loc.gov/ead/eaddev.html>).

EAD is designed to support individual texts or collections of texts. EAD divides the text into a header, which houses the metadata, and a body, which contains the description of the text itself. The EAD header focuses squarely on bibliographic and some descriptive information and it is modeled on the Text Encoding Initiative (TEI) header in order to encourage uniformity in the use of metadata across different fields and genres of text. In other words, the EAD header is TEI compliant. EAD elements can also be made to adhere to Dublin Core and other standards, thus increase the flexibility of the protocol.

The header is identified by `<eadheader>`. Like all metadata headers, the `<eadheader>` is a required element. Four subelements are available, which must occur in the following order:

`<eadid>` which is a required element, that assigns a unique id to the text establishing it as compliant with the EAD protocol.

`<fileDesc>` like its TEI counterpart, this describes a complete bibliographical description of the text. This element should ultimately serve as a catalog entry in an archival-like capacity. The file description also includes information about the source or sources from which the file was obtained.

<**profileDesc**> slightly different from its TEI counterpart, this contains information about the creation of the encoded version of the text, such as the name of the person or persons responsible for encoding the data, the place and date of the data preparation, and the version of encoding protocol used.

<**revisionDesc**> like its TEI counterpart, this element allows the researcher to document the history of changes and the editing process during the development of the text.

Within <**fileDesc**>, there is one required subelement and four optional ones:

<**titlestmt**> is required and may include the title and subtitle of the text, the name of the author, and the name of the sponsor, respectively.

<**editionstmt**> contains information concerning the edition of a text.

<**publicationstmt**> contains information about the publication or distribution of a text.

<**seriesstmt**> contains information about a series in which a text may belong.

<**notestmt**> assembles notes providing supplementary information about a text

The <**titlestmt**> declaration requires the subelement <**titleproper**>. It should be used to identify a formal title for the text. Other than the ead header declaration and the difference in the <**profileDesc**> element, the EAD header elements behave in the same manner as their TEI counterparts. Thus they provide unique identification code for locating and retrieving, bibliographic information, information about the encoding procedure, and statements about significant revision to the text.

The previous discussion of TEI metadata applies to the EAD metadata. The main differences in the two exist in element attributes, which EAD has developed advanced ways of finding and retrieving texts as well as detailing other bibliographic information important in this process.

The following is an example of a basic <**eadheader**>.

```

<eadheader>
  <eadid>[...]</eadid>
  <filedesc>
    <titlestmt>
      <titleproper>[...]</titleproper>
    </titlestmt>
  </filedesc>
</eadheader>

```

This is an example of a full EAD header:

```

<eadheader audience="internal" langencoding="ISO 639-2"
findaidstatus="edited-full-draft">
<eadid systemid="dlc" encodinganalog="856">
loc.mss/eadmss.ms996001 </eadid>
<filedesc>
  <titlestmt>
    <titleproper>Shirley Jackson</titleproper>
    <subtitle>A Register of Her Papers in the Library
of Congress</subtitle>
    <author>Prepared by Grover Batts. Revised and
      expanded by Michael McElderry with the assistance
      of Scott McLemee</author>
  </titlestmt>
  <publicationstmt>
<date>1993</date>
<publisher>Manuscript Division, Library of Congress
</publisher>
<address>
  <addressline>Washington, D.C. 20540-4860
  </addressline>
</address>
  </publicationstmt>
  <notestmt>
    <note> <p>Edited full draft</p></note>
  </notestmt>
</filedesc>
<profiledesc>
  <creation>Finding aid encoded by Library of Congress
  Manuscript Division
    <date>1996.</date>
  </creation>
  <language>Finding aid written in
    <language>English.</language>
  </language>

```

```

</profiledesc>
<revisiondesc>
  <change>
    <date>1997</date>
    <item>Encoding revised</item>
  </change>
</revisiondesc>
</eadheader>

```

One advantage of the EAD header over the TEI header on which it is based, is that it makes use of the higher level elements and subelements. Because it is mainly focused on archiving concerns, it does not necessarily encourage nor require the exhaustive use of recommended categories that a TEI compliant body of text-based data requires. The focus of EAD is on metadata. It also does a commendable job of endeavoring to promote like standards to create comparable bodies of data. Still, it offers a similar array of choices like TEI, but unlike TEI, EAD does exhibit the judicious use of the standards, which demonstrates a TEI compliant protocol at its most exclusive. This is a valuable task in general, but it does not satisfy the task of describing a body of transcribed spoken discourse in the model developed here.

### 5.1.3 DCMI AND OLAC

Open Language Archive Community or OLAC is based on Dublin Core metadata initiative (hereafter DCMI), which is focused mainly on bibliographic functions of metadata. The DCMI is also committed to ease of implementation and maintenance. Its priority is to organize and categorize bodies of data in such a way as to allow language researchers to locate and retrieve them. It is based on fifteen elements that can be “refined” to add layers of complexity to an otherwise minimal encoding scheme. OLAC has expanded a few of DCMI core elements to cater to texts and collections of texts of natural language data for the purposes of linguistic research and thus added a few more elements into the procedure,

while maintaining the underlying goal of ease of implementation (<http://www.language-archives.org/OLAC/olacms.html>).

DCMI and OLAC compliant material centers on metadata and, accordingly, the header is the center piece in producing DCMI compliant material. The OLAC header is made up of twenty three elements. All of the elements are contained in one unifying element: `<olac>`. Of the twenty three elements, only sixteen are relevant to natural language converted into a text transcript. Eight of them deal with formatting, software, and other technical issues associated with a resource that would fall outside of the data typical of Language Variation research. Every OLAC header element is optional and every element is repeatable, allowing for maximum flexibility and varying levels of depth when describing a text. Every element has the option of the same set of attributes, which will be discussed en masse after discussing the header elements. OLAC does not try to be TEI compliant, although under TEI guidelines, which include a conversion DTD, almost any encoding schema can be made TEI compliant.

The first header element is the `<contributor>` element and it is used to record the person or institution responsible for contributing to the text in a secondary capacity, such as by providing funding or sponsorship that resulted in the creation of the text. The element does not identify primary authorship. In order to specify the nature of a contributor, a “refine” attribute is recommended.

The second header element is `<coverage>` which identifies a geographical range, relevant timeline, or jurisdiction for a resource. The next header element is `<creator>` which identifies a person or institution responsible for the creation of the text. This element is reserved for primary authorship of a text, as opposed to the `<contributor>`, which refers to persons or entities who play a secondary role in creating a text. This element has the option of a “refine” attribute.

The next header element is `<date>` which identifies a relevant date of a resource. In natural language texts this may correspond to the date the spoken event was recorded, or the date it was transformed into text whatever spatial event the date references, it is

recommended that a coded value be used, since it guarantees that a collections of texts can be arranged in chronological order.

The fifth header element is **<description>** which allows for a description of the content of the text. This element can contain a prose description or abstract, a table of content, or any other format that represents an account of the text or resource.

The next header element is **<format>** which identifies media specific formatting of a resource, as well as physical dimensions.

The next header element is **<format.markup>** which identifies a specific markup scheme referenced by a XML encoded text.

The next header element is **<identifier>** which gives a text or resource a unique, unambiguous identifying means of indexing and referencing.

The next header element is **<language>** which identifies the language of the content of a resource or text.

The next header element is **<publisher>** which identifies the individual or institution is responsible for making the text available.

The next header element is **<relation>** which identifies the relationship between texts, such as a part-whole relationship in which a transcription is a part of a larger corpus of transcribed spoken language, for example. This element is also used for version information.

The next header element is **<rights>** which identifies the individual or institution who has rights over the resource. This includes intellectual property, copyright or trademark information.

The next header element is **<source>** which identifies the origin of the resource from which a text is derived. For example, if a text is a transcript derived from an audio recording, then this element is used to convey that information.

The next header element is **<subject>** which provides a description of the content using keywords or key phrases.

The next header element is `<type>` which supplies a text or resource with a formal name. And the last element, which is related to `<type>` is `<type.linguistic>`. This element describes the content of the text or resource from a linguistic perspective.

All of the metadata elements can use the same four attributes throughout the OLAC encoding protocol: refine, code, scheme, and lang. A fifth attribute, langs, is used on complete metadata records. The refine attribute is used to more narrowly define or precisely qualify an element. The code attribute is a specific value given to a metadata element, which relates in a meaningful way to the encoding schema as a whole. Similarly, the scheme attribute identifies an XML encoding schema that constrains how the text in the metadata element is encoded. Both the code and the scheme attributes must be validated with OLAC to ensure resource or text transparency, which is essential in data sharing guidelines. The fourth attribute used in OLAC metadata elements is lang. By default, this attribute has the value of 'en' but whenever the language of the content of the text is something other than English, the lang attribute is recommended. Finally, the langs element is given to the whole text or resource in the `<olac>` element, the container of all the metadata elements in an OLAC compliant text. Again, the default value for this is 'en'. A basic OLAC header takes this form:

```
<olac>
  <contributor> [...] </contributor>
  <coverage> [...] </coverage>
  <creator> [...] </creator>
  <date> [...] </date>
  <description> [...] </description>
  <format> [...] </format>
  <format.markup> [...] </format.markup>
  <identifier> [...] </identifier>
  <language> [...] </language>
  <publisher> [...] </publisher>
  <relation>>[...] </relation>
  <rights> [...] </rights>
  <source> [...] </source>
  <subject> [...] </subject>
  <type> [...] </type>
  <type.linguistic> [...] </type.linguistic>
</olac>
```

```

<olac xsi:schemaLocation="http://www.language-archives.
org/OLAC/0.4/ http://www.language-archives.org/OLAC/0.4/olac.xsd"
langs="en x-sil-LLU">
<contributor>Smith, John L.</contributor>
<contributor refine="sponsor">National Science Foundation</contributor>
<coverage>India</coverage>
<coverage>19th century</coverage>
<creator>Bloomfield, Leonard</creator>
<creator>Linguistic Society of America</creator>
<creator refine="editor">Sapir, Edward</creator>
<date>circa 1950</date>
<description>The CALLHOME Japanese corpus of telephone
speech consists of 120 unscripted telephone conversations
between native speakers of Japanese. All calls, which lasted
up to 30 minutes, originated in North America and were placed
to locations overseas (typically Japan). Most participants
called family members or close friends. This corpus contains
speech data files ONLY, along with the minimal amount of
documentation needed to describe the contents and format of
the speech files and the software packages needed to uncompress
the speech data.</description>
<description>http://www ldc.upenn.edu/Catalog/LDC96S37.html
</description>
<format code="text/xml">5,237 entries in a 1.2M XML
file.</format>
<format code="audio/wav">Duration: 153 seconds. Size: 3.3M.
Sampling: 1 channel, 22 KHz, 8 bits.</format>
  <format.cpu code="ppc"/>
  <format.cpu code="x86">At least 64M memory</format.cpu>
  <format.os code="OS2"/>
  <format.os code="MSWindows">NT 4.0 higher</format.os>
  <format.sourcecode code="C"/>
  <format.sourcecode code="Java">Version 1.2 library </format.sourcecode>
<identifier>http://arxiv.org/abs/cs.CL/0010033</identifier>
<identifier>Shelf 12, Box 7</identifier>
<language code="en"/>
<subject.language code="x-sil-SKY"/>
<language code="fr"/>
<subject.language code="x-sil-BAN">Dschang</subject.language>
<subject.language>Ancient Sumerian</subject.language>
<publisher>Oxford University Press</publisher>
<publisher>http://www.oup.com/</publisher>
<relationrefine="requires">oai:sil:software/ipafont
</relation>
<relation refine="hasPart">oai:somearchive:holding126
</relation>
<relation refine="hasPart">oai:somearchive:holding127 </relation>

```



```

<relation refine="hasPart">oai:somearchive:holding128 </relation>
<relation refine="hasPart">oai:somearchive:holding129 </relation>
<relation refine="hasPart">oai:somearchive:holding130 </relation>
<relation refine="isPartOf"> In Joel Sherzer and Greg Urban (eds.),
Native South American discourse , 237--306. Berlin: Mouton.</relation>
<source>oai:somearchive:holding1023</source>
<source>Kwara'ae flora vocabulary extracted from Guide to the Forests
of the British Solomon Islands, by T. C. Whitmore. Oxford University
Press, 1966. </source>
<subject scheme="LCSH">African languages</subject>
<title>A Dictionary of the Nggela Language</title>
<title lang="x-sil-LLU">Na tala 'uria na idulaa diana</title>
<title refine="alternative" lang="en">The road to good reading</title>
<type code="Image"/>
<type.linguistic code="description/grammar"/>
<type.linguistic code="transcription/orthographic"/>
</olac>

```

(<http://www.language-archives.org/OLAC/0.4/olac.xml>)

This metadata header is a step in the right direction in terms of manageability. There is a manageable quantity of metadata elements and very little recursive nesting of subelements within each element in this protocol. It is the exact opposite of TEI. However, what it has going for it in terms of quantity, it lacks in quality. The categories themselves are very focused on the bibliographic aspect of categorizing a body of text-based data and are somewhat deficient in other areas such as content description. For example, there is only one metadata category that is available to describe a body of language data, the <description> element. For a body of transcribed spoken data in Language Variation research, it is necessary to provide a nominal amount of detail about the spoken discourse in order to provide some extralinguistic context necessary for empirical research endeavors. What kind of interaction does the body of transcribed data represent? Who are the individuals contributing to the spoken language? When and where did this conversation take place? All of these variables are important to document in describing the content of a body of transcribed speech. These metadata categories are also deficient in the area of editorial and revision information. Clear categories for defining who the researcher is, who prepared the transcription, as well as revision histories are important information in data transformation processes. This standard

does not accommodate this sort of information. All in all, this metadata protocol does not include the types of categories that are helpful in describing archival material, as much as it is in creating a bibliographic citation for a body of text.

#### 5.1.4 LINGUISTIC ATLAS PROTOCOL

In addition to the previously discussed standards, the Linguistic Atlas has created relevant metadata categories to suit the type of research the atlas projects are historically known for, and also to accommodate the type of extralinguistic information that Language Variationists require of their data. This metadata protocol follows the Dublin Core standards, which are based on fifteen elements designed to be both generic enough and broad enough to accommodate a wide swath of encoding endeavors and text types. The description of the metadata is consistent with that of Dublin Core and OLAC and thus the previous explanation of fields applies here. The recommended metadata categories for this transcription protocol are as follows:

```
<atlasheader>
<title> [...] </title>
  <creator> [...] </creator>
  <subject/keywords> [...] </subject/keywords>
  <description> [...] </description>
  <publisher> [...] </publisher>
  <contributor> [...] </contributor>
  <date> [...] </date>
  <resourcetype> [...] </ resourcetype >
  <format> [...] </source>
  <resourceID> [...] </ Format >
  <source> [...] </source>
  <language> [...] </ language >
  <relation> [...] </ relation >
  <coverage> [...] </coverage >
  <rights> [...] <rights >
</atlasheader>
```

A full version of the header looks like this:

```
<atlasheader>
```

```

<title>Atlanta Survey Pilot</title>
<creator>Iyabo Osipem, Nikki Kong</creator>
<subject/keywords>African American English, Atlanta speech
<subject/keywords>
<description>good</description>
<publisher>Kretzschmar, Lanehart</publisher>
<contributor>K B</>
<date>September 4, 2003</date>
<resourcetype>transcribed audio recording</resourcetype>
<format>atlas text encoding protocol</format>
<resourceID>1234</resourceID>
<source>Linguistic Atlas Project</source>
<language>English, idiolect from Atlanta</language>
<relation>interview related to early section of Western States
work sheets</relation>
<coverage>approximately 60 minute interviews in three parts:
conversation, direct elicitation, and perceptual elicitation</coverage>
<rights>copyright UGA, Linguistic Atlas Project</rights>
</atlasheader>

```

Like OLAC, this header provides mostly bibliographic information and is lacking in general categories that contextualize the body of spoken data as a whole. There is no description of the research, sociolinguistic information about the informant, or details concerning the audio recording, or revision history of the data transformation. As with OLAC, this header is too restrictive to fully describe the body of transcribed speech that follows.

### 5.1.5 LV HEADER

As illustrated in this survey, existing metadata standards operate on two ends of the spectrum: highly complex and fairly basic. Complex standards like TEI have incorporated a system of categories designed to dissect information into the most nuanced levels of organization, whereas simple standards like Dublin Core attempt to create overarching categories into a single-tiered system that includes the only the most basic necessities in terms of categorizing information for archival purposes. On the surface, it would seem as if the complex standards place excessive expectations on the researcher, but in reality both types of standards offer potential drawbacks for the typical Language Variationist interested in standard

metadata practices. While standards like TEI tend to overwhelm the user with the magnitude of categories, this profusion of optional metadata categories and refining attributes do offer choices for describing and customizing a header in a way that best fits a body of data. The question is: Does the average researcher have the time or faculty to pick and choose through the extensive inventory of elements necessary to best customize a header for their data?

More basic standards like Dublin Core may force a user to impose categories that do not prioritize the types of information that best describe traditional sociolinguistic data, or even worse, that do not seem to apply to the data at all. Thus, the other end of the spectrum can be just as frustrating as it neglects certain descriptive elements needed to represent their data in a meaningful, yet useful manner.

The core of the problem resides in the fact that none of the above systems was developed for the type of natural language data most common in the field of Language Variation: transcripts of spoken discourse. Even a metadata system that caters primarily to archiving information needs to take into consideration the unique aspects of a corpus of natural language data. After all, a transcribed sociolinguistic interview is a different type of text than a grammar on the Icelandic language. Optimally, a metadata standard should accommodate both types of text. Over the years, standards like TEI have revised and broadened themselves to include a very wide range of texts, but the results have been the creation of an ever-expanding, ever-complex and multi-layered system that on the surface looks reasonable but very quickly morphs into a awkward roadmap that loses you by taking a very intricate, drawn-out route to your destination. DCMI, on the other hand, is mostly concerned with making sure you find the route that will take you to a particular destination, although it may not be exactly where you want to go. The marriage of quality metadata in a reasonable quantity should be the objective here. The best place to start is by outlining the roles that metadata fulfills in the first place. Next it is necessary to translate these roles into a reason-

able number of categories that are maximally useful but that do not place too much burden on the researcher preparing the data.

Metadata is a central in the practice of sharing data, or preparing data in a systematic and transparent manner which gives primary and secondary researchers the ability to readily locate, access and utilize said data. Thus, metadata is used for locating archived information as well as for searching and retrieving information. It is used to describe and contextualize a body of information in meaningful ways and finally, it is used as a method of cataloging the treatment and preparation of a body of information. Simply put, these goals need to translate clearly into data preparation practices in Language Variation community. In order for this to be possible, it must be explicitly shown how the theoretical and practical goals of metadata intersect with the theoretical and practical goals of Language Variation research and analyses.

One practical consideration for the metadata protocol developed here has to do with the means of encoding information. In the spirit of converging toward the idea of adhering to standard practices, I have chosen to implement all encoding of information, including metadata, using XML. It makes sense to use an established system such as XML, as it requires a systematic and consistent treatment of encoding information, a necessity in the practices put forth in this dissertation. Another important consideration is the issue of quantity versus quality. It is important to provide good quality, descriptive metadata categories constrained to realistic quantity guidelines so as not to overwhelm and overburden researchers with front matter on one hand, but while also providing adequate coverage of meaningful information on the other. I recommend dividing the metadata into 4 elemental categories, each with 3–5 sub elements among them. The top tier elements are transparent and many of the sub elements will require little or no extra clarification or extended description in satisfying their objectives. The sub elements that deserve more extensive descriptions of information can include as little or as detailed an explanation as the researcher desires. Any constraints on quantity are at the metadata element and sub element levels only, and do not apply

to the descriptive information provided by the researcher to fulfill these elemental divisions. Descriptions and explanations at the metadata level should be as simple or as comprehensive as needed to adequately convey the information at hand, information essential for fully characterizing a body of data.

After establishing a quantity-based guideline, the next necessity is to look at the actual categories themselves, beginning with three categories that are not necessarily metadata proper, but necessary in declaring the header encoded text, satisfying the function of a header in a corpus of information that may become a part in a larger collection, and establishing an encoding schema in a text, respectively. The first category is used for identifying the header of a corpus by way of an overarching category used as a protocol declaration. It is the highest level of organization and as it encapsulates all of the matter in the header of a text. This is not so unlike the declaration methods used in both TEI and DCMI. As for the categories nested within header declaration itself, the first one needed is an identification value, which uniquely identifies the text as a whole unit of information. The second category is a traditional XML header link that declares and indexes the DTD and a style sheet or encoding schema, if one is used in the creation of the text. Including a style sheet or schema is traditionally a feature of electronic text creation, a secondary data transformation process that is discussed in Chapter 7; however, because a style sheet offers an explanation of an encoding protocol and guidelines for acceptable encoding tags within a text in a display capacity, it is a good idea to include a place to reference such a file for future transformation processes.

The next four categories describe the text in various capacities: First, there needs to be a category which includes basic bibliographic information. Following the bibliographic information, there must be a complete description of what the content of the text consists of, then a description of the text as data, and finally a description of the editorial history of the text. All of these elements are required. The metadata header in its simplest form looks like this:

```

<lvheader>
  <id> [...] </id>
  <link> [...] </link>
  <bibdesc> [...] </bibdesc>
  <contentdesc> [...] </contentdesc>
  <datadesc> [...] </datadesc>
  <editdesc> [...] </editdesc>
</lvheader>

```

The **<bibdesc>** element is central in creating a comprehensive archival record of the text. The record takes two things into consideration. First, it situates the archive in private context, creating an organizational structure for an individual research environment. Secondly, it readies a body of data to be a part of a larger network of archived information. The practice of making data available to other researchers for not only review, but to encourage data sharing as well, underscores the need for the inclusion of this type of information on a standard basis. Creating a bibliographic reference for data accommodates this vision.

This element includes several sub elements that provide essential bibliographic information: The individuals or institution responsible for creating the body of data, a general overview of what the body of data consists of, a date indicating when the file was created or made available and details about location and format of the data file. This information closely resembles copyright information, or other information that identifies a body of data as an individual or an institution's intellectual property. From a Variationists' point of view, this information is most often reserved for preparing research for publication or presentation. The subelements of the bibliographic description are the following:

**<creator>** individual(s) or institution responsible for creating the text.

**<overview>** brief summary of the body of data.

**<date>** a date indicating when the data file was created.

**<location>** the physical location of the data file.

**<format>** the format of the data file.

This metadata group is not so unlike the required bibliographic meta categories found in both TEI and DCMI. Both protocols adequately respond to the importance of creating a bibliographic reference for a body of information, and the <**bibdesc**> category put forth here is a merger between both protocols, providing more transparent subcategories to accommodate natural language data, as well as a manner in which to naturally group the bibliographic information as its own entity within the header.

The second element in the metadata header, or the <**contentdesc**> element, provides information about the actual content of the body of data. This important aspect of the metadata should be a more detailed account of what type of linguistic event the data represents, who, other than the researcher, is responsible for contributing to the linguistic events depicted in the text, a date pinpointing when was the linguistic event took place, the environment and locale where the data was collected. This section also provides general information about the methods used to gather the linguistic data including how individual contributors were chosen in both theory and practice. These metadata categories are specifically designed to accommodate Variationists' research. They depict the social information and extralinguistic information as well as methodological explanations that are important in contextualizing research and analyses, as well as provide details for other researchers looking for like data to make cross-linguistic comparisons. The <**contentdesc**> element must have the following subelements:

<**type**> indicates the type of language event depicted in the text, such as a sociolinguistic interview.

<**contributor**> identifies the individuals or individuals responsible for the language event depicted in the text, as well as any relevant social or extralinguistic information describing the individual.<sup>4</sup>

<**date**> pinpoints a date when the language event depicted in the text occurred.

---

<sup>4</sup>In order to protect the identity of the informant, it is recommended that initials or something other than the individual's name be used in the contributor field in the metadata header.



<**locale**> gives the physical location of the contributor at the time of the language event depicted in the text.

<**supplementary\_info**> provides an opportunity to relate any other relevant information concerning the contributor, conditions surrounding the language event, or other information needed for contextualization.

The third aspect of metadata deals with the technical issues surrounding the collection and transformation the data, or the <**datadesc**> element. Sub elements include a category identifying the nature of the data that makes up the body of the text, such as transcribed spoken English, while also including a category of sampling methods behind selecting contributors to the collection of data. Another subelement in this data section covers the means used for recording a linguistic event, such as digital recorder, analog recorder, etc. This section needs to outline how data was transformed into a digital medium after the initial recording for archiving and processing purposes. It should also catalog any application used as an intermediary step in the transfer of data from recorder to computer and the specifications therein. Finally, it is also important to provide an overview of the transcription conventions employed as well as any encoding conventions included to mark-up the data. This portion of the metadata would include information about the use of phonetic spellings and other intersections between exercise of the linguistic treatment of language data and the practical conventions employed in transforming data from speech to sound and from sound to text. Like the previous section of metadata, the use of more thorough explanations here is necessary to encourage methodological transparency and accommodate researchers seeking comparable data for their own research. <**datadesc**> contains the following subelements:

<**type**> identifies the original form of the language data.

<**sampling**> outlines any sampling procedure used in gathering data.

<**capture**> gives the details of the medium used to record or capture language data.

<**conversion**> provides information about the transformation of language data from sound to text, to include any transfer from recorder to computer, from one medium to another.

<**transcription**> provides a synopsis of the transcription or encoding conventions used to transform spoken language into a text based representation.

The final metadata category should document the process of data preparation in terms of creation of the text, and the editing and revision processes. This is the <**editdesc**> element. It is important to keep accurate records indicating who made additions, revisions and basic editorial decisions during the transformation of sound to text. This record keeping should also include dates that indicate when editorial modifications were made. This element contains the following subelements:

<**editor**> name of individual or individuals responsible for the editing the body of text.

<**revision\_dates**> list of revisions and dates made to a text during the course of editing processing.

The guiding principals behind the selection of these metadata categories are archiving and creating comparable data to accommodating data sharing in a Language Variation research environment. To accomplish this, a high degree of transparency is both needed and expected by a researcher using data not gathered and processed by anybody other than themselves. For language variationists, this principally entails knowing how information was gathered, what data conversions took place, what the general content of the body of language data is, and what the extralinguistic variables are that characterized the content as well. Focusing on practicality and economy, the metadata categories developed here provide the most useful information in order to set the stage for the language data that follows. The following example is the LV metadata header. A full demonstration of the LV metadata header, along with a body of transcribed data is available in Appendix 2: LV Transcript.

```

<lvheader>
  <id>atl9</id>
  <link> lvdttd; lvstylesheet </link>
  <bibdesc>
    <creator>Betsy Barry, Nikki Kong </creator>
    <overview> atlas style interview for Atlanta
    Survey Pilot under the direction of William
    Kretzschmar</overview>
    <date>February 14, 2003 </date>
    <location> Linguistic Atlas, University of
    Georgia </location>
    <format>transcript: text file, audio file: wav</format>
  </bibdesc>
  <contentdesc>
    <type> atlas style interview, one hour interview
    abridged Western worksheet, fixed-format
    elicitations, perceptual information</type>
    <contributor> JW, female residing in the suburbs
    of Atlanta, Georgia</contributor>
    <date> February 7, 2003 </date>
    <locale> Tucker, Georgia</locale>
    <supplementary_info> interview took place in interviewee's
    home and lasted approx 1 hour</supplementary_info>
  </contentdesc>
  <datadesc>
    <type>cda audio file</type>
    <sampling>pilot informants located via cold call
    from phone list</sampling>
    <capture> digital recorder</capture>
    <conversion> cda cd produced on recorder transferred
    to wav on computer via Sound Forge</conversion>
    <transcription> Lv protocol used with XML based tagging
    scheme, see DTD. Can be made TEI compliant</transcription>
  </datadesc>
  <editdesc>
    <editor> Nikki Kong, Betsy Barry</editor>
    <revision_dates> first edit: scribe
    normalizations editing; May 16, 2008</revision_dates>
  </editdesc>
</lvheader>

```

Like the metadata standards surveyed in this chapter, the protocol developed in this dissertation utilizes XML. The process of creating a transcript requires a method of encoding information. In the initial data transformation of sound-to-text, methods of representing

information extend beyond visually signifying words on a page, but also to representing necessary information that has no physical index with the audio recording of primary linguistic data, such as metadata for example. As a mark-up language, XML is a good choice for the task at hand for many reasons, the most important being that the practice of encoding information in XML, both metadata and body alike, must be approached in a clear, systematic manner and every encoding element must be instantiated in a valid way. There is no room for inconsistent, erratic treatment of information in XML as the mark-up language requires identifying information in a semantically clear manner. XML tags are defined according to their relationship with the text, their relationship to other tags, and what attributes and values they are allowed. XML allows for the creation of straightforward and meaningful ways in which to identify and organize information. Furthermore, XML is an established standard in the area of mark-up languages. It is not a novel system nor is it particular to one area of scholarship, a particular field or industry. It traverses all technology-driven areas of interest and has a large and diverse community of practitioners and experts. It is another means of employing unifying techniques in data preparation practices in empirical language research.

One important characteristic of XML is that it requires a formal declaration of all encoding elements and attributes. This formal declaration is an established protocol that outlines what information is encoded, as well as how that information must be represented via XML. This declared protocol is referred to as a Document Type Definition, or DTD. The DTD is a set of directions physically linked to a body of encoded natural language detailing what was done over the course of the transformation processes, it specifies the required and optional encoding elements that can be used in the metadata header and body of data, and the attributes allowable therein, and any notes that elucidate the process for the user. The DTD is the framework for the physical manifestation of the theoretical intent behind the data preparation principles. It is essentially the “guidelines” aspect of the standard practices put to use. The DTD is required of all XML compliant data, and it is referenced in the header of a body of language data; however, the DTD itself is a separate file that exists in

tandem with a corpus of annotated information. It must exist in the same general location as the corpus in order to be properly referenced.

In practical terms, a DTD should be all of the instructions that make up a set of guidelines. It should also be a transparent source of the theoretical suppositions that justify the creation of the guidelines. It should house notes that detail not only what methods were adopted, but why were adopted and how this effects the process. It details the specifics of the mark-up protocol used, what is “legal” in terms of particular tagging information, and how to make a valid corpus of language data that adheres to the proposed standards. It is essentially the point of convergence where theory meets practice in the realm of data transformation and data preparation practices. The DTD itself is a technical document and can appear complex. It has a particular syntax used when defining tags and their component parts that is unlike XML, although it is considered an integrated part of an XML document. As complex as the syntax seems, the basic element, attribute declaration is relatively simple. For example, here is the DTD portion of the body of this protocol:

```
<!ELEMENT lvheader      (ID+, TS+)>
<!ELEMENT ID            (#PCDATA | TS)>
<!ATTLIST who           (ID #REQUIRED)>
```

The left hand part of the declarations are the tag elements and the right hand part lists the tags that you can expect to find housed within that particular tag. Likewise, in an attribute declaration, the left hand side represents the declaration itself, while the right hand side represents the expected value of the attribute. In addition to actual declarations, the DTD created for this protocol contains an extensive amount of user notes, which offer condensed explanations of the guidelines. It describes what is being done and why it is being done. It also contains “template” structures that will allow the user to define their own basic set of encoding tags and attributes.

Once the discussion of the treatment of the body of language data concludes in the next chapter, I will then demonstrate all of the aspects of transcription protocol referenced in the

DTD. The DTD itself is included in Appendix 3: LV DTD. As all of the elements of the metadata header have been systematically described, it is time to examine the actual corpus of the transcript, the body of transformed spoken data itself.

## CHAPTER 6

### ANATOMY OF A TRANSCRIPT PART 2: TEXT BODY

Now that the header has been established, complete with required metadata for characterizing a body of spoken language, it is necessary to focus on developing a protocol for the body of transcribed speech. The manner in which the actual body of the transcript is prepared must be met in the same utilitarian manner that propelled the organization of metadata header: straightforward yet able to handle the complexities of dealing with representing spoken language. Furthermore, the body of the transcript must be prepared in a way that is theoretically transparent so that there is no question to the underlying factors that motivate this aspect of data transformation.

Just as forming and testing a hypothesis is the foundation of good research, formulating a clear definition of a transcript and exemplifying that definition in practice is essential in good data preparation. Thus before exploring methods of organizing transcribed linguistic data, it is important to reaffirm the definition of a transcribed linguistic event as a snapshot of language at a moment in time, and one that has cultural and historical import, as well as linguistic significance. As such, it is necessary to recognize that a transcript has several key purposes: First, it is one half of an archive, the other half being an audio recording, which houses culturally relevant information in both historical and linguistic contexts. The audio recording of a linguistic event and its written representation must be preserved in a manner that is amenable to any and all who value socio-cultural and socio-historical artifacts in their research. Secondly, it is an index that adds a visual layer to an otherwise aural experience. This is an extremely important fact to remember: The transcript does not usurp the audio recording of spoken language as the point of reference in a linguistic study,

but rather complements it in a manner that expedites the analytical process in research endeavors. Thirdly, as a tool that aids discovery in a research setting, a transcript must be suitable to both human eyes and computational processing. This is especially important when one considers their language data to be a part of a larger, communal corpus of language data, available to a wide variety of researchers in various fields. As this corpus of natural language data grows, so does the need for automated methods of processing, from searching to information retrieval, thus motivating data preparation practices and production to support such methods at the outset. In sum, the conventions that propel the creation of the guidelines developed here are that a transcript represents a record of spoken language at a moment in time, it also represents a body of text created as an index to an aural data source, should not be predisposed to linguistic inquiries only, and must accommodate both human and computational processing.

When preparing natural language data, the importance of clearly laying out objectives cannot be underestimated. Clearly defining methodological intentions is at the heart of establishing good consistent practices and guidelines. In creating a transcript, or any text based body of analyzable data according to a set of guidelines, such practices must be institutionalized in a way that allows others to clearly follow set procedures. In Chapter 4, I laid out some very general guidelines, accompanied by a theoretical discussion meant to inform the process of creating a transcript. From that foundation, it is now necessary to take the next step towards the actual transformation from an audio recording to a text-based representation of spoken discourse. It is time to move from theory to practice.

The first area that needs revisiting and expanding is the means by which recorded language materializes into graphical form: standard orthography. Standard orthography is a whole package of conventions including spelling, grammatical guidelines and punctuation, each of which must be taken into consideration when representing a speech event in writing. The first aspect of standard orthography I want to take into account is spelling. I have touched on the idea of alternate spelling in previous chapters but will now do so in outlining



how to specifically deal with it when transforming sound into text. In short, conventions of standard spelling should be followed in all instances. Importantly, transcribers should not attempt to approximate pronunciation with alternate spellings. After all, words, not units of sound, are the smallest denominator of a transcript and any attempt to represent speech via spelling is inconsistent with that tenet. In fact, there are several reasons why not to take liberties with alternate spelling qua pronunciation: First, it introduces an erratic element into the transcript that diminishes consistent treatment of the data. Secondly, it encourages an impressionistic treatment of the data that goes against the transcript as archive rather than linguistic inventory model proposed here. These types of judgments should really be reserved for another stage in the data transformation process. Finally, it risks imparting judgment on the interviewee whose speech is being transcribed as there is really no other reason to use an alternate spelling for a word that has a standard spelling counterpart other than to “mark” the speaker’s pronunciation as quaint or nonstandard (see Preston 1992). Consider the now familiar example:

So Margaret said, “I’m not talking to you,” she said “Celeste S., I’m talking to John Santorini.” So I says, “I don’t care who you’re talkin’ to. That angel has to hang part on my property, and I don’t want that goddamn angel.” So she says, “Well John, you do what you wanna do.” So John was afraid to say, “Yah, I’ll put the angel up,” ’cause he knows what woulda happen. So uh– she comes down walkin’ with the angel. So I says, “Listen, you want the angel?” She says, “Yeah,” I said, “Well you stand and hold it for Christmas. ’Cause there’s no angel gonna be put on my property.” So she got the angel and she put it on my pavement, and I got the angel and threw it after her” (Labov 2001, 390).

This example of transcribed speech was first referenced in Chapter 2, again in Chapter 4, and is being scrutinized again in this context because it fully demonstrates all sides of the problematic nature of alternate spellings. First, the transcription incorporates what linguists refer to as “g–dropping.” In reality, g–dropping is actually a phenomenon that occurs when

individuals replace the [ʔ]’ sound in a word final position with [n]. In this selection, the transcriber alternates between the forms *talking* and *talkin’* without any indication as to why. The reader is left to assume that the latter instance is a case of *g*-dropping, but this is never made clear. Also, the individual telling the story appears to have it in her own speech, but not when she is quoting somebody else. Again, this is speculation as Labov never clarifies the alternating spellings of the same form. He never indicates whether its inclusion is identifying phonological phenomenon or some sort of discourse characteristic the speaker demonstrates when telling a personal narrative. As mentioned previously, opting for this alternative spelling convention is little more than a source of confusion here, as it introduces the inconsistent treatment of a form but never properly orients the reader as to its function. The *g*-dropping feature, plus the other alternative spellings such as “*woulda*,” leaves the impression that the speaker’s speech is different, some particular dialect that is set apart, as it requires being made visually distinctive by way of nonstandard spelling conventions, the conclusion being that nonstandard spelling conventions must indicate nonstandard speech here. Because Labov never discloses any transcription conventions used in preparing his data, you are left to wonder not only about the accuracy of his impressions of the data, but the intentions behind his methods. In fact, in the Labov’s text, this selection is provided to demonstrate the vernacular of a woman that Labov has identified as a “leader of linguistic change” in her working class neighborhood, Celeste S. With examples of Celeste’s speech, Labov provides a socio-economic description of this individual, her working class background, a commentary on her vernacular and her ability to influence others in her social network (389–391). This taken in consideration with the transcribed speech Labov provided, the next logical leap is to make assumptions about the speaker. The use of nonstandard means of orthography here begs the question: “Is it really necessary?” The answer is no, it is not necessary nor is it constructive.

Granted, as spoken language is a spontaneous event, it is subject to dynamic phonological, morphological, and syntactic phenomena that make using standard orthography a challenge.

The Labov example offers one glance at what can happen when a transcriber takes liberties in order to remedy the limitations of standard orthography. There are other situations, however, that will present even more challenges during the transformation process. For example, there will be a number of instances during the transcription process where word-level utterances do not lend themselves to standard spelling representation. Some utterances are inaudible while others represent commonly occurring speech sounds that are not necessarily considered words and thus not subject to standard spelling conventions. Other utterances may represent neologisms, slang, borrowed word forms, jargon or lexical variants unique to an informant or group of informants. Plus, there are various word forms that are common in spoken language but not written language. All of these can be troublesome for the transcriber, but these potentially difficult situations must be dealt within the framework of standard spelling conventions. Data transformation must not be a process in selective representation. The use of standard orthography and standard spelling conventions in many ways exonerates the transcriber from having to have to decide “what’s legal” with respect to a word and its representation; however, it does require addressing these potential tricky situations and making decisions on the onset of the transcription process.

Utterances that are inaudible should be labeled as such after all attempts to transcribe them have failed. The practices recommended here marking inaudibilia with a descriptive tag: <inaudible>. Some speech sounds that are considered non word utterances, or interjections to be more concise, but are common in spoken language include: hmm, um, huh, uh, duh, and ah. Some transcription systems choose to leave out these types of utterances but for this protocol, all audible speech sounds must be represented. This requirement will discourage random “normalizing” on the part of the scribe, which would involve a selective process of representation. The goal in these guidelines is to minimize or eliminate selective representation. As far as audible utterances such as hmm, um, ah, huh, and the like are concerned, these forms have recently made appearances as dictionary entries in Random House, Websters and American Heritage, and they are considered legal; however, alternate spellings

are offered for some, and since consistency is central in transcriptions, a spelling convention must be predetermined and adhered to in all instances. Here are a few examples from the Atlanta pilot to illustrate<sup>1</sup>:

“Gas heaters and all of the the upstairs they were was uh downstairs there were the gas heaters and upstairs was a floor furnace.” “Elberton is just a small town um but I’m sure that it was either that or midwives uh or they did it on their own you know.”

“Oh just the um a stove, refrigerator, um.um.just regular pots and pans.”

Other sounds that frequently occur during the course of a speech event but are not considered speech sounds, such as coughing or laughing, can be included by simple tags that index their occurrences: <cough> and <laugh> respectively. Other types of sounds that fall into this category are grunts, mumbles, exhaling, throat clearing, sobbing, clicks, or vocalizations people might make while thinking or formulating a response to a question. In essence, any type of non speech sound is fair game for a unique tag that identifies it as such. The key is to be consistent throughout the body of the transcribed data. Some Atlanta examples include<sup>2</sup>:

“I mean it wasn’t like he <cough> was real mean to her or anything. he was just real funny.”

“I guess when central heat started coming along or gas heaters people thought ‘well oh we don’t need the fireplace anymore so we don’t use it’ I guess I don’t who knows <laugh>”

“What’s the matter with you people up here? <laugh> “Ain’t you ever heard of Southerner beforej’

---

<sup>1</sup>The following 3 examples are from an interview with JW, an informant in the Atlanta Pilot study, interviewed in 2003 the author and transcribed by Nikki Kong.

<sup>2</sup>The following 3 examples are from an interview with JW, an informant in the Atlanta pilot study, interviewed in 2003 by the author and transcribed by Nikki Kong.

Neologisms, slang, jargon, borrowed forms or otherwise unfamiliar and potentially undocumented words should be represented by creating a simple written form that adheres to the rules of standard spelling. Exact phonetic rendering can occur at later stages of analysis. Consider the following lines from a corpus of spoken language of informants from the state of Colorado.<sup>3</sup>

“Shipotes. What do they call them? Herons, maybe.”

“Some places call them a leppie.”

“What they called pottywurst.”

“Billy goats are just gutterers”

“But criminy ??? <inaudible> Can you afford to fly?”

“We had those up the wazoo.”

The first two examples are regional variants, neither of which has a standard spelling representation in dictionaries or encyclopedias. In these examples, the transcriber simply had to come up with a reasonable spelling that fit into the realm of standard spelling conventions. The third example described a German food product and is more than likely based on a borrowing from the language, but again, as no standard representation exists, a practical spelling was the best choice in order to transcribe the word onto the page. The fourth example depicts some type of specialized slang, probably regional as the fifth and sixth examples depict more widely recognized slang forms. All illustrate reasonable written representations of forms that accommodate standard spelling conventions despite their lack of standardization in most modern English language reference texts. Because there is no standard treatment of these forms however, it is recommended that the form is marked as such with a <ns> tag for

---

<sup>3</sup>Western Atlas data was graciously provided by Dr. Lamont Antieau and is featured in his dissertation “Colorado English A Distributional Analysis of Rural Colorado English.” Examples 1, 3, 4, 5, and 6 are from interviews conducted by Dr. Antieau in Alamosa, Colorado in 2001–2002. The 2nd example is from an interview conducted by Josephine Preston in Kremmling, Colorado in 1991. All were transcribed by Dr. Antieau.

two reasons: First, it identifies the form as being a product of questionable representation in the face of having no empirically verifiable resource or evidence to substantiate a standard spelling. Secondly, if at a point in time such a standard emerges, it allows the researcher to locate the form in question and replace it with an acceptable spelling. The treatment of these mostly unfamiliar forms is consistent with the commitment to consistent method of representation developed here.

Other forms to consider during the transcription process are contractions. Contractions are common word forms in both spoken and written language. Seemingly unproblematic, there are a number of contractions that are considered perfectly standard in the realm of conventional orthography. For example, modals and auxiliaries typically contract with “not” and forms such as “can’t” are more frequent in casual spoken exchanges than its counterpart “cannot.” There are other forms, however, typical of casual spoken speech events, that are extremely common but considered less standard and therefore need to be justified before including or excluding them from consideration. Examples of these are forms are “gonna” and “gotta” as representing “going to” and “got to,” respectively. In the transcription protocol developed here, the litmus test for determining each contraction’s status as legal or off limits is whether or not it appears as an entry in a recent edition of at least two the major English language dictionaries. Tables 6.1, 6.2, 6.3, and 6.4 present a comprehensive list of common contracted forms found in most dialects of American English and whether or not the form is an entry in Merriam Websters, American Heritage, Random House, Oxford American Dictionary, or Dictionary.com, respectively.

In addition to these contractions, all abbreviations such as Dr., Ms., and St. must be spelled out as doctor, miss, and street or saint, respectively. Numbers need to be spelled out as well.

Just as the use of standard orthography requires setting some spelling ground rules, it also necessitates the mention of grammatical conventions and suggested guidelines. In general, when we speak of grammatical conventions, we are speaking of the grammar of

Table 6.1: Positive Contractions.

Short form	Long form	MW	AH	RH	OA	OD
I'm	I am	Y	Y	Y	Y	Y
I've	I have	Y	Y	Y	Y	Y
I'll	I will, shall	Y	Y	Y	Y	Y
I'd	I would, should, had	Y	Y	Y	Y	Y
you're	you are	Y	Y	Y	Y	Y
you've	you have	Y	Y	Y	Y	Y
you'll	you will	Y	Y	Y	Y	Y
you'd	you had, would	Y	Y	Y	Y	Y
he's	he has, is	Y	Y	Y	Y	Y
he'll	he will	Y	Y	Y	Y	Y
he'd	he had, would	Y	Y	Y	Y	Y
she's	she is	Y	Y	Y	Y	Y
she'll	she will	Y	Y	Y	Y	Y
she'd	she had, would	Y	Y	Y	Y	Y
it's	it is	Y	Y	Y	Y	Y
it'll	it will	Y	Y	Y	Y	Y
we're	we are	Y	Y	Y	Y	Y
we've	we have	Y	Y	Y	Y	Y
we'll	we will	Y	Y	Y	Y	Y
they're	they are	Y	Y	Y	Y	Y
they've	they have	Y	Y	Y	Y	Y
they'll	they will	Y	Y	Y	Y	Y
they'd	they had, would	Y	Y	Y	Y	Y

written language. Until written language came along, the units of spoken language were subconscious aspects of speech event and the “structure of language” something innate. When written language was developed, the structure of language became an object of study (see Halliday 1989, Saussure 1915,1959). In theoretical linguistic circles, the structure of a language is its grammar and it is defined as the innate system of rules that allow users of a language to interpret words and phrases in their native language (O’Grady et al 2001). In this respect, a grammatical sentence is one that a speaker of a language deems a possible combination of words and thus well formed. This also implies that the notion of grammar

Table 6.2: Negative Contractions.

Short form	Long form	MW	AH	RH	OA	OD
aren't	are not	Y	Y	Y	Y	Y
can't	can not, cannot	Y	Y	Y	Y	Y
couldn't	could not	Y	Y	Y	Y	Y
daren't	dare not	Y	N	Y	N	Y
didn't	did not	Y	Y	Y	Y	Y
doesn't	does not	Y	Y	Y	Y	Y
don't	do not	Y	Y	Y	Y	Y
hasn't	has not	Y	Y	Y	Y	Y
haven't	have not	Y	Y	Y	Y	Y
hadn't	had not	Y	Y	Y	Y	Y
isn't	is not	Y	Y	Y	Y	Y
mayn't	may not	Y	Y	Y	Y	Y
mightn't	might not	Y	Y	Y	Y	Y
mustn't	must not	Y	Y	Y	Y	Y
needn't	need not	Y	Y	Y	Y	Y
oughtn't	out not	Y	Y	Y	Y	Y
shan't	shall not	Y	Y	Y	Y	Y
shouldn't	should not	Y	Y	Y	Y	Y
wasn't	was not	Y	Y	Y	Y	Y
weren't	were not	Y	Y	Y	Y	Y
won't	will not	Y	Y	Y	Y	Y
wouldn't	would not	Y	Y	Y	Y	Y

resides with the individual or individuals interpreting language, as well as those producing it. Importantly, in the linguist's model of grammar, the practitioners of a language sense what works and what does not with respect to combining words and phrases. Another facet of the linguist's model of grammar is that different groups of speakers exhibit variation in the realm of what is acceptable language usage. So, what is the norm for one group may exist outside the norm for another. Crucially, these aspects of grammar refer to descriptions of language use.



Table 6.3: Miscellaneous Contractions.

Short form	Long form	MW	AH	RH	OA	OD
here's	here is	N	N	Y	Y	Y
there'll	there will	Y	N	N	N	N
there'd	there would	Y	N	N	N	N
there's	there is	Y	N	Y	Y	Y
that's	that is	N	N	Y	Y	Y
that'll	that will	N	N	N	N	N
how's	how is	N	N	N	Y	Y
what'll	what will	N	N	N	Y	Y
what is	what is	N	N	Y	N	Y
when's	when is	N	N	N	N	Y
where's	where is	N	N	N	N	Y
who's	who is	Y	Y	Y	N	Y
who'd	who would	Y	Y	Y	Y	Y
who'll	who will	N	Y	Y	N	Y

The term “grammar” also refers to a codified set of conventions mostly viewed as the “rules” of language, which in effect prescribe correct terms language usage. This is quite different than the linguist’s model of grammar. This notion of grammar is not related to describing what is the norm in terms of language usage but rather prescribing a set of standards for what is acceptable and what is not. In this prescriptive model, the notion of what is well formed for spoken and written language is usually handed down by grammar books and the educational establishment, a few language “authorities,” rather than by the practitioners of a language. The prescriptive model of grammar is predicated on the idea of correct and incorrect language usage. In short, these two ideas about grammar and grammatical usage must be kept separate when transforming spoken language in to a written representation. The only definition of grammar that needs to be taken into consideration when transcribing is one that characterizes grammar as a set of subconscious expectations that speakers of a lan-

Table 6.4: Informal Contractions.

Short form	Long form	MW	AH	RH	OA	OD
ain't	am, are, is not	Y	Y	Y	Y	Y
gimme	give me	Y	Y	Y	Y	Y
gonna	going to	N	Y	Y	Y	Y
gotta	got to, got a	N	Y	N	Y	Y
gotcha	got you	Y	Y	Y	Y	Y
kinda	kind of	N	N	N	Y	Y
lemme	let me	N	N	N	Y	Y
wanna	want to	N	Y	Y	Y	Y
whatcha	what are you	Y	Y	N	Y	N
y'all	you all	Y	Y	Y	Y	Y

guage have that enables them to negotiate language and all its complexities in a competent manner. As with all things however, this is easier said than done.

In reality, underlying expectations can vary from speaker to speaker. Invariably, this can present challenges during the transcription process. If the transcriber's expectations are even the slightest bit different than the individual whose speech they are transcribing, this can tempt the transcriber into using their own underlying grammar as a barometer in which to record linguistic information. If a speaker splits an infinitive during the course of an exchange, then it should be represented as such regardless of whether or not the transcriber deems this as "grammatical." Or if a speaker says an utterance like "The baby sick," then it must be represented that way. Consider the following excerpt from a transcription from a speaker from Atlanta, Georgia<sup>4</sup>:

"Right now we trying to start our own bank."

In an example like this, a transcriber must resist the urge to insert an auxiliary, thus changing the utterance to "Right now we are trying to start our own bank," which would

---

<sup>4</sup>The following 3 examples are from an interview with GB, an informant in the Atlanta Pilot study interviewed in 2003 by Dr. Iyabo Osiapem and transcribed by Nikki Kong.

imply that the speaker failed to adhere to some prescribed grammatical convention. Likewise, consider this next selection from the same speaker:

“I told her that, but she don’t agree.”

Again, a transcriber must not take liberties with an utterance such as this and alternate “she doesn’t agree” in order to satisfy some prescribed standard. In instances where a speaker does not necessarily meet the transcriber’s expectations of language usage, the transcriber must accept and convey the natural usage without imparting personal judgment as to what they believe the utterance should be, rather than what it actually is. This would be an instance of the transcriber prescribing a grammatical rule onto the speaker, rather than authentically representing the speech event. A good rule to abide by is to represent only speech sounds and forms that are clearly audible in the recording and nothing else. Again, this is intended to minimize normalizing effects on the data by way of selective representation.

The third and perhaps most complex aspect of standard orthography is punctuation. Over the centuries, a number of symbols and other conventions have been introduced into writing and have come to characterize standards of punctuation. These include innovations such as spaces in between words, marking sentences boundaries, having upper case letters and lower case letters function systematically within a text, marking the status of a sentence, etc (Halliday 1989, 32–33). Various types of written texts make different use of punctuation conventions, but there are some conventions that hold steady across most genres. Words are separated by spaces. Reading sequences proceeds from left to right. Not maintaining these two conventions would all but render a text unreadable. Other basic punctuation practices include upper case letters used in sentence initial positions as well as to delimit the first person singular pronoun, place names, and other proper nouns. Sentence boundaries are marked with periods. Interrogatives are marked with question marks. Sentence–initial words, proper nouns, are capitalized as well as the first person singular pronoun. In American English, single quotations are used to indicate possession and contracted forms. Double quotations are used to indicate something someone other than the speaker said. These conventions

are essential in creating readable language, as eliminating them considerably diminishes the readability of a text as illustrated with the following passage<sup>5</sup>:

clarkston high school in clarkston georgia which is right outside of stone mountain  
 you can actually see the mountain from that whole area i got an afa degree an afa  
 certification a personal training certification from afa step reebok so i personal  
 trained and worked in gyms when i was growing well not growing up when i got  
 out of high school just to kind of figure out what i was going to do and i waited  
 tables and i made great money doing that kind of stuff.

Essentially, this passage challenges the reader's expectations of written language conventions. It is difficult to parse and interpret. Thus, to include no punctuation at all during the transformation process would not be a reasonable practice. As punctuation is an expected aspect of written language and required to facilitate readability, it is necessary to include it in the transcription process; however, we must establish appropriate punctuation conventions with respect to representing speech events in writing. In theory, this seems like an easy enough task as punctuation is such an inextricable aspect of written language. In practice, however, it is especially difficult to negotiate the intersection between punctuation, a custom that is so uniquely associated with written language, and transcribed speech, the graphical representation of a transient phenomenon that exists outside of the boundaries of such practices. The use of punctuation is further complicated by the fact that it tends to fluctuate according to personal taste and habit, as well as stylistic preferences. Therefore, putting forth some basic principles of punctuation is necessary in that it will expedite the transcription process considerably.

There are several aspects of standard punctuation that are problematic when transforming spoken language into written form. For example, rarely do speakers impose boundaries for neatly delimited sentences. And in writing, sentence boundaries are wherever there

---

<sup>5</sup>The following example is from an interview with CS, an informant in the Atlanta pilot study, interviewed in 2003 by the author and transcribed by Nikki Kong.

is a period. Reconciling this is can be difficult. Often times, a boundary simply indicates an utterance has stopped, rather than indicating a completed train of thought, or something that translates into a sentence as we would recognize it in writing. Another potentially problematic practice is using commas when transcribing sound to text. Traditionally, commas have been used as boundary markers in the case of clauses; however, when representing speech in writing there is a tendency to use commas to indicate a perceived pause. This practice can lead to overuse. Consider this example from a speaker in Wyoming<sup>6</sup>: We had a <hesitation> self-pollinating pear and a peach but we got them out of synchronization because we put them out in the summertime and then they'd go up and then we didn't let them get <false\_start> <hesitation> let them go dormant and then we'd bring them in and then that would force them into bloom.

This type utterance is fairly typical of spoken language. It includes the speaker hesitating in some instances, a false start in another instance where the speaker begins a thread and then shifts, a common occurrence that often happens mid-word, and a series of associated phrases appended together in order to complete a response to the interviewer's question. The only use of punctuation the transcriber includes is to indicate contractions and at the end of the interviewee's completed turn, which is marked as a sentence boundary. Sometimes during the course of an exchange, the end of an interlocutor's turn is the only thing that is easily identified as an unambiguous sentence boundary, and punctuating it as such is a perfectly fine practice when transforming spoken language into written language. Another thing to notice in this example is that the transcriber did not employ commas to indicate hesitations or pauses, but rather identified them as such. In the above example the transcriber rightly abstained from inserting commas to correspond to a perceived pause, or ellipsis to indicate a false start, both of which are also sound practices.

Other potentially problematic pieces of punctuation include the hyphen, which is used when forming a compound, dashes, which are used to demonstrate apposition, and ellipses,

---

<sup>6</sup>Western Atlas data from an interview in Ten Sleep, Wyoming conducted by Michael Madsen in 1988 and transcribed by Dr. Antieau.

which are used to indicate omission. The problematic nature of these are due to the fact that folks tend to use them according to personal habits or taste, use them erratically, or misuse them altogether. The key to punctuation guidelines developed here is to avoid extraneous punctuation by limiting use to what is only completely necessary and what is the least obtrusive when transforming sound to text. The fact is, including punctuation in a transcription of spoken language is always an act in pure interpretation on the part of the scribe. This should not present an obstacle in these guidelines however, as punctuation is a practice aimed at making a body of text readable. It is meant to facilitate human readability, easing reading and perusing text for human eyes. In text-to-text transformation processes which convert transcribed text into electronic corpora, the body of text is tokenized and punctuation is taken out altogether. Taking this into consideration, the following guidelines are suggested:

- Mark sentence boundaries where necessary to promote readability. It is not possible to equate pauses with punctuation, especially periods. In some cases, a pause followed by a shift in topic is a reasonable opportunity to insert a period but in most cases however, sentence boundaries should be inserted to give the reader an appropriate interpretation of the chunk of discourse. In the case of inserting question marks as sentence boundaries, rising intonation and the use of unambiguous grammatical forms, such as tag questions, *who*, *what*, *where* forms, and subject-auxiliary inversion present reasonable evidence of an interrogative.
- Use capitalization conventions for sentence-initial words, proper nouns, and the first person singular pronominal. Do not use capitalization to emphasize forms that are perceived being emphasized by the speaker.
- Use grammatical commas such as serial commas to promote readability, and to indicate city/state combinations, dates, and other instances where commas are traditionally inserted. Do not mark every overt pause with a comma. As with sentence boundaries,

there is no reliable link between a pause and punctuation, only a perceived index between the two.

- Do not use hyphens for compound words or otherwise.
- Do not use dashes or parenthesis.
- Do not use colons or semicolons.
- Do not use exclamation points.
- In keeping with American English practices, use single quotes for possessives and contractions only.
- Use double quotes when speaker is relaying a statement made by another individual to facilitate readability.

One final note on punctuation: There are a number of common occurrences in speech events which often tempt transcribers into creative uses of various punctuation marks. Like false starts and discernible hesitations, both of which have been previously demonstrated, interruptions often inspire some kind of graphical demonstration via punctuation. Rather than indicate these phenomena with dashes, a succession of commas, or ellipses, they should be marked as such with semantically transparent tags that accurately index their occurrences such as: <false\_start>, <hesitation>, <interruption>, respectively. By identifying these occurrences outright, it will discourage the practice of intermittently inserting some sort of correlating punctuation in order to depict them graphically. Identifying the occurrences will ensure a more consistent, transparent treatment of these type of phenomena. In creating a transcript using standard orthography, or any means of representation, an appropriate approach is “tread lightly.” This especially holds true in the protocol developed here, whereby a transcript is primarily a means of indexation to an audio recording of a data source, rather than attempting to be an “exact” record and thus replacement the original

speech event. This is not a reasonable objective anyway. A transcript is simply one version, one aspect of the recorded language data that facilitates readability and indexes information in a manner that is useful to both the researcher and to automated means of processing. In order to minimize the transcriber's effect on the transformation of sound to writing, it is necessary to try to identify the fundamental elements of a transcript, establish what information needs to be highlighted by the transcriber and what information needs to be supplied by the transcriber, and subsequently develop a modest protocol that incorporates this information without imparting too much bias onto the transcribed data.

The following offers a recap of the transcription conventions suggested in the protocol developed here, all of which are incorporated into the DTD:

- Use standard spelling conventions when transcribing speech. Do not attempt to approximate pronunciation with alternate spellings.
- In the case that you encounter unfamiliar word forms, apply standard spelling practices to come up with a likely spelling and mark the form with a <ns> tag.
- All audible speech sounds must be represented and inaudibilia should be labeled as such.
- Audible vocalizations must be represented with a transparent tag labeling them as such, for example <cough>, <laugh>, <exhale>, etc.
- Contractions should be transcribed as they are spoken in accordance with recognized forms that are considered “legal,” for example, gonna, gotta, etc.
- Apostrophes are acceptable for contractions and possessive nouns.
- Double quotes are acceptable as quotative punctuation. Avoid using double quotes to add emphasis.



- Spell out typically abbreviated forms, for example, if informant says “doctor,” write doctor instead of Dr., street instead of St., Avenue instead of Ave, Mister instead of Mr., Missus instead of Mrs., Miss instead of Ms.
- In the case of acronyms, if an informant says “tv,” “pta,” or “cia” specifically using that form in speech instead of the unabbreviated form, then include the acronym. Do not include periods to indicate acronyms such as “t.v.” or “p.t.a.” or “c.i.a.,” respectively.
- Spell out numbers, both digits and ordinals.
- Transcribe verbatim, without “correcting” perceived grammatical errors.
- Punctuate with care. For example, do not include commas to indicate perceived pauses. Commas are acceptable in serial lists, city/state collocates, date constructions. Mark sentence boundaries with periods to promote readability. Mark sentence boundaries with question marks where there is grammatical evidence of an interrogative: subject–auxiliary inversion, “wh” words, tag questions. Avoid using question marks to indicate perceived rising intonation.
- Include phenomena such as false starts, hesitations and speakers engaging in simultaneous speech with a tag identifying them as such. Do not use hyphens, ellipses, parentheses or other creative uses punctuation to indicate these types of phenomena.

These principles are put forth specifically to satisfy the theory of a transcript developed here while offering practical guidelines for transforming sound to text; So far however, these guidelines only address how to react to the audible speech sounds during the transcription process. There are other elements, however, that are required in order to satisfy the creation of a valid XML based corpus. Just as a declaration must be made identifying the metadata header with an appropriate <header> tag and the embedded element tags therein, so the body of the text must be labeled as such with a corresponding <text> tag that houses the

entire contents of the transcribed language. The basic structure of the corpus in its simplest form will look like this:

```
<lvheader>

    <id> [...] </id>
    <link> [...] </link>
    <bibdesc> [...] </bibdesc>
    <contentdesc> [...] </contentdesc>
    <datadesc> [...] </datadesc>
    <editdesc> [...] </editdesc>

</lvheader>
<text>
</text>
```

As the surveyed guidelines in the metadata chapter suggest, there are a myriad of approaches that can be adopted when representing the actually “body” part of a text-based body of data. In short, there is a large degree of variation in different approaches assigned to this task. For example, TEI focuses on a more structural organization of a transcription. This method is based on the assumption that language is a structural, systematic entity and therefore any encoding protocol should reflect that by offering as many encoding options as there are structural features of a language. This is not the approach that this protocol encourages. In fact, other than the guidelines already mentioned for dealing with vocalizations and utterances that lie outside of the realm of standard orthographic representation, there are only two additional encoding elements that figure into preparing the body of transcribed text. Of these two elements, one is accounted for in TEI and the other one is not. The first one is speaker identification. In an interview setting or in a speech event where there are different individuals are engaged in turn taking, it is necessary to attribute a string of speech to its speaker. This is accomplished by an identifying tag surrounding a string of text in order to attribute who is talking at a particular moment in time during the course of the speech event. The utterance is identified with a <u> tag and a ‘who’ attribute identifies the speaker of the utterance: <u who = “interviewer”>, <u who = “informant1”>, <u who

= “informant2”>. Traditionally, interviewer prompts and the elicited responses, known as turns, offer a discourse level of structural organization during the interview process. For this protocol, strings of speech must be attributed to the individual speakers from which they originate, but there is no emphasis necessarily placed on the prompt–response structural aspect of the interaction at this level of preparing the data.<sup>7</sup>

The identification and tagging of speakers is an essential element in transcribing a body of spoken text, but what is missing now is the element that provides the foundation for creating an enduring index between sound and text, and one that does not rely on categorizing or tokenizing a written linguistic form and mapping said form back to its recorded counterpart. This is one of the most important aspects of the protocol developed here, as it allows the researcher to refer back to a primary data source repeatedly, both before and during the analytical processes. Before talking about the method of indexing sound to text itself, a brief discussion on sound recordings is warranted. In order to create an effective, physical index between sound and text, audio recordings of distinct speech events should exist as independent sound files and these sound files should be in the same location or at least at an accessible location with respect to the transcribed language data. The key is to be able to access both sound files and text files from the same location, which can mean that they are either both housed on the same machine or that one or both are available on a server that is accessible via internet or intranet networking. Having the sound file and text together ensures that audio is fully available both during the transcription process and after its completion. Ideally, both the text and the sound should be archived in the same location, ensuring a physical link between the two mediums. Until now, the audio recording and the transcript have been treated like two separate entities, but creating and adhering to a good method of indexation fuses the two mediums together in a way that will hopefully ensure that the transcript is not treated as the primary data source in subsequent analyses. The purpose

---

<sup>7</sup>To protect the informant’s identity, it is necessary to delete their name and it may be necessary to delete any other identifying information, such as street addresses and the like.

of the index developed here is to make the audio recording an integrated feature of the transcript so that the transcript truly becomes an accessory to the primary language data.

In this protocol, sound-to-text indexation is accomplished by identifying and tagging regular time intervals during the transcription process of the recorded speech event, referred to here as time stamps. As the recorded speech event is a snapshot of a moment in time, and because the duration of the event exists on a timeline, using time stamps as indexing agents provides an effective, non-linguistically biased manner of preserving the relationship between sound and text. Because time is automatically tracked in an audio recording, providing a necessary navigational feature in an audio file, this makes timestamps a natural choice for establishing a robust link between sound and text. Using timestamps as indexes does not privilege the use of linguistically derived categories, so this method of indexing avoids predisposing the body of natural language to elements that would bias it for other types of research at this early stage in data transforming process. Consequently, types of data preparation methods that focus on forms as indexes, which would possibly include processes such as parsing or tokenizing, can be applied after a transcript is developed as another step in transforming natural language data, and a step that is geared towards more specific analytical purposes. Although narrow, analytically driven means of preparing data lie outside of the realm of the objectives stated for this preliminary stage in the process, the manner in which the transcription is prepared is still the necessary initial step in readying language data and indexing with timestamps will not hinder further data transformations that focus on more linguistically driven encoding techniques. Consider the following example from the Atlanta pilot:

“He started out in theological school and started out as a preacher, and then for some unknown reason to all of us he decided he didn’t want to do that anymore and never talked about it again.”

This selection represents approximately 12 seconds of spoken English. It is a small slice of speech in the context of the entire interview that can easily be indexed by the fact that

it begins at 00:01:17:00 and ends at 00:01:29:07. Of course, this is one of dozens of ways this selection could be indexed. It could be parsed and indexed according to traditional parts-of-speech. Now, consider the same selection<sup>8</sup>:

```
(S (NP He)
  (VP (VP started
      (PRT out)
      (PP in
        (NP theological school)))
    and
    (VP stated
      (ADVP out
        (PP as
          (NP a preacher))
        , and
        (PP then))
      (PP for
        (NP some))
      (NP unknown reason))))
```

The selection now is segmented into grammatical categories, which is yet another manner of representing the same string of speech; however, this represents another transformation altogether and does not belong in the initial process, the transcription process as now this selection seems primarily poised for some sort of traditional grammatical analysis only. In short, the method of time stamping provides a seamless point of reference to a primary language data source by relying on the temporal contour of a speech event rather than institutionalizing qualitatively created linguistic categories in the transcription process. Such categories should be reserved for secondary or tertiary stages in data transformation and in relation to some working hypothesis driving the analytical stages of research.

In theoretical terms, having timestamps as primary indexes offers a transparent, linguistically neutral way in which to satisfy linking two important mediums of language data within an empirical framework: Sound and its text-based counterpart. In practical terms, having

---

<sup>8</sup>This data is from JW interview and (see footnotes 1 and 2) was parsed using the Link Grammar Parser, which is a syntactic parser of found at <http://www.link.cs.cmu.edu/link/>, courtesy of Davy Temperley, Daniel Sleator, and John Lafferty.

timestamps as primary indexes between sound and text can alleviate potentially problematic situations that arise when transcribing a speech event, such as overlapping speech from multiple speakers, which can pose difficulties in protocols that rely on form-based indexes. When basing index on time, the occurrence of a phenomenon such as overlapping or indecipherable speech would not create an anomaly in the index, as the speech sounds themselves are not anchoring the protocol, but rather the temporal interval itself in which the forms occur. This timeline is providing the foundation. Thus, transcribing such a phenomenon requires ordered management in identifying speakers and best attempts at deciphering their respective speech, not to maintain the integrity of the index protocol, but to ensure consistent treatment of the data throughout this stage in the transformation process. Tags identifying overlapping and inaudible speech should be used to indicate when more than one speaker is talking at a given point in time and when it is not possible to understand an utterance, respectively. The inclusion of a timestamp will provide the key point of reference, taking the burden off of the transcriber or the researcher to organize the occurrence of multiple speakers engaging in speech simultaneously.

The signature for indexing an audio-to-text speech event with timestamps focuses on time in relation to a starting point, an ending point, and the duration of time that elapses on the timeline in between these two points. The starting point is aligned in the audio recording and text and usually is given the value 00:00:00:00 for each respective speech event. This necessitates that a single, complete speech event, such as one sociolinguistic interview for example, be treated as a corpus of language data. The speech event itself is what exists on a timeline, thus an audio recording, or recordings, for the speech event must reflect this by being arranged in a linear fashion, temporally speaking. If interviewing an informant spans several days and results in several audio recordings, then each interview segment represents a single, complete speech event correlating to a single timeline. Timestamps illustrate elapsing time in terms of hours, minutes, seconds, and milliseconds, respectively. They can be inserted at fixed intervals, such as 3 minutes for example, or by relying on opportune pauses in speech,

based on an approximate amount of time. A time stamp is identified as such by the tag <ts> and it takes a ‘when’ attribute that takes a value of a moment in time: <ts when = “00:00:00:00”>. The timestamp and the time designated in the attribute, aligns the audio recording and text in the transcript, situating both mediums on the same timeline.

The inclusion of speaker identification tags and timestamps are the only two other standard elements featured in the body of the language data itself. All other required tags included in the body of data have some relationship to an actual speech sound or vocalization present in the audio recording, such as inaudibilia or laughter, and their treatment has been covered in previously. At the transcript level, no other linguistic or non-linguistic features are identified and marked up. In keeping with the tenet that a transcript should exist in an archival capacity, and as a non-linguistically biased body of text, no other features are highlighted or marked up during this initial stage in the transformation process. In transforming large bodies of natural language data, the cost of identifying and annotating a large amount of extraneous information can possibly outweigh any benefits derived from excessive encoding. Not only does this practice make ensuring consistent treatment of the data extremely difficult, it puts inherent limitations on the research potential of a body of language data by imposing the scribe’s theoretical assessment onto the text. For these reasons, marking up extralinguistic and linguistic elements in the transcript are kept to a minimum. Further data preparation is recommended to transform the transcript from the archival stage into a body of linguistic data primed for various analyses. Any further identifying and encoding linguistic or extralinguistic elements can be added to the prepared transcript as another layer of encoding.

Unlike the metadata elements, the body elements devised here do not necessarily build on other encoding protocols but rather offer a way in which to minimally encode a transcript in an initial transformation from sound to text. As metadata standards are predicated on creating unifying ways to identify and describe data as a part of a comprehensive, ever evolving corpus of data, a survey of various guidelines was presented in order to situate

these guidelines in a larger research context. The idea to take other metadata standards into consideration was to employ useful existing elements and build on them, but also to make up for shortcomings in order to accommodate empirical researchers, namely Language Variationists, whose research focuses on natural language data. Thus the driving force in creating unifying metadata elements is to encourage data sharing amongst the linguistic community and beyond, which is largely facilitated by and dependent on strong metadata categories that will allow others to not only locate a corpus of language data, but to assess whether or not the data are compatible with their research agenda. Metadata elements should largely transcend various fields and research areas in order to simply characterize the body of text as a whole in a meaningful and useful way, which is a key element in preparing a transcript that will be made a public record, rather than the sole possession of an individual researcher. Also, metadata makes no attempts to impose a structural element onto a body of data.

On the other hand, methods of preparing the body of data itself remain distinctive in nature because they typically offer ways in which to impart structural organization onto a body of text. As the number of ways one can impose structure onto a body of natural language is extensive, so is the number of encoding categories created to facilitate this endeavor. Furthermore, for empirical researchers such as Language Variationists, there is another complicated part of the equation in the already complicated process of transforming spoken language into a text based representation: Standard transcription annotation systems and standard encoding protocols are not necessarily two sides of the same coin. In fact, these two annotation practices can be seen as existing at odds with one another. Transcription systems are predicated on the assumption that a transcript is a text-based structural account of a spoken linguistic event. Transcription systems typically render a body of natural language a very distinctive type of research data. By their invention, transcription systems are specific to a type of application that can read and interpret them, or individuals trained to employ, read and interpret them. They are not necessarily systems meant for wide, general



consumption. Generally speaking, modern encoding protocols such as TEI were developed as a method for the representation of texts in digital form, detailing encoding methods for machine-readable texts. Importantly, TEI was created for general consumption, not specific to any one application or particular individual or group of individuals. Also, transcription systems focus on transforming sound to text, while encoding protocols focus on text-to-text transformations. As it has evolved, standards such as TEI have set out to accommodate the same types of structural and categorical divisions that more linguistic-based transcription systems impose onto language data; however, there is an underlying assumption that any transformation that data undergoes will be at a text-to-text level. In the guidelines developed here, it is extremely important to understand that the encoding protocol is meant to accommodate sound-to-text transformations first and foremost, to have a protocol in place before a recorded language event ever transfers mediums, also to provide a foundation for text-to-text transformations. In short, these guidelines seek to unify the two systems by providing a text encoding protocol for the transformation of sound to text via transcription. Essentially, it is to be used to transcribe a body of data from sound to text, and as the foundation for further text-to-text transformations. Not only that, but in adhering to these guidelines empirical linguists are encouraged to rethink more traditional definitions of spoken language data, especially transcripts, by not conceiving of them as a highly particular form, meant to cater to a few specialists but rather a system that is developed to be wide-reaching, available to and employable by all.

Now that both the header and the body elements of the protocol have been presented and discussed, I will now provide an illustration of them used together in a transcript. This is a selection of an interview from the Atlanta survey project outlined in Chapter 3, representing about 10 minutes of a wav file. For a complete transcript, see Appendix 2: LV Transcript.

```
<lvheader>
  <id>atl9</id>
  <link> lvdtd; lvstylesheet </link>
  <bibdesc>
    <creator>Betsy Barry, Nikki Kong </creator>
```

```

<overview> atlas style interview for Atlanta
Survey Pilot under the direction of William
Kretzschmar</overview>
<date>February 14, 2003 </date>
<location> Linguistic Atlas, University of
Georgia </location>
<format>transcript: text file, audio file:
wav</format>
</bibdesc>
<contentdesc>
  <type> atlas style interview.</type>
  <contributor> JW, female residing in the
  suburbs of Atlanta, Georgia</contributor>
  <date> February 7, 2003 </date>
  <locale> Tucker, Georgia</locale>
  <supplementary_info> interview took place
  in interviewee's home and lasted approx.
  1 hour</supplementary_info>
</contentdesc>
<datadesc>
  <type>cda audio file</type>
  <sampling>pilot informants located via cold
  call from phone list</sampling>
  <capture> digital recorder</capture>
  <conversion> cda cd produced on recorder
  transferred to wav on computer via Sound
  Forge</conversion>
  <transcription> Lv protocol used with XML
  based tagging scheme, see DTD. Can be made
  TEI compliant</transcription>
</datadesc>
<editdesc>
  <editor> Nikki Kong, Betsy Barry</editor>
  <revisions> first edit: scribe normalizations
  editing first date: May 16, 2008 </revisions>
</editdesc>
</lvheader>

<ts when = "00.00.00">

<u who = "interviewer">Will you state your name and the date into the
microphone so that I can kind of just see if it's working? You don't
have to if you can just sit back it's got pretty good pick up.</u>

<u who = "jw">JW, February eighth, two thousand and three <cough> </u>

<u who = "interviewer">Wonderful. And this community is?</u>

```

<u who = "jw">Tucker.</u>

<u who = "interviewer">In the state of?</u>

<u who = "jw">Georgia</u>

<u who = "interviewer">And Tucker is um is a suburb of?</u>

<u who = "jw">It a suburb of Atlanta</u>

<u who = "interviewer">ok will you tell me maybe just about you family, about where you are from</u>

<u who = "jw">Ok umm well I'm from South Carolina, from Greenwood. It's a very small town and a cotton mill town. Uh we moved here when I was about six years old uh to Atlanta. My father was going to college in Greenville and after he left there uh he came to Atlanta to work and brought us here just a couple of months after he moved here and he got set up and got us here and I've been here ever since.</u>

<u who = "interviewer"> What does your dad do?</u>

<u who = "jw">It's really funny. He started out in theological school and stated out as a preacher and then for some unknown reason to all of us he decided he didn't want to do that anymore and never talked about it again. Um <cough> started work here at Riches and worked for Riches for oh I don't know maybe ten or twelve years and he worked in their carpet department at downtown Riches that was the only Riches there was and uh they had a really big carpet department and he worked there. Uh when he left there he stayed in the carpet business he went to work for uh an independent owner and he stayed in the carpet business until he retired <cough> which was some years ago he is seventy eight and now he lives in Columbia, South Carolina.</u>

<u who = "interviewer">What about your mom?</u>

<u who = "jw">My mother um she's dead. She died nineteen eighty three uh but she was basically uh a house wife she never took to my knowledge she may have had one job while I was growing up but I was very very small and really don't remember even what she did but she did not graduate from high school. She worked in the cotton mills in Green <false\_start> in Greenwood. Um she uh she married my daddy when she was seventeen he was in the navy and uh um they met through a cousin of hers and they got married and then came to Atlanta. Uh but no she was a house wife all those years.</u>

<u who = "interviewer"> What about your grandparents?</u>

<ts when = "00.03.07.03">

<u who = "jw"> My grandparents on my mother's side um her father was uh he worked in the cotton mills all her side worked in the cotton mills in Green <false\_start> in Greenwood. Um he was uh from Georgia in Elberton which is northeast, which is right over the South Carolina line, and uh my grandmother also was from Elberton. My mother was born in Elberton and then they moved into Greenwood later but uh my grandfather he'd <false\_start> he worked in the cotton mills in Greenwood and he worked some at Fulton Cotton Mills in Atlanta. Uh when she was young and uh my grandmother stayed at home. She was a housewife also. And on my father's side his uh his mother and father were from Virginia and they lived there for a while then moved into North Carolina into Rocksbourgh. It's a small town and uh a lot of tobacco farming up there and uh that was about it. My grandmother didn't work she was at home and my grandfather. Tell you the truth I'm not sure what all he did <laugh>.</u>

<u who = "interviewer"> So did they have a lot of kids?</u>

<u who = "jw">Uh yeah. On my father's side that's uh he has three brothers and one sister uh all living. And on my mother's side she has one sister and three brothers and she has two brothers left living and they're the only two left on my mother's side of the family.</u>

<u who = "interviewer">What about you and your siblings?</u>

<u who = "jw">Yes I have one sister and one brother.</u>

<u who = "interviewer"> and are you <false\_start> where do you fall?</u>

<u who = "jw">In the middle <laugh></u>

<u who = "interviewer">Oh in the middle.</u>

<u who = "jw">The middle child syndrome. <laugh></u>

<u who = "interviewer"> To <laugh> well um so you where exactly were you born in Greenwood did you say?</u>

<u who = "jw">In Greenwood</u>

<u who = "interviewer">In Greenwood. I'm sorry.</u>

<u who = "jw">That's ok.</u>

<u who = "interviewer">But you were pretty much you think that you grew up pretty much in Atlanta.</u>

<u who = "jw">Oh, from six years old. Yeah I've been here forty eight years. So I would say yes.</u>

<u who = "interviewer">Yeah, so would you consider yourself from here?</u>

<u who = "jw">Most of my formative years yes</u>

<u who = "interviewer">When you have a lot of people in South Carolina which is</u>

<u who = "jw">I still have a lot of family in South Carolina and North Carolina.</u>

<u who = "interviewer">And what about around here?</u>

<u who = "jw">Around here? Around here is just, is my immediate family. My uh all of my kids live here and all of my grandchildren live here.</u>

<ts when = "00.05.53.10">

<u who = "interviewer">So how many kids do you have?</u>

<u who = "jw">I have four children and five grandchildren.</u>

<u who = "interviewer">Wow. so you were married?</u>

<u who = "jw">Uh yea I was. I was married twice actually. My first husband we were married for about six years. We married right out of high school, and he uh he's kind of disappeared one day and just never came back again and we had three kids at the time we already had so I had three by the time I was twenty three. And then a couple of years later I remarried and we were married ten years and he died in eighty three. So and we had one more child so.</u>

<u who = "interviewer">That's a lot of kids and grand kids</u>

<u who = "jw">And then since then I said forget it <laugh>. Too busy for anything else.</u>

<u who = "interviewer">What about your kids what do they do?</u>

<u who = "jw">I have uh my oldest daughter works in a <false\_start> she's an analyst in programming, programs for uh hotels uh special promotions. Those kinds of things. Um my son oldest son is in landscaping. He likes to do his own thing keep his own business. He doesn't like to work for anybody. Then my next one is a boy too, well they are not boys anymore but he works for a company that pours cement. He drives a truck. He loves it. I know. He likes that kind of work. And then my youngest daughter is a programming assistant at Turner South so.</u>

<u who = "interviewer">oh. wow. <inaudible>.</u>

<u who = "jw"> Yes it's nice</u>

<u who = "interviewer"> Did they go to school?</u>

<u who = "jw"> Do they go to school?</u>

<u who = "interviewer"> Did they?</u>

<u who = "jw"> They all graduated they're all graduated high school. My youngest one was the only one that ever had any interest in going to college though and she went to Berry for four years.</u>

<u who = "interviewer">And what about you?</u>

<u who = "jw"> I graduated from high school. I did not go to college. I took some art courses and that was it.</u>

<u who = "interviewer"> You mentioned that your dad was possibly going to go into be a preacher</u>

<u who = "jw"> He was. He was. He was going to go into ministry </u>

<u who = "interviewer"> Did you go to church at all?</u>

<ts when = "00.08.39.05">

<u who = "jw">I did but not because of him. It was really interesting because like I said he when I was very young is when he decided he wasn't going to do that anymore and I just have vague vague memories of being in the church with my mother and father but after that he just stopped going at all. And I think the only time he every goes to church now if there's a wedding.</u>

<u who = "interviewer">Right</u>

<u who = "jw"> Yeah. So uh but I went on my own. My aunt, I spent time with my aunt in the summers, my mother's sister they were very close, and my aunt was about eleven years older than my mother so they were more like she more like treated her like her daughter. So we spent a lot of time in Elberton during the summers and holidays and I spent a lot of time vacationing my summers out from school in Elberton and in Greenwood with my grandmother, with my aunt, and my aunt was very big in the Methodist church. There were little small country churches but we always went with her. My sister and I spent a lot of time together up there with her. She always made us sing. She make us get up and sing. She played the piano in the churches, and sometimes we traveled around the different churches where they needed a pianist. So it was a lot of fun. Mostly the Methodist church growing up. Later I got involved in the Baptist church. My husband and I stayed in the Baptist church for about eighteen years, and then I just decided it wasn't for me anymore and in eighty five I became Episcopalian. </u>

<ts when = "00.10.28.00">

...

</text>

Although this selection does not demonstrate the entire range of markup possibilities that could present themselves over the course of a transcription, it is a solid example of the type of speech event that characterizes empirical linguistic research. It offers a glimpse of this type transcribed speech event using these guidelines. Although the transcript is encoded, it is still readable. It contains contextualizing information that may prove helpful in a variety of research settings. It is not a linguistic-centric body of information. It is not overwhelmed with highlighted linguistic information that would potentially add a layer of bias, excluding future analytical approaches to the data. It has undergone an initial transformation from sound to text in a manner that will accommodate more extensive, automated methods of data preparation. Once a transcript such as this has been established, the initial archive is in place and a researcher can go on to subsequent data transformations in order to showcase a lexical or syntactic feature or some other linguistic category to explore during the analysis stage of research. Using XML tags suited for the task of identifying features, an electronic text can be easily generated from a transcript using this protocol. The DTD referenced

in the transcript header and provided in the appendix should be copied and augmented to accommodate any further text-to-text transformations that are XML based. Again, the DTD provides basic template structures for adding such tags and attributes, as well as a location for documenting new and relevant encoding schemas for the expressed purpose of future text-to-text transformations.

The protocol for transcription demonstrates a viable method of transforming sound to text in an efficient manner that is manageable in terms of time and economy, important factors to consider in empirically based research. Since spoken language data is the driving force behind all empirical linguistic studies, it stands to reason that the handling and treatment of it in every stage of research of is carried out with purpose and consistency. These general guidelines for transforming spoken language into graphical form are meant to support an updated model of a transcription developed in this dissertation. Importantly, these guidelines are meant to steer the transcription process in a manner that best attends to the basic research needs of the Language Variation community and beyond, providing a practical method of handling language data in the post-collection, pre-analysis stages of empirical research. Again, the focus here is on providing a practical, consistent means for establishing a graphical representation of a spoken linguistic event that has previously existed only as an audio recording. It is not meant to provide a written inventory of linguistic features that characterizes a spoken linguistic event, an undertaking which would reach far beyond the function of a transcript in general and one that should be approached as a subsequent stage of data transformation, not an initial one. These guidelines are meant to bridge the gap between theory and practice, or more specifically, take into consideration the theoretical component of the physical task of preparing data, especially with respect to transforming spoken language into a transcript in a manner that does not unduly bias the empirical research process. In devising these basic principles, it is important to temper the abstract vision of a transcript with the reality of transforming empirical language, and importantly data typical of Language Variation studies in general.



## CHAPTER 7

### CONCLUSION

The research behind this dissertation began as practical exercise in context of a pilot project studying speech in a strict methodological framework. During the second phase of the Atlanta pilot project, the phase characterized by transcribing data and preparing it for various analyses, it became clear that the protocol we had developed for this task was not necessarily providing a clear, comprehensive set of instructions for the scribes to follow. This in turn prompted a personal investigation to see how other linguists in the field transcribed their spoken data, and importantly, why they chose to employ the methods they did. In short, I consulted the scholarship to find out “how” and “why” with respect to taking an audio recording of spoken language and transforming it into words on a page, or computer screen as it were. In my attempts to understand data handling and preparation practices in the field of Language Variation, I quickly noticed that the space in Variationist research between collecting data and analyzing it, the space that involves what I now refer to as data transformation processes, was almost completely absent in the literature. There was basically no methods-centered treatment of transcription processes, or data transformation processes of any kind, at the field wide level. Despite this fact, it was abundantly clear that transcribed data played a major role in the study of spoken language, especially in a Variationist framework. In the introduction of Sali Tagliamonte’s text *Analysing Sociolinguistic Variation*, the author states:

Most knowledge and learning in variation theory has been acquired like this, passed on through word of mouth, from one research to the next. In fact, it has

often been noted that the practical details of how to actually do variation analysis are arcane, largely unwritten and, for the most part, undocumented (ix).

Tagliamonte goes on to say that this inspired her to write a reference text documenting Variationist methods, to offer a solution for the fact that Variationist methodology, for the most part, had existed as a loose collection of practices and procedures developed over the past several decades for studying patterns of variation in spoken language. Similarly, this dissertation has sought to engage in a formal discussion on one of the least documented, yet pervasive practices in Variationist methodology: transcribing spoken language. Specifically, this dissertation is a response to the lack of field-wide attention given to a highly complex process, one that is a staple in Variationist data preparation practices. How we linguists operating in the Variationist tradition handle and prepare our data, the manner in which we literally transform our data from one medium to another in order to support our analytical and investigative efforts, has not only been “largely unwritten” and “undocumented,” as Tagliamonte puts it, but not even fully considered in a methodological context. This dissertation has sought to remedy that. This dissertation is meant as a general reference for Variationist research, a set of guidelines documenting and assisting with the complex task of data transformation in a consistent and transparent manner. At the very least, it is a contribution to a more principled treatment of data transformation practices, which have long been overlooked in the context Variationist methodology and research.

In this dissertation I have redefined the role of the transcript, taking into consideration the current technological environment which now wholly influences both the quality and quantity of empirical linguistic research. In redefining the role of a transcript, I also outlined general objectives of this data transformation process. With this, I then set out to devise a practical method for accomplishing a highly complex task, transforming sound to text. Now I want to conclude this dissertation by presenting the set of data transformation guidelines developed here as an organized protocol, a check list or package of sorts, which can be easily interpreted, referred to and applied by fellow empirical researchers, particularly those in the

field of Language Variation. I will start with the foundation representation practices that characterize the sound-to-text transformation guidelines:

- The unit of organization for transforming a speech signal to transcription is the word.
- The method of representing words, strings of speech, and chunks of discourse is standard orthography.
- When using standard orthography, use standard spelling conventions, consulting dictionaries when needed. If a form is represented in two or more major reference dictionaries, such as Random House or the Oxford American Dictionary, then it is considered “legal.” Do not use “eye dialect” or other novel representations to mark variation in the body of the transcript.
- All speech sounds must be represented, including forms such as 'uh', 'um' and the like. These forms are now attested in most major dictionaries, although spelling does vary. The important thing is to pick a spelling and stick to it.
- Unattested forms should be spelled with a reasonable approximation using conventional means.
- Abbreviations and numbers must be spelled out.
- Do not normalize grammatical information to adhere to a prescribed notion of standard language.
- Punctuation should accommodate human readability and not to provide a grammatical parse of a chunk of discourse to accommodate a prescriptive system of written language. Periods should be incorporated at reasonable sentence boundaries. Likewise, question marks should be incorporated at reasonable interrogative boundaries, marked with unambiguous interrogative forms. The use of grammatical commas is fine, but caution is required in comma use in general, as a pause does not indicate an opportunity for

this punctuation. The use of double quotes to indicate quoted speech is allowed. Do not use hyphens, dashes, parentheses, ellipses, or any other form of gratuitous punctuation as human readability does not warrant it.

- Use appropriate conventions of capitalization.

These conventions are meant to facilitate the consistent and practical treatment of representing a speech signal using standard orthographic means; however, in the model of the transcript developed here, there are instances where additional information must be provided by way of XML encoding. The first encoded information is the metadata categories found in the header, which is appended to the transcript ahead of the body of transcribed speech. Metadata encoding guidelines are as follows:

- The metadata header must be labeled with the root category tag `<lvheader>`. An id tag `<id>` with a unique identifying value is the first element within the metadata header.
- The second element within the metadata header is a `<link>` tag, which references any outside files, including the audio file using the attribute "audio" with the value being the path and file name to the primary audio data, and any other reference material being identified as such with a "reference" attribute with a value being the path to a file.
- The first metadata category is the `<bibdesc>` element, central in creating a comprehensive archival record of the text. The subelements are as follows: The subelements of the bibliographic description are the following:

`<creator>` individual(s) or institution responsible for creating the text.

`<overview>` brief summary of the body of data.

`<date>` a date indicating when the data file was created.

`<location>` the physical location of the data file.

<format> the format of the data file.

- The second element in the metadata header, or the <contentdesc> element, provides information about the actual content of the body of data. The subelements depict the social information and extralinguistic information as well as methodological explanations important in contextualizing research and analyses, as well as providing details for other researchers looking for like data to make cross-linguistic comparisons.

<type> indicates the type of language event depicted in the text, such as a sociolinguistic interview.

<contributor> identifies the individuals or individuals responsible for the language event depicted in the text, as well as any relevant social or extralinguistic information describing the individual.

<date> pinpoints a date when the language event depicted in the text occurred.

<locale> gives the physical location of the contributor at the time of the language event depicted in the text.

<supplementary\_info> provides an opportunity to relate any other relevant information concerning the contributor, conditions surrounding the language event, or other information needed for contextualization.

- The third aspect of metadata deals with the technical issues surrounding the collection and transformation of the data, or the <datadesc> element. The use of the following subelements require more thorough explanations, necessary to encourage methodological transparency and accommodate researchers seeking comparable data for their own research.

<type> identifies the original form of the language data.

<sampling> outlines any sampling procedure used in gathering data.

<capture> gives the details of the medium used to record or capture language data.

<conversion> provides information about the transformation of language data from sound to text, to include any transfer from recorder to computer, from one medium to another.

<transcription> provides a synopsis of the transcription or encoding conventions used to transform spoken language into a text based representation.

- The final metadata category is the <editdesc> element should document the process of data preparation in terms of creation of the text, and the editing and revision processes. These subelements include names and dates that indicate who made editorial revisions and when editorial modifications were made.

<editor> name of individual or individuals responsible for the editing the body of text.

<revision\_dates> list of revisions and dates made to a text during the course of editing processing.

The guidelines for encoding information in the body of transcribed speech are as follows:

- The body of speech must be labeled as such with a root category tag <text>
- Any inaudible speech sound must be labeled as such, with an <inaudible> tag.
- Any audible sound other than a speech sound, such as a cough, sneeze, laugh, must be consistently labeled with semantically transparent tags identifying them: <cough>, <sneeze>, and <laugh>, respectively.
- Certain phenomena that characterize spoken language, such as false starts, palpable hesitations, interruptions, should be marked consistently as such with semantically transparent tags that index their occurrences such as: <>false\_start>, <hesitation>, <interruption>, respectively.

- Any word form whose spelling is not empirically verifiable using a dictionary, neologisms, local place names and the like, should be represented with a reasonable representation within standard orthographic spelling conventions and should be labeled with the a <ns> tag.
- Speaker turns are identified with a <u> tag and a ‘who’ attribute identifies the speaker of the utterance: <u who = “interviewer”>, <u who = “informant1”>, <u who = “informant2”>.
- Time intervals should be established and properly indexed to the audio file of recorded speech using the tag <ts>, with a ‘when’ attribute that takes a value of a moment in time: <ts when = “00:00:00:00”>.

Generally speaking, these guidelines are a non-technical interpretation of the DTD; however, these guidelines should be consulted in tandem with the DTD, which is found in its entirety in Appendix 2: DTD.

One of the goals with providing this reference protocol is to assist and guide the decision making process in data transformation, and to demonstrate the breadth of issues that need to be addressed in the context of transforming sound into a body of transcribed speech in order to advance solid practices in the field and provide a proper foundation for quality research. These guidelines entail viewing a transcript not as a body of empirical linguistic data, not as analyzable text meant as a replacement the need to reference primary audio, but rather as an index between audio and a graphical representation that facilitates continual referencing of primary spoken data. It is important to remember that a primary goal of these guidelines is to provide a method for the creation of an enduring archive, one that will add a visual layer to the spoken event in a manner that situates it in a larger, ever growing repository of natural language data poised for general accessibility and future research. As such, these guidelines encompass the initial transformation process. While transforming audio into a graphical representation is a crucial transformation process, it is by no means the only transformation

process that supports the study of spoken language. In today's research, researchers also desire a body of analyzable text, a body of text-based data that will be the subject of high level computational processing or that will represent a more narrowly focused, linguistic-centered body of information that will be treated to some sort of quantitative analytical endeavor. In these instances, another data transformation process is required via the creation of an electronic text, or the creation of a corpus of information meant to accommodate such analytical endeavors. In fact, the model of a transcript developed here compels a second transformation process to create such a body of analyzable data. The initial data transformation process described in these pages sets the stage for others to come by laying the proper foundation for another tier of data transformation processes, the cornerstones of which are consistency and transparency.

The importance of consistency and transparency in data transformation processes should not be underestimated. Consistent and transparent practices contribute not only to a clear system of indexing audio and text, but both will expedite any future text-to-text transformations. The consistent practical methods employed in the first tier of transformation processes ensure a seamless move towards corpus creation, or the creation of any body of electronic data, much like the data preparation process described by Tagliamonte 2007. The transcription process is crucial in establishing an intimate relationship with one's data, as it paves the way for repeated examination of primary audio data necessary for moving forward to the next level of data transformation. In the event that second tier, text-to-text transformation entails identifying and coding particular linguistic features for example, a researcher would have to consult both the audio of primary data and text-based representation, using the transcript as an aid in navigating the audio file, in what would likely be multiple passes through the primary data. In the end, the repeated attention spent accessing the primary spoken data will afford a greater level of familiarity and understanding with respect to your primary data, which can only lead to overall improved data preparation practices, strengthening your research endeavors and broadening your understanding of spoken language.



As far as technical aspects of text-to-text transformations are concerned, it is recommended that the XML encoding schema in these guidelines be expanded and refined to assist in the development of electronic corpora. Consistency and transparency allow for the basic XML coding used in these guidelines to be employed for transforming transcripts into all manner of electronic texts, including TEI compliant corpora, or a general electronic database poised for tokenization or any other level of automated processing. Additionally, the DTD is easily modifiable to accommodate future text-to-text transformations.

The obvious next step in the research presented here is to examine these second tier transformation processes, to support moving toward creating a comprehensive picture of our data preparation practices and creating an inclusive point of reference at the field-wide level. This dissertation has taken a small, first step in the direction of defining and clarifying data transformation processes in order to properly acknowledge the importance of these practices not only in our methodology, but in the theories born out of a principled examination of our subject matter, spoken language.

## BIBLIOGRAPHY

- Bailey, Guy, and Cynthia Bernstein. 1989. Methodology for a phonological survey of Texas. *Journal of English Linguistics*. 22: 6–16.
- Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge UP.
- British Academic Spoken English (BASE) Corpus. <http://www2.warwick.ac.uk/fac/soc/al/research/projects/resources/base> (accessed December 11, 2007).
- Burnard, Lou. <http://users.ox.ac.uk/~lou/wip/metadata.html> (accessed July 8, 2008).
- Burnard, Lou. 1995. The Text Encoding Initiative: an Overview. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 63–68. London: Longman.
- Chafe, Wallace. 1993. Prosodic and Functional Units of Language. In *Talking Data: Transcription and Coding in Discourse Research*, ed. Jane Edwards and Martin D. Lampert, 33–44. New Jersey: Lawrence Erlbaum Associates.
- Chafe, Wallace. 1995. Adequacy, user-friendliness, and practicality in transcribing. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 35–53. London: Longman.
- Chambers, J.K. 2003. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Oxford: Blackwell Publishers.

- Cook, Guy. 1995. Theoretical Issues: Transcribing the Untranscribable. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 19–34. London: Longman.
- EAD Encoded Archive Description. <http://www.loc.gov/ead/eaddev.html> (accessed July 8, 2008).
- Edwards, Jane. 1995. Principles and alternative systems in the transcription, coding and Mark-up of spoken discourse. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 15–18. London: Longman.
- Everett, Daniel L. 2004. Coherent Fieldwork. In *Linguistics Today—Facing a Greater Challenge*, ed. Piet van Sterkenburg, 141–162. Amsterdam: John Benjamins.
- Halliday, MAK. 1989. *Spoken and Written Language*. Oxford: Oxford UP.
- Harold, Elliott Rusty. 2001. *The XML Bible*. 2nd ed. New York: Hungry Minds, Inc.
- Ives, Edward D. 1995. *The Tape-Recorded Interview: A Manual for Field Workers in Folklore and Oral History*. Knoxville: University of Tennessee Press.
- Jefferson, Gail. 2004. Glossary of transcript symbols with an introduction. In *Conversation Analysis: Studies from the first generation*, ed. Gene Lerner, 13–31. Amsterdam: John Benjamins.
- Johnstone, Barbara. 1999. *Qualitative Methods in Sociolinguistics*. Oxford: Oxford UP.
- Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. New Jersey: Prentice-Hall.
- Kennedy, Graeme. 1999. *An introduction to Corpus Linguistics*. Harlow, Essex: Addison, Wesley, Longman, Ltd.

- Kretzschmar, William, Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa Lena Opas-Hanninen, and Bartłomiej Plichta. 2006. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics*. 34: 172–205.
- Kurath, Hans and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.
- Labov, William. 2001. *Principles of Linguistic Change: Social Factors*. 2nd ed. Oxford: Blackwell Publishing.
- Lazinger, Susan. 2001. *Digital Preservation and Metadata: History, Theory, Practice*. Greenwood, Colorado: Greenwood Publishing Group, Inc.
- Leech, Geoffrey, Greg Myers, and Jenny Thomas, eds. 1995. *Spoken English on Computer: Transcription, Mark-up, and Application*. London: Longman.
- Linguistic Atlas of the United States and Canada. <http://us.english.uga.edu/> (accessed August 26, 2008).
- Linguistic Data Consortium. <http://www ldc.upenn.edu/> (accessed August 26, 2008).
- Maclay H. and C.E. Osgood. 1959. Hesitation Phenomenon in spontaneous English speech. *Word*. 15: 19–44.
- MacWhinney, Brian. 2000. *The Childes Project: Tools for analyzing Talk*. New Jersey: Lawrence Erlbaum Associates.
- MARC Understanding MARC Bibliographic Machine-Readable Cataloging <http://www.loc.gov/marc/umb/> (accessed August 11, 2008).
- Markee, Numa. 2000. *Conversation Analysis*. New Jersey: Lawrence Erlbaum Associates.
- McLemore, Cynthia A. 1991. The Pragmatic Interpretation of English Intonation: Sorority Speech. PhD diss., University of Texas at Austin.

- Michigan Corpus of Academic Spoken English (MICASE). <http://quod.lib.umich.edu/m/micase/> (accessed June 1, 2007).
- Milroy, Leslie and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.
- Monaghan, James. 1995. Whole-text analyses in computerized spoken discourse. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 54–61. London: Longman.
- Newman, Paul and Martha Ratliff, eds. 2001. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.
- Nofsinger, Robert. 1991. *Everyday Conversation*. Newbury Park, California: Sage.
- O’Connell, Daniel C. and Sabine Kowal. 1999. Transcription and the Issue of Standardization. *Journal of Psycholinguistic Research*. 28: 103–120.
- O’Grady, William and J. Archibald. 2001. *Contemporary Linguistic Analysis: An Introduction*. 6th ed. Toronto: Pearson–Longman.
- Ochs, Elinor. 1979. Transcription as Theory. In *Developmental Pragmatics*, ed. Elinor Ochs and B. Schieffelin, 43–72. New York: Academic Press.
- Open Language Archives. <http://www.language-archives.org/> (accessed August 26, 2008).
- Payne, J. 1995. The COBUILD spoken corpus: transcription conventions. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 203–207. London: Longman.
- Pederson, Lee. 1974. *A Manual for Dialect Research in the Southern States*. 2nd ed. Tuscaloosa: University of Alabama Press.

- Pederson, Lee. 1986–92. *Linguistic Atlas of the Gulf States*. 7 Vols. Athens: University of Georgia Press.
- Pederson et al. 1986. *Linguistic Atlas of the Gulf States: Concordance to the Basic Materials*. Ann Arbor: University Microfilms Int.
- Pederson, Lee. 1996. LAWCU project worksheets. *Journal of English Linguistics* 24: 52–60.
- Preston, Dennis. 1992. Ritin' Fowklower Daun 'Rong: Folklorists' Failures in Phonology. *Journal of American Folklore*. 95: 304–326.
- Preston, Dennis. 1993. *American Dialect Research*. Amsterdam: John Benjamins.
- Saussure, Ferdinand de. 1916, 1959. *Course in General Linguistics*. New York: McGraw Hill.
- Sinclair, John. 1995. From theory to Practice. In *Spoken English on Computer: Transcription, Mark-up and Application*, ed. Geoffrey Leech, Greg Myers, and Jenny Thomas, 99–112. London: Longman.
- Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Svartvik, Jan, ed. 1990. *The London Corpus of Spoken English: Description and Research*. Sweden: Lund University Press.
- Switchboard Corpus. [http://www ldc.upenn.edu/Catalog/readme\\_files/switchboard.readme.html](http://www ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html) (accessed August 26, 2008).
- Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge UP.
- Talkbank. <http://talkbank.org/> (accessed June 1, 2007).
- TEI Guidelines for Electronic Text Encoding and Interchange. <http://www.tei-c.org/Guidelines/> (accessed August 26, 2008).
- Telsur Project. [http://www.ling.upenn.edu/phono\\_atlas/home.html](http://www.ling.upenn.edu/phono_atlas/home.html) (accessed June 1, 2007).

## APPENDIX A

### APPENDIX 1: ORIGINAL ATLANTA PILOT TRANSCRIPT

<title>Atlanta Survey Pilot</title>  
<creator> Betsy Barry, Nikki Kong </creator>  
<keywords>LEAVE BLANK FOR NOW</keywords>  
<description>DESCRIBE QUALITY OF INTERVIEW HERE</description>  
<publisher>Kretzschmar, Lanehart</publisher>  
<contributer>JW, ADDRESS, PHONE HERE AS WELL AS  
AUXILIARY SPEAKERS</contributer>  
<date>DATE OF INTERVIEW, DATE OF TRANSCRIPTION HERE</date>  
<resource\_type>transcribed audio recording</resource\_type>  
<format>atlas text encoding protocol</format>  
<resource\_id>LEAVE BLANK FOR NOW</resource\_id>  
<source>Linguistic Atlas Project</source>  
<language>English, idelect from Atlanta</language>  
<relation>interview related to early section of Western States  
work sheets</relation>  
<coverage>approximately 60 minute interviews in three parts: conversation,  
direct elicitation, and perceptual elicitation</coverage>  
<rights>copyright UGA, Linguistic Atlas Project</rights>

I: State your name and the date into the microphone--

R: JW, February eighth, two-thousand-and-three (Cough).

I: And this community is?

R: Tucker.

I: In the state of?

R: Georgia

I: And X um, is a suburb of?

R: It-A suburb of Atlanta

I: ok will you tell me maybe just about you family, about where you are from

R: Ok..umm.. well I'm from X, from X. It's a very small town, and a cotton mill town. Uh, we moved here when I was about six years old uh, to Atlanta. My father was going to college in Greenville, and after he left there uh, he came to Atlanta to work and brought us here just a couple of months after he moved here, and he got set up and got us here, and I've been here ever since.

I: What does your dad do?

R: It's really funny. He started out in theological school and stated out as a preacher, and then for some unknown reason to all of us he decided he didn't want to do that anymore and never talked about it again. Um (clear throat), started work here at Riches and worked for Riches for oh I don't know maybe ten or twelve years, and he worked in their carpet department at downtown Riches that was the only Riches there was and uh they had a really big carpet department and he worked there. Uh, when he left there he stayed in the carpet business he went to work for uh an independent owner and he stayed in the carpet business until he retired which was some years ago he is seventy-eight and now he lives in Columbia, South Carolina.

I: What about your mom?

R: My mother, um, she's dead. She died nineteen-eight-three uh but she was basically uh a house wife she never took to my knowledge she may have had one job while I was growing up, but I was very very small and really don't remember even what she did, but she did not graduate from high school. She worked in the cotton mills in Green- in X. Um, she uh, she married my daddy when she was seventeen he was in the navy and uh, um, they met through a cousin of hers and they got married and then came to Atlanta. Uh, but no she was a house wife all those years.

I: What about your grandparents?

R: My grandparents on my mother's side um, her father was uh he worked in the cotton mills all her side worked in the cotton mills in Green- in X. Um, he was uh from Georgia in X which is northeast, which is right over the South Carolina line, and uh my grandmother also was from X. My mother was born in X, and then they moved into X later, but uh my grandfather he'd- he worked in the cotton mills in



X, and he worked some at Fulton Cotton Mills in Atlanta.  
Uh, when she was young and uh my grandmother stayed at home. She was a housewife also. And on my father's side his uh his mother and father were from Virginia and they lived there for a while then moved into North Carolina into Rocksbourgh. It's a small town and uh a lot of tobacco farming up there and uh-- that was about it. My grandmother didn't work she was at home and my grandfather-- tell you the truth I'm not sure what all he did (laugh).

I: So did they have a lot of kids?

R: Uh, yeah. On my father's side that's uh-- he has three brothers and one sister, uh-- all living, and on my mother's side she has one sister and three brothers, and she has two brothers left living and they're the only two left on my mother's side of the family.

I: What about you and your siblings?

R: Yes, I have one sister and one brother.

I: Huh, and are you-- where do you fall?

R: In the middle (laugh)

I: Oh in the middle.

R: The middle child syndrome. (laugh)

I: Toh (laugh)well um, so you where exactly were you born in X did you say?

R: In X

I: In X. I'm sorry.

R: That's ok.

I: But you were pretty much you think that you grew up pretty much in Atlanta.

R: Oh, from six years old. Yeah I've been here forty-eight years. So I would say yes.

I: Yeah, so would you consider yourself from here?

R: Most of my formative years yes

I: When you have a lot of people in South Carolina which is-

R: I still have a lot of family in South Carolina and North Carolina.

I: And what about around here?

R: Around here? Around here is just, is my immediate family. My uh, all of my kids live here and all of my grandchildren live here.

I: So how many kids do you have?

R: I have four children and five grandchildren.

I: Wow. so you were married?

R: Uh, yea I was. I was married twice actually. My first husband we were married for about six years. We married right out of high school, and he uh, he's kind of disappeared one day and just never came back again, and we had three kids at the time we already had so I had three by the time I was twenty-three, and then a couple of years later I remarried and we were married ten years and he died in 'eighty-three. So and we had one more child so.

I: That's a lot of kids and grand kids

R: And then since then I said forget it (laugh). Too busy for anything else.

I: What about your kids what do they do?

R: I have uh-- my oldest daughter works in a..she's an analyst in programming-- programs for uh hotels, uh special promotions. Those kinds of things. Um, my son, oldest son is in landscaping. He likes to do his own thing, keep his own business. He doesn't like to work for anybody. Then my next one is a boy too-- well they are not boys anymore but he works for a company that pours cement. He drives a truck. He loves it. I know. He likes that kind of work, and then my youngest daughter is a programming

assistant at Turner South so.

I: oh. wow. (I can't make out the rest of the sentence here).

R: Yes it's nice

I: Did they go to school?

R: Do they go to school?

I: Did they?

R: They all graduated they're all graduated high school. My youngest one was the only one that ever had any interest in going to college though and she went to Berry for four years.

I: And what about you?

R: I graduated from high school. I did not go to college. I took some art courses and that was it.

I: You mentioned that your dad was possibly going to go into be a preacher

R: He was. He was. He was going to go into ministry-

I: Did you go to church at all?

R: I did, but not because of him. It was really interesting because like I said he when I was very young is when he decided he wasn't going to do that anymore, and I just have vague vague memories of being in the church with my mother and father but after that he just stopped going at all, and I think the only time he every goes to church now if there's a wedding.

I: Right

R: Yeah. So uh but I went on my own. My aunt-- I spent time with my aunt in the summers, my mother's sister, they were very close, and my aunt was about eleven years older than my

mother so they were more like- she more like- treated her like her daughter. So we spent a lot of time in X during the summers and holidays, and I spent a lot of time vacationing my summers out from school in X and in X with my grandmother, with my aunt, and my aunt was very big in the Methodist church. there were little small country churches but we always went with her. My sister and I spent a lot of time together up there with her. She always made us sing. She make us get up and sing-- she played the piano in the churches, and sometimes we traveled around the different churches where they needed a pianist. So it was a lot of fun. Mostly the Methodist church growing up. Later I got involved in the Baptist church. My husband and I stayed in the Baptist church for about eighteen years, and then I just decided it wasn't for me anymore, and in 'eighty-five I became Episcopalian.

I: Hmm.

R: So--

I: So lets see, do you remember anything about maybe like your first wedding?

R: My first wedding?

I: Yeah

R: YES!

I: Ok are you going to tell me about it?

R: Yeah I remember. Actually we to uh-- we decided-- we started to planning a bigger-- uh, a big wedding and then decided we didn't want to do that so we ended up with just sort of a family thing but we went to Greenville, South Carolina, and we got married in a Baptist church in Greenville, South Carolina. We just took the family and went over there. It was not-- Greenville is not that bad of a drive from Atlanta and--

I: Did you have people up there with ?

R: Yes, my sister-- my sister was with us and her husband um, and you know our parents were there so. It was mainly just us. Its just this very small, but it wa- it was nice. It was a nice little wedding. That was-- That one! The other one, the second one where I got married the second time, my husband had never been married. Uh, and he and I decided we're going to have a church wedding. And so we had a larger type-- it wasn't hug,e but it was more with the bridesmaids and the grooms men and so we did the whole formal thing.

I: Right, right.

R: Which my kids refer to as the hippy wedding. We were married in 'seventy-three and it-- all the guys had long hair and the big moustaches-- even my husband had a (laugh)

I: Great pictures

R: Oh great-- (laugh) great pictures.

I: Ok, so I am going to talk a little now about ( thump)some female household stuff.

R: Ok.

I:It might be easier-- did you said that you visited your grandparents when you were real little?

R: Uh hmm.

I: Why don't you tell me a little about their place?

R: Ok. well, well my grandmother was by herself. My grandfather died the ones that-- on my mother's side the ones that I stayed with most. my grandfather died when I was four years years old. so I didn't-- I remembered him, but I don't remember their house that well, and after he died he'd been sick for a long time. he had cancer and they eventually moved in with my aunt and uncle. so my grandmother remained there. but the house was um-- it was um-- it's in X and it sat way back off the road and there was a smaller house in front of it that my aunt and uncle owned. they owned both houses, and my uncles' brother lived in the house in the front

so we were just back and forth. plus he had another sister who lived across the street so-- X is such a small town you know you just-- all you do is run next door you are-- right where you think you see everybody but the house was a two story house in the back and it had um, a small front porch but the thing that I remember that I always loved about it the most was that it had a front porch that went across the entire width of the house upstairs and it had the outside stairs that went up.

I: how neat

R: yeah it and it was-- it was great.

I: what about (clear throat) some of the rooms in the house maybe where everyone hung out

R: well mainly the living room. the house was separated into two areas because my grandparents lived downstairs so there was there was a living room and uh a smaller room and I can't remember what it served as then when they were there. And there was a kitchen on the back side and probably the smaller room served as their bedroom I'm sure and bath and then upstairs there was a room-- a front room like a living room and two bedrooms and one big room that was a kitchen and a back porch and the stairs that came in the back. So it was separated because of the way they-- because my grandparents were there but after my grandfather died and after my grandmother eventually moved back to X and it stayed with one of her sons there and my aunt and uncle took over the whole house and had the whole thing renovated so it became quite a large two story house but we hung out mainly in the living room downstairs and the kids mainly upstairs cause that was what we called the piano room. my aunt's piano was up there and we loved to bang on it and play and that was where we always slept on palates and pull out sofas and whatever and the attic fan was up there too so it was a lot cooler and the porch. and we love the porch so (laugh).

I: Right, what about some of the kinds of furniture you might find in the big living room?

R: oh, gosh um she had uh-- she just had regular furniture: sofa, a recliner, and the tv, and but there were radiators you know in every room or little gas heaters.

I: No central?

R: No central. No-- you come in from the cold and everybody was fighting to stand over the little gas heaters. we stood there all the time because you could turn them up

I: Is that the only way the house was heated?

R: uh hu

I: any other

R: gas heaters and all of the-- the upstairs they were was uh-- downstairs there were the gas heaters and upstairs was a floor furnace. you had to learn in the winter time. you got up in the middle of the night you had to remember to walk around the furnace because if you stepped on it. it was hot. the grates on top so.

I: what about in the family room where there any furnaces?

R: little gas heaters

I: that's interesting. what about like a fireplace?

R: no they didn't have any fireplaces in that house. The front house had fireplace in the living room. I can't ever remember it being used. seems like it was always blocked up. a lot of times back then in older houses they were blocking out the fireplaces with those tin things that go-- were kind of decorative. People just-- I don't know why. I guess when central heat started coming along or gas heaters people thought "well oh we don't need the fireplace anymore so we don't use it" I guess-- I don't-- who knows! (chuckle)

I: what about to cool it-- the place? was it hot?

R: you open the windows and turn the fans on  
(laugh) (cough)

I: what about the bedrooms? what kind of things would you find in there?

R: oh the bedrooms were there (clear throat) they were the most fun for us because the front bedroom had all of my aunt's things in it and she had uh what we call a chifforobe and one side of it was a mirror and the other had all of her hats. and that was the most fun because she would let us get into her shoes and hats and played our hearts' content so I think I tried on every hat she had. and she let us go though her jewelry box because too. and she wore-- because she was so big in the church you know she dressed up every Sunday. She always had a hat and she always had jewelry so she had lots of costume jewelry and lots of hats.

I: that's fun.

R: it was a lot of fun.

I: what about what you slept on in there.

R: if we were lucky we got a bed. as I got older and it was just my aunt and uncle there my sister and I slept in the front room where the porch was because the windows, uh the windows were down-- the big windows. You could just raise them up. They didn't have screens on them so you could climb in and out of the windows to get on the porch so that was lots of fun. So we always slept on the front, uh in the bed. otherwise if we-- if there was a lot of us there, we ended up sleeping on the floor. Just pallets-- just-- she had stacks and stacks of quilts. and if we stayed at my um-- my great uncle was a caretaker at the cemetery. and all of our family we have a huge amount of family in Elmhurst, in X. and my-- that house was the most fun because um, my great aunt and great uncle lived there and they lived back of the cemetery and it was a little, it was like three room house: uh a living room, a bedroom, and a kitchen and they-- she kept a bucket of well water on the back porch with a dipper and that's with no running water. we had an out house. there was one cow that wandered around out back. and um, you swept the yard-- didn't have grass um, because nobody had lawnmowers and they didn't upkeep, so the yard was



always dirt and the house stood up um, little uh pillars or-- st, st-- and you know you could crawl under the house. it was always cool underneath the house. but she had all these, this straw brooms tied together you know and we sweep the yard and in her house when we spent the night there it big feather bed and we would all sleep on it but you would roll towards the middle because everybody just-- so we all end up on top of each other But I can remember as many as five of us being in that bed in one time. Five of us kids.

I: That's neat

R: yeah that was really, really neat (laugh)

I: What about the kitchen in that house? how did they cook?

R: She had a wood stove-- a wood burning stove.

I: huh.

R: with a you know little things on the top of it with the handles you know you just open it up stuff wood back down inside of it and that's how she baked and uh they lived there until they died. and back then you didn't go to funeral homes either. whoever died was laid-- laid out right there in the living room usually uh or a bedroom, another room in the house and everybody came there and to view the body and that's where um the service was and that's what it was with my grandfather. I remember when my grandfather died he was laid out in the downstairs of the big house.

I: things are so different now. what about how they were born?

R: I'm sure they never went to a hospital. I'm sure never. Um, I'm-- I'm my mother-- I don't know if my mother was-- I don't even know if she was born in a hospital. I'm sure we had the conversation at some time and I can't remember.

I: did you know how that worked at all?

R: probably just a-- a doctor, if they could get a doctor, came to the house. X is just a small town

um-- but I'm sure that it was either that or midwives uh, or they did it on their own you know.

I: that's unbelievable.

R: I know it is. I can't even imagine not having a hospital to go to have a baby (laugh)

I: That's all very, very different. ok, what about let's talk about the kitchen a little bit. um, so uh what kinds of things might you find in the kitchen besides maybe you'd cook on ?

R: oh just the-- um a stove, refrigerator, um-- um-- just regular pots and pans and--

I: what kinds of things you remember uh, like your grandparents or your aunt and uncle cooking?

R: what kinds of food? well-- uh-- we grew up on beans (laugh). dried beans uh, and biscuits. my aunt cooked biscuits all the time-- every meal. my aunt and My mother. We didn't had a meal that we didn't either have biscuits or cornbread.

I: Homemade?

R: uh hu. oh absolutely. And I still make them, but don't do it all the time but my -- every once in a while I will. Yeah, and always tea. you always had sweet tea. never a meal without tea. um breakfast was always fried eggs and bacon and biscuits or toast for the most part. Sometimes oatmeal. We ate a lot of oatmeal. grits, ah-- of course grits! (chuckle) and always jelly on the table or preserves you know. preserves were a big treat because there was always somebody in the family who would put out preserves. And my grandmother, the one in North Carolina, she canned all the time more than my-- I remember my aunts doing here in X. But she canned all the time. She would put up all these kinds of beans and tomatoes and uh preserves, made preserves. She made her own butter.

I: Did she have a garden?

R: She did. they had a farm for a long time and

it just got smaller and smaller. They butchered hogs and had their own hams. they had a smokehouse in the back.

I: Really?

R: yea

I: huh. did you visit that and -- ?

R: Oh absolutely! yeah. I loved it. I loved to go to my grandmother's in North Carolina. She was so much fun.

I: what other kinds of things did she have in her garden?

R: oh she had okra, and tomatoes, and corn and beans, uh collard, um anything she could get-- squash anything she could get planted. she had quite a large garden. um and the smokehouse always had hams hanging in it. But that's where my grandfather hid. He hid in the smokehouse. He drank a lot and he drank--

I: What would he drink?

R: Moonshine. And um my grandmother would-- she was totally against alcohol. she didn't-- um didn't-- because he drank excessively, but he was really funny when he drank. I mean it wasn't like he (clear throat) was real mean to her or anything. He was just real funny. But she didn't like him drinking. He hung out in the pool hall a lot too. (clear throat) She use to have to retrieve him from the pool hall. But he did. and he would hide his bottles in different places and his smokehouse was his favorite place. He got there and drank. (laugh and clear throat). But in her kitchen we're talking about finding things in kitchen, she had an older kitchen too and converted over the years you know to-- to be more modernized but she had a safe-- a pie safe-- and I don't know if you know what a pie safe--

I: ummhmm.

R: ok. uh, and it had a flower bed in it and to me that was the most interesting thing because

living in Atlanta and growing up in Atlanta we would be more and more modernized over the years but my grandmother stayed the same and she always had that pie saving and always had the big flower bend in it. and it was really-- and she had some old things-- old dishes and old wooden rolling pins which I still have a old wooden rolling pin.

I: What about what would they drink in the morning?

R: Coffee.

I: How would they prepare that?

R: Um, different ways. Um, just one of those old coffee pots just a little you know had the little filter thing-- the metal? I don't know if you remember those and I don't see those anymore but just put water in it and the coffee went in on the top and I don't know how it did-- I hate coffee so I don't ever make it. And then later my mother bought a percolator which she thought that was wonderful (laugh). But those are the only two ways I can remember them making coffee unless we were on the lake and then God knows how they made coffee that, some weird, I mean-- threw it in the pot I guess. I don't know.

I: But what about you said that you always had fresh hams. What other kinds of meats do you remember eating?

R: Um, oh everything! we had every Sunday we either had uh fried chicken or roast-- roast beef. uh, a roast. and that was mainly yeah on Sundays. It was, it was uh-- I grew up on traditional Sunday dinners. Uh big. Everybody at the table. All kinds of vegetables, um, roast beef or chicken, or ham. Sometimes barbecue chicken. Um, and always a dessert and biscuits. Always. Always. My grandmother too. she did that. My grandmother would go out back and kill chickens and and cook.

I: What about eggs?

R: Uh huh.

I: Fresh?

R: Um Hm. Fresh eggs. Um, when, when we're at my grandparents cause they had chickens but, um, no in Atlanta no! (laugh).

I: right. what about dairy?

R: Um, well my great aunt and uncle had a cow out back so they always milked the cow. My grandmother in North Carolina probably did for a while until they sized down and then no longer had that ability but mostly the milkman brought the milk.

I: Do you remember drinking milk from the cow?

R: Uh, not straight from the cow no i don't remember that. I may have but I don't remember it.

I: Do you remember if the milk is different back then to what it is now?

R: Yeah, the milk had a different taste. Um, the milk that was delivered-- to me the milk was delivered by the milkman in the glass bottles was always the best milk cause we had milk delivered until I can't remember how long milk was delivered to our house. Even in Atlanta we had milk delivered. And I always loved going out and getting milk off the front porch because the milkman just left it there you know. And it was real treat when you got chocolate milk. But milk in a glass bottles that came in the glass bottles always tastes different and now it's all in plastic. Um, so it was-- yeah it was different.

I: Any other kind of dairy stuff you use to eat when you were a kid?

R: Just-- you know, just cheese. milk and cheese and nobody ate cottage cheese. If there was cottage cheese I don't know

if they ever had it then but nobody ate yogurt and cottage cheese (laugh) and now I don't even drink whole milk. I drink skim milk. I can't even stand the taste of whole milk anymore.

I: Right. Ok what about you said that your granddaddy worked at a smokehouse? What do you remember him wearing to go back there and work?

R: Oh, granddaddy always had on the same thing that I can remember. He just had on uh like um, what would now like be khaki pants but they were always darker, like darker work pants and a shirt and he always kept the button-- his button buttoned up at the top and they would be uh like a felt kind of shirt or flannel so I always saw the shirts and he always wore a hat.

I: what about in the winter to keep warm?

R: Oh, well I don't know what he wore I guess but-- uh, but, probably some kind of jacket but we always had jackets and coats. Mostly um, growing up you know we had-- you had a jacket to wear outside to play in and then a dress coat. You always had dress coat to go shopping or to go to church so--

I: Ok, what I think if you don't mind we'll do is this task that I have. And then we'll talk a little bit more it's going to seem kind of weird but if you could read-- I'm going to have you read the words in these cards. I'm going to shuffle them.

R: Ok

I: Get you to read them. Shuffle them again so--

R: (laugh)

I: we're going to do three tests and I'll put them there so you can see. Alright.

R: bit

R: hid

R: tech

R: hod

R: boot

R: toke

R: hood

R: but

R: tick

R: hate

R: duke

R: boat

R: took

R: head

R: tock

R: bet

R: put

R: hut

R: bake

R: bat

R: tack

R: teek

R: tuck

R: hoot

R: take

R: hode

R: had

R: heed

R: bought

R: beat (clear throat) (cough)

I: ok

R: uh bait

R: bought

R: heed

R: had

R: hode

R: take

R: head

R: took

R: boat

R: tuck

R: duke

R: hate

R: tick

R: but

R: hood

R: toke

R: boot



R: hod

R: tech

R: hid

R: bit

R: hoot

R: hut

R: put

R: bet

R: tock

R: teek

R: beat

R: tack

R: bat

I: ok last time.

R: hate

R: duke

R: tuck

R: boat

R: took

R: head

R: heed

R: take

R: tack

R: beat

R: hode

R: had

R: boot

R: bat

R: toke

R: hood

R: but

R: tick

R: teek

R: tock

R: bet

R: put

R: hut

R: hoot

R: bit

R: hid

R: tech

R: bought

R: hod

R: bait

I: Three??? ok. Alright now I want to talk about a little bit about how people talk or how you hear people talk. The differences.

R: Um,

I: What about the way people talk around Tucker vs. inside Atlanta? Do you notice the difference?

R: No, um I don't really because Atlanta has spread out in all directions and I-- when I was growing up, I would have said yes I could tell more of a difference the further out you get, but Tucker is so-- I mean we're twelve miles from Atlanta so it's not a little city by itself anymore. Um, but there are a lot of different cultures in Tucker. Um, I guess maybe more so than some other areas you know, areas are different-- so different around Atlanta but we have a lot of Orientals, Asians. Primarily Orientals and Asians so.

I: How do you differentiate between the two?

R: Orien-- well orientals-- no-- uh Chinese and you know and their words are more clipped, I guess. um, I don't notice any-- I don't even know if I can even be able to tell the difference if they were Japanese, but-- but maybe I'm not real sure. But a lot of Chinese. Um, and Asian as far as like middle eastern people and I don't always know where they come from. Sometimes I 'll ask them you know if I-- if we're in a store enough there's a-- right here you said you were at the Citgo a minute ago the man who owns it he and his parents own it and they're, they're middle eastern and I uh, I don't really talk to him a lot, but uh, um, my son does. My son goes in there and has regular conversations with-- he knows all about him um, which I think is really funny, but um, but that be how they differentiate I mean I know the difference in the way the people look as well as the way they talk what about-- I can just about tell where they're from you know Middle East, China or--

I: What about something like the way people talk in Atlanta vs. the way they talk in X?

R: Ha, ha now that-- now there you got a difference. Atlanta uh, I've noticed over the years we've gotten so many Yankeesuh, that uh, now that people now, even my-- even my own children which God forbid that everybody is loosing their Southern accent and it's a shame uh, I haven't lost mine and never will I'm sure, but uh, my kids even growing up in school they've been exposed

to so many different culture,s which is a good thing because I never was, but uh, they've been exposed to so many different cultures that I've noticed their language changes. they still have some southern accent, but, but not uh, like mine um, but if you but in X, you know people still have that slow southern talk, and sometimes it just drives you crazy. I mean my cousin still says "grass" which is grass, not like mine or "glass" you know she just drags out and she's still in X, but they've-- I've noticed since I've been in Atlanta that is bad as mine accent is, theirs is just way drawn out and then um, coming from downtown I can work downtown all day I, half the time I can't tell you where anybody is from anymore.

I: So you work downtown?

R: I do. I work in midtown.

I: Ahhhhhh.

R: so I go a lot, and we have um, a lot of uh, uh in our office-- our office is not real huge but everyone there is uh, um from the South except one women but we call her our token Yankee.

I: Where's she from?

R: She's from Ohio. And we just love to rib her. But her, her talk is-- to me her, her I guess the way she sounds, the inflections that she puts on words it's really different. But yeah you get quite a difference in the speech from say Atlanta to, to X. But they still talk like that in Elber-- you don't find many Yankees in X. (laugh)

I: So you think you can hear a difference I thought that you said in the way younger people younger people talk today?

R: Yes! Absolutely. Absolutely. Uh, they're not uh, the more they're exposed to the different cultures the more their language the way they sound changes. Um, I uh, it's really funny to me tha-- that uh I guess it's like I said I was exposed only to Southerners so we all talk the same and nobody knew any different you-- you didn't think you sound-- until-- unless you went up North and which was a rare occasion. I never visited the North until I was older.

I: Can people tell you're Southern?

R: Ha! I have cousins who live in Philadelphia. One of my aunts, my, my father's sister. when she married they moved to-- to-- Phili-- to Pennsylvania. Then they've been in Philadelphia all their lives and my cousins were all born there and I communicate regularly with one of them we take turns calling each other just because we don't get to visit. So we try to stay in touch and uh, I'll call in-- I'll call his work mostly because I'll be at work and that's when we communicate the most and um, even the people who answer the phone they-- they-- if I leave a message on the recorder they save the messages. He said, "They get such a kick out of listening to you talk." So they will save the messages so they can listen to my voice. (laugh)

I: What do you think of the way he talks?

R: Sounds like a Yankee. (laugh)

I: What is that you can pick up on? You think when you hear someone you can whether they're from the South or?

R: But most of the time, yeah, uh, sometimes uh, it's a little harder to tell especially people who have been through college. Uh, you notice them dropping more and more of that Southern accent. Um, I--

I: What do you think that is?

R: Well, I don't know. I work, I work with a girl uh, she's, uh, twenty-seven and she's from Mississippi and grew up in Mississippi and she tells me how hard, how hard it was for her to drop that Mississippi twain and I said, "W-- you know did you hav-- did you think you needed to?" and she felt like that she was-- she could communicate better or um, I don't know why-- articulate better? I, I don't know why. but it never occurred to me to drop my or to try to. Although they've tried to get me to up North they've, they've-- he said-- my cousin used-- (?) I came spent about three months up there uh several years back, and he said "Before you leave here, we're going to have you talking like a Yankee (laugh).

I: Did people notice a change when you came back here at all?

R: oh are you kidding?! (laugh) I didn't even get near a Northern accent. But uh, I don't no why I guess she just felt like that. she didn't want to be viewed as a-- and which is a huge Southern stigma I think that they view Southerners as rednecks or uh, just you know old country people which I think there's a big difference even in Southerners who are raised in Atlanta um-- when I grew up on the south side of town, so people talked more like I do and as I've gotten older I have friends who grew up in Atlanta, in the city, uh like in the Brookhaven-Buckhead area which that was still a classy area back then or what we use to call highfalutin'. uh, but I have a friend who grew up there and she has that more of what we called "Buckhead Betty" talk.

I: what's that like?

R: Ha! you know she talks like this. (chuckle) Um, but she lives in chester, south Carolina which is-- talk about po(I can't understand what she is saying there) little towns, but uh when she calls me, you know she still talks like that so uh it very-- it's very proper southern lady talk and not so, it's not so twangy as it is just drawn out. uh, its, it's really difficult to explain unless you've heard it. (chuckle)

I: Bbut you can pick up on it?

R: You can pick up on-- and her brother talks the same way.

I: Can you think of anything specific that's chewing you that maybe that's-- ?

R: Um-- I don't know maybe, maybe that you know like I said-- it's more-- I know that I have a draw but hers is more drawn out than a draw and, and the words are still very proper you know you speak very proper. They don't use a lot of uh, like little colloquialisms, little slangs, like uh, if you grew up on the south side of town or in a little bitty town, you know a little small town but it's more, I don't know, more of a, suppose to be more of a educated, proper way to speak in Atlanta when you grew up with money or grew up with the right kind of people.

I: Um hm.

R: Atlanta has very severe class distinctions.

I: So you think that causes a difference to education and money and speech?

R: I think so. absolutely.

I: So you think you can pick up on that?

R: Um hm.

I: What is it about your uh--

R: I can tell you the country club set from the ones of us who grew up playing in the ball fields.

I: Huh? really?

R: Uh huh

I: What about what you were saying um, they save your messages when you call-- who is it that you call?

R: Uh, uh, in Philadelphia. My cousin which the store he works in they say if I leave-- sometimes I've called there and the store is just closed or something or so I'll leave a message for him. They have an answering machine in their store and they save the messages so they thinking (chuckle)-- but they know who it is that called they say, "It's your cousin. Your cousin called."

I: and what is it about his voice do you think that's so different from--?

R: Yankee. Well I say water and he says wuter. (laugh) He says when are you ever going to learn how to talk? (laugh) I said, "well, not from you," (laugh). but just different words. I mean words are so different. And when uh I had the hardest time getting a glass of ice tea up there. It's the most ridiculous thing you've ever seen. Uh, I would go in and ask for tea and, and in, in a restaurant-- we've been in a restaurant and

he would look at me said, "they don't know what you're talking about." I said " they drink teas." He said "hot tea. they don't know what ice tea is or sweet tea is a sweet tea." I said, "Geeze, you grew up with a Southern mom." He said, "Well I know." you know?

I: So what did you did you get when you ordered it?

R: they never knew what I was talking bout. Nobody could ever make you may understand(?) and if I wanted a Coke, you don't get a coke up. There you get a soda. (laugh)

I: So what--

R: And nobody eats grits (laugh)

I: What do they have instead?

R: Oh uh, well I don't know like cream of wheat and cream of rice and those kinds of things you know. (laugh) I don't know.

I: Everything's different.

R: The food is different, but thank goodness, I mean my aunt was, like I said, she was raised in a Southern household so she still cooks Southern. (chuckle)

I: So he thinks the way you talk and the way Southerners talk-- he's says we're going to teach to how to talk. What do you think his ideas are about Southern speech?

R: Oh well, he's, he's just-- mainly he's just messin' with me. but um, he, he just uh he'll make funny me all the time. He just says we don't know how to talk. "You have to talk like a Yankee." But just different things I'll say. And he'll "hi" and I'll say "hey" uh, stuff like he always makes a distinction where I don't really notice-- I don't pay that much attention to it but uh, he pays more attention to those kinds of-- the differences in the words and my other cousins do too up there.

I: Is that the only Northern-- ?

R: That's the only Northern relatives I have.



Bless their hearts.

I: Interesting. They like the sound of your voice enough to say that--

R: Oh yeah he's take me out. Like I said I spent about three months up there and we would go out to different places or we would go out to some of his favorite places to hang out and uh, all the guys whe, when the-- standin' around and made me talk forever just cause they wanted to listen to me talk and I was-- "What's the matter with you people up here? (laugh) Ain't you ever heard of Southerner before?"

I: Do you know the English language?

R: I guess not.

I: Do you notice in the way that men and women talk?

R: Um-- in the South?

I: Sure.

R: Um-- it's just-- some, you know, sometime-- I don't know um, it's, it's hard to say I guess it just depends on where they were raised, uh, where you were born, where you grew up but the uh, i still get together with all of my high school buddies, and they all sound the same to me. They sound the same like i do. (chuckle) I haven't noticed any difference in their speech.

I: where did they move off? so they still live here?

R: uh huh. And most of them are still livin' on the South side of town. I'm the only one that came north.

I: What is the South Side?

R: South Side? South Side-- everyone knows South Side. you could grew up here-- uh, it's all of the uh, um, Southeast Atlanta, Hapeful, College Park, East Point area. We had a uh-- it was like one big huge family down there. Um, I went to school-- I was in at an Atlanta school but um you know we, we always hung out with people from College Park High School, and Hapeful High School, Russell and East Point all area, so that was we consider the south side.

Course then we didn't know it was South Side. We just thought we were in our area until I-- you know until people started growing up and getting older then it's always referred to as the South Side.

I: And everyone you went to high school with sound the same to you?

R: Yeah yeah. Yep, they all sound the same. I haven't notice any changes in any of them.

I: What about the way um, do you notice people with ethnic backgrounds-- can you tell the difference in the way they talk? You spoke of being able to tell someone of Middle Eastern what about black folks and white folks? Can you tell the difference?

R: Uh--

I: So if someone were to call you on the phone you'd think you'd be able to?

R: Most of the time yes, especially if they grew up in Atlanta. I've noticed um, the black people over the years uh, there's no differen-- the older black people that I still come in contact with they still have the ol' South talk and talk like they always did. Um, then ther-- the-- then you have the younger ones who have their own language going: the Ebonics and um, half the time you don't know what anybody's saying. and then I, I work in a building where we do have a lot of black people and they're very well educated and their speech is uh, um, um you can tell that they've been to college or that they grew up in the North. Not so much of a Yankee accent as just more uh, more cultured. um, I don't think that you can get that in the South.

I: Where do you work?

R: Um, I work um, at one GA center um, it's right-- do you know where the Varsity is?

I: Oh yeah.

R: It's right there

I: What kind of work is it?

R: Oh I work for a small industrial real estate firm. We do-- we deal with warehouses. We have a uh, a lot of brokers. And I'm the business manager.

I: So you have a lot of different types of people. You said you had some from Ohio.

R: (giggle) Yeah, everybody there though is from the South we have some South Georgia boys. We one South Georgia boy who grew up around Columbus and went to school in Alabama.

I: Do you notice a difference in how he talks?

R: He's um, he's, he's a good ol' Southern boy, um, but he's very well educated and uh, and you can tell you know and he uh but's he's, he's, he still has the Souther-- ness-- he has the Southern draw to some extent.

I: So maybe you can tell where if people were from souther states you can think that I mean you obviously can hear in Southern speech: the buckhead betty vs--

R: (Chuckle) That's from growin' up here.

I: Do you think you can tell differences in-- is Southern speech just one big thing to you?

R: I-- eh-- yeah uh, pretty much. It seems like, seems like Tennessee. Um because I grew up in Atlanta it's, it's a little, I think it's a little different if I still-- if I-- if it had grown up in a small town and we're still there I probably able to notice the difference even more but, you know I'm exposed to the different cultures everyday that um, to me it's kind of hard to tell anymore. In Tennessee, different parts of Tennessee. Um, people who come from larger cities there-- there speech is becoming more refined and more cultured um, the-- the older the kids get-- I mean you know if the kids grew up in a larger city I think they'd become more and more cultured, and, and like I said exposed to the different cultures so they're speech changes. Um, so it's a little harder to tell anymore, but, but all I have to do is go to South Carolina and visit relatives

or, or ju-- you know small towns in south Carolina to, to (snap fingers) catch it again, to catch the difference in the speech.

I: Do you find yourself more southern when you go to those places?

R: Oh absolutely. Oh, I go to-- I can spend a week in X and come back to work and they'll probably look at me and say, "WHAT? What are you talking about?" Cause it, cause, you know they already, they already know at work that I have (clear throat)I have probably a more southern draw than anybody in the office because most of the people in the office are co-- ed-- they went to college so they're, they are college educated and uh, and, and then we uh, uh have pretty much of a younger crowd in out office some of the older people um, but like our owner went to Kentucky and he'd, he was, uh, he's from New Orleans but you cannot even tell there's no way (clear throat).

I: Cause New Orleans--

R: New Orleans is, is very distinct.

I: What did it sound like to you? What does New Orleans speech sound like?

R: Oh my gosh it's a , it's a cross between uh southern, uh southern white, black, and French and-- (laugh) and God knows else was thrown in there but uh, it's, it's uh, uh, half the time I can't-- a real New Orleans person I can't understand a word they're saying uh or if you watch a movie on tv and, and it's you know feelin' the New Orleans and your-- it's suppose to be New Orleans people and can't understand them and, and I don't know where they-- it's just such a mixture they whole Cajun thing but I think it's fascinating. I just love to hear them talk

I: I do too. It's one of my favorites. Have you ever been around New Orleans?

R: not--

I: To (a?) little bit?

R: No, not, not really out of New Orleans. I've only been

to New Orleans a few times and uh, Biloxi um, d-- you know that area. Only Biloxi once and Biloxi I mean it's such a big gambling place now. You-- who would know if anybody was from Biloxi or not. Um, but uh, New Orleans that-- no I've never gotten really out into-- to be able to experience that-- I bet it is (laugh). I bet it is (laugh).

I: Well I guess that's just about it. We have about a minute and a half left on the tape.

R: Oh, ok.

end conversation

## APPENDIX B

### APPENDIX 2: LV TRANSCRIPT

```
<lvroot>
  <lvheader>
    <id>atl9</id>
    <link> lvdtd; lvstylesheet </link>
    <bibdesc>
      <creator>Betsy Barry, Nikki Kong </creator>
      <overview> atlas style interview for Atlanta Survey Pilot
      under the direction of William Kretzschmar</overview>
      <date>February 14, 2003 </date>
      <location> Linguistic Atlas, University of Georgia </location>
      <format>transcript: text file, audio file: wav</format>
    </bibdesc>
    <contentdesc>
      <type> atlas style interview.</type>
      <contributor> JW, female residing in the suburbs of
      Atlanta, Georgia</contributor>
      <date> February 7, 2003 </date>
      <locale> X, Georgia</locale>
      <supplementary_info> interview took place in interviewee's home
      and lasted approx 1 hour</supplementary_info>
    </contentdesc>
    <datadesc>
      <type>cda audio file</type>
      <sampling>pilot informants located via cold call from phone list</type>
      <capture> digital recorder</capture>
      <conversion> cda cd produced on recorder transfered to wav on computer
      via Sound Forge</conversion>
      <transcription> Lv protocol used with XML based tagging scheme, see DTD.
      Can be made TEI compliant</transcription>
    </datadesc>
    <editdesc>
      <editor> Nikki Kong, Betsy Barry</editor>
      <revisions> first edit: scribe normalizations editing </revisions>
      <date> first date: May 16, 2008</date>
    </editdesc>
  </lvheader>
```

<text>

<ts when = "00.00.00">

<u who = "interviewer">Will you state your name and the date into the microphone so that I can kind of just see if it's working? You don't have to if you can just sit back it's got pretty good pick up.</u>

<u who = "jw">JW, February eighth, two thousand and three <cough> </u>

<u who = "interviewer">Wonderful. And this community is?</u>

<u who = "jw">Tucker.</u>

<u who = "interviewer">In the state of?</u>

<u who = "jw">Georgia</u>

<u who = "interviewer">And Tucker is um is a suburb of?</u>

<u who = "jw">It a suburb of Atlanta</u>

<u who = "interviewer">ok will you tell me maybe just about your family, about where you are from</u>

<u who = "jw">Ok um well I'm from X, from X.

It's a very small town and a cotton mill town. Uh we moved here when I was about six years old uh to Atlanta. My father was going to college in Greenville and after he left there uh he came to Atlanta to work and brought us here just a couple of months after he moved here and he got set up and got us here and I've been here ever since.</u>

<u who = "interviewer"> What does your dad do?</u>

<u who = "jw">Um It's really funny. He started out in theological school and stated out as a preacher and then for some unknown reason to all of us he decided he didn't want to do that anymore and never talked about it again. Um <cough> started work here at Riches and worked for Riches for oh I don't know maybe ten or twelve years and he worked in their carpet department at downtown Riches that was the only Riches there was and they uh they had a really big carpet department and he worked there. Uh when he left there he stayed in the carpet business he went to work for uh an independent owner and he stayed in the carpet business until he retired <cough> which was some years ago he' about seventy eight and now he lives in

Columbia, South Carolina.</u>

<u who = "interviewer">What about your mom?</u>

<u who = "jw">My mother um she's dead. She died nineteen eighty three uh but she was basically uh a house wife. She never took to my knowledge she may have had one job while I was growing up but I was very very small and really don't remember even what she did but she did not graduate from high school. She worked in the cotton mills in X <false\_start> in X. Um she uh she married <cough>my daddy when she was seventeen he was in the navy and uh um they met through a cousin of hers and they got married and then came to Atlanta. Uh uh but no she was a house wife all those years.</u>

<u who = "interviewer"> What about your grandparents?</u>

<ts when = "00.03.07.03">

<u who = "jw"> My grandparents on my mother's side um her father was uh he worked in the cotton mills all <false\_start> all her side worked in the cotton mills in X <false\_start> in X. Um he was uh from Georgia in El <false\_start> in X which is northeast, which is right over the South Carolina line, and uh my grandmother also was from X. My mother was born in X and then they moved into X later but uh my grandfather he <false\_start> he worked in uh the cotton mills in X and he worked some at Fulton Cotton Mills in Atlanta. Uh when when she was young and uh my grandmother stayed at home. She was a housewife also. And on my father's side his uh his mother and father were from Virginia and they lived there for a while then moved into North Carolina into Rocksborough. It's a small town and h a lot of tobacco uh farming up there and uh that was about it. My grandmother didn't work she was at home and my grandfather. <hesitation>Tell you the truth I'm not sure what all he did <laugh>.</u>

<u who = "interviewer"> So did they have a lot of kids?</u>

<u who = "jw">Uh yea. On my father's side that's uh he has three brothers and one sister um all living. And on my mother's side she has one sister and three brothers and she has two brothers left living and they're the only two left on my mother's side of the family.</u>

<u who = "interviewer">What about you and your siblings?</u>

<u who = "jw">Yes I have one sister and one brother.</u>



<u who = "interviewer">Huh and are you <false\_start> where do you fall?</u>

<u who = "jw">In the middle <laugh></u>

<u who = "interviewer">Oh in the middle.</u>

<u who = "jw">The middle child syndrome. <laugh></u>

<u who = "interviewer"> To <laugh> well um so you where exactly were you born in X did you say?</u>

<u who = "jw"> Green<false\_start> X</u>

<u who = "interviewer">In X. I'm sorry.</u>

<u who = "jw">That's ok.</u>

<u who = "interviewer">But you were pretty much you think that you grew up pretty much in Atlanta.</u>

<u who = "jw">Oh, from six years old, yea. I've been here forty eight years. So I would say yes.</u>

<u who = "interviewer">Yea, so would you consider yourself from here?</u>

<u who = "jw">Most of my formative years yes</u>

<u who = "interviewer">When you have a lot of people in South Carolina which is-</u>

<u who = "jw">I still have a lot of family in South Carolina and North Carolina.</u>

<u who = "interviewer">And what about around here?</u>

<u who = "jw">Around here around here is just, is my immediate family. My uh all of my kids live here and all of my grandchildren live here. So.</u>

<ts when = "00.05.53.10">

<u who = "interviewer">So how many kids do you have?</u>

<u who = "jw">I have four children and five grandchildren.</u>

<u who = "interviewer">Wow. so you were married?</u>

<u who = "jw">Uh yea I was. I was married twice actually. My first husband we were married<cough> for about six years. We married right out of high school, and he uh he just kind of disappeared one day and just never came back again and we had three kids at the time we already had so I had three by the time I was twenty three. And then a couple of years later I remarried and we were married ten years and he died in eighty three. So and we had one more child so.</u>

<u who = "interviewer">That's a lot of kids and grand kids</u>

<u who = "jw">And then since then I said forget it <laugh>. Too busy for anything else.</u>

<u who = "interviewer">What about your kids what do they do?</u>

<u who = "jw">I have um my oldest daughter works in a <false\_start> she's uh uh an analyst in uh programming, uh programs for uh hotels uh special promotions. Those kinds of things. Um my my son oldest son is in landscaping. He likes to do his own thing keep his own business. He doesn't like to work for anybody. Then my next one is a boy too, well they are not boys anymore but he works for um a company that um pours cement. He drives a truck. And he loves it. I know, he likes that kind of work. And then my youngest daughter is a um a programming assistant at Turner South so.</u>

<u who = "interviewer">Oh. wow.<inaudible>.</u>

<u who = "jw"> Yes it's nice</u>

<u who = "interviewer"> Did they go to school?</u>

<u who = "jw"> Do they go to school?</u>

<u who = "interviewer"> Did they?</u>

<u who = "jw"> They all graduated. They're all graduated high school. And my youngest one was the only one that ever had any interest in going to college though and she went to Berry for four years.</u>

<u who = "interviewer">And what about you?<inaudible></u>

<u who = "jw"> I graduated from high school. I did not go to

college. I took some art courses and but that was it.</u>

<u who = "interviewer"> You mentioned that your dad was um possibly going to go into be a preacher</u>

<u who = "jw"> He was. He was. He did go into ministry</u>

<u who = "interviewer"> Did you go to church at all?</u>

<ts when = "00.08.39.05">

<u who = "jw"><overlap>I did uh huh but not because of him. It was really interesting because like I said he when I was very young is when he decided he wasn't going to do that anymore and I just have vague vague memories of being in the church with my mother and father but after that he just stopped going at all. And I think the only time he every goes to church now if there's a wedding.</u>

<u who = "interviewer">Right</u>

<u who = "jw"> Yea. So uh but I went on my own. My aunt, I spent time with my aunt in the summers, my mother's sister they were very close, and my aunt was about eleven years older than my mother so they were more like she more like treated her like her daughter. So we spent a lot of time in X during the summers and holidays and I spent a lot of time vacationing my my summers out from school in X and in X with my grandmother, with my aunt, and my aunt was uh very big in the Methodist church. There were little small country churches uh but uh we always went with her. My sister and I spent a lot of time together up there with her. She always made us sing. She make us get up and sing. She played the piano in the churches, and sometimes we traveled around the different churches where they needed a pianist. So it was a lot of fun. But mostly the Methodist church growing up. Later I got involved in the Baptist church. My husband and I stayed in the Baptist church for about eighteen years and and then I just decided it wasn't for me anymore and in eighty five I became Episcopalian. So</u>

<ts when = "00.10.28.00">

<u who = "interviewer">Hmm.</u>

<u who = "jw">So.</u>

<u who = "interviewer">So lets see, do you remember anything about maybe like your

first wedding?</u>

<u who = "jw">My first wedding?</u>

<u who = "interviewer">Yea</u>

<u who = "jw">yes<laughs></u>

<u who = "interviewer">Ok are you going to tell me about it?</u>

<u who = "jw"><laugh>Yea I remember. Actually we to uh we decided <false\_start> we started to planning uh a bigger a bigger a big wedding and then decided we didn't want to do that um so we ended up with just sort of a family thing but we went to Greenville, South Carolina, and we got married in uh a Baptist church in Greenville, South Carolina. Uh we just took the family and went over there. It was not <false\_start> you know Greenville is not that bad of a drive from Atlanta and </u>

<ts when = "00.11.13.08">

<u who = "interviewer">Did you have people up there with ? </u>

<u who = "jw">Yes, my sister, my sister was with us and her husband um, and uh you know our parents were there. So, it was just mainly just us, just this very small, but uh it was it was nice. It was a nice little wedding. That was that one. The other one, the second one, when I got married a second time, my husband had never been married. Uh and he and I decided we were gonna have a church wedding. And so we had a um larger type one. It wasn't huge, but it was more with the bridesmaids and the groomsmen and so we did the whole formal thing. </u>

<u who = "interviewer">Right, right.</u>

<u who = "jw">Which my kids refer to as the hippy wedding. We were married in seventy three and and all the all the guys had long hair and the big moustaches. Even my husband did <laugh></u>

<u who = "interviewer">Great pictures</u>

<u who = "jw">Oh great <laugh> great pictures.</u>

<u who = "interviewer">Ok, so I am going to talk a little now about <noise> some female household stuff.</u>

<u who = "jw">Ok.</u>

<u who = "interviewer">It might be easier, did you said that you visited your grandparents when you were real little?</u>

<u who = "jw">Um hm.</u>

<u who = "interviewer">Why don't you tell me a little about their place?</u>

<u who = "jw">Well my, my grandmother was by herself. My grandfather died the ones that <false\_start> on my mother's side the ones that I stayed with most, my grandfather died when I was four years old so I didn't <false\_start> I remember him but I don't remember their house that well. And after he died, he'd been sick for a long time he had cancer, and they eventually moved in with my aunt and uncle. So my grandmother remained there um but the house was um it was um it it was in X and it sat way back off the road and then there was a smaller house in front of it that my aunt and uncle owned. They owned both houses and my uncle's brother lived in the house in the front so we were just back and forth plus he had another sister who lived across the street so and X is such a small town you know you just all you do is run next door you were right where <false\_start>you know you could see everybody but the house was a two story house in the back and it had um a a small front porch but the thing that I remember that I always loved about it the most was that it had a front porch that went across the entire width of the house upstairs and it had the outside stairs that went up.</u>

<u who = "interviewer">how neat</u>

<u who = "jw">yea it and it was it just <false\_start> it was really great.</u>

<u who = "interviewer"> what about <cough> some of the rooms in the house maybe where everyone hung out</u>

<ts when = "00.14.15.00">

<u who = "jw"> Well <false\_start> mainly the living room. The house was separated into two areas because my grandparents lived downstairs so uh there was there was a living room and uh a smaller room and I can't remember what it served as then when they were there. And there was a a kitchen on the back side and uh uh probably the smaller room served as their bedroom I'm not sure and bath and then upstairs there was uh uh a room,

a a front room like a living room and two bedrooms and uh one big room that was a kitchen and a back porch and the stairs that came down the back. So it was separated because of the way they <false\_start>because my grandparents were there but after after my grandfather died and after my my grandmother eventually moved back to X and stayed with one of her sons there and my aunt and uncle took over the whole house and had the whole thing renovated so it became quite a large two story house but we hung out mainly in the living room downstairs and the kids mainly upstairs cause that was what we called the piano room, my aunt's piano was up there uh and we loved to bang on it and play and that's where we always slept on palates and pull out sofas and whatever and then the attic fan was up there too so it was a lot cooler and the porch. And we love the porch so <laugh>.</u>

<u who = "interviewer"> Right, what about some of the kinds of furniture you might find in the big living room? </u>

<u who = "jw">Oh <cough> gosh um well she had uh she just had regular furniture, a sofa, a recliner, and the tv, and <false\_start>but there were radiators you know in every room or little gas heaters. </u>

<u who = "interviewer">No central? </u>

<u who = "jw">No central. No because<false\_start> you come in from the cold and you, everybody was fighting to stand over the little gas heaters. We stood there all the time because you could turn them up</u>

<u who = "interviewer">Is that the only way the house was heated?</u>

<u who = "jw">Uh huh</u>

<u who = "interviewer">any other</u>

<u who = "jw"> gas heaters and all of the <false\_start> the upstairs they was um <false\_start> downstairs there were the gas heaters and upstairs was a floor furnace. A big<false\_start> so you had to learn in the winter time if you got up in the middle of the night you had to remember to walk around the furnace because if you stepped on it was hot.

<u who = "interviewer"><overlap>right</u>

<u who = "jw"> The grates on top so. </u>

<u who = "interviewer">what about in the family room where there any furnaces? </u>

<u who = "jw">little gas heaters</u>

<u who = "interviewer">that's interesting. what about like a fireplace? </u>

<u who = "jw">Uh no they didn't have any fireplaces in that house. Uh the front house had a fireplace in the living room but I can't ever remember it being used. Seems like it was always blocked up. A lot of times back then in older houses they were blocking up the fireplaces with those tin things that go they were kind of decorative. People just<false\_start>I don't know why I guess guess when central heat started coming along or gas heaters people thought "well oh we don't need the fireplace anymore so we don't use it" I guess<false\_start>I don't<false\_start> who knows!<laugh></u>

<u who = "interviewer">What about to cool it, the place? Was it hot? </u>

<u who = "jw">You opened the windows and turn the fans on <laugh> <cough> </u>

<u who = "interviewer"> So, um let's see. What about the bedrooms? What kind of things would you find in there? </u>

<ts when = "00.17.35.14">

<u who = "jw"> Oh<cough> the bedrooms that were there <cough> they were the most fun for us because the front bedroom had all of my aunt's things in it and she had uh what we call a chifforobe um and one side of it was a mirror and the other had all of her hats. And that was the most fun because she would let us get into her shoes and her hats and play to our hearts' content so <cough> I think I tried on every hat she had and and she let us go though her jewelry box because too and she wore<false\_start>because she was so big in the church you know she dressed up every Sunday. She always had a hat and she always had jewelry so she had lots of costume jewelry and lots of hats. </u>

<u who = "interviewer"> that's fun. </u>

<u who = "jw">oh it was a lot of fun. </u>

<u who = "interviewer"><overlap>what about what you slept on in there. </u>

<u who = "jw">Uh if we were lucky we got a bed. As as I got older and it was just my aunt and uncle there uh my sister and I slept in the in the front room uh where the porch was cause the windows, uh the windows were down, the big windows. You could just raise them up. They didn't have screens on them so you could climb in and out of the windows to get on the porch so that was lots of fun. So we always slept on the front, uh in the bed. Otherwise if we had uh if there was a lot of us there, we ended up sleeping on the floor. On just pallets, just<false\_start>she had stacks and stacks of quilts. And if we stayed at my um <false\_start> my great uncle was a caretaker at the cemetery and all of our family, we have a huge amount of family in Elmhurst, in X. And my uh <false\_start>that now that house was the most fun because um, my great aunt and great uncle lived there and they lived back of the cemetery and it was a little, it was like three room house uh a living room, a bedroom, and a kitchen and they <false\_start> she kept a bucket of well water on the back porch with a dipper and that's with no running water. We had an out house. There was one cow that wandered around out back and um you swept the yard<false\_start> didn't have grass um because nobody had lawnmowers and they didn't upkeep, so the yard was always dirt and the house stood up um little uh pillars or <false\_start> st, st<false\_start> and you know you could crawl under the house. It was always real cool underneath the house. But she had all these this uh straw brooms tied together you know. We'd sweep the yard and every and in her house when we spent the night there it big feather bed and we would all sleep on it but you would roll towards the middle because everybody just<false\_start>so we all end up on top of each other. But I can remember as many as five of us being in that bed at one time. Five of us kids. </u>

<u who = "interviewer"> that's neat</u>

<u who = "jw">Yea that was really, really, really neat<laugh></u>

<u who = "interviewer"> What about the kitchen in that house? How did they cook?</u>

<u who = "jw"> She had a wood stove. A wood burning stove</u>

<u who = "interviewer">huh</u>

<u who = "jw">Um with the you know little things on the top of it with the handles you know you just open it up stuff wood back down inside of it and that's how she baked and uh they lived there until they died. And back then you didn't go to funeral homes either. Whoever died was laid laid out right there in the living



room usually uh or a bedroom, another room in the house and everybody came there and to view the body and that's where um the service was and that's the way it was with my grandfather. I remember when my grandfather died he was laid out in the downstairs of the big house</u>  
<ts when = "00.21.28.02">

<u who = "interviewer">Things are so different now. What about how they were born? </u>

<u who = "jw">I'm sure they never went to a hospital. I'm sure never um, I'm<false\_start>I'm my mo<false\_start>I don't know if my mother was. I don't even know if she as born in a hospital. I'm sure we had the conversation at some time and I can't remember</u>

<u who = "interviewer">Did you know how that worked at all? </u>

<u who = "jw">Probably just a a doctor, if they could get a doctor, came to the house. X is just a small town um but I'm sure that it was either that or midwives uh, or they did it on their own you know. </u>

<u who = "interviewer">That's unbelievable. </u>

<u who = "jw">I know it is. I can't even imagine not having a hospital to go to have a baby <laugh> </u>

<u who = "interviewer">That's all very, very different. Ok, what about let's talk about the kitchen a little bit. um, so uh what kinds of things might you find in the kitchen besides maybe you'd cook on ? </u>

<u who = "jw">Oh just the um a stove, refrigerator, um um just regular pots and pans and<overlap></u>

<u who = "interviewer"> what kinds of things you remember uh, like your grandparents or your aunt and uncle cooking? </u>

<u who = "jw"> what kinds of food?

<u who = "interviewer">Um hm</u>

<u who = "jw">Well um we grew up on beans <laugh>. Dried beans um and biscuits, my aunt cooked biscuits all the time, every meal, my aunt and my mother. We never had a meal that we didn't either have biscuits or cornbread. </u>

<u who = "interviewer">Homemade? </u>

<u who = "jw">Uh huh. Oh absolutely. Uh and I still make them, but don't do it all the time but my<false\_start>every once in a while I will. But um yea and always tea. You always had sweet tea. Never a meal without tea. Um breakfast was always fried eggs and bacon and biscuits or toast uh for the most part. Sometimes oatmeal. We ate a lot of oatmeal. Grits, ah of course grits <laugh>. Um and always jelly was on the table or preserves you know. Preserves were a big treat because there was always somebody in the family who would put up preserves. And my grandmother, the one in North Carolina, she she canned all the time uh more than my<false\_start>I remember my aunts doing here in X. But she canned all the time. She would put up <cough>all these kinds of beans and tomatoes and uh preserves, made preserves. She made her own butter. </u>  
<u who = "interviewer"> did she have a garden? </u>

<u who = "jw"> She did. They had a farm for a long time and it just got smaller and smaller. They butchered hogs and had their own hams and and they had a smokehouse in the back. </u>

<u who = "interviewer"> Really? </u>

<u who = "jw"> yea</u>

<u who = "interviewer"> huh. did you visit that and?</u>

<u who = "jw"> Oh absolutely. Yea, I loved it. I loved to go to my grandmother's in North Carolina. She was so much fun. </u>

<u who = "interviewer"> Huh. what other kinds of things did she have in her garden?</u>

<u who = "jw"> Oh she had okra, and tomatoes, and corn and beans, uh collards, um anything she could get, squash anything she could get planted. She had quite a large garden. Um and then the smokehouse always had hams hanging in it. But that's where my grandfather hid. He hid in the smokehouse. He drank a lot and he drank</u>

<u who = "interviewer"> <overlap>What would he drink? </u>

<u who = "jw"> Moonshine. And um my grandmother would she was totally against alcohol. She didn't um didn't <false\_start> because he drank excessively, but he was really funny when he drank. I mean it wasn't like he was<cough> really mean to her

or anything. He was just real funny. But she didn't like him drinking. He hung out in the pool hall a lot too. <cough> She use to have to retrieve him from the pool hall. </u>

<ts when = "00.25.28.09">

<u who = "jw"> But he did. And he would hide his bottles in different places and his smokehouse was his favorite place. He'd go out there and drink.<laugh><cough>But in her kitchen we're talking about finding things in kitchens, she had an older kitchen too um and and converted over the years you know to to be more modernized but she had a safe, a pie safe, and I don't know if you know what a pie safe</u>

<u who = "interviewer">um hm. </u>

<u who = "jw"> Ok. uh, and it had a flour bin in it and to me that was the most interesting thing because living in Atlanta and growing up in Atlanta we would be more and more modernized over the years but my grandmother stayed the same and she always had that pie safe and always had the big flour bin in it. And it was really<false\_start> and she had some old things, old dishes and old wooden rolling pins which I still have a old wooden rolling pin. </u>

<u who = "interviewer"> What about what would they drink in the morning? </u>

<u who = "jw">Coffee.</u>

<u who = "interviewer"> How would they prepare that? </u>

<u who = "jw"> Um <cough>different ways. Um just one of those old coffee pots just a little you know had the little filter thing, the metal. I don't know if you remember those and I don't see those anymore but just put water in it and the coffee went in on the top and I don't know how it did it. I hate coffee so I don't ever make it. Uh and then later my mother bought a percolator which she thought that was wonderful <laugh>. But those were the only two ways I can remember them making coffee unless we were on the lake and then God knows how they made coffee that, some weird <false\_start> threw it in the pot I guess. I don't know. </u>

<u who = "interviewer"> But what about you said that you always had fresh hams. What other kinds of meats do you remember eating? </u>

<u who = "jw"> Um oh everything. We had every Sunday you either had uh fried chicken or roast roast beef. Um a roast and that was mainly it on Sundays. It was, it was uh <false\_start>I grew up on traditional Sunday dinners. Uh big. Everybody at the table. All kinds of vegetables um roast beef or chicken or ham.<cough>Sometimes barbecue chicken. Um and always a dessert and biscuits. Always. Always. My grandmother too. She did that. My grandmother would go out back and kill chickens and and cook. </u>

<ts when = "00.28.12.04">

<u who = "interviewer">What about eggs? </u>

<u who = "jw">Uh huh</u>

<u who = "interviewer">Fresh? </u>

<u who = "jw">Um hm. Fresh eggs. Um when, when we were at my grandparents cause they had chickens but um no in Atlanta no. <laugh>. </u>

<u who = "interviewer"> Right. what about dairy? </u>

<u who = "jw"> Um well my great aunt and uncle had a cow out back so they always milked the cow. My grandmother in North Carolina <cough>probably did for a while until they sized down and then no longer had that ability. Ah but mostly the milkman brought the milk. </u>

<u who = "interviewer"> Do you remember drinking milk from the cow? </u>

<u who = "jw"> Uh, not straight from the cow no I don't remember that. I may have but I don't remember it. </u>

<u who = "interviewer"> do you remember if the milk is different back then to what it is now? </u>

<u who = "jw"> Yea, the milk had a different taste. Um, the milk that was delivered to me the milk was delivered by the milkman in the glass bottles was always the best milk cause we had milk delivered up until I can't remember how long milk was delivered to our house. Even in Atlanta we had milk delivered. I always loved going out and getting milk off the front porch because the milkman just left it there you know. And it was real treat when you got chocolate milk. But milk in a glass bottles that came in the glass bottles always tastes different and now it's

all in plastic. Um, so it was yea it was different. </u>

<u who = "interviewer"> Any other kind of dairy stuff you use to eat when you were a kid? </u>

<u who = "jw"> Just you know just cheese. Milk and cheese and nobody ate cottage cheese, if there was cottage cheese I don't even know if they ever had it then but nobody ate yogurt and cottage cheese <laugh>. And now I don't even drink whole milk. I drink skim milk. I won't I can't even stand the taste of whole milk anymore. </u>

<u who = "interviewer"> Right. ok what about um you said that your granddaddy worked at a smokehouse? What do you remember him wearing to go back there and work? </u>

<u who = "jw"> Oh granddaddy always had on the same thing that I can remember. He just had on uh like um um what would ow like be khaki pants but they were always darker like darker work pants and a shirt and he always kept the buttons buttoned buttoned up at the top and uh they would be uh like a felt kind of shirt or flannel soft always soft shirts and he always wore a hat. </u>

<u who = "interviewer"> what about in the winter to keep warm? </u>

<u who = "jw">Oh well I don't know what he wore I guess but uh but probably some kind of jacket but we always had jackets and coats. Mostly um growing up you know we had you had a jacket to wear outside to play in and then a dress coat. You always had dress coat to go shopping or to go to church so</u>

<u who = "interviewer">Ok, what I think if you don't mind we'll do is this task that I have. And then we'll talk a little bit more it's going to seem kind of weird but if you could read I'm going to have you read the words in these cards. I'm going to shuffle them. </u>

<u who = "jw"> Ok</u>

<u who = "interviewer"> Get you to read them. Shuffle them again so</u>

<u who = "jw"><laugh>

<u who = "interviewer"> we're going to do three tests and I'll put them there so you can see. Alright. </u>

<ts when = "00.31.44.03">  
<u who = "jw"> bit</u>  
<u who = "jw">hid </u>

<u who = "jw">tech</u>  
<u who = "jw">hod</u>  
<u who = "jw">boot </u>

<u who = "jw">toke </u>

<u who = "jw">hood</u>  
<u who = "jw">but </u>

<u who = "jw">tick</u>  
<u who = "jw">hate</u>  
<u who = "jw">duke</u>  
<u who = "jw">boat </u>

<u who = "jw">took</u>  
<u who = "jw">head</u>  
<u who = "jw">tock</u>  
<u who = "jw">bet</u>  
<u who = "jw">put</u>  
<u who = "jw">hut</u>  
<u who = "jw">bake</u>

<u who = "jw">bat </u>

<u who = "jw">tack</u>  
<u who = "jw">teek</u>  
<u who = "jw">tuck</u>  
<u who = "jw">hoot</u>  
<u who = "jw">take</u>  
<u who = "jw">hode </u>

<u who = "jw">had</u>  
<u who = "jw">heed </u>

<u who = "jw">bought</u>  
<u who = "jw">beat <cough> </u>

<u who = "interviewer">ok</u>

<u who = "jw">uh bait</u>  
<u who = "jw">bought</u>  
<u who = "jw">heed </u>

<u who = "jw">had </u>

<u who = "jw">hode </u>

<u who = "jw">take </u>

<u who = "jw">head </u>

<u who = "jw">took</u>

<u who = "jw">boat </u>

<u who = "jw">tuck</u>

<u who = "jw">duke</u>

<u who = "jw">hate</u>

<u who = "jw">tick</u>

<u who = "jw">but </u>

<u who = "jw">hood</u>

<u who = "jw">toke</u>

<u who = "jw">boot </u>

<u who = "jw">hod </u>

<u who = "jw">tech </u>

<u who = "jw">hid</u>

<u who = "jw">bit </u>

<u who = "jw">hoot</u>

<u who = "jw">hut</u>

<u who = "jw">put</u>

<u who = "jw">bet </u>

<u who = "jw">tock</u>

<u who = "jw">teek </u>

<u who = "jw">beat </u>

<u who = "jw">tack</u>

<u who = "jw">bat </u>

<u who = "interviewer">ok last time. </u>

<u who = "jw">hate </u>

<u who = "jw">duke </u>

<u who = "jw">tuck</u>  
<u who = "jw">boat</u>  
<u who = "jw">took </u>  
  
<u who = "jw">head </u>  
  
<u who = "jw">heed </u>  
  
<u who = "jw">take</u>  
<u who = "jw">tack </u>  
  
<u who = "jw">beat </u>  
  
<u who = "jw">hode </u>  
  
<u who = "jw">had </u>  
  
<u who = "jw">boot </u>  
  
<u who = "jw">bat </u>  
  
<u who = "jw">toke</u>  
<u who = "jw">hood </u>  
  
<u who = "jw">but</u>  
<u who = "jw">tick </u>  
  
<u who = "jw">teek </u>  
  
<u who = "jw">tock </u>  
  
<u who = "jw">bet </u>  
  
<u who = "jw">put </u>  
  
<u who = "jw">hut</u>  
<u who = "jw">hoot </u>  
  
<u who = "jw">bit</u>  
<u who = "jw">hid </u>  
  
<u who = "jw">tech</u>  
  
<u who = "jw">bought</u>  
<u who = "jw">hod </u>  
  
<u who = "jw">bait </u>



<ts when = "00.35.11.09">

<u who = "interviewer"> Alright now I want to talk about a little bit about how people talk or how you hear people talk. The differences.</u>

<u who = "jw">Um</u>

<u who = "interviewer"> What about the way people talk around Tucker versus inside Atlanta? Do you notice the difference? </u>

<u who = "jw"> No um I don't really because Atlanta has spread out in all directions and I when I was growing up I would have said yes I could tell more of a difference the further out you get but Tucker is so<false\_start>I mean we're twelve miles from Atlanta so it's not a little city by itself anymore um but there there are a lot of different cultures in Tucker. Um I guess uh maybe more so than some other areas you know areas are different so different around Atlanta but we have a lot of um Orientals, uh Asians. Primarily Orientals and Asians so<inaudible></u>

<u who = "interviewer"><overlap>How do you differentiate between the two? </u>

<u who = "jw"> Orien<false\_start>well oriental you know Chinese and you know and their words are more clipped, I guess. Um, I don't notice any<false\_start>I don't even know if I can even be able to tell the difference if they were Japanese, but but maybe I'm not real sure. But a lot of Chinese. Um, and Asian as far as like Middle Eastern people and I don't always know where they come from. Sometimes I'll ask them you know if I<false\_start>if we're in a store enough there's a<false\_start>right here you said you were at the Citgo a minute ago the man who owns it he and his parents own it and they're, they're Middle Eastern and I uh I don't really talk to him a lot but uh um my son does. My son goes in there and has regular conversations with<false\_start>he knows all about him um which is I think is really funny. But um but that be how they differentiate I mean I know the difference in the way the people look as well as the way they talk what about I can just about tell where they're from you know Middle East, China or</u>

<u who = "interviewer"><overlap> What about something like the way people talk in Atlanta versus the way they talk in X? </u>

<u who = "jw"><laugh>Now that now there you got a difference. Atlanta um I've noticed over the years we've gotten so many Yankees uh that uh that people now even my even my own children which God forbid that everybody is losing their Southern accent and it's a shame uh I haven't

lost mine and never will I'm sure but uh my kids even growing up in school they've been exposed to so many different cultures which is a good thing because I never was but uh um they've been exposed to so many different cultures that I've noticed their language changes. They still have some southern accent but but not uh like mine um but if you but in X uh the you know people still have that slow southern talk and sometimes it just drives you crazy. Uh I mean my cousin still says "grass" which is grass, or "glass" you know she just drags out and she's still in X, but they've<false\_start>I've noticed since I've been in Atlanta that is bad as mine accent is, theirs is just way drawn out and then um coming from downtown I can work downtown all day I half the time I can't tell you where anybody is from anymore. </u>

<ts when = "00.39.16.01">

<u who = "interviewer"> So you work downtown? </u>

<u who = "jw">I do. I work in midtown. </u>

<u who = "interviewer">Ah. </u>

<u who = "jw"> So I go a lot and we have um a lot of uh uh in our office our office is not real huge but everyone there is um uh from the South except one one women we have we call her our token Yankee. </u>

<u who = "interviewer"> Where's she from? </u>

<u who = "jw"> She's from Ohio and we just love to rib her. But her her talk is I mean her her I guess the the way she sounds the the inflections that she puts on words is it's really different. But but yea you get quite a difference in the speech from say Atlanta to to X. But they still talk like that in Elber <false\_start>you don't find many Yankees in X. <laugh> </u>

<u who = "interviewer">So you think you can hear a difference I thought that you said in the way younger people younger people talk today? </u>

<u who = "jw">Yes. Absolutely. Absolutely. Um they're not uh the more they're exposed to the different cultures the more their language the way they sound changes. Um I uh it's really funny to me that that uh I I guess it's like I said when I was growing up I was exposed only to Southerners so we all talk the same and nobody knew any different you know you didn't think you sound until unless you went up North and which was a rare occasion. I never visited the North until I was older. Uh</u>

<u who = "interviewer"> Can people tell you're Southern? </u>

<u who = "jw"><laugh>My I have cousins who live in Philadelphia. My one of my aunts my my father's sister when she married they moved to to Phili to Pennsylvania.<cough> Then they've been in Philadelphia all their lives and my cousins were all born there and <cough> I communicate regularly with one of them. We we take turns calling each other just because we don't get to visit. So we try to stay in touch and and uh I'll call in I'll call his work mostly because I'll be at work and that's when we communicate the most and um even the people who answer the phone they they <false\_start>if I leave a message on the recorder they save the messages. He said, "They get such a kick out of listening to you talk." So they will save the messages so they can listen to my voice.<laugh> </u>

<u who = "interviewer"> What do you think of the way he talks? </u>

<u who = "jw">Sounds like a Yankee. <laugh> </u>

<u who = "interviewer">What is that you can pick up on? You think when you hear someone you can whether they're from the South or? </u>

<u who = "jw"> But most of the time yea uh sometimes uh it's a little harder to tell especially people who have been through college. Uh you notice them dropping more and more of that Southern accent. Um I</u>

<u who = "interviewer"><overlap> What do you think that is? </u>

<u who = "jw"> Well I don't know. I work I work with a girl uh she's uh twenty-seven and she's from Mississippi and grew up in Mississippi and she tells me how hard how hard it was for her to drop that Mississippi twang and I said, "Well you know did you have<false\_start>did you think you needed to?" And she felt like that she was<false\_start> she could communicate better or um I don't know uh articulate better. I I don't know why. But it never occurred to me to drop mine or to try to. Although they've tried to get me to up North they've they've<false\_start>he said my cousin used to come <false\_start>I spent about three months up there uh several years back and he said you know "Before you l eave here, we're going to have you talking like a Yankee" and I said "It'll never happen <laugh> </u>

<u who = "interviewer"> Did people notice a change when you came back here at all? </u>

<ts when = "00.43.15.08">

<u who = "jw"> Oh are you kidding <laugh> I didn't even get near a

Northern accent. But uh uh I don't no why I guess she just felt like that she didn't want to be viewed as a and which is a a huge Southern stigma I think that they view Southerners as rednecks or uh just you know just old country people which I think there's a big difference even in even in Southerners who are raised in Atlanta um when I grew up on the south side of town so people talked more like I do and as I've gotten older I have friends who grew up in Atlanta in the city uh like in the Brookhaven, Buckhead area which that was still a classy area back then or what we use to call highfalutin. Uh but I have a friend who grew up there and she has that more what we called Buckhead Betty talk. </u>

<u who = "interviewer">What's that like? </u>

<u who = "jw"><laugh>You know she talks like this. <laugh> Um, but she lives in Chester, South Carolina which is talk about <ns>podunk</ns> little towns but uh when she calls me you know she still talks like that so uh it very it's very proper southern lady talk and and not so it's not so twangy as it is just drawn out. Uh it's it's really difficult to explain unless you've heard it. <laugh> </u>

<u who = "interviewer"> But you can pick up on it? </u>

<u who = "jw"> You can pick up on and her brother talks the same way. </u>

<u who = "interviewer"> Can you think of anything specific that's cueing you that maybe that's</u>

<u who = "jw">Um I don't know maybe maybe that's you know like I said it's more I know that I have a drawl but hers is more drawn out than a drawl and and the words are still very proper you know you speak very proper. They don't use a lot of uh like little colloquialisms little slangs like uh if you grew up on the south side of town or in a little bitty town you know a little small town but it's more I don't know more of<false\_start> a suppose to be more of a educated proper way to speak in Atlanta when you grew up with money or grew up with the right kind of people. </u>

<u who = "interviewer"> Um Hm. </u>

<u who = "jw"> Atlanta has very very very severe class distinctions. </u>

<u who = "interviewer"> So you think that causes a difference too education and

money</u?

<u who = "jw"> <overlap>Um hum. </u>

<u who = "interviewer"> and speech </u>

<u who = "jw"> I think so. Absolutely. </u>

<u who = "interviewer"><overlap>So you think you can pick up on that?</u>

<u who = "jw">Um hm. </u>

<u who = "interviewer"><overlap>What is it about your uh</u>

<u who = "jw"><overlap> I can tell you the country club set from the ones of us who grew up playing in the ball fields.<laughs> </u>

<u who = "interviewer"><overlap>Huh. Really </u>

<u who = "jw">Uh huh</u>

<u who = "interviewer"> What about what you were saying um they save your messages when you call who is it that you call? </u>

<u who = "jw"> Uh uh, in Philadelphia. My cousin the the store he works in they say if I leave sometimes I've called there and the store is just closed or something or I'll so I'll leave a message for him. They have an answering machine in their store and and they save the messages so they thinking <laugh>but they know who it is that called they say, "It's your cousin. Your cousin called." </u>

<u who = "interviewer">And what is it about his voice do you think that's so different from?</u>

<u who = "jw"> Yankee. Well I say water and he says water. <laugh> He says when are you ever going to learn how to talk? <laugh> I said, "well, not from you" <laugh> But just different words. I mean words are so different and when uh uh I had the hardest time getting a glass of ice tea up there. It's the most ridiculous thing you've ever seen. Uh I would go in and ask for tea and and in in a restaurant, we'd be in a restaurant and he would look at me and say, "they don't know what you're talking about." I said "Don't they drink tea?" He said "hot tea." They don't know what ice tea is or sweet tea if it's sweet tea." I said, "<ns>Jeez</ns>", you grew up with a Southern mom." He said, "Well I know." But you know </u>

<u who = "interviewer"><overlap> So what did you did you get when

you ordered it?</u>

<ts when = "00.47.45.11">

<u who = "jw"> They never knew what I was talking about. Nobody I never could make anybody understand. And if I wanted a Coke, you don't get a coke up. There you get a soda. <laugh> </u>

<u who = "interviewer"> So what</u>

<u who = "jw"> <overlap> And nobody eats grits <laugh> </u>

<u who = "interviewer"> What do they have instead? </u>

<u who = "jw"> Oh uh well I don't know like cream of wheat and cream of rice and those kinds of things you know. <laugh> I don't know. </u>

<u who = "interviewer"> Everything's different</u>

<u who = "jw"> The food is different but thank goodness i mean my aunt was like I said, she was raised in a Southern household so she still cooks Southern. <laugh> </u>

<u who = "interviewer"> So he thinks the way you talk and the way Southerners talk he's says "we're going to teach to how to talk," what do you think his ideas are about Southern speech? </u>

<u who = "jw"> Oh well he's he's just messing mainly he's just messing with me but um he, he just uh he'll make fun of me all the time. He just says we we you know we don't know how to talk. "Southerners don't know how to talk. You have to talk like a Yankee." But just just different things I'll say<false\_start>and he says, he'll "hi" and I'll say "hey" uh stuff like he always makes a distinction where I don't really notice I don't pay that much attention to it but uh he pays more attention to those kinds of the differences in the words and my other cousins do too up there. </u>

<u who = "interviewer"> Is that the only Northern?</u>

<u who = "jw">That's the only Northern relatives I have. Bless their hearts.

<laugh> </u>

<u who = "interviewer"> Interesting but they like the sound of your voice enough to say that </u>

<u who = "jw"> Oh yea he would take me out. Like I said I spent about

three months up there and we would go out to different places or we would go out to some of his favorite places to hang out and uh all the guys whe<false\_start>when they'd stand around and made me talk they just listen to me talk forever just cause they wanted to listen to me talk and I was all "What's the matter with you people up here? <laugh> Ain't you ever heard a Southerner before?"

</u>

<u who = "interviewer"> Do you know the English language? </u>

<u who = "jw"> I guess not. </u>

<u who = "interviewer"> Um what about do you notice in the way that men and women talk? </u>

<u who = "jw"> Um in the South? </u>

<u who = "interviewer">Sure</u>

<u who = "jw"> Um it's just some you know sometime<false\_start>I don't know um it's it's hard to say I guess it just depends on where they were raised uh where you were born, where you grew up but the uh i still get together with all of my high school buddies and they all sound the same to me. They sound the same like i do. <laugh> I haven't noticed any difference in their speech. </u>

<u who = "interviewer"> Where did they move off? So they still live here?</u>

<u who = "jw"> Uh huh. And most of them are still living on the South side of town. I'm the only one that came north </u>

<u who = "interviewer"> What is the South Side? </u>

<u who = "jw">South Side. South Side everybody knows the South Side if you grew up here<false\_start>uh it's all of the uh um Southeast Atlanta, Hapeful, College Park, East Point area. It we had a uh it was like uh one big huge family down there. Um I went to school <false\_start>I was in at an Atlanta school but um you know we we always hung out with people from College Park High School, and Hapeful High School, and Russell and East Point all that area so that was what we consider the South Side. Course then we didn't know it was South Side. <laugh> We just thought we were in our area until I <false\_start>you know until people started growing up and getting older then it's always referred to as the South Side. </u>

<ts when = "00.51.34.03">

<u who = "interviewer">And everyone you went to high school with sound the same to you? </u>

<u who = "jw">Yea yea yea they all sound the same. I haven't notice any changes in any of them. </u>

<u who = "interviewer"><overlap>Wow. What about the way um, do you notice people with ethnic backgrounds. Can you tell the difference in the way they talk? You spoke of being able to tell someone of Middle Eastern what about black folks and white folks? Can you tell the difference? </u>

<u who = "jw"><overlap> uh </u>

<u who = "interviewer"> So if someone were to call you on the phone you'd think you'd be able to? </u>

<u who = "jw"> Most of the time yes especially if they grew up in Atlanta. Um<cough>I've noticed um the black people over the years uh there's no different <false\_start> the older older black people that I still come in contact with they still have the old South talk and talk like they always did. Um then there<false\_start> the then you have the younger ones who have their own language going the Ebonics and um half the time you don't know what anybody's saying. And then I I work in a building where we do have a lot of black people and they're very well educated and their speech is uh um um you can tell that they've been to college or that they grew up in the North. Not so much of a Yankee accent as just more uh more cultured.

<u who = "interviewer">where do you work? </u>

<u who = "jw"><overlap>Um I don't think that you can get that in the South.</u>

<u who = "jw">Um I work um at One Georgia Center um it's right<false\_start> do you know where the Varsity is? </u>

<u who = "interviewer"> oh yea. </u>

<u who = "jw">It's right there</u>

<u who = "interviewer"> what kind of work is it? </u>

<u who = "jw"> Oh I work for um a small industrial real estate firm. We do we deal with warehouses. We have uh a lot of brokers and I'm the business manager. </u>

<u who = "interviewer">So you have a lot of different types of people. You said you had some from Ohio. </u>

<u who = "jw"><laugh> Yea everybody there though is from the South we have some South Georgia boys. We one South Georgia boy who grew up around Columbus



and went to school in uh Alabama. </u>

<u who = "interviewer"> Do you notice a difference in how he talks? </u>

<u who = "jw">He's um he's he's a good ole Southern boy um but he's very well educated and uh and you can tell you know and he uh but's he's he's he still has the Southernness he has the Southern draw to some extent. </u>

<u who = "interviewer"> So maybe you can tell where if people were from Southern states you can think that I mean you obviously can hear in Southern speech, the buckhead betty versus</u>

<u who = "jw"> <laugh> That's from growing up here. </u>

<u who = "interviewer"><overlap>Do you think you can tell differences in is Southern speech just one big thing to you? </u>

<u who = "jw"> Uh yea uh pretty much. It seems like seems like Tennessee um <false\_start>because I grew up in Atlanta it's it's a little I think it's a little different if I still if I if it had grown up in a small town and we're still there I probably able to notice the difference even more but I'm you know I'm exposed to the different cultures everyday that um to me it's kind of hard to tell anymore. Tennessee different parts of Tennessee um people who come from larger cities their their speech is becoming more refined and more cultured um the the older the kids get<false\_start>I mean you know if the kids grew up in a larger city I think they'd become more and more cultured and and like I said exposed to the different cultures so their speech changes. Um so it's a little harder to tell anymore but but all I have to do is go to South Carolina and visit relatives or or you know small towns in South Carolina to to <snap>catch it again, to catch the difference in the speech. </u>

<u who = "interviewer">Do you find yourself more southern when you go to those places?</u>

<u who = "jw">Oh absolutely. Oh I go to I can spend a week in X and come back to work and they'll probably look at me and say, "What? What are you talking about?"<cough>Cause it cause you know they already they already know at work that I have<cough>I have probably a more southern draw than anybody in the office because most of the people in the office are <false\_start> they went to college so they're they are college educated and uh and and then we uh uh have pretty much of a younger crowd in out office some of the older people um but like our owner went to Kentucky and he he was uh he's from New Orleans but you cannot even tell. There's no way <cough>. </u>

<ts when = "00.56.46.08">

<u who = "interviewer"> Cause New Orleans</u>

<u who = "jw"> New Orleans is is very distinct</u>

<u who = "interviewer"> What did it sound like to you? What does New Orleans speech sound like? </u>

<u who = "jw"> Oh my gosh it's a it's a cross between uh southern uh southern white, black, and French, and<laugh> and God knows else was thrown in there but uh it's it's uh uh half the time I can't a real New Orleans person I can't understand a word they're saying uh or if you watch a movie on tv and and it's you know filmed in New Orleans and your<false\_start>it's suppose to be New Orleans people and can't understand them and and I don't know where they <false\_start>it's just such a mixture they whole Cajun thing but I think it's fascinating. I just love to hear them talk</u>

<u who = "interviewer"> I do too. It's one of my favorites. Have you ever been around New Orleans? </u>

<u who = "jw"> not</u>

<u who = "interviewer"><overlap>To little bit? </u>

<u who = "jw"> No not not really out of New Orleans I've only been to New Orleans a few times and uh Biloxi um d <false\_start>you know that area. Only Biloxi once and Biloxi I mean it's such a big gambling place now. You <false\_start> who would know if anybody was from Biloxi or not. Um but uh New Orleans that<false\_start>no I've never gotten really out into be able to experience that</u>

<u who = "interviewer">It's fun</u>

<u who = "jw"><overlap>I bet it is <laugh>. I bet it is <laugh> </u>

<u who = "interviewer"> Well I guess that's just about it. We have about a minute and a half left on the tape. </u>

<u who = "jw"> Oh ok. </u>

<ts when = "00.58.26.14">

</text>

</lvroot>

## APPENDIX C

### APPENDIX 3: LV DTD

```
<!ELEMENT id (#CDATA)>
<!ELEMENT link (#PCDATA)>

<!ELEMENT bibdesc (creator,overview,date,location,format)>
<!ELEMENT creator (#PCDATA)>
<!ELEMENT overview (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT location (#PCDATA)>
<!ELEMENT format (#PCDATA)>

<!ELEMENT contentdesc (type,contributor,date,locale,supplementary_info)>
<!ELEMENT type (#PCDATA)>
<!ELEMENT contributor (#PCDATA)>
<!ELEMENT supplementary_info (#PCDATA)>
<!ELEMENT locale (#PCDATA)>

<!ELEMENT datadesc (sampling,capture,conversion,transcription)>
<!ELEMENT sampling (#PCDATA)>
<!ELEMENT capture (#PCDATA)>
<!ELEMENT conversion (#PCDATA)>
<!ELEMENT transcription (#PCDATA)>

<!ELEMENT editdesc (editor,revisions,date)>
<!ELEMENT editor (#PCDATA)>
<!ELEMENT revisions (#PCDATA)>

#This list will have to be modified depending on what tags are included
#for non speech sound representation, such as coughs, noise, and for
#paralinguistic information such as false starts, overlapping speech,
#and the like. The * by the tag names indicates that this could happen
#0 or more #times as I have included a few typical tags that occur in
#the transcript available here, and the "could" occur as they are typical
#features of spoken discourse.

<!ELEMENT text (#PCDATA,ts,u,inaudibilia,laugh*,false_start*,ns*,cough*,
hesitation*,overlap*,snap*)>
```

```
<!ELEMENT ts EMPTY>
<!ELEMENT u (#PCDATA)>
<!ELEMENT inaudibilia EMPTY>
<!ELEMENT laugh EMPTY>
<!ELEMENT false_start EMPTY>
<!ELEMENT hesitation EMPTY>
<!ELEMENT overlap EMPTY>

<!ELEMENT ns (#PCDATA)>
<!ELEMENT cough EMPTY>

<!ATTLIST ts when CDATA #REQUIRED>
<!ATTLIST u who PCDATA #REQUIRED>
```