

# Transcriptional Analyses of the LENA Natural Language Corpus

**Jill Gilkerson, Kimberly K. Coulter, & Jeffrey A. Richards**

LENA Foundation, Boulder, CO

LTR-06-2

September 2008

Software Version: V3.1.0

## ABSTRACT

---

The developing of the LENA™ language environment analysis software V3.1.0 as well as quantifying its accuracy and reliability is dependent on the accurate and reliable transcription of the audio recording files by professional transcribers. Here, we discuss the analytical procedure designed and implemented by the LENA Foundation's professional transcription team to identify and code speakers. We also reveal the degree to which inter-rater reliability was achieved between the criterion rater and four to seven secondary raters in terms of agreement for segment classification and adult word counts. Significant accuracy in the transcriptions was critical to train the processing models used to identify segments in audio recordings automatically and subsequently to test the accuracy and reliability of the segmentation process.

## Keywords

Classification agreement, Digital Language Processor, inter-rater reliability, segmentation, speaker identification, transcribe, transcription.

## 1.0 INTRODUCTION

---

The LENA language environment analysis software V3.1.0 software is a compilation of statistical models and algorithms trained to extract and segment features of the audio data. The statistical models were trained using professional transcriptions. Thus, the reliability of the professional transcripts was essential to model development and refinement. In this report, we describe the transcription process and discuss the level of inter-rater reliability that was achieved.

## 2.0 DESCRIPTION OF THE DATABASE

---

Starting in 2005, scientists at the LENA Foundation have conducted a large-scale study to assess the language environments of 2-month to 48-month-old children from families of varying socioeconomic status (SES) backgrounds. The result of this effort is a large database of quantifiable adult-child speech phenomena in a natural home environment, known officially as the LENA Natural Language Corpus. The current database consists of over 45,000 hours of audio from 329 diverse families whose demographic information closely matches the demographics of the United States with respect to mother's attained education as reported by the US Census (US Census, 2005). It was necessary to transcribe a subset of the LENA Natural Language Corpus to develop and test the speech processing algorithms that contribute to the LENA software V3.1.0.

### 2.1 Methodology

The professional transcription team at the LENA Foundation consisted of one criterion rater and seven secondary raters. Six of the eight raters performed the transcriptions off-site; off-site transcribers received the audio files by courier. The full-day audio files that the transcribers received contained notation regarding which segments to transcribe. Only transcribed segments were analyzed (typically 30 minutes per file).

The primary purpose of the LENA Foundation transcriptions was to optimize the LENA software V3.1.0 by training the speech processing algorithms to identify and segment a variety of features from the audio samples accurately and reliably. For example, it was necessary for the speech processing algorithms to differentiate adult speech from child speech and to differentiate the speech of the key child from the speech of other children or non-speech sounds (e.g. cries or vegetative sounds). The segmentation of the LENA audio processing algorithms was compared to the segmentation of professional transcribers to assess the accuracy and reliability of the algorithms. See Technical Report LTR-05-2 for information about the reliability of the LENA System.

## 2.2 Segment Identification

### *Overview*

For the purposes of speaker identification, audio data recorded by the LENA DLP were segmented into categories that included adult male and adult female, key child, noise, overlapping speech, other child, and electronic noise. The transcribers segmented instances of  $\geq 300$  ms of silence or transient noise (e.g. bumps and thuds) and “media” noise (any TV or radio sounds) to select and eliminate extraneous sounds that did not contribute meaningfully to the child’s language environment.

### *General Speaker Codes*

In order to identify the source of each audio segment consistently, a set of speaker codes was developed by the transcription team and used by all LENA Foundation transcribers throughout the transcription process. These speaker codes were designed to identify voices of children and adults as well as non-human sounds. For example, one particular speaker code referred to the key child; a separate code was assigned to children other than the key child. There are five general categories of speaker codes, including: 1) clear human speakers (e.g., key child, mother, father, etc.); 2) unidentifiable human speakers; 3) overlapping sounds (e.g., human + human, human + noise, noise + noise); 4) ambient sounds (e.g., sounds resulting from crowd gatherings, transient noises, etc.); and 5) other media noise (e.g., television, radio, telephone, electronic toys, etc.).

### *Speaker Code Subcategories*

Many of these speaker identifications were subcategorized. For example, overlapping voices in a speech segment occurring between the key child and another speaker, or between two other speakers, were assigned a specific code. Other types of overlap were categorized as well. In all, the LENA Foundation transcription team assigned 31 general codes for audio file analyses. Each of the 31 unique speaker identifications were subject to further classification based on sound wave amplitude. Low amplitude regions were identified as unclear or faint, indicating that the transcriber believed the speaker to be distant from the key child. Segments containing faint signals were not included in the word count processing since the meaningful contribution of these segments to the child's language environment was assumed to be minimal. Once the clear adult segments were identified and segmented, LENA Foundation transcribers counted the number of words spoken in each segment. The reliability of the adult word count estimates are described in detail in Technical Report LTR-05-2. To make it possible to test the reliability of Conversational Turns, LENA Foundation transcribers noted whether adult speech was directed toward the child.

### *Key Child Speech*

For each child speech segment, the transcribers further coded the key child segments as usable child vocalizations or child non-speech sounds. Sounds that were considered usable speech included words, babbles, and pre-speech communicative sounds or "protophones" such as squeals, growls, or raspberries (see Oller, 2000 for more detail on protophones). Sounds classified as non-speech were further subdivided into either fixed signals or vegetative sounds. Fixed signals include sounds that are instinctive emotional reactions to the environment (e.g. crying, screaming, laughing). Vegetative sounds are those resulting from respiration (e.g. breathing) or digestion (e.g. burping).

### *Event Reports*

In addition to coding the audio files, transcribers provided reports within the transcripts to note specific events occurring during the child's day. For example, the transcribers noted periods of time during which the child was in a car or at daycare.



## 2.3 File selection

There were two major transcription datasets that the engineers used either to train or to test their audio processing algorithms. These datasets are described below.

### *Training Set*

Large-scale training sets derived from the LENA Natural Language Corpus were used to train statistical models designed to assess the language environment of infants and toddlers. The training set is a compilation of 309 independent 30-minute long audio files selected for transcription. The files were selected on the basis of age and gender. Approximately five male and five female children ranging in age from 1 month to 42 months were selected per each month of age. The final sample included 157 male and 152 female children.

### *Test set*

The primary purpose of the transcription test set generation was to assess the accuracy and reliability of the LENA software V3.1.0 using a representative sub-sample of the LENA Natural Language Corpus. The test set transcription database is a compilation of 70 independent audio files selected for transcription. The files were selected on the basis of a block randomization scheme to ensure a test sample representative of the entire dataset with respect to age (2 months-36 months), gender, maternal SES, and independent standardized assessments of the recorded child's language ability (e.g., Preschool Language Scale, 4th Edition (Zimmerman, Lee, Steiner, & Pond, 2002)). Two children were selected per each month age group (i.e., two four-month-olds, two five-month-olds, etc.), one from a relatively higher SES bracket and the other from a comparatively lower SES bracket as determined by mother's education level. Most age groups contained one male and one female participant; there were 32 male and 38 female children in the test set sample.

An algorithm developed by the LENA Foundation Research and Development software engineers was used to select six ten-minute segments from each of the 70 audio recordings. The algorithm was programmed to detect high levels of back-and-forth alternation between key child and adult segments. For later analyses, these six ten-minute segments were concatenated to create one hour of transcription per file.

### 3.0 INTER-RATER RELIABILITY

---

Inter-rater reliability was assessed using two audio files obtained from one 11-month-old female and one 25-month-old male comparing a criterion rater with seven secondary raters. All transcribers transcribed three 10-minute segments from each file. The six 10-minute segments comprised 4,836 unique criterion-identified speaker segments that further included 916 criterion-identified key child speech segments and 716 criterion-identified adult male and female speech segments. Three separate classification analyses were conducted to assess inter-rater agreement with respect to: adult versus non-adult speech; key child versus non key child speech; and, for the key child, speech versus non-speech. Adult word counts for six ten-minute segments from a third file also were compared for the criterion rater and four of the secondary raters.<sup>1</sup>

#### 3.1 Segment Boundaries

To assess inter-rater reliability, the criterion rater's segment identification established the official segments of interest. Secondary raters' segments were matched to criterion rater segments using time codes that identified the start and stop points of each segment. Initial segment boundaries were set by the LENA software processing algorithm, but because each rater was permitted to adjust segment boundaries, it was sometimes the case that one criterion-based segment could span several secondary rater segments.<sup>2</sup> For these analyses, segments were considered to match when any portion of a secondary rater's segment overlapped with the criterion segment.

#### 3.2 Classification Agreement

Classification agreements (hits) were defined as occurring when a secondary rater's segment identification matched the criterion rater's and any portion of the secondary rater's segment overlapped with the corresponding criterion rater segment. Classification agreement was computed separately for identification of adult versus non-adult and key child versus non-key child speakers. Within criterion-identified key child speaker segments, agreement was computed for identification of key child speech versus key child non-speech, as described in Section 2.2 of this technical report. Agreement ratings were computed both as total (unadjusted) agreement and as the more conservative coefficient kappa,  $\kappa$  (Cohen, 1960). Results of the agreement rating analyses are shown in Table 1.

---

1 Adult word counts for three of the secondary raters were unavailable for this analysis.

2 In these data over 78% of criterion-based segment boundaries matched those of secondary raters exactly, over 95% spanned one or two segments, and over 99% spanned three or fewer segments.

**Table 1: Kappa ( $\kappa$ ) and Total Percent Agreement Between Criterion and Secondary Raters for: Adult versus Non-Adult Speaker; Key Child versus Non-Key Child Speaker; and Key Child Speech versus Non-Speech.**

Rater	Adult vs. Non-Adult Speaker (N=4836)		Key Child vs. Non-Key Child Speaker (N=4836)		Key Child: Speech vs. Non-Speech (N=916)		
	$\kappa$	Percent	$\kappa$	Percent	N <sup>a</sup>	$\kappa$	Percent
1	.64	.89	.74	.91	825	.69	.87
2	.64	.89	.75	.92	785	.81	.93
3	.61	.88	.69	.90	756	.73	.91
4	.61	.88	.76	.93	785	.74	.91
5	.62	.88	.78	.93	808	.77	.91
6	.58	.87	.71	.90	801	.73	.88
7	.63	.89	.71	.90	823	.76	.90
<b>Mean</b>	.62	.88	.74	.91	798	.75	.90

<sup>a</sup> There were 916 criterion-identified key child segments. Secondary rater agreement is based on the number of segments for which they matched the key child identification.

As Table 1 demonstrates, secondary raters' identification of adult versus non-adult segments matched the criterion rater's identification on average 88% of the time. Similarly, secondary raters' identification of key child versus non-key child speakers typically matched the criterion rater's 91% of the time. When the raters identified segments containing key child speech versus key child non-speech, on average they were in agreement approximately 90% of the time. Thus, the level of agreement among raters across the categories of interest indicated sufficient accuracy for these transcriptions to be used for model training and testing.

### 3.3 Word Count Agreement

To determine a more representative estimate of inter-rater agreement for Adult Word Count, the criterion rater selected six 10-minute audio segments from a third recording and estimated adult word counts for each segment. Four of the secondary raters were instructed not to adjust segment identifications or boundaries but only to estimate adult word counts. Table 2 and Figure 1 detail total AWC for each ten-minute segment for each rater. Average differences in total word counts for five of the six segments ranged from approximately one to twelve words (0.2% – 1.5%); differences for the remaining segment were somewhat larger (from 32 to 54 words) for an average difference of 4.9%. For all segments combined, the average difference from the criterion rater was 1.3%.

**Table 2: Adult Word Counts for Six 10-Minute Audio Segments.**

Rater	Adult Word Count (10-Minute Segment)						Total
	1	2	3	4	5	6	
<b>Criterion</b>	689	464	756	937	501	607	3954
<b>Mean (1-4)</b>	686	464	755	891	501	607	3903
1	685	459	768	883	501	612	3908
2	685	457	736	891	499	600	3868
3	698	473	766	905	502	606	3950
4	675	465	751	886	500	610	3887